Orientation Selectivity and Contrast Gain Control in Representations of Natural Images

Dissertation

zur Erlangung des Grades eines Doktors
der Naturwissenschaften

der Mathematisch-Naturwissenschaftlichen Fakultät
und
der Medizinischen Fakultät
der Eberhard-Karls-Universität Tübingen

vorgelegt
von

Fabian H. Sinz
aus Straubing, Deutschland

Dezember 2011

Tag der mündlichen Prüfung:	12. 6. 2012
Dekan der math.-nat. Fakultät	Prof. Dr. W. Rosenstiel
Dekan der med. Fakultät	Prof. Dr. I. B. Autenrieth
1. Berichterstatter:	Prof. Dr. M. Bethge
2. Berichterstatter:	Prof. Dr. M. Giese
3. Berichterstatter:	Prof. R. W. Fleming, PhD
Prüfungskomission:	Prof. Dr. M. Bethge
	Prof. Dr. A. Schilling
	Prof. Dr. M. Giese
	Prof. R. W. Fleming, PhD

I hereby declare that I have produced the work entitled: "Orientation Selectivity and Contrast Gain Control in Representations of Natural Images", submitted for the award of a doctorate, on my own (without external help), have used only the sources and aids indicated and have marked passages included from other works, whether verbatim or in content, as such. I swear upon oath that these statements are true and that I have not concealed anything. I am aware that making a false declaration under oath is punishable by a term of imprisonment of up to three years or by a fine.

Tübingen, _____ _____

Date Signature

Acknowledgements During my time at the Max Planck Institute of Biological Cybernetics, which not only covers the time as a PhD student but also the time when I worked there as a student, I met many people who greatly impressed and influenced me on a scientific and personal level. Those people I would like to thank here. It is almost impossible to mention all of them, so I stick to the most important ones and assure my gratitude to all the others.

First and foremost, I would like to thank my supervisor Matthias Bethge, whose sharp thinking, brilliant intuition, and deep care for his lab members created an scientific and personal environment I consider myself lucky to have enjoyed during my time as a PhD student. A good deal of this environment was also due to the other members of the Bethge lab who were great company, and who willingly and passionately engaged in discussions about science and life in general which I consider one of the most important parts in the creative process of science. In particular, for hours of whiteboard freestyling, I would like to thank Sebastian Gerwinn, Lucas Theis, and Reshad Hosseini. For great discussions and advice about science, life, politics and philosophy, I am particularly grateful to Philipp Berens and Ralf Häfner.

Beyond the Bethge lab, there are several persons who work or have worked at the Max Planck Institutes, the Centre for Integrative Neuroscience or other research institutions whose company I enjoy and whose opinion I appreciate. In no particular order, these are: Moritz Grosse-Wentrup, Lisa Smith, Dilan Görür, Frank Jäkel, Elisabeth Hopp, Suvrit Sra, Roland Fleming, Daniel Holtmann-Rice, Steffi Jegelka, Florian Steinke, Matthias Hein, Olivier Chapelle, Gunnar Rätsch, Jason Weston, Ronan Collobert, Laura Busse, Steffen Katzner, Joaquin Quiñonero-Candela, Peter Gehler, Jeremy Hill, Nicole Fröhlich, Matthias Franz, Alekh Agarwal, Leon Bottou, and Arthur Gretton. I especially want to thank Bernhard Schölkopf because he hired and supported me as a student assistant and thereby opened this great research environment to me.

Furthermore, I would like to acknowledge the German National Academic Foundation (Studienstiftung des deutschen Volkes) for financial support and its academies where I met many nice and inspiring people. Also, I would like to thank everyone at the Graduate School for Neural and Behavioural Sciences who were always friendly and very supportive.

Finally, I would like to thank all my friends and my family for their support and for keeping my life balanced and happy.

I once attended a talk by nobel laureate Olivier Smithies who told us that everyone needs a companion in life. Because it is so hard to find the right words to properly express what should be said, I just want to thank mine the simplest way possible and be sure that she knows what I intend to convey: Thank you Lisa.

Abstract This thesis explores the role of orientation selectivity and contrast gain control with respect to Barlow's normative redundancy reduction hypothesis in simple models of the early visual system. Our general approach uses the fact that—under the goal of redundancy reduction—early vision models are density models on natural images. We identify and develop new classes of probabilistic models for natural image patches that contain these early vision models. We use those classes to quantitatively explore their parameter space around the early vision models statistically and information theoretically with respect to the influence of filter shapes and contrast transforms on redundancy reduction. We identify an optimal contrast gain control transform and compare it to the standard model of cortical divisive contrast gain control, divsive normalization. We also identify a new estimation method for the true redundancy of natural images.

Our main findings are that, in contrast to divisive contrast gain control, orientation selectivity plays a minor role for redundancy reduction in the models investigated, and that the cortical model of divisive contrast normalization is not the optimal redundancy reducing contrast transformation on static image patches. However, we are able to specify a dynamical model of cortical contrast gain control with strong redundancy reduction, through extending the static model by adaptation to temporal correlations between consecutive contrasts caused by fixations under natural viewing conditions.

Contents

1 Introduction **1**

2 Results **11**
2.1 Natural Image Coding in V1: How Much Use Is Orientation Selectivity? 11
2.2 Divisive Normalization and Orientation Selectivity 13
2.3 Characterization of the p-generalized normal distribution 15
2.4 Hierarchical Modeling of Local Image Features through L_p-Nested Symmetric Distributions 15
2.5 L_p-nested symmetric distributions 17
2.6 Lower bounds on the redundancy of natural images 18
2.7 Temporal adaptation enhances efficient contrast gain control on natural images 19

3 Discussion and Conclusion **21**

4 Appendix **39**
4.1 Coding of Natural Images in V1: Original Article 39
4.2 Coding of Natural Images in V1: Supplementary Material 57
4.3 Divisive Normalization and Orientation Selectivity: Original Article . . 67
4.4 Divisive Normalization and Orientation Selectivity: Supplementary Material 77
4.5 Characterization of the p-generalized normal distribution: Original Article 95
4.6 Hierarchical Modeling of Local Image Features: Original Article 101
4.7 Hierarchical Modeling of Local Image Features: Supplementary Material 111
4.8 L_p-Nested Symmetric Distributions: Original Article 151
4.9 Lower bounds on the redundancy of natural images: Original Article . . 195
4.10 Temporal adaptation: Submitted Draft 207

The difference in behaviour in different species reflect different ways of coping with the environment, or with distinct niches of the environment. These different behaviours have their material counterpart in different brains. Therefore the peculiar nature of any animal brain cannot be explained from the physiological components alone, the explanation necessarily involving causes residing outside the animal, i.e. information derived from the environment.

from *Manifesto of Brain Science* by Valentino Braitenberg (1926-2011)

1 Introduction

In his book *Vision* David Marr emphasized that the visual system should be seen as an information processing machine, and that its understanding requires knowledge of the machine as well as the information processing task itself [Marr, 1983]. He distinguishes three levels of understanding: the *computational goal* which determines what is to be computed, the *representation* of the data and the *algorithm* achieving that goal, and the specific neural *implementation* of the algorithm.

Representations and normative models It is not clear whether there is a single computational goal that includes all the capabilities of the visual system like object recognition, figure ground segregation, or stereo vision, and others. However, since all higher visual areas in cortex obtain the visual information via the so called *early visual system*, which is the pathway from retina to primary visual cortex, it appears as if all algorithms in the visual system start from the same cortical representation of data. Since all higher visual areas get their signals via this pathway, there is hope that understanding the principles behind information representation in the early visual system also reveals insights about the algorithms which use that representation. David Marr stressed that the representation of information and the algorithm processing it are not independent, because certain representations will make the computations easier, others harder. Therefore, one would expect that an efficient algorithm uses a representation that is especially tailored to it. For example, the Arabic representation of numbers is better suited for addition or multiplication than the Roman.

Since the representation of visual information in primary visual cortex must serve many goals further up in the visual system, it has been hypothesized that there might be a general computational principle governing the representation of information in that pathway. In particular, it has been proposed that the visual system is adapted to the statistics of natural images in an information theoretic sense via the so called *redundancy reduction* or *efficient coding hypothesis* [Barlow, 1961, Attneave, 1954]. Although the two are closely related, they differ in important aspects. In the following, we will briefly introduce both of them, work out the main differences, and motivate why we focus on redundancy reduction here.

Efficient coding and redundancy reduction The number of possible input patterns to the visual system is enormous and if each input pattern were equally likely, the amount of neurons needed to represent visual patterns would be immense [Simoncelli and Olshausen, 2003]. Fortunately, natural visual signals are highly structured and only make up a very small fraction of all possible patterns. Due to that structure, certain parts of the signal can be predicted from others. For example, a simple but effective

strategy for predicting the grayscale value of a missing pixel in an image is to use the mean grayscale value of its neighbors. This strategy works because large parts of natural images are surfaces with similar grayscale. It would not work for white noise, for instance. The structure in the signal can be used to design a representation that carries as much information about the signal as possible. The idea is to spend little resources on the part of the signal that can be predicted and spend more resources on the unpredictable part. The efficient coding hypothesis acknowledges this fact and postulates that the visual system makes best use of its resources and transmits information as efficiently as possible [Barlow, 1961, Linsker, 1988, Atick, 1992, Nadal and Parga, 1994].

Information theory behind efficient coding Efficient coding can be cast in information theoretic terms by considering the visual input and the neural response to be random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. According to efficient coding, the visual system tries to choose the representation $\boldsymbol{Y}$ such that the transmitted information about $\boldsymbol{X}$ is maximized. The amount of information a neural response $\boldsymbol{Y}$ conveys about an input pattern $\boldsymbol{X}$ is captured by the mutual information between $\boldsymbol{X}$ and $\boldsymbol{Y}$. The mutual information represents the average reduction in uncertainty about an input $\boldsymbol{X}$ if the corresponding neural response $\boldsymbol{Y}$ is observed, or vice versa [Cover and Thomas, 2006]. The uncertainty about $\boldsymbol{X}$ is expressed in the joint entropy $H[\boldsymbol{X}]$, the uncertainty after observing $\boldsymbol{Y}$ by the conditional entropy $H[\boldsymbol{X}|\boldsymbol{Y}]$. The difference between the two yields the mutual information

$$I[\boldsymbol{Y};\boldsymbol{X}] \quad = \quad H[\boldsymbol{X}] - H[\boldsymbol{X}|\boldsymbol{Y}] = H[\boldsymbol{Y}] - H[\boldsymbol{Y}|\boldsymbol{X}]. \tag{1.0.1}$$

The mutual information of $\boldsymbol{X}$ and $\boldsymbol{Y}$ depends on their joint distribution. If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent, then $H[\boldsymbol{Y}|\boldsymbol{X}] = H[\boldsymbol{Y}]$ and $H[\boldsymbol{X}|\boldsymbol{Y}] = H[\boldsymbol{X}]$, and the mutual information attains zero, its lowest possible value [Cover and Thomas, 2006]. As soon as $\boldsymbol{X}$ and $\boldsymbol{Y}$ are dependent, $I[\boldsymbol{X};\boldsymbol{Y}]$ becomes positive. If $\boldsymbol{Y}$ is an invertible deterministic function of $\boldsymbol{X}$, the mutual information is maximal. In reality, however, the relation between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is probably neither of both since information might be discarded, a single input might be represented by several neural signals, or simply due to noise. After making certain assumptions about the relation between $\boldsymbol{X}$ and $\boldsymbol{Y}$, the mutual information can be maximized via the choice of representation $\boldsymbol{Y}$ by maximizing entropy in the responses $H[\boldsymbol{Y}]$ and minimizing the noise entropy $H[\boldsymbol{Y}|\boldsymbol{X}]$ at the same time.

Relation between efficient coding and redundancy reduction If the uncertainty in the response $\boldsymbol{Y}$ for given $\boldsymbol{X}$ does not depend on $\boldsymbol{X}$, for example when each neural response is distorted with independent additive noise, then $H[\boldsymbol{Y}|\boldsymbol{X}]$ is constant, and the maximization of the mutual information is equivalent to maximizing $H[\boldsymbol{Y}]$. Under certain technical conditions that exclude trivial maximizations of $H[\boldsymbol{Y}]$, for example by just increasing the signal variance, $H[\boldsymbol{Y}]$ is maximized by making its single components $Y_1, ..., Y_n$ statistically independent [Bell and Sejnowski, 1997]. For neural populations this means that $H[\boldsymbol{Y}]$ can be maximized by making the single neural responses

Y_i statistically as independent, or, equivalently, as non-redundant as possible. The *redundancy reduction hypothesis* postulates that this is the goal of populations of sensory neurons. Since $\boldsymbol{Y}$ is thought to be a stochastic function of $\boldsymbol{X}$, redundancy reduction depends on the statistics of $\boldsymbol{X}$. The information theoretic measure for redundancy is the *multi-information* $I[\boldsymbol{Y}] = \sum_i H[Y_i] - H[\boldsymbol{Y}]$ [Perez, 1977].

In general, *redundancy reduction* and *efficient coding* are not the same. In particular, redundancy reduction does not take into account the role of noise or intrinsic uncertainty. For instance, the value of $H[\boldsymbol{Y}|\boldsymbol{X}]$ might depend on $\boldsymbol{X}$, or the dimensionality of $\boldsymbol{Y}$ is larger than the dimensionality of $\boldsymbol{X}$ which generates an intrinsic uncertainty since several values of $\boldsymbol{Y}$ correspond to a single value of $\boldsymbol{X}$. In that case, maximization of $I[\boldsymbol{Y};\boldsymbol{X}]$ will require a trade-off between maximizing $H[\boldsymbol{Y}]$ and minimizing $H[\boldsymbol{Y}|\boldsymbol{X}]$.

Reasons for studying redundancy reduction However, there are good reasons to focus on redundancy reduction hypothesis for the visual system over efficient coding. In efficient coding, the maximization of transmitted information $I[\boldsymbol{X};\boldsymbol{Y}] = H[\boldsymbol{Y}] - H[\boldsymbol{Y}|\boldsymbol{X}]$ is a trade-off between maximizing the information contained in $\boldsymbol{Y}$ via $H[\boldsymbol{Y}]$ and minimizing the influence of noise via $H[\boldsymbol{Y}|\boldsymbol{X}]$. Intuitively, however, choosing a representation $\boldsymbol{Y}$ that yields enough information to reliably decode the state of the outside world with a limited number of neurons seems a much harder problem than dealing with internal noise. In other words, when presented with a specific visual input, e.g. a door, the hard problem is to find out that—among all possible things—it is a door which is facing us at the moment, and not to deal with the noise that got into the signal while it was transmitted from retina to cortex. In that sense, redundancy reduction concentrates on the more crucial problem by ignoring the noise $H[\boldsymbol{Y}|\boldsymbol{X}]$ and focusing on the maximization of $H[\boldsymbol{Y}]$. Apart from that, redundancy reduction by itself can be used as a strategy to achieve many potential goals the visual system might have [Barlow, 1961, 1985, 1989, 2001, 2002] . For example, redundancy reduction could in principle enable the visual system to learn the hidden causes for the sensory input [Barlow, 1989, Bell and Sejnowski, 1997]: Redundancies in the sensory input are often due to regularities in the objects causing it. For instance, one can think of a rigid object as a collection of redundant points in space and time since their spatial configuration is fixed. If the visual system is able to detect and remove those redundancies it effectively has learnt a model of rigid objects and obtained an efficient representation of it. Another motivation for redundancy reduction is that it can be seen as a way to build a probabilistic model of the sensory input [Barlow, 1985]: The idea is related to a density estimation algorithm known as *projection pursuit* [Friedman et al., 1984] in which a random variable $\boldsymbol{X}$ with an unknown source density is iteratively remapped into a random variable $\boldsymbol{Y}$ that becomes more and more Gaussian after each iteration. Knowing that, after enough iterations, the distribution of $\boldsymbol{Y}$ is Gaussian and knowing the individual mappings effectively yields a density model for the input. If the Gaussian target distribution is replaced with an arbitrary *factorial* distribution, i.e. one that has independent marginals, then projection pursuit and redundancy reduction become equivalent. For those reasons, this thesis focuses on redundancy reduction in the representation of

visual input in the early visual system.

Scientific status of normative principles How can *normative principles* like the redundancy reduction hypothesis be tested? As mentioned by Simoncelli and Olshausen [2003], it is difficult to establish a firm link between neurophysiological response properties and natural image statistics.

Testing normative hypotheses in vivo The most direct way of testing the redundancy reduction hypothesis would be to measure the statistical dependencies of neural responses at the different stages of the visual pathway. Despite the large experimental difficulties there have been attempts to do this in the retina [Puchalla et al., 2005], thalamus [Dan et al., 1996], and the primary as well as inferior temporal cortex [Baddeley et al., 1997, Vinje and Gallant, 2000]. Other studies also measured redundancies or coding efficiency in the auditory pathway [Rieke et al., 1995, Chechik et al., 2002] or insect visual systems [Laughlin, 1981]. While these studies indicate that the responses of different neurons indeed become increasingly independent along sensory pathways, neurophysiological tests still struggle with the fact that only a small portion of the entire population can be observed and that the amount of data to estimate the information theoretic measures is limited.

Testing normative hypotheses in silico Another way to test normative hypotheses is to use models of the early visual system and optimize their free parameters on large collections of natural images with respect to a statistical optimality criterion defined by the normative principle [Simoncelli and Olshausen, 2003, Simoncelli, 2003]. If the normative principle and the model are correct, then one would expect to find neurophysiologically plausible features of the model at the optimum. This approach has been very fruitful for understanding the interplay between the visual system and the redundancy reduction hypothesis. Buchsbaum and Gottschalk [1983] as well as Ruderman et al. [1998] demonstrated that decorrelation of the three color channels of natural images leads to blue-yellow, red-green and dark-bright color opponency as observed in retinal ganglion cells. Atick and coworkers as well as van Hateren showed that spatial and spatio-temporal decorrelation of natural images yields band-pass filters as observed in the retina and thalamus [Atick and Redlich, 1990, 1992, Dong and Atick, 1995, van Hateren, 1992, Van Hateren, 1993]. Since removing second order correlations does not uniquely specify the linear receptive fields of the model neurons, later studies introduced neurophysiological constraints in order to obtain localized and oriented band-pass filters similar to the receptive fields of simple cells in primary visual cortex [Sanger, 1989, Hancock et al., 1992, Shouval et al., 1997, Li and Atick, 1994]. Only after the reduction of higher order redundancies was incorporated into the objective was it possible to also obtain orientation selective filters without additional constraints [Olshausen and Field, 1996, Bell and Sejnowski, 1997, Van Hateren and Van Der Schaaf, 1998, Lewicki and Olshausen, 1999]. By optimizing for independent groups of neurons, Hyvärinen and coworkers reproduced orientation selective but phase invariant

groups of filters like in the energy model of complex cells [Hyvärinen and Hoyer, 2000, Hyvärinen and Koester, 2007, Adelson and Bergen, 1985, Pollen and Ronner, 1983]. Finally, Schwartz and coworkers showed that *divisive normalization*, which is one of the prominent non-linear functional properties of primary visual cortex [Albrecht and Hamilton, 1982, Bonds, 1989, Heeger, 1992, Geisler and Albrecht, 1992, Carandini et al., 1997], reduces higher order statistical dependencies of natural images [Schwartz and Simoncelli, 2001, Wainwright et al., 2002].

Consistency conditions for tests in silico These results are encouraging evidence in favor of redundancy reduction. However, in order for this evidence to be resilient, further criteria must be met. First of all, the models used to reproduce neurophysiological features and response properties from natural images must be realistic enough to allow firm conclusions. It is possible that neural features arise in a model which is too simple or not adequate, but these features would not be optimal in terms of redundancy reduction in a more realistic model. Second, the results must be discriminative: If there is a whole set of model parameters that performs well in terms of redundancy reduction of which only a part is neurophysiologically reasonable, then redundancy reduction is not a very strong explanation for the neurophysiological features. Third, the assumptions about the statistics of natural images entering the model and the optimization should be correct. Investigating whether these criteria are met is particularly important for higher order redundancy reduction, since modeling and measuring them is more difficult and subtle. Higher order redundancy reduction results mainly concern features of the primary visual cortex, orientation selectivity and divisive normalization, and are the main focus of this thesis.

Relation between neural population models and density models on natural images under redundancy reduction In terms of redundancy reduction, the algorithms used in previous studies reproducing orientation selective filters similar to simple and complex cells are equivalent to independent component analysis (ICA) and independent subspace analysis (ISA) [Comon, 1994, Bell and Sejnowski, 1997, Hyvärinen and Hoyer, 2000]. These algorithms are in turn equivalent to minimizing the redundancy in a population of independent linear-nonlinear (LN) neurons [Chichilnisky, 2001], since an invertible element-wise nonlinearity like some of the ones used in LN-neurons for turning the filter output into a firing rate does not change the redundancy, and it is therefore sufficient to directly look at the redundancy of the filter outputs. However, neurons in cortex are not independently wired units but ones that interact. One of the most prominent interactions between neurons is divisive normalization [Heeger, 1992]. Over-complete linear models, like the one in the study by Olshausen and Field [1996], also nonlinearily couple the neural response by a *maximum a posteriori* (MAP) estimate of the neural response given the visual input. This nonlinearity can resemble certain cortical features, like end-stopping, but it does not reproduce divisive normalization. Additionally, there is also no guarantee that these MAP estimates yield statistically independent neural responses which would be necessary to agree with the redundancy

reduction hypothesis.

Further need for quantitative evaluation of early vision models and natural image statistics Concerning model discriminability, Bethge [2006] showed that the optimum around orientation selective filters in linear ICA models for redundancy reduction is very shallow. He demonstrated that after whitening, which is ascribed to stages earlier than the cortex [Atick and Redlich, 1990, 1992, Dong and Atick, 1995, van Hateren, 1992], the particular filter shape only makes up for about 5% of the total redundancy reduction. The small contribution of linear filters to higher order redundancy reduction is mainly caused by the fact that natural image patches are not well modeled by a linear ICA model [Simoncelli, 1997, Eichhorn et al., 2009]. This means that the amount of higher order redundancies removed by linear filters in these models is small and that random whitening filters and orientation selective filters perform almost equally well.

The studies on the redundancy reducing effect of divisive normalization use a fixed filter bank to model the receptive fields of simple cells [Schwartz and Simoncelli, 2001, Wainwright et al., 2002]. It is not clear, however, whether optimizing a model which includes divisive normalization still yields orientation selective filters as the optimal filter shape for redundancy reduction and what the quantitative contributions of the filter shape would then be. Additionally, previous work on divisive normalization and higher order redundancy reduction visualized the higher order statistical dependencies via so called *bow-tie plot*s. A bow-tie plot shows the conditional distributions of one filter response given the response of a neighboring filter [Schwartz and Simoncelli, 2001]. From these plots one can see that the variance of the conditional distribution depends on the absolute value of the response on which it is conditioned. This leads to the typical bow-tie shape of the plots and demonstrates the presence of variance correlations. After divisive normalization, the bow-tie plots become flat which indicates that variance correlations have been removed [Schwartz and Simoncelli, 2001]. However, bow-tie plots depend on binning of the signals and only depict one certain type of higher order correlation. Although one can show that the underlying *Gaussian scale mixture model* used in these studies has non-decreasing variance correlations as soon as the distribution has higher order correlations [Wainwright and Simoncelli, 2000, Cambanis et al., 2000, Kac, 1939], these correlations might be subtle and not be apparent from the bow-tie plot. Since there was no quantitative evaluation of the multi-information, it is not clear how much redundancy is left after divisive normalization. Furthermore, it is not clear what the maximal amount of redundancies is that can be removed by transformations like divisive normalization.

Contributions of this thesis The studies contained in this thesis address the aforementioned consistency issues of

- discriminability of redundancy reduction for certain features
- adequacy of the statistical assumptions about natural images made by these early vision models

- quantitative assessment of the influence of features on redundancy reduction.

Building upon the work of Bethge [2006], the shortcomings of linear ICA models on natural image patches are investigated, and objectives other than redundancy reduction are assessed for which filters resembling receptive fields in primary visual cortex show a clear advantage. A major objective in all the studies is to obtain *quantitative* measurements which, in the end, will hopefully enable us to rule out certain models in favor of others. Unfortunately, quantitative measurements of probabilistic and information theoretic quantities on natural images are difficult to obtain. Therefore, the models and their extensions developed in this thesis have to make a trade-off between fully capturing the complexity of cortical neural networks and allowing for quantitative measurements at the same time. However, by thoroughly analyzing simpler models first, it is easier to disentangle the essential mechanisms and their interplay.

The general methodological approach The common scheme in addressing the above mentioned questions in a quantitative manner is to use the fact that, under the goal of redundancy reduction, different neural models correspond to different statistical models on natural image patches (see Figure 1.0.1). We embed these models into a larger class of probability distributions which allows us to explore the parameter space of these models with information theoretic and probabilistic measures with respect to features like filter shapes or contrast gain control. To this end, the studies in this thesis identify and develop new classes of probability distributions that better match the regularities found in natural images. These models not only form a better basis for linking natural image statistics to neural response properties, but also are a contribution to the field of natural image statistics themselves.

From these classes of distributions, a unique divisive normalization mechanism is derived that is optimal with respect to redundancy reduction. A link between the likelihood of natural image models and the amount of redundancy reduction they achieve is established and used for quantitative model comparison between models with or without divisive normalization mechanisms. A new class of probability models is developed that allows for the quantification of the relative influence of orientation selective filters on redundancy reduction in complex cell models as proposed by Hyvärinen and coworkers [Hyvärinen and Hoyer, 2000, Hyvärinen and Koester, 2007]. By developing a new information theoretic estimation method, we get better estimates of the true redundancy of natural images and take a first step in developing models for whole images instead of image patches. Finally, we explore how the physiologically plausible divisive normalization model compares to the optimal transformation in terms of redundancy reduction. It is demonstrated that a static model of cortical divisive normalization is not sufficient for strong redundancy reduction but that a simple dynamic adaptive mechanism which uses temporal correlations in the images as induced by eye movements can significantly enhance the performance.

By developing new probability models for natural images, characterizing them mathematically, using them to investigate normative hypotheses for the early visual system, and by developing new information theoretic estimation methods, this thesis provides

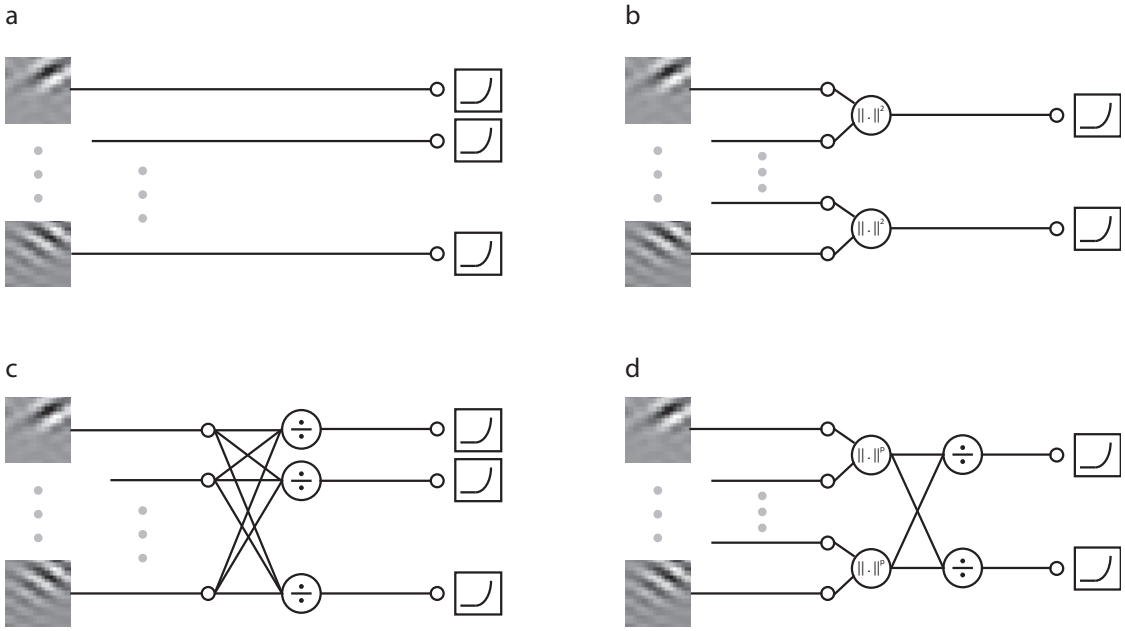


Figure 1.0.1: Neural population models and their probabilistic counterpart. Since an invertible element-wise non-linearity on the outputs does not change the redundancy, it has no influence on the redundancy reduction results and can also be ignored in all models. **a**: A population of linear-nonlinear simple cell models for which the output is required to be statistically independent is equivalent to an independent component analysis (ICA) model. Models like this are the subject of Eichhorn et al. [2009]. **b**: A population in which the response of oriented filters are squared and grouped by summation correspond to the energy model of complex cells. If the outputs after summation are required to be independent, the model corresponds to an independent subspace analysis (ISA) model. Models like this are the subject of Sinz et al. [2009b]. **c**: A population in which the linear filter responses are transformed by a (divisive) mechanism on the L_p-norm correspond to a population of simple cells with contrast gain control mechanism. If the outputs are required to be statistically independent, the corresponding probabilistic model belongs to the class of L_p-spherically symmetric distributions. Such models are the subject of Sinz and Bethge [2009], Sinz et al. [2009a], and Sinz and Bethge [*submitted*]. **d**: A population in which the response of oriented filters are raised to a positive power, grouped by summation, and transformed by a (divisive) mechanism on the L_p-norm of the grouped responses corresponds to an energy model of complex cells with contrast gain control. If the outputs are required to be statistically independent, the corresponding probabilistic model belongs to the class of L_p-nested symmetric distributions. Such models are the subject of Sinz et al. [2009b] and Sinz and Bethge [2010].

contributions to the field of computational vision, natural image statistics, mathematical statistics and information theory.

2 Results

This section briefly presents the research questions and the results of every article included in this thesis.

2.1 Natural Image Coding in V1: How Much Use Is Orientation Selectivity?

Motivation Several previous studies reported that orientation selective filters yield an additional reduction of higher order redundancies of about 20% for gray value images and over 100% for color images when compared to decorrelating filters like the ones obtained from PCA [Lewicki and Olshausen, 1999, Wachtler et al., 2001, Lee et al., 2002]. If these findings were correct, it would mean that, in contrast to what was reported in [Bethge, 2006], the shape of the filter which is determined by the choice of an orthogonal transformation after whitening makes a significant difference for the reduction of higher order redundancies.

The goal of the study was to carry out a thorough quantitative analysis of how much higher order redundancy reduction can be achieved with orientation selective filters resulting from ICA. To this end, we compared the ICA filters to filters from principal component analysis (PCA) and random whitening filters which only aim at removing second-order correlations. In addition to redundancy measured via the multi-information, we also evaluated two other objective functions for which orientation selective receptive fields might be an advantage: the average log-loss and rate distortion curves [Bernardo, 1979]. The average log-loss is the negative average log-likelihood. The higher this loss is, the less a probabilistic model fits the data. Its lowest value is the entropy of the true data distribution in the case the model matches the true distribution. The use of the average log-loss is motivated by the density estimation view on redundancy reduction (see Introduction).

Finally, we also evalutated the potential advantage of orientation selective filters in a rate-distortion curve setting. The mere maximization of the amount of information that is transmitted, as in the information maximization framework of redundancy reduction [Linsker, 1988, Atick, 1992, Nadal and Parga, 1994], is agnostic to the information that is relevant [Simoncelli and Olshausen, 2003]. Rate-distortion curves represent not only the information that can be transmitted, but also take into account what information is relevant through the use of a loss function. In order to evaluate whether orientation selective filters might be superior in transmitting the relevant information, we resorted to rate-distortion curves in a linear transform coding framework with mean squared error loss [Lewicki and Olshausen, 1999, Lewicki and Sejnowski, 2000]. The mean squared

error is strongly related to image compression and its choice is a trade-off between computational feasibility and capturing perceptually relevant image features.

Results For all measures evaluated, we did not find a clear advantage of ICA over other models. We carried out the analysis on the same dataset as the studies reporting high reductions of higher order redundancies [Wachtler et al., 2001, Lee et al., 2002]. In terms of multi-information reduction, we found that the amount of higher order redundancies removed by ICA makes up for about 3% of the total redundancy reduction achieved. With 3.20% the reduction for color images was a bit higher than for monochromatic images with 2.39%. We could not reproduce the previously reported gains of 20% to 100%.

In terms of density estimation, ICA filters perform best among other factorial density models of natural image patches which differed in the choice of filters. This is no surprise since ICA filters are optimized to yield statistically independent responses. However, the difference compared to factorial models on other filter responses was small. In particular, a simple spherically symmetric model with less degrees of freedom performed significantly better. This strongly indicates that higher order correlations in natural image patches are not well removed by any linear transformations, because spherically symmetric models are agnostic with respect to the specific shape of the filters. Additionally, the only factorial spherically symmetric density is the Gaussian, and it is well known that filter responses to natural images do not exhibit a Gaussian distribution [Field, 1987] but shares its spherical symmetry [Zetzsche et al., 1999]. The better performance of the non-factorial spherically symmetric density over factorial distributions in terms of average log-loss shows that (i) there are still higher order correlations left which are not removed by ICA filters and (ii) there is no choice of filters that can remove those. If there were, then the density model should exhibit specific axes for which the distribution becomes factorial. Those should be found by ICA whose density model should then have a clear advantage over spherically symmetric models. However, the good performance of the spherically symmetric model suggests that there are no such axes and hence, that the choice of filter shape will probably not have much impact on redundancy reduction or on density modeling, respectively.

The performance of orientation selective filters for rate distortion curves is even worse than the performance of PCA filters. We carefully analyzed why this is the case and found that the determining factor for the performance is the change of the metric induced by different filter choices. Intuitively, conserving the metric means that a square (hypercube) in the input will be mapped into a square (hypercube) in the output. Only orthogonal linear transformations are metric preserving. Due to that, PCA filters are the only decorrelating transform that leave the metric invariant. ICA filters are not orthogonal, which means that the square becomes diamond shaped, and all whitening transforms at least change the side length of the square. For that reason, PCA filters outperform ICA filters.

Conclusions Based on the results of the study we were not able to find a single objective for which orientation selective filter shapes yield a clear advantage. The good performance of the spherically symmetric model demonstrates the presence of higher order redundancies that cannot be removed by a complete set of linear filters. This result holds for grayscale and for color images. This, and the insufficient performance of the ICA model in terms of log-likelihood, demonstrates that the underlying assumptions about natural images that led to orientation selective filters [Bell and Sejnowski, 1997] are not justified. Therefore, strong higher order redundancy reduction requires a more powerful nonlinear mechanism.

2.2 The Conjoint Effect of Divisive Normalization and Orientation Selectivity on Redundancy Reduction

Motivation Based on the previous superior performance of the spherically symmetric model, the goal of this study was to further explore the statistics of filter responses to natural images, analyze whether there are certain filter shapes that are better suited for density modeling, and develop nonlinear mechanisms for redundancy reduction. Since spherically symmetric models are agnostic with respect to the particular filter shape, we used the class of L_p-spherically symmetric models as a basis for our exploration [Gupta and Song, 1997, Song and Gupta, 1997], which contain the spherically symmetric distributions but also other models that have iso-density contours shaped like the unit sphere in other norms (therefore the name L_p-spherically symmetric). These other iso-density contours exhibit a symmetry breaking with respect to the filter shapes which enables us to compare different filters via the log-likelihood.

Results We modeled the responses of whitening filters to natural image patches with L_p-spherically symmetric models. While spherically symmetric models have iso-density contours of constant Euclidean norm, L_p-spherically symmetric models have constant density along the contours of the L_p-norm $\|\boldsymbol{x}\|_p = (\sum |x_i|^p)^{1/p}$. For $p \neq 2$ the corresponding density is not invariant with respect to the filter shapes. Importantly, for a fixed value of p, an L_p-spherically symmetric distribution is completely characterized by a univariate radial distribution. For each value of p, there is a single type of radial distribution that corresponds to a joint factorial model called *p-generalized Normal distribution* [Goodman and Kotz, 1973] (see Section 2.3). We used this property to derive a non-linear transformation on the L_p-norm which transforms the radial distribution of the data into the radial distribution of the p-generalized Normal. This mechanism is the *optimal* non-linear redundancy reduction transform for L_p-spherically symmetric distributed data. Most importantly, this mechanism, called *radial factorization*, can turn any L_p-spherically symmetric source into statistically independent signals. This nonlinear redundancy reduction mechanism can now be compared to physiologically known nonlinear mechanisms. Furthermore, we were able to assess the relative influence of the receptive field shapes on redundancy reduction in models with or without

the non-linear mechanism, respectively.

We evaluated the likelihood for different values of p using a flexible radial distribution. Consistent with our findings for the spherically symmetric distribution (see Section 2.1), we found that non-factorial L_p-spherically symmetric models exhibit a higher likelihood. The L_p-spherically symmetric model with ICA filters yielded the best likelihood and a value of $p < 2$. This means that the filter shape makes a difference in performance. We also optimized the filters with respect to the log-likelihood under a L_p-spherically symmetric model and found that orientation selective filters are a stable optimum.

For different filter shapes ranging from orientation selective to random, we computed the radial factorization transform and evaluated its redundancy reduction performance. To this end, we derived a connection between the log-likelihood of the L_p-spherically symmetric model and its redundancy reduction performance. We found that radial factorization is a more powerful redundancy reduction mechanism than ICA. Even with random filters it outperforms the ICA model. The highest redundancy reduction was achieved with orientation selective ICA filters in combination with radial factorization. However, when comparing the influence of the filter shape on redundancy reduction, we found that the relative contribution decreases from about 5% in models without radial factorization to less than 2% in models with radial factorization. As in the studies by Bethge [2006] and Eichhorn et al. [2009] this indicates that orientation selectivity does not play a prominent role in these models in terms of higher order redundancy reduction.

Since the L_p-norm of the filter responses is proportional to the contrast of the image patch and since rescaling the data radii in the L_p-norm involves normalizing the filter responses and rescaling them with a transformed radius, radial factorization bears similarity to divisive normalization contrast gain control [Albrecht and Hamilton, 1982, Heeger, 1992, Geisler and Albrecht, 1992]. Even though the two mechanisms are mathematically not equivalent, we found that in the range of values for natural image patches both transformations are qualitatively very similar. Previous studies already motivated a link between divisive normalization and redundancy reduction [Schwartz and Simoncelli, 2001, Wainwright et al., 2002]. However, so far, there was no reference transformation for comparison. Since radial factorization is optimal for L_p-spherically symmetric sources and since L_p-spherically symmetric distributions yield a good fit to the statistics of natural image patches, the empirical similarity motivates a link between divisive normalization and strong redundancy reduction, and offers the possibility to assess how close to optimal the divisive normalization mechanism is (see Section 2.7).

Conclusions In a statistically more adequate model for natural images, we demonstrated that orientation selective filters are at the optimum of the log-likelihood. However, like Bethge [2006] and Eichhorn et al. [2009], we also found that the difference compared to other filter shapes is small. Consistent with previous results, the optimum in the class of L_p-spherically symmetric models was not a factorial model. Using properties of the L_p-spherically symmetric class, we were able to derive an optimal redun-

dancy reduction mechanism, called radial factorization. This mechanism exhibits similarities to divisive normalization which is a prominent nonlinear mechanism throughout neural sensory systems, including primary visual cortex. We also found that in a model including the radial factorization mechanism, the relative influence of filter shapes on redundancy reduction and likelihood becomes even less.

2.3 Characterization of the p-generalized normal distribution

Motivation Radial factorization is the optimal non-linear redundancy reduction mechanism for L_p-spherically symmetric distributed sources. It was not clear, however, whether this transformation is unique. This is an important question for the comparison of physiological divisive normalization mechanisms to radial factorization. If it were not unique, the visual system might just implement another strategy which performs equally well.

Results We showed that for a fixed value of p, the p-generalized Normal distribution is the only distribution with independent marginals. This theoretical result generalized the well known theorem that the Gaussian is the only factorial spherically symmetric distribution. The original proof for this special case is ascribed to Maxwell [Kac, 1939]. The generalization of the theorem is not a straightforward extension of the spherically symmetric case and needs completely different proof techniques. Since radial factorization maps any radial distribution into the radial distribution of the p-generalized Normal distribution, this result implies that radial factorization is unique up to the output scale.

Conclusions Up to scaling, radial factorization is the unique optimal redundancy reducing mechanism for L_p-spherically symmetric distributed random variables.

2.4 Hierarchical Modeling of Local Image Features through L_p-Nested Symmetric Distributions

Motivation The key idea in the analysis of Sinz and Bethge [2009] was to enlarge the class of probabilistic models and determine the joint optimum with respect to the filter shape and the use of a nonlinear contrast gain control mechanism (see Section 2.2). The class of L_p-spherically symmetric distributions used in that study, however, does not contain the independent subspace analysis (ISA) model which was used by Hyvärinen and coworkers to derive a redundancy reduction complex cell model similar to the complex cell energy model [Adelson and Bergen, 1985, Hyvärinen and Hoyer, 2000, Hyvärinen and Koester, 2007, Pollen and Ronner, 1983]. From a natural image statistics point of view, two observations indicate that an ISA model is better suited to model the statistics of natural images than an ICA or L_p-spherically symmetric model. First, L_p-spherically symmetric models are permutation invariant in the filter responses. This

means that all image patches that arise from a permutation of the filter coefficients are equally likely under the model. This is certainly not true for natural image patches. Second, one can observe that the pairwise iso-density contour of two filter responses changes from almost spherical for adjacent filters to L_p-spherical with $p < 2$ for more distant filters. Neither ICA nor an L_p-spherically symmetric model can capture this change in contour shape. An ISA model, which can have different L_p-spherically symmetric models on each subspace, however, can reproduce this property to some extent. Therefore, it could be the right interpolation between a completely factorial model like ICA and non-factorial models like most L_p-spherically symmetric distributions.

The motivation for this study, therefore, was to carry out an analysis similar to the one for simple cells in Sinz and Bethge [2009]: using a larger class of probability models, embedding the existing model in it, and exploring the parameter space with respect to redundancy reduction and likelihood. While the larger class of probability distributions, in which existing models could be integrated, was already at hand in Sinz and Bethge [2009], we had to develop a new class of distributions that contained the complex cell models (the technical details were published in an additional study which is discussed in Section 2.5). With this new class, which we called *L_p-nested symmetric distributions*, we were able to quantitatively evaluate how well the ISA model fits to the statistics of natural images, how much a nonlinear mechanism like in the L_p-spherically symmetric case can influence redundancy reduction, and whether orientation selective filters are also at the optimum of a model that includes a nonlinear contrast gain control mechanism.

Results The main finding of this paper is that ISA models with L_p-spherically symmetric subspaces is not a good model on natural image patches. These ISA models are a special case of L_p-nested symmetric models. Within the class of L_p-nested models, distributions with independent subspaces yield a significantly lower likelihood than the model that allows for dependent subspaces. In fact, the subspaces are even more dependent than the single filter responses themselves, which is exactly opposite to the assumptions made by ISA. We demonstrated this by deriving the marginal responses over subspaces in the L_p-nested symmetric model, which we called *Dirichlet Scale Mixture*. However, when optimizing filter shapes with respect to the likelihood on natural image patches, localized orientation selective filters are again at the optimum. As in the ISA case, the filters in the maximum likelihood solution naturally split up into groups of similar spatial frequency and orientation, but different phase.

For L_p-nested symmetric distributions, one can derive a similar nonlinear redundancy reduction mechanism like radial factorization, called *nested radial factorization* (see Section 2.5). Since the likelihood of the model is again proportional to the redundancies that can be removed by this nonlinear mechanism, the better performance of L_p-nested symmetric distributions with dependent subspaces demonstrates again that linear filters cannot remove higher order redundancies well, even when the redundancy reducing requirement is relaxed to groups of filters. A nonlinear mechanism like nested radial factorization can significantly increase the redundancy reduction perfor-

mance. This again highlights the importance of a divisive normalization type of mechanism for redundancy reduction.

Conclusions The density model used by Hyvärinen and coworkers to derive complex cell properties from redundancy reduction on natural images does not capture the statistical regularities of the patches. In fact, the opposite of the assumptions made by the ISA model is true for natural images: the filters are less dependent than the subspaces. This means that ISA is not a good interpolation between non-factorial L_p-spherically symmetric models and factorial ICA models for natural images.

2.5 L_p-nested symmetric distributions

Motivation As mentioned in Section 2.4, the observed limitations of previous models suggested the developement of a generalized class of distributions that contained the independent subspace analysis model for natural images as a special case. L_p-spherically symmetric models are a special case of so called ν-spherically symmetric models [Fernández et al., 1995]. The density of ν-spherically symmetric models can be written as $\rho(\boldsymbol{y}) = \varrho(\nu(\boldsymbol{y}))$, where ν is a positively homogeneous function of degree one, which means that it has the property $\nu(a\boldsymbol{y}) = a\nu(\boldsymbol{y})$. The normalization constant of ν-spherically symmetric distributions, which is of key importance for a quantitative evaluation, depends on the surface area of the ν-unit sphere $\{\boldsymbol{y} \in \mathbb{R}^n | \nu(\boldsymbol{y}) = 1\}$. In general, an analytical expression for the surface area is infeasible. For the special case where ν is a cascade of L_p-norms we were able to compute the surface area and, therefore, the normalization constant for any ν-spherically symmetric distribution of that form. Since ν was chosen to have the form of a nested cascade of L_p-norms, we called this class L_p-nested symmetric distributions. This class is the first generalization that contains the Gaussian, L_2- as well as L_p-spherically symmetric models, and the relevant ISA model for natural images. The goal of this study was to theoretically characterize L_p-nested symmetric distributions and derive its most important properties.

Results We derived the general form of L_p-nested symmetric distributions, the uniform distribution on the L_p-nested unit sphere, and the general form of the joint distribution between variables and subspaces in an L_p-nested symmetric function. While sampling from L_p-spherically symmetric distributions is straightforward, sampling from L_p-nested symmetric distributions is not and we needed to specify an efficient and exact sampling scheme.

We derived the nested generalization of radial factorization. Importantly, we showed that every factorial L_p-nested symmetric distribution must be L_p-spherically symmetric, and, therefore, a p-generalized Normal distribution. This shows that the nested generalization of radial factorization necessarily needs the iterative nested structure and cannot be resembled by a simple one-step algorithm.

2.6 Lower bounds on the redundancy of natural images

Motivation All models discussed so far are limited to patches of natural images. While they offer advantages in terms of analytical tractability, patch based models also have several drawbacks. Two of them were addressed in this study. The first is that, in terms of redundancy reduction, we are interested in the true redundancy, i.e. the multi-information rate of natural images. Natural images clearly have dependencies beyond the distance which is usually covered by patch based models. Therefore, in order to get a more realistic estimate of the true redundancy, it is necessary to develop models that cover long range dependencies beyond the extensions of patches. The second is that receptive fields of cortical neurons are not restricted to rectangular patches but cover the whole visual field. Therefore, an important question is how to extend models for natural images and redundancy reducing representations of them to whole images.

This study investigates the estimation of the multi-information rate of natural images with conditional probability models. It explicitly makes use of the stationarity of natural image statistics which yields a tighter lower bound on the multi-information rate of natural images that can additionally be estimated with less pixels compared to a corresponding method on the joint distribution. Additionally, our empirical results suggest that, for the parametric models we used, describing natural images via conditional distributions is superior to modeling the joint distribution of larger and larger patches.

Results For one-dimensional stationary signals, it is well known that the conditional entropy converges to the entropy rate for increasing neighborhood size from above [Shannon, 1948, Cover and Thomas, 2006]. This holds for stationary random fields of arbitrary dimensions [Föllmer, 1973]. One can show that the conditional entropy of one random variable given a neighborhood of other random variables converges faster to the true entropy rate in the number of variables in the neighborhood than the joint entropy divided by the number of random variables [Cover and Thomas, 2006, p. 76]. This result can be transferred to multi-information rates, which means that the mutual information between one random variable and its neighborhood converges faster to the multi-information rate than the multi-information per random variable. Using the method relying on the mutual information between a pixel and its neighborhood, we were able to obtain conservative estimates for the multi-information rate which exceeded the estimate of the patch based method using less pixels. Our estimated multi-information rate exceeds the estimate by Petrov and Zhaoping [2003] by more than 20%, but is similar to the results by Chandler and Field [2007]. It also slightly outperforms the estimates obtained with L_p-spherically symmetric distributions.

We used the negative average log-likelihood of a (conditional) Gaussian scale mixture model [Wainwright and Simoncelli, 2000] to approximate the joint or conditional entropy needed for multi- and mutual information estimates, respectively. Since the parameters found with maximum likelihood for the joint model are not necessarily the maximum likelihood parameters for the conditional model, we carried out separate op-

timizations for both of them. When computing the multi-information rates with those two models, we noticed that the estimates with the conditional method are better and that this difference cannot be explained solely by the tighter bound that holds for the method relying on the conditional entropy. Instead, the difference was due to a better fit of the conditional model. This suggests that using conditional Gaussian scale mixtures yields a better model of natural images than increasing the dimensionality of a joint Gaussian Scale mixture. Follow-up studies by Hosseini and Theis seem to confirm this conjecture.

Conclusions We derived a generic method that yields a tighter lower bound on the multi-information rate of a stationary stochastic process on a discrete lattice and also converges faster to that bound in the number of pixels. Furthermore, our experiments suggest that conditional Gaussian scale mixture models of one pixel given an appropriate neighborhood are a better description for natural images than joint models of increasing size. In terms of redundancy reduction representations of natural images one can also use conditional models to derive redundancy reduction schemes. The idea is to transform one pixel given its neighbors into a signal with a standardized Gaussian distribution. In this way, the pixel becomes independent of its neighborhood and redundancies are removed. This is investigated in more detail in follow-up studies by Hosseini and Theis.

2.7 Temporal adaptation enhances efficient contrast gain control on natural images

Motivation Radial factorization is a very efficient nonlinear mechanism for redundancy reduction which bears similarity to divisive normalization in primary visual cortex [Sinz and Bethge, 2009, Heeger, 1992]. Previous studies have linked divisive normalization to redundancy reduction, but not measured the residual redundancies [Schwartz and Simoncelli, 2001, Wainwright et al., 2002]. Although the two transforms are generally not equivalent, they might be almost identical for the empirical distribution of natural image patches. Furthermore, radial factorization represents the optimal redundancy reduction mechanism on the Euclidean norm which allows one to evaluate the effectiveness of divisive normalization for redundancy reduction. The goal of this study was to compare divisive normalization and radial factorization and to specify conditions for which the two coincide which means that divisive normalization performs almost optimal in terms of redundancy reduction.

Results We used the assumption of L_p-spherical symmetry to robustly measure the residual amount of redundancies left in filter responses after divisive normalization. We found a substantial amount of redundancies in the model responses after divisive normalization. In order to understand this in more detail, we derived the distribution that natural images would have if divisive normalization was the optimal redundancy

reducing mechanism and called it *Naka-Rushton distribution*. After fitting that distribution to natural image patches, we still found a substantial mismatch between the model and the empirical distribution, which explained the residual redundancies after divisive normalization. The main reason for the suboptimal performance of divisive normalization is that the Naka-Rushton distribution expects most of the responses to fall into a much narrower range than responses to natural images do in reality.

We investigated two possibilities to increase the degrees of freedom in divisive normalization that would enhance its redundancy reduction capabilities. The first was to introduce more parameters in a static divisive normalization transform. This leads to a substantial increase in redundancy reduction. However, the resulting contrast response function had a physiologically implausible shape. As an alternative option we allowed for a temporal adaptation of the semi-saturation constant in divisive normalization. Such an adaptation shifts the contrast response function along the log-contrast axis. Such shifts are well known from physiology [Bonds, 1991], where it is thought to adapt the dynamic range of the response curve to the ambient contrast level. In order to choose a strategy for the adaptation of the semi-saturation constant, we used the fact that consecutively viewed patches during natural viewing conditions are correlated. We simulated eye movements on natural images and found a simple strategy to adapt divisive normalization which substantially increased the redundancy reduction performance.

Conclusions The standard model of divisive normalization does not offer enough degrees of freedom for efficient redundancy reduction on natural images. Redundancy reduction performance can be substantially increased by introducing an adaptation to the current contrast level by using temporal correlations caused by eye movements. This offers a possible functional significance of the adaptation of the contrast response curve to the ambient contrast level in terms of redundancy reduction.

The analysis of this study did not commit to a certain physiological implementation or biophysical constraints. However, it demonstrated that the redundancies imposed by the contrast statistics of natural images cannot be removed by simple static divisive normalization. From our simulations, we postulate a measure for the spread of the joint population response that gives a general signature for the performance of early vision models in terms of redundancy reduction. This signature can be used for future physiological experiments to test the suggested link between redundancy reduction and contrast gain control.

3 Discussion and Conclusion

This thesis explored the influence of orientation selectivity and contrast gain control on redundancy reduction in simple models of the early visual system. An important objective for the investigations were quantitative measurements of their effectiveness for redundancy reduction. We developed new statistical models for filter responses of whitening filters to natural image patches and used these to explore the parameter space around orientation selective filters and non-linear divisive transformations of contrast. For the quantitative evaluations we used information theoretic and probabilistic measures like multi-information and log-likelihood. We also developed new estimation methods for the multi-information rate of natural images, which led to the insight that conditional modeling of natural images might be a particularly useful approach for obaining models for entire images.

In summary, our results for the role of orientation selectivity and contrast gain control in natural image representations are

- Orientation selectivity plays a minor role for redundancy reduction. We also did not find other objectives for which orientation selectivity would yield a clear advantage (Sections 2.1, 2.2 and 2.4).
- A non-linear rescaling of the Euclidean or L_p-norm of filter responses to natural image patches is much more effective in reducing redundancy than the choice of filters. Since the aforementioned norms are proportional to the contrast of image patches, this highlights the role of contrast gain control for redundancy reduction (Sections 2.2 and 2.4).
- There is a unique and optimal contrast gain control mechanism with respect to the class of L_p-spherically and L_p-nested symmetric models. We called that mechanism (nested) radial factorization (Sections 2.2, 2.3, and 2.7).
- The relatively small difference between the amount of redundancies estimated with L_p-spherically symmetric models, which is usually competely removed by radial factorization, and the estimates of our newly developed conditional information rate estimator indicate that radial factorization is a powerful redundancy reduction mechanism which can remove a major part of higher order redundancies that can be captured with natural image models so far (Section 2.6). Since the true redundancy of natural images is not known so far, it is currently not possible to say what percentage of the total redundancy can be removed by radial factorization. With the developement of better models for (entire) natural images, the relative contribution will necessarily shrink. A promising way to obtain better models of natural images is via conditional models (Section 2.6).

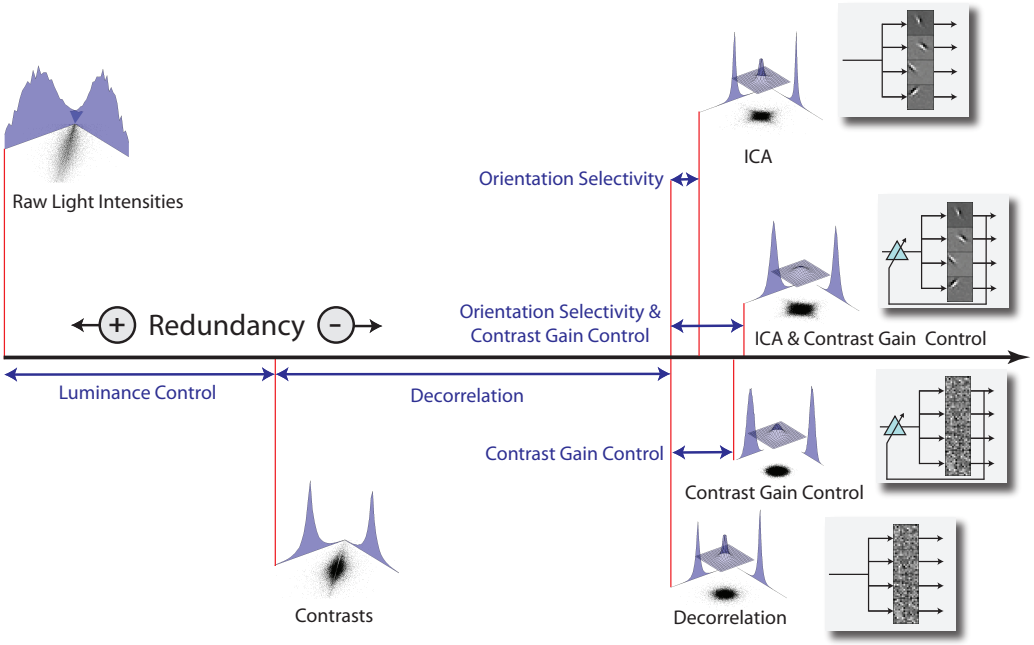


Figure 3.0.1: Relative contribution of different mechanisms in the investigated early vision models to redundancy reduction: The redundancy decreases from left to right. Setting the raw pixel representation (leftmost model) to 0% and the best performing model with orientation selectivity and contrast gain control (rightmost model) to 100%, the removal of the DC component (luminance control) and 2nd order decorrelation can already account for about 90% of the redundancy. Decorrelation together with contrast gain control accounts for 98.5%. Therefore, the maximally possible contribution of orientation selectivity is less than 5% in the former case and less than 2% in the more general cases including contrast gain control.

- Radial factorization is similar to the standard model of cortical contrast gain control, divisive normalization [Albrecht and Hamilton, 1982, Heeger, 1992]. The two mechanisms are mathematically not equivalent. In fact, when comparing them, we found that static divisive normalization leaves a significant amount of residual higher order redundancies. However, by making divisive normalization dynamically adapt to the current contrast level through using temporal correlations caused by fixations in natural viewing conditions, we demonstrated that the redundancy reduction performance can be greatly increased (Section 2.7).

In the following we discuss the details of our findings. Figure 3.0.1 gives an overview of the redundancy reduction achieved with the different mechanisms.

Orientation selectivity We found that the assumptions in the models previously used to obtain orientation selective filters from the statistics of natural image patches do not reflect their complex regularities and dependencies. In particular, the assumptions that natural images can be separated in statistically independent signals or groups of signals by linear filters, which is made by ICA and ISA models, is not well rooted in their statistical nature. The fact that ICA does not yield statistically independent filter responses on natural images had already been reported before [Simoncelli, 1997]. Bethge [2006] performed quantitative model comparison and showed that the bulk of reduced redundancies by linear filters are second order correlations. Our first study strengthened these previous findings by showing that a spherically symmetric model has a substantially better likelihood than an ICA model. Since the spherically symmetric model is agnostic with respect to the filter shapes and since natural images cannot be spherically symmetric and statistically independent at the same time, this means that orientation selectivity does not show a clear advantage in terms of redundancy reduction in linear models. Other studies, however, claimed to have found a strong influence of orientation selectivity on redundancy reduction of color image patches [Wachtler et al., 2001, Lee et al., 2002]. Even when using the same dataset as those studies, we could not reproduce their results. Another study by Lewicki and Olshausen [1999] arrives at an estimate of approximately 20% higher order redundancy reduction using over-complete sparse linear models. However, the images in this study were pre-whitentened. This initial loss of second order correlations leads to an overestimation of the relative portion of higher order redundancies. Since we do not know the amount of second order correlations that was removed by the pre-whitening step, we cannot properly compare the results of Lewicki and Olshausen to ours. We also investigated whether orientation selectivity is advantageous in terms of rate distortion curves, but found that it actually performs worse than global PCA filters.

As discussed in more detail below, augmenting the linear model by a nonlinear contrast gain control step significantly increases the redundancy reduction performance. The relative importance of filter shape, however, becomes even less [Sinz and Bethge, 2009]. Interestingly, a recent study in computer vision has also recognized the importance of nonlinear steps and the relative unimportance of filter shapes for object recognition [Jarrett et al., 2009].

We also investigated whether redundancy reduction can yield an explanation for orientation selectivity in complex cell models in which several filters are grouped and only the groups are required to be independent. Here the statistics of natural images and the assumptions by the model were even contradictory: groups of filters were even more dependent than the individual filter responses within a group [Sinz et al., 2009b]. This is exactly opposite to the assumption of the ISA model used for a redundancy reduction explanation of complex cell properties [Hyvärinen and Hoyer, 2000, Hyvärinen and Koester, 2007].

A striking finding, however, is that even though orientation selectivity does not have a great influence on redundancy reduction, it is the optimal filter shape for linear models with or without contrast gain control [Sinz and Bethge, 2009, Sinz et al., 2009b]. Therefore, it remains an open question how important orientation selectivity is for re-

dundancy reduction. There are several ways how this could be resolved in the future. It is possible that more realistic models of the early visual system yield a strong vote in favor or against orientation selectivity, although we do not expect that unless these models exhibit genuinely different nonlinear mechanisms. It might also be that there are other objectives of the visual system which clearly favor a localized orientation selective filter shape. Finally, the fact that localized oriented band-pass filters arise in so many models might also hint at a more fundamental statistical reason for the shape of those filters. If this is indeed the case, understanding this fundamental principle will bring us closer to understanding the computational principles governing receptive fields in primary visual cortex.

Contrast gain control Based on the good performance of a spherically symmetric model in terms of likelihood we started to explore the class of L_p-spherically symmetric models for modeling the response of whitening filters to natural image patches. L_p-spherically symmetric models contain the spherically symmetric as well as the ICA case for natural image patches. However, the optimum in terms of likelihood was located at neither of them: Optimal L_p-spherically symmetric models on orientation selective filters were non-factorial and exhibited a p of approximately 1.3. When optimizing for the filters as well, orientation selective filters also arose as the optimal shape for this class of distributions. However, the difference compared to other filter shapes in terms of redundancy reduction was again marginal.

By using fundamental statistical properties of this class of distributions, we derived an optimal mechanism for redundancy reduction based on a non-linear rescaling of the L_p-norm, called radial factorization. Since L_p-norms are proportional to the root mean square contrast, this mechanism is a contrast gain control transformation. We demonstrated that radial factorization significantly outperforms the redundancy reduction performance of linear filters. We showed that under the assumption of L_p-spherical symmetry, radial factorization is the unique and optimal mechanism for redundancy reduction.

By introducing L_p-nested symmetric distributions, we were able to find an even more general class of distributions that included all the models before, but also the relevant ISA models. This allowed us to demonstrate that in this larger class there is again a unique contrast gain control mechanism which is again of crucial importance for redundancy reduction in complex cell ISA models, and that even in this larger class of L_p-nested models orientation selectivity is again at the optimum in terms of likelihood and, therefore, redundancy reduction.

Radial factorization and nested radial factorization are the optimal redundancy reduction mechanisms with respect to L_p-spherically and L_p-nested symmetric distributed sources. One obvious question is, of course, how close to optimal their performance is for natural images in general. In a separate study, we developed a new estimation method for the multi-information rate of natural images. Comparing the estimates from that method to estimates with L_p-spherically symmetric distributions, we find that there is but a small difference, which indicates that radial factorization is a power-

ful redundancy reduction mechanism even in a context which is not restricted to image patches. However, the conditional estimation method developed by us [Hosseini et al., 2010] is generic in the sense that better conditional models of natural images will improve the lower bound on the multi-information rate. When that happens, the relative performance of radial factorization will decrease. But from our current measurements, the redundancy removed by radial factorization represents a substantial part of the known higher order dependencies in natural images.

There are previous studies that have pursued a quantitative assessment redundancy in natural images. Using non-parametric statistical methods several studies have measured the amount of higher order dependencies in natural images [Schreiber, 1956, Li and Atick, 1994, Petrov and Zhaoping, 2003, Chandler and Field, 2007]. Our estimates are larger than those of Petrov and Zhaoping [2003] and comparable to Chandler and Field [2007]. The likely reason why Petrov and Zhaoping [2003] underestimated the amount of higher order redundancies is that they restricted themselves to a very small neighborhood which lead to an overestimation of the entropy and, therefore, an underestimation of the multi-information.

Radial factorization and divisive normalization—the standard model of cortical contrast gain control—share many properties. Since previous studies have highlighted the possible role for divisive normalization for redundancy reduction on natural signals [Schwartz and Simoncelli, 2001, Wainwright et al., 2002], we investigated the interrelation between radial factorization and divisive normalization. We found that filter responses transformed with divisive normalization still exhibit a substantial amount of higher order redundancies, and that the contrast distribution for which divisive normalization equals radial factorization does not fit the contrast distribution of natural images well. Nevertheless, our results are not at odds with previous work on divisive normalization and redundancy reduction. First of all, we do find a reducing effect of divisive normalization. Second, previous studies either looked at surrogate measures of statistical independence [Schwartz and Simoncelli, 2001, Wainwright et al., 2002], considered pairwise measures only [Malo et al., 2006], or were based on theoretical analyses of probability distribution and not real data [Lyu, 2011].

The studies by Schwartz, Wainwright and Simoncelli used so called bow-tie plots to visualize redundancy or the reduction thereof. After divisive normalization, the bow-tie plots become flat indicating that this type of dependency has been removed. In principle, bow-tie plots should capture the right dependencies: For the class of Gaussian scale mixtures around which the work of Schwartz, Wainwright and Simoncelli is built, one can prove that the conditional variance is non-decreasing [Cambanis et al., 2000]. For the relevant class of models the conditional variance is even increasing which means that higher order dependencies should show up in the bow-tie plots. The problem is, however, that bow-tie plots are not very sensitive. Both radial factorization and divisive normalization produce a flat bow-tie plot, but significantly differ in the amount of residual redundancies.

The study by Lyu is based on the multivariate t-model [Kotz and Nadarajah, 2004] for which he shows mathematically that divisive normalization is an approximate radial factorization transform. Not surprisingly, the multivariate t-model and the distribution

for which divisive normalization and radial factorization coincide are very similar. This means that the multivariate t-model, like the corresponding distribution for divisive normalization (see Section 2.7), also exhibits an inferior fit to the statistics of natural images.

Does this mean that cortical divisive contrast gain control and redundancy reduction disagree? Not necessarily. All our previous investigations were based on models of static image patches, randomly sampled from a large collection of images, but this is not the way in which visual information enters the visual system under free viewing conditions. In particular, contrasts during the fixation between two saccades are very correlated. It is also known that the contrast response curve of simple and complex cells shifts along the log-contrast axis to adapt to the current ambient contrast level [Ohzawa et al., 1982, Bonds, 1991]. Using this fact, we demonstrated that the disagreement of redundancy reduction with cortical models of contrast gain control can be resolved by allowing divisive normalization to adapt to the ambient contrast between two saccades. This substantially decreased the amount of residual redundancies to almost the level of radial factorization and suggests a potential role of contrast response curve adaptation for redundancy reduction. Our mechanism works on the short time scale between two saccades. The adaptation mechanism via shifts of the contrast response curve is usually thought to happen on larger time scales [Bonds, 1991]. However, those studies have been perfomed in anasthetized cats with drifting gratings. Hence, it is not very clear how representative these results are for natural viewing conditions. Additionally, a recent study claims that this might have been an artifact from the stimulation protocol and that shifts can actually occur on a much shorter time scale [Hu and Wang, 2011].

Contributions to the understanding of natural image statistics While the statistical models in this thesis were important tools to understand the contribution of different mechanisms to redundancy reduction, they also contributed to the understanding of the statistics of natural image patches by developing state-of-the-art density models for them (see Table 3.1). Popular models for natural image patches are, for instance, ICA, Gaussian scale mixtures (GSM) [Wainwright and Simoncelli, 2000], and ISA [Hyvärinen and Hoyer, 2000, Hyvärinen and Koester, 2007]. The models used and developed in this thesis generalize all of them: the class of L_p-spherically symmetric models contains ICA for natural images and Gaussian scale mixtures; the class of L_p-nested symmetric distributions contains the L_p-spherically symmetric ones and ISA on natural images. In general, ICA and ISA are not fully contained in the L_p-spherically symmetric or L_p-nested symmetric class, but due to the special form of the marginal distributions of filter responses to natural image patches, the relevant cases are part of the L_p-spherical and L_p-nested classes.

While it is apparent that ICA is not a very good model for natural image patches [Simoncelli, 1997, Bethge, 2006, Eichhorn et al., 2009], the GSMs are among the state-of-the-art models. They model natural image patches as a mixture of—possibly infinitely many—Gaussians with the same mean but different scales. Since the underlying distribution is spherically symmetric, the class of GSMs belongs to the spherically symmet-

Table 3.1: Performance of different natural image models in terms of log-likelihood compared to a factorial model on raw pixels (courtesy of all authors involved and Lucas Theis who produced the table). L_p-spherically and L_p-nested symmetric models rank among the state-of-the-art models. Hierarchical deep belief networks (DBN) perform even worse than ICA. The currently best performing model is a mixture of Gaussian scale mixtures.

	bandpass filtering	orientation selectivity	divisive normalization	complex cell pooling	hierarchical	references	Δ log-likelihood [bits/pixel]
PCA / Whitening	✓					Bethge [2006] Eichhorn et al. [2009]	2.7
ICA	✓	✓				Bethge [2006] Eichhorn et al. [2009]	2.92
L_2-spherical model	✓		✓			Eichhorn et al. [2009] Sinz and Bethge [2009]	3.05
L_p-spherical model	✓	✓	✓			Sinz et al. [2009b]	3.17
L_p-nested model	✓	✓	✓	✓	(✓)	Sinz et al. [2009b] Sinz and Bethge [2010]	3.2
Hierarchical ICA	✓	✓			✓	Hosseini and Bethge [2009]	3.0
Deep Belief Networks	✓	✓			✓	Theis et al. [2010]	2.9
Mixture of GSMs	✓	✓	✓	✓		Bethge and Hosseini [2008]	3.3

ric class of distributions. They are, however, only a subset since not every spherically symmetric distribution is also a GSM. GSMs perform well at modeling the variance correlations of natural images captured by the bow-tie plots mentioned before. Since spherically symmetric models are invariant under orthogonal transformations of the input, GSMs are agnostic to the particular shape of the whitening filters. Our investigations with the L_p-spherically symmetric models demonstrate, however, that the filter shape matters and that the distribution of natural image patches is not exactly spherically symmetric but rather $L_{1.3}$-spherically symmetric [Sinz and Bethge, 2009]. L_p-spherically symmetric models capture variance correlations as well (see Section 4.10) and allow a straightforward evaluation of the log-likelihood. Table 3.1 lists the improvement of the log-likelihood over a factorial model on raw pixels for several natural image models. L_p-spherically symmetric models rank third.

It is not only the case that whitening filter responses to natural image patches deviate from spherical symmetry, but also that the contour lines of the distribution can vary depending on different features of the filters involved [Sinz et al., 2009b]. L_p-nested symmetric distributions solve that problem by composing the contour shapes from different L_p-spheres while maintaining a straightforward evaluation of the log-likelihood [Sinz and Bethge, 2010]. Among the models in Table 3.1, L_p-nested symmetric distributions rank second.

The only class of models that currently outperforms L_p-spherical and L_p-nested models on natural images are mixtures of GSMs [Bethge and Hosseini, 2008] which belongs to the class of Gaussian mixture models. These models allow for modeling different covariance structures of natural image patches, while the L_p-spherical and L_p-nested models always assume the same covariance structure in the whitening step. Gaussian mixture models for natural image statistics with a rich hidden structure have also been proposed for natural image modeling by other authors [Karklin and Lewicki, 2008, Ranzato and Hinton, 2010, Ranzato et al., 2010]. However, these studies did not evaluate the likelihood of their models.

An important insight that arose from the quantitative comparisons between statistical models of natural images is that hierarchical deep belief networks (DBNs) do not rank among state-of-the-art models. In a separate study, which is not included in this thesis, we developed an estimator for the likelihood in DBNs and used it to evaluate the likelihood of DBNs on image patches [Theis et al., 2010, 2011]. Osindero and Hinton [2008] reported promising results from training deep belief networks (DBN) on natural images by presenting random samples from the model that looked very encouraging. However, one great challenge in natural image patch modeling is to find a model that achieves a high likelihood on unseen patches and produces realistically looking samples. While each single goal is relatively easy to accomplish, there is currently no model that achieves both. Judging from the random samples, DBNs seemed to be a promising density model. Unfortunately, likelihood evaluation in DBNs is very hard. The results obtained from our likelihood estimator demonstrated that DBNs perform significantly worse than L_p-spherical or L_p-nested models. In fact, DBNs even perform worse than ICA which was already established to be a suboptimal model on natural image patches (see Table 3.1). These results emphasize the importance of quantitative model compar-

ison via the log-likelihood.

How important is redundancy reduction for the early visual system? Finally, the question remains how important the redundancy reduction hypothesis is and how it can take us further in understanding the early visual system. Many issues concerning redundancy reduction and sensory systems have been discussed in excellent papers elsewhere [Barlow, 2001, Simoncelli, 2003, Simoncelli and Olshausen, 2003]. Here, we want to highlight a few points relevant to the current work. A frequently raised argument against redundancy reduction in cortex is the relative overcompleteness of the number of cortical neurons compared to the number of retinal ganglion cells. This, at first sight, indicates that the redundancy in cortex should increase instead of decrease. However, as noted by Barlow [2001] and Simoncelli [2003], this assumes that the coding capacity (entropy) of cortical and retinal neurons is the same. Since cortical neurons do have lower firing rates it is well possible that the coding capacity (entropy) per neuron is smaller which renders their relative overcompleteness and redundancy reduction compatible.

Another objection might be that the brain only cares about behaviorally relevant information which does not necessarily have to coincide with the information measured in bits. Unfortunately, it is difficult to clearly state what "behaviorally relevant" exactly means. On the other hand, it might be that an object can be defined as a collection of incoming signals that exhibit very specific correlations in space and time. As already mentioned in the introduction, redundancy reduction could extract exactly those correlations. This means that it could serve as a feature extractor for important aspects about the ouside world which could later be classified as behaviorally relevant or not. Recent psychophysical studies indicate that human observers are indeed sensitive to higher order statistical dependencies already in natural image patches of 3×3 pixels [Gerhard et al., 2012, in preparation]. Even though it is clearly not obvious to talk about behavioral relevance at this scale, it indicates that the visual system cares about the statistical regularities even in small natural image patches. In this study, subjects were asked to discriminate between patches from natural images and samples that incorporated key assumptions of statistical models for natural image patches. The likelihood of the statistical models predicted the rank order of the discrimination performance of subjects for the different models. This indicates that the statistical regularities captured by the models are sufficiently relevant to the visual system.

Another issue is whether the visual system should really aim at removing *all* redundancies from natural images or whether it should just decrease them enough to achieve an efficient information transmission but still be robust against internal noise. In a system with internal uncertainty, a small amount of redundancy can be helpful to counteract information loss due to noise. If the uncertainty is independent additive noise, efficient coding and redundancy reduction coincide. However, even if that is not the case, there are reasons to believe that strong redundancy reduction is a goal worth pursuing: First, it seems reasonable to assume that the influence of internal noise should be small compared to the uncertainty about the stimulus. In information theoretic terms,

this means that the mutual information $I[\boldsymbol{X};\boldsymbol{Y}] = \langle H[\boldsymbol{Y}] - H[\boldsymbol{Y}|\boldsymbol{x}]\rangle_{\boldsymbol{X}}$ should be large and not be dominated by the structure of the uncertainty $H[\boldsymbol{Y}|\boldsymbol{x}]$ of the neural response given a fixed stimulus, i.e. $H[\boldsymbol{Y}] \gg H[\boldsymbol{Y}|\boldsymbol{X}]$. While it might be beneficial to maintain a small amount of redundancy to account for noise, it still seems reasonable to decrease the redundancy as much as possible in order to hit the maximum of $H[\boldsymbol{Y}]$. Secondly, it is clear that the higher order dependencies in natural images, which—so far—seem to make up for the minor part of the total dependencies [Chandler and Field, 2007, Hosseini et al., 2010], are the perceptually relevant part. While the content of a whitened image with no second order correlations left can still be perceived, this is not the case for a phase scrambled image for which all higher order correlations have been destroyed. The visual system must have a way to extract those higher order dependencies, which in a redundancy reduction framework means to separate them into statistically independent signals.

Conclusion The studies in this thesis developed new statistical models, statistical theory, and information theoretic estimation methods in order to explore the statistics of natural image patches around the question of how important orientation selective filters and contrast gain control mechanisms are for a factorial representation. One major objective behind all work presented was to base our findings on a firm quantitative ground. Our statistical models are among the state-of-the-art for natural image patches in terms of likelihood and built the basis for state-of-the-art models on whole natural images.

Our patch based models allowed us to disentangle the contributions of filter shape and contrast gain control to density modeling and redundancy reduction on natural images. We found that orientation selectivity does not play a crucial role for redundancy reduction in the current standard model of visual neurons, while contrast gain control does. The model classes also provided a framework in which we could demonstrate that a dynamical component in the physiological standard model for contrast gain control is crucial to achieve strong redundancy reduction on natural images.

Bibliography

E. H. Adelson and J. R. Bergen. Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A*, 2(2):284–299, February 1985. doi: 10.1364/JOSAA.2.000284.

D. G. Albrecht and D. B. Hamilton. Striate cortex of monkey and cat: contrast response function. *Journal of Neurophysiology*, 48(1):217–237, 1982.

J. J. Atick. Could information theory provide an ecological theory of sensory processing? *Network: Computation in Neural Systems*, 3(2):213–251, 1992. ISSN 0954-898X. doi: 10.1088/0954-898X/3/2/009.

J. J. Atick and A. N. Redlich. Towards a Theory of Early Visual Processing. *Neural Computation*, 2(3):308–320, 1990. doi: 10.1162/neco.1990.2.3.308.

J. J. Atick and A. N. Redlich. What Does the Retina Know about Natural Scenes? *Neural Computation*, 4(2):196–210, March 1992. doi: 10.1162/neco.1992.4.2.196.

F. Attneave. Informational aspects of visual perception. *Psychological Review*, 61:183–193, 1954.

R. Baddeley, L. F. Abbott, M. C. Booth, F. Sengpiel, T. Freeman, E. A. Wakeman, and E. T. Rolls. Responses of neurons in primary and inferior temporal visual cortices to natural scenes. *Proceedings of the Royal Society B Biological Sciences*, 264(1389):1775–1783, 1997.

H. B. Barlow. Possible Principles Underlying the Transformations of Sensory Messages. In W. A. Rosenblith, editor, *Sensory Communication*, pages 217–234. MIT Press, Cambridge, MA., 1961.

H. B. Barlow. Cerebral cortex as a model builder. In D Rose and V G Dobson, editors, *Models of the Visual Cortex*, pages 37–46. John Wiley & Sons Ltd, 1985.

H. B. Barlow. Unsupervised Learning. *Neural Computation*, 1(3):295–311, 1989. doi: 10.1162/neco.1989.1.3.295.

H. B. Barlow. Redundancy reduction revisited. *Network: Computation in Neural Systems*, 12(3):241–253, 2001.

H. B. Barlow. The exploitation of regularities in the environment by the brain. *Behavioral and Brain Sciences*, 24(04):602–607, 2002.

A. J. Bell and T. J. Sejnowski. The "independent components" of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, December 1997. ISSN 0042-6989. doi: 10.1016/S0042-6989(97)00121-1.

J. M. Bernardo. Expected information as expected utility. *Annals of Statistics*, 7(3):686–690, 1979. ISSN 00905364. doi: 10.1214/aos/1176344689.

M. Bethge. Factorial coding of natural images: how effective are linear models in removing higher-order dependencies? *Journal of the Optical Society of America A*, 23(6): 1253–1268, June 2006. doi: 10.1364/JOSAA.23.001253.

M. Bethge and R. Hosseini. Method and Device for Image Compression, 2008. URL `https://register.epo.org/espacenet/application?number=EP08010343`.

A. B. Bonds. Role of Inhibition in the Specification of Orientation Selectivity of Cells in the Cat Striate Cortex. *Visual Neuroscience*, 2(01):41–55, 1989. doi: 10.1017/S0952523800004314.

A. B. Bonds. Temporal dynamics of contrast gain in single cells of the cat striate cortex. *Vis Neurosci*, 6(3):239–255, 1991. ISSN 09525238.

G. Buchsbaum and A. Gottschalk. Trichromacy, opponent colours coding and optimum colour information transmission in the retina. *Proceedings of the Royal Society of London. Series B, Containing papers of a Biological character. Royal Society (Great Britain)*, 220 (1218):89–113, November 1983. ISSN 0080-4649.

S. Cambanis, S.P. Fotopoulos, and L. He. On the Conditional Variance for Scale Mixtures of Normal Distributions. *Journal of Multivariate Analysis*, 74(2):163–192, August 2000. ISSN 0047259X. doi: 10.1006/jmva.1999.1888.

M. Carandini, D. J. Heeger, and J. A. Movshon. Linearity and Normalization in Simple Cells of the Macaque Primary Visual Cortex. *J. Neurosci.*, 17(21):8621–8644, November 1997.

D. M. Chandler and D. J. Field. Estimates of the information content and dimensionality of natural scenes from proximity distributions. *Journal of the Optical Society of America A*, 24(4):922–941, April 2007. doi: 10.1364/JOSAA.24.000922.

G. Chechik, A. Globerson, M. J. Anderson, E. D. Young, I. Nelken, and N Tishby. Group Redundancy Measures Reveal Redundancy Reduction in the Auditory Pathway. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, volume 14, pages 173–180. MIT Press, 2002.

E. J. Chichilnisky. A simple white noise analysis of neuronal light responses. *Network: Comput. Neural Syst*, 12:199–213, 2001.

P. Comon. Independent component analysis, A new concept? *Signal Processing*, 36(3): 287–314, 1994. ISSN 01651684. doi: 10.1016/0165-1684(94)90029-9.

T. M. Cover and J. A. Thomas. *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006. ISBN 0471241954.

Y. Dan, J. J. Atick, and R. C. Reid. Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory. *Journal of Neuroscience*, 16 (10):3351–3362, 1996.

D. W. Dong and J. J. Atick. Statistics of natural time-varying images. *Network: Computation in Neural Systems*, 6:345–358, 1995. doi: 10.1.1.39.1878.

J. Eichhorn, F. Sinz, and M. Bethge. Natural Image Coding in V1: How Much Use Is Orientation Selectivity? *PLoS Comput Biol*, 5(4), April 2009.

C. Fernández, J. Osiewalski, and M. F. J. Steel. Modeling and Inference with v-Spherical Distributions. *Journal of the American Statistical Association*, 90(432):1331–1340, 1995.

D. J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, 4(12):2379–2394, December 1987. doi: 10.1364/JOSAA.4.002379.

H. Föllmer. On entropy and information gain in random fields. *Probability Theory and Related Fields*, 26(3):207–217, 1973. doi: 10.1007/BF00532723.

J. H. Friedman, W. Stuetzle, and A. Schroeder. Projection Pursuit Density Estimation. *Journal of the American Statistical Association*, 79(387):599–608, September 1984. ISSN 01621459.

W. S. Geisler and D. G. Albrecht. Cortical neurons: Isolation of contrast gain control. *Vision Research*, 32(8):1409–1410, August 1992. ISSN 0042-6989. doi: 10.1016/0042-6989(92)90196-P.

H. Gerhard, F. A. Wichmann, and M. Bethge. How sensitive is the visual system to the local statistics of natural images?, 2012.

I. R. Goodman and S. Kotz. Multivariate [theta]-generalized normal distributions. *Journal of Multivariate Analysis*, 3(2):204–219, June 1973. ISSN 0047-259X. doi: 10.1016/0047-259X(73)90023-7.

A. K. Gupta and D. Song. Lp-norm spherical distribution. *Journal of Statistical Planning and Inference*, 60(2):241–260, May 1997. ISSN 03783758.

P. J. B. Hancock, R. Baddeley, and L. S. Smith. The principal components of natural images. *Network Computation in Neural Systems*, 3(1):61–70, 1992. ISSN 0954898X. doi: 10.1088/0954-898X/3/1/008.

D. J. Heeger. Normalization of cell responses in cat striate cortex. *Vis Neurosci*, 9(2): 181–197, 1992. ISSN 09525238.

R. Hosseini and M. Bethge. Hierachical Models of Natural Images, 2009.

R. Hosseini, F. Sinz, and M. Bethge. Lower bounds on the redundancy of natural images. *Vision Research*, 50(22):2213–2222, October 2010. ISSN 0042-6989. doi: 10.1016/j.visres.2010.07.025.

M. Hu and Y. Wang. Rapid Dynamics of Contrast Responses in the Cat Primary Visual Cortex. *PLoS ONE*, 6(10):e25410, October 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0025410.

A. Hyvärinen and P. Hoyer. Emergence of Phase- and Shift-Invariant Features by Decomposition of Natural Images into Independent Feature Subspaces. *Neural Computation*, 12(7):1705–1720, July 2000.

A. Hyvärinen and U. Koester. Complex cell pooling and the statistics of natural images. *Network: Computation in Neural Systems*, 18(2):81–100, 2007.

K. Jarrett, K. Kavukcuoglu, M. A. Ranzato, and Y. LeCun. What is the best multi-stage architecture for object recognition? *2009 IEEE 12th International Conference on Computer Vision*, pages 2146–2153, 2009. doi: 10.1109/ICCV.2009.5459469.

M. Kac. On a Characterization of the Normal Distribution. *American Journal of Mathematics*, 61(3):726–728, July 1939. ISSN 00029327. doi: 10.2307/2371328.

Y. Karklin and M. S. Lewicki. Emergence of complex cell properties by learning to generalize in natural scenes. *Nature*, November 2008. ISSN 0028-0836. doi: 10.1038/nature07481.

S. Kotz and S. Nadarajah. *Multivariate T-Distributions and Their Applications*. Cambridge University Press, Cambridge, 1 edition, 2004. ISBN 9780511550683. doi: 10.1017/CBO9780511550683.

S. Laughlin. A simple coding procedure enhances a neuron's information capacity. *Z Naturforschung*, 36(9-10):910–912, 1981. ISSN 03410382.

T.-W. Lee, T. Wachtler, and T. J. Sejnowski. Color opponency is an efficient representation of spectral properties in natural scenes. *Vision Research*, 42(17):2095–2103, August 2002. ISSN 0042-6989. doi: 10.1016/S0042-6989(02)00122-0.

M. S. Lewicki and B. A. Olshausen. Probabilistic framework for the adaptation and comparison of image codes. *Journal of the Optical Society of America A*, 16:1587–1601, July 1999.

M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12(2):337–365, 2000.

Z. Li and J. J. Atick. Toward a Theory of the Striate Cortex. *Neural Computation*, 6(1): 127–146, January 1994. doi: 10.1162/neco.1994.6.1.127.

R Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988. ISSN 00189162. doi: 10.1109/2.36.

S. Lyu. Dependency Reduction with Divisive Normalization: Justification and Effectiveness. *Neural Computation*, 23(11):2942–2973, August 2011. ISSN 0899-7667. doi: 10.1162/NECO_a_00197.

J. Malo, I. Epifanio, R. Navarro, and E. P. Simoncelli. Nonlinear image representation for efficient perceptual coding. *IEEE Transactions on Image Processing*, 15(1):68–80, 2006.

D. Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman, 1983. ISBN 0716715678.

J.-P. Nadal and N. Parga. Nonlinear neurons in the low-noise limit: a factorial code maximizes information transfer. *Network: Computation in Neural Systems*, 5(4):565–581, 1994. ISSN 0954-898X. doi: 10.1088/0954-898X/5/4/008.

I. Ohzawa, G. Sclar, and R. D. Freeman. Contrast gain control in the cat visual cortex. *Nature*, 298(5871):266–268, July 1982. doi: 10.1038/298266a0.

B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, June 1996. doi: 10.1038/381607a0.

S. Osindero and G. E. Hinton. Modeling image patches with a directed hierarchy of Markov random fields. In J C Platt, D Koller, Y Singer, and S Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1121–1128. MIT Press, Cambridge, MA, 2008.

A. Perez. ϵ-admissible simplification of the dependence structure of a set of random variables. *Kybernetika*, 13:439–444, 1977.

Y. Petrov and L. Zhaoping. Local correlations, information redundancy, and sufficient pixel depth in natural images. *Journal of the Optical Society of America A*, 20(1):56–66, January 2003. doi: 10.1364/JOSAA.20.000056.

D. A. Pollen and S. F. Ronner. Visual cortical neurons as localized spatial frequency filters. *IEEE Trans on Systems Man and Cybernetics*, 13(5):907–916, 1983.

J. L. Puchalla, E. Schneidman, R. A. Harris, and M. J. Berry. Redundancy in the population code of the retina. *Neuron*, 46(3):493–504, 2005.

M. A. Ranzato and G. E. Hinton. Modeling Pixel Means and Covariances Using Factorized Third-Order Boltzmann Machines. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 9, 2010.

M. A. Ranzato, A. Krizhevsky, and G. E. Hinton. Factored 3-Way Restricted Boltzmann Machines For Modeling Natural Images. In *Proc. Thirteenth International Conference on Artificial Intelligence and Statistics.*, volume 9, pages 621–628. JMLR: W&CP, 2010.

F. Rieke, D. A. Bodnar, and W. Bialek. Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. *Proceedings of the Royal Society B: Biological Sciences*, 262(1365):259–265, 1995. ISSN 09628452. doi: 10.1098/rspb.1995.0204.

D. L. Ruderman, T. W. Cronin, and C.-C. Chiao. Statistics of cone responses to natural images: implications for visual coding. *Journal of the Optical Society of America A*, 15 (8):2036–2045, 1998. ISSN 10847529. doi: 10.1364/JOSAA.15.002036.

T. D. Sanger. Optimal unsupervised learning in a single-layer linear feedforward network. *Neural Networks*, 2:459–473, 1989.

W. Schreiber. The measurement of third order probability distributions of television signals. *Information Theory, IRE Transactions on*, 2(3):94–105, 1956. ISSN 0096-1000. doi: 10.1109/TIT.1956.1056811.

O. Schwartz and E. P. Simoncelli. Natural signal statistics and sensory gain control. *Nat Neurosci*, 4(8):819–825, 2001. ISSN 1097-6256. doi: 10.1038/90526.

C. E. Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27(1):379–423, 1948. ISSN 07246811. doi: 10.1145/584091.584093.

H Shouval, N Intrator, and L N Cooper. BCM network develops orientation selectivity and ocular dominance in natural scene environment. *Vision Research*, 37(23):3339–3342, 1997.

E. P. Simoncelli. Statistical models for images: compression, restoration and synthesis. In *Conference Record of the Thirty First Asilomar Conference on Signals Systems and Computers Cat No97CB36136*, volume 1, pages 673–678. IEEE Comput. Soc, 1997. ISBN 0818683163. doi: 10.1109/ACSSC.1997.680530.

E. P. Simoncelli. Vision and the statistics of the visual environment. *Current Opinion in Neurobiology*, 13(2):144–149, April 2003. doi: 10.1016/S0959-4388(03)00047-3.

E. P. Simoncelli and B. A. Olshausen. Natural Image Statistics and Neural Representation. *Annual Review of Neuroscience*, 24:1193–1216, November 2003. doi: 10.1146/annurev.neuro.24.1.1193.

F. Sinz and M. Bethge. The Conjoint Effect of Divisive Normalization and Orientation Selectivity on Redundancy Reduction. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in neural information processing systems 21 : 22nd Annual Conference on Neural Information Processing Systems 2008*, pages 1521–1528, Red Hook, NY, USA, 2009. Curran.

F. Sinz and M. Bethge. Lp -Nested Symmetric Distributions. *Journal of Machine Learning Research*, 11:3409–3451, 2010.

F. Sinz, S. Gerwinn, and M. Bethge. Characterization of the p-generalized normal distribution. *Journal of Multivariate Analysis*, 100:817–820, 2009a.

F. Sinz, E. P. Simoncelli, and M. Bethge. Hierarchical Modeling of Local Image Features through Lp-Nested Symmetric Distributions. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009*, pages 1696–1704, Red Hook, NY, USA, 2009b. Curran.

D. Song and A. K. Gupta. Lp-Norm Uniform Distribution. *Proceedings of the American Mathematical Society*, 125(2):595–601, 1997.

L. Theis, S. Gerwinn, F. Sinz, and M. Bethge. Likelihood Estimation in Deep Belief Networks, 2010.

L. Theis, S. Gerwinn, F. Sinz, and M. Bethge. In All Likelihood, Deep Belief Is Not Enough. *Journal of Machine Learning Research*, 12:3071–3096, November 2011.

J. H. van Hateren. Real and optimal neural images in early vision. *Nature*, 360(6399): 68–70, November 1992. doi: 10.1038/360068a0.

J. H. Van Hateren. Spatiotemporal contrast sensitivity of early vision. *Vision Research*, 33(2):257–267, 1993.

J. H. Van Hateren and A. Van Der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society B Biological Sciences*, 265(1394):359–366, 1998.

W. E. Vinje and J. L. Gallant. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287(5456):1273–1276, 2000.

T. Wachtler, T.-W. Lee, and T. J. Sejnowski. Chromatic structure of natural scenes. *Journal of the Optical Society of America A*, 18(1):65–77, January 2001. doi: 10.1364/JOSAA.18.000065.

M. J. Wainwright and E. P. Simoncelli. Scale mixtures of Gaussians and the statistics of natural images. *Neural Information Processing Systems*, 12(1):855–861, 2000.

M. J. Wainwright, O. Schwartz, and E. P. Simoncelli. Natural image statistics and divisive normalization: modeling nonlinearities and adaptation in cortical neurons. In *Statistical theories of the brain*, pages 203–222. MIT Press, 2002.

C. Zetzsche, G. Krieger, and B. Wegmann. The atoms of vision: Cartesian or polar? *Journal of the Optical Society of America A*, 16(7):1554–1565, July 1999. doi: 10.1364/JOSAA.16.001554.

4 Appendix

4.1 Natural Image Coding in V1: How Much Use Is Orientation Selectivity: Original Article

OPEN ACCESS Freely available online PLoS COMPUTATIONAL BIOLOGY

Natural Image Coding in V1: How Much Use Is Orientation Selectivity?

Jan Eichhorn[•], **Fabian Sinz**[•], **Matthias Bethge***

Max Planck Institute for Biological Cybernetics, Tübingen, Germany

Abstract

Orientation selectivity is the most striking feature of simple cell coding in V1 that has been shown to emerge from the reduction of higher-order correlations in natural images in a large variety of statistical image models. The most parsimonious one among these models is linear Independent Component Analysis (ICA), whereas second-order decorrelation transformations such as Principal Component Analysis (PCA) do not yield oriented filters. Because of this finding, it has been suggested that the emergence of orientation selectivity may be explained by higher-order redundancy reduction. To assess the tenability of this hypothesis, it is an important empirical question how much more redundancy can be removed with ICA in comparison to PCA or other second-order decorrelation methods. Although some previous studies have concluded that the amount of higher-order correlation in natural images is generally insignificant, other studies reported an extra gain for ICA of more than 100%. A consistent conclusion about the role of higher-order correlations in natural images can be reached only by the development of reliable quantitative evaluation methods. Here, we present a very careful and comprehensive analysis using three evaluation criteria related to redundancy reduction: In addition to the multi-information and the average log-loss, we compute complete rate–distortion curves for ICA in comparison with PCA. Without exception, we find that the advantage of the ICA filters is small. At the same time, we show that a simple spherically symmetric distribution with only two parameters can fit the data significantly better than the probabilistic model underlying ICA. This finding suggests that, although the amount of higher-order correlation in natural images can in fact be significant, the feature of orientation selectivity does not yield a large contribution to redundancy reduction within the linear filter bank models of V1 simple cells.

Citation: Eichhorn J, Sinz F, Bethge M (2009) Natural Image Coding in V1: How Much Use Is Orientation Selectivity?. PLoS Comput Biol 5(4): e1000336. doi:10.1371/journal.pcbi.1000336

Editor: Li Zhaoping, University College London, United Kingdom

Received April 21, 2008; **Accepted** February 18, 2009; **Published** April 3, 2009

Funding: This study was financially supported by the German Ministry of Education, Science, Research and Technology through the Bernstein award (BMBF; FKZ: 01GQ0601) and a scholarship by the German National Academic Foundation. The design and conduct of the study was independent of the ministry.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: mbethge@tuebingen.mpg.de

• These authors contributed equally to this work.

Introduction

It is a long standing hypothesis that neural representations in sensory systems are adapted to the statistical regularities of the environment [1,2]. Despite widespread agreement that neural processing in the early visual system must be influenced by the statistics of natural images, there are many different viewpoints on how to precisely formulate the computational goal the system is trying to achieve. At the same time, different goals might be achieved by the same optimization criterion or learning principle. Redundancy reduction [2], the most prominent example of such a principle, can be beneficial in various ways: it can help to maximize the information to be sent through a channel of limited capacity [3,4], it can be used to learn the statistics of the input [5] or to facilitate pattern recognition [6].

Besides redundancy reduction, a variety of other interesting criteria such as *sparseness* [7,8], *temporal coherence* [9], *predictive information* [10,11] , or *bottom-up saliency* [12] have been formulated. An important commonality among all these ideas is the tight link to density estimation of the input signal.

At the level of primary visual cortex there is a large increase in the number of neurons. Hence, at this stage the idea of redundancy reduction cannot be motivated by a need for compression. However, the redundancy reduction principle is not limited to be useful for compression only. More generally, it can be interpreted as a special form of density estimation where the goal is to model the statistics of the input by finding a mapping which transforms the data into a representation with statistically independent coefficients [5]. In statistics, this idea is known as projection pursuit density estimation [13] where density estimation is carried out by optimizing over a set of possible transformations in order to match the statistics of the transformed signal as good as possible to a pre-specified target distribution. Once the distribution has been matched, applying the inverse transformation effectively yields a density model for the original data. From a neurobiological point of view, we may think of the neural response properties as an implementation of such transformations. Accordingly, we here think of redundancy reduction mainly in terms of projection pursuit density estimation.

A crucial aspect of this kind of approach is the class of transformations over which to optimize. From a statistician's point of view it is important to choose a regularized function space in order to avoid overfitting. On the other hand, if the class of possible transformations is too restricted, it may be impossible to

Author Summary

Since the Nobel Prize winning work of Hubel and Wiesel it has been known that orientation selectivity is an important feature of simple cells in the primary visual cortex. The standard description of this stage of visual processing is that of a linear filter bank where each neuron responds to an oriented edge at a certain location within the visual field. From a vision scientist's point of view, we would like to understand why an orientation selective filter bank provides a *useful* image representation. Several previous studies have shown that orientation selectivity arises when the individual filter shapes are optimized according to the statistics of natural images. Here, we investigate quantitatively how critical the feature of orientation selectivity is for this optimization. We find that there is a large range of non-oriented filter shapes that perform nearly as well as the optimal orientation selective filters. We conclude that the standard filter bank model is not suitable to reveal a strong link between orientation selectivity and the statistics of natural images. Thus, to understand the role of orientation selectivity in the primary visual cortex, we will have to develop more sophisticated, nonlinear models of natural images.

find a good match to the target distribution. From a visual neuroscientist's point of view, the choice of transformations should be related to the class of possible computations in the early visual system. Here we assume the simplest case of linear transformations, optionally followed by a pointwise nonlinearity.

Intriguingly, a number of response properties of visual neurons have been reproduced by optimizing over the class of linear transformations on natural images for redundancy reduction (for a review see [12,14]). For instance, Buchsbaum and Gottschalk as well as Ruderman et al. revealed a link between the second-order statistics of color images and opponent color coding of retinal ganglion cells by demonstrating that decorrelating natural images in the trichromatic color space with Principal Component Analysis (PCA) yields the luminance, the red-green, and the blue-yellow channel [15,16]. Atick and Redlich derived the center-surround receptive fields by optimizing a symmetric decorrelation transformation [17]. Later, also spatio-temporal correlations in natural images or sequences of natural images were linked to the receptive field properties in the retina and the lateral geniculate nucleus (LGN) [18–20].

On the way from LGN to primary visual cortex, orientation selectivity emerges as a striking new receptive field property. A number of researchers (e.g., [21,22]) have used the covariance properties of natural images to derive linear basis functions that exhibit similar properties. Decorrelation alone, however, was not sufficient to achieve this goal. Rather, additional constraints were necessary, such as spatial locality or symmetry.

It was not until the reduction of higher-order correlations were taken into account that the derivation of localized and oriented bandpass filters—resembling orientation selective receptive fields in V1—was achieved without the necessity to assume any further constraints. Those filters were derived with Independent Component Analysis (ICA), a generalization of Principal Component Analysis (PCA), which aims at reducing higher-order correlations as well [8,23].

This finding suggests that within the linear filter model, orientation selectivity can serve as a further mechanism for redundancy reduction. The tenability of this hypothesis can be tested by measuring how large the advantage of orientation selective filters is over non-oriented filter shapes. The importance of such a *quantitative* assessment has first been pointed out by Li and Atick [22] and are the main focus of several publications [12,22,24–29]. Generally speaking, two different approaches have been taken in the past: In the first approach, nonparametric methods such as histograms or nearest neighbor statistics have been used with the goal to estimate the total redundancy of natural images [22,27,29]. While this approach seeks to answer the more difficult question how large the total redundancy of natural images is, the second approach compares the importance of orientation selectivity for redundancy reduction only within the class of models that are commonly used to describe V1 simple cell responses [24–26,28].

Using histogram estimators, Zhaoping and coworkers [22,27] argued that the contribution of higher-order correlations to the redundancy of natural images is five times smaller than the amount of second-order correlations. They concluded that this amount is so small that higher-order redundancy minimization is unlikely to be the main principle in shaping the cortical receptive fields.

Two objections may be raised against this conclusion: First, it is not clear how generally valid the result of [27] is. The study relies on the assumption that higher-order dependencies at distances beyond three pixels are negligible. More recent work based on nearest neighbor methods [29], however, finds a substantially larger amount of higher-order correlations when taking dependencies over longer distances into account. Secondly, even if the contribution of higher-order correlation was only 20% of the amount of second-order correlations, this contribution is not necessarily negligible. Several previous studies report that the redundancy reduction achieved with ICA for gray level images is at the same level at about 20% [24–26]. Taken together these two findings suggest that orientation selective ICA filters can account for virtually all higher-order correlations in natural images. If this was true, it would strongly support the idea that redundancy reduction could be the main principle in shaping the cortical receptive fields.

In general, however, density estimation in high dimensions is a hard problem and the results reported in the literature do not fit into a consistent view. Therefore, the crucial challenge is to control for all technical issues in order to allow for safe conclusions about the effect of orientation selectivity on redundancy reduction. Here, we address many such issues that have not been addressed before. In our study, we take the second approach and focus on "linear redundancy reduction"—the removal of statistical dependencies that can be achieved by linear filtering. While most studies have been carried out for gray level images the two studies on color images find the advantage of ICA over PCA to be many times larger for color images than for gray level images with an improvement of more than 100% [25,26]. Since it is not clear how to explain the large difference between color and gray value images, we reinvestigate the comparison between the orientation selective ICA filters and the PCA filters for color images using the same data set as in [25,26].

Our goal is to establish a reliable reference against which more sophisticated image models can be compared to in the future. We elaborate on our own previous work [28] by optimizing the ICA algorithm for the multi-information estimators used in the comparison. Additionally, we now test the advantage of the resulting orientation selective ICA filters comprehensively with three different types of analyses that are related to the notion of redundancy reduction, density estimation, and coding efficiency: (A) multi-information reduction, (B) average log-likelihood, and (C) rate-distortion curves.

Our results show that orientation selective ICA filters do not excel in any of these measures: We find that the gain of ICA in

redundancy reduction over a random decorrelation method is only about 3% for color and gray-value images. In terms of rate-distortion curves, ICA performs even worse than PCA. Furthermore, we demonstrate that a simple spherically symmetric model with only two parameters fits the filter responses significantly better than a model that assumes marginal independence . Since in this model the specific shape of the filters is ignored, we conclude that it is unlikely that orientation selectivity plays a critical role for redundancy reduction even if the class of transformations is extended to include contrast gain control mechanisms [30,31]. While many of the previous studies do not provide enough detail in order to explain their different outcomes, we provide our code and the dataset online (http://www.kyb.tuebingen.mpg.de/bethge/code/QICA/) in order to ensure the reproducibility and verifiability of our results.

Materials and Methods

An important difficulty in setting up a quantitative comparison originates from the fact that it bears several issues that may be critical for the results. In particular, choices have to be made regarding the *evaluation criteria*, the *image data*, the *estimation methods*, which *linear transformations* to include in the comparison, and which *particular implementation of ICA* to use. The significance of the outcome of the comparison will depend on how careful these choices have been made. The most relevant issues will be addressed in the following.

Notation and Nomenclature

For both, color and gray-value data, we write $\mathbf{x}$ to refer to single vectors which contain the raw pixel intensities. Vectors are indicated by bold font while the same letter in normal font with a subindex denotes one of its components. Vectors without subindices usually denote random variables, while subindices indicate specific examples. In some cases it is convenient to define the corresponding data matrix $X=(\mathbf{x}_1,\ldots,\mathbf{x}_N)$ which holds single images patches in its columns. The letter N denotes the number of examples in the dataset, while n is used for the dimension of a single data point.

Transformations are denoted by W, oftentimes with a subindex to distinguish different types. The result of a transformation to either a vector $\mathbf{x}$ or a data matrix X will be written as $\mathbf{y}=W\mathbf{x}$ or $Y=WX$, respectively.

Probability densities are denoted with the letters p and q, sometimes with a subindex to indicate differences between distributions whenever it seems necessary for clarity. In general, we use the hat symbol to distinguish between true entities and their empirical estimates. For instance, $p_{\mathbf{y}}(\mathbf{y})=p_{\mathbf{x}}(W^{-1}\mathbf{y})\cdot|\det W|^{-1}$ is the true probability density of $\mathbf{y}$ after applying a fixed transformation W, while $\hat{p}_{\mathbf{y}}(\mathbf{y})$ refers to the corresponding empirical estimate.

A distribution $p(\mathbf{y})$ is called *factorial*, or *marginally independent*, if it can be written as a product of its marginals, i.e., $p(\mathbf{y})=\Pi_{i=1}^{n}p_i(y_i)$ where $p_i(y_i)$ is obtained by integrating $p(\mathbf{y})$ over all components but y_i.

Finally, the expectation over some entity f with respect to $\mathbf{y}$ is written as $\mathbb{E}_{\mathbf{y}}[f(\mathbf{y})]=\int p(\mathbf{y})f(\mathbf{y})d\mathbf{y}$. Sometimes, we use the density instead of the random variable in the subindex to indicate the distribution, over which the expectation is taken. If there is no risk for confusion we drop the subindex. Just as above, the empirical expectation is marked with a hat symbol, i.e., $\hat{\mathbb{E}}[f(\mathbf{y})]=\frac{1}{N}\sum_{k=1}^{N}f(\mathbf{y}_k)$.

How to Compare Early Vision Models?

A principal complicacy in low-level vision is the lack of a clearly defined task. Therefore, it is difficult to compare different image representations as it is not obvious *a priori* what measure should be used.

Multi-information. The first measure we consider is the *multi-information* [32], which is the original objective function that is minimized by ICA over the choice of filters W. The multi-information assesses the total amount of statistical dependencies between the components y_i of a filtered patch $\mathbf{y}=W\mathbf{x}$:

$$I[p(\mathbf{y})]=D_{\mathrm{KL}}\left[p(\mathbf{y})||\prod_j p_j(y_j).\right]=\mathop{\mathbb{E}}_{p}\left[\log\frac{p(\mathbf{y})}{\Pi_j p_j(y_j)}\right]=\sum_{j=1}^{n}h[p_j(y_j)]-h[p(\mathbf{y})]. \quad (1)$$

The terms $h[p_j(y_j)]$ and $h[p(\mathbf{y})]$ denote the marginal and the joint entropies of the true distribution, respectively. The *Kullback-Leibler-Divergence* or *Relative Entropy*

$$D_{\mathrm{KL}}[p||q]=\mathop{\mathbb{E}}_{p}\left[\log\frac{p(\mathbf{y})}{q(\mathbf{y})}\right]$$

is an information theoretic dissimilarity measure between two distributions p and q [33]. It is always non-negative and zero if and only if p equals q. If the redundancy reduction hypothesis is taken literally, the multi-information is the right measure to minimize, since it measures how close to factorial the true distribution of the image patches in the representation $\mathbf{y}$ really is.

The application of linear ICA algorithms to ensembles of natural images reliably yields transformations consisting of localized and oriented bandpass filters similar to the receptive fields of neurons in V1. It is less clear, however, whether these filter properties also critical to the minimization of the multi-information? In order to assess the tenability of the idea that a V1 simple cell is adjusted to the purpose of redundancy reduction, it is important to know whether such a tuning can—*in principle*—result in a large reduction of the multi-information. One way to address this question is to measure *how much* more the multi-information is actually reduced by the ICA filters in comparison to others such as PCA filters. This approach has been taken in [28].

One problem with estimating multi-information is that it involves the joint entropy $h[p(\mathbf{y})]$ of the true distribution which is generally hard to estimate. In certain cases, however, the problem can be bypassed by evaluating the difference in the multi-information between two representations $\mathbf{x}$ and $\mathbf{y}$. In particular, if $\mathbf{y}$ is related to $\mathbf{x}$ by the linear transformation $\mathbf{y}=W\mathbf{x}$ it follows from definition (1) and the transformation theorem for probability densities

$$p_{\mathbf{y}}(\mathbf{y})=p_{\mathbf{x}}(\mathbf{x})\left|\det\left(\frac{\partial y}{\partial x}\right)\right|^{-1}=p_{\mathbf{x}}(W^{-1}\mathbf{y})\cdot|\det W|^{-1}$$

that difference in multi-information can be expressed as

$$\begin{aligned}I[p(\mathbf{y})]-I[p(\mathbf{x})]&=\sum_k h[p_k(y_k)]-h[p(y)]-\left(\sum_k h[p_k(x_k)]-h[p(x)]\right)\\&=\sum_k h[p_k(y_k)]-\sum_k h[p_k(x_k)]-\log|\det W|.\end{aligned}$$

For convenience, we chose a volume-conserving gauge [28] where all linear decorrelation transforms are of determinant one, and

hence $\log|\det W|=0$. This means that differences in multi-information are equal to differences of marginal entropies which can be estimated robustly. Thus, our empirical estimates of the multi-information differences are given by:

$$\Delta I \approx \sum_k h[\hat{\mathbf{p}}_k(y_k)] - \sum_k h[\hat{\mathbf{p}}_k(x_k)] \quad s.t. |\det(W)|=1 \quad (2)$$

For estimating the entropy of the univariate marginal distributions, we employ the OPT estimator introduced in [28] which uses the exponential power family to fit the marginal distributions by OPTimizing over the shape parameter. This estimator has been shown to give highly reliable results for natural images. In particular, it is much more robust than entropy estimators based on the sample kurtosis which easily overestimate the multi-information.

Average log loss (ALL). As mentioned earlier, redundancy reduction can be interpreted as a special form of density estimation where the goal is to find a mapping which transforms the data into a representation with statistically independent coefficients. This means that any given transformation specifies a density model over the data. Our second measure, the average log-loss (ALL), evaluates the agreement of this density model with the actual distribution of the data:

$$\mathop{\mathbb{E}}_{p}[-\log\hat{\mathbf{p}}(\mathbf{y})] = -\int p(\mathbf{y})\log\hat{\mathbf{p}}(\mathbf{y})d\mathbf{y} = H[p] + D_{\mathrm{KL}}[p\|\hat{\mathbf{p}}] \quad (3)$$

The average log-loss is a principled measure quantifying how different the model density $\hat{\mathbf{p}}(\mathbf{y})$ is from the true density $p(\mathbf{y})$ [34]. Since the KL-divergence is positive and zero if and only if $\hat{\mathbf{p}}=p$ the ALL is minimal only if $\hat{\mathbf{p}}$ matches the true density. Furthermore, differences in the average log-loss correspond to differences in the coding cost (i.e., information rate) in the case of sufficiently fine quantization. For natural images, different image representations have been compared with respect to this measure in [24–26].

For the estimation of the average log-loss, we compute the empirical average

$$\mathop{\mathbb{E}}_{p}[-\log\hat{\mathbf{p}}(\mathbf{y})] \approx \mathop{\hat{\mathbb{E}}}_{\mathbf{y}}[-\log\hat{\mathbf{p}}(\mathbf{y})] = -\frac{1}{N}\sum_{k=1}^{N}\log\hat{\mathbf{p}}(\mathbf{y}_k). \quad (4)$$

This estimator is equivalent to the first method in Lewicki et al. [24,35] apart from an extra term $N\log\sigma$ in their defining equation. This extra term is only necessary if one aims at relating the result to a discrete entropy obtained for a particular bin width σ.

While the empirical average in Eq. 4 in principle can be prone to overfitting, we control for this risk by evaluating all estimates on an independent test set, whose data has not been used during the parameter fit. Furthermore, we compare the average log-loss to the parametric entropy estimates $h[\hat{\mathbf{p}}]$ that we use in (A) for estimating the multi-information changes (see Eq. 2). The difference between both quantities has been named *differential log-likelihood* [36] and can be used to assess the goodness of fit of a model distribution:

$$\hat{\mathbb{E}}[-\log\hat{\mathbf{p}}] - h[\hat{\mathbf{p}}] = \mathop{\mathbb{E}}_{\hat{\mathbf{p}}}[\log\hat{\mathbf{p}}] - \hat{\mathbb{E}}[\log\hat{\mathbf{p}}].$$

The shape of the parametric model is well matched to the actual distribution if the differential log-likelihood converges to zero with increasing number of data points.

Rate-distortion curves. Finally, we consider *efficient coding* or *minimum mean square error reconstruction* as a third objective. In contrast to the previous objectives, it is now assumed that there is some limitation of the amount of information that can be transmitted, and the goal is to maximize the amount of *relevant* information transmitted about the image. In the context of neural coding, the redundancy reduction hypothesis has oftentimes been motivated in terms of coding efficiency. In fact, instead of minimizing the multi-information one can equivalently ask for the linear transformation W which maximizes the mutual information between its input $\mathbf{x}$ and its output $W\mathbf{x}+\xi$ when additive noise ξ is added to the output [3,37,38]. It is important to note, however, that this minimalist approach of "information maximization" is ignorant with respect to how useful or *relevant* the information is that has been transmitted [14].

For natural images, the source signal $\mathbf{x}$ is a continuous random variable which requires infinitely many bits to be specified with unlimited precision. In reality, however, the precision is always limited so that only a finite amount of bits can be represented. Both, the multi-information and the average log-loss do not take into account the problem what information should be encoded and what information can be discarded. Therefore, it is interesting to compare the redundancy reduction of the linear transforms with respect to the relevant image information (while the irrelevant information can be discarded anyway). To this end, we here resort to the framework of linear transform coding as it has been developed in the field of image compression [39,40], and which constitutes the theoretical foundation of the JPEG standard.

It is clear that at the level of V1 the number of neurons, encoding the retinal image, is substantially larger than the number of fibers in the optic nerve. Therefore, it is not the need for compression that makes rate distortion theory interesting at this stage. However, Barlow's redundancy reduction hypothesis must not be equated with compression. In more recent work, Barlow introduced the term 'redundancy exploitation' instead of 'redundancy reduction' in order to avoid this misunderstanding [41]. But also if we think in terms of density estimation rather than compression, it is still important to take into account that not all possible changes in the image pixels may be of equal importance for inferring the content of an image. Therefore, we here want to combine the notion of redundancy reduction with a measure for the quality with which the image can be reconstructed from the information that is preserved by the representation. Following Lewicki and coworkers (method 2 in [24,35]) we will consider the mean squared error reconstruction that can be achieved at a certain quantization level of the transformed representation. This objective is in fact very much related to the task of image compression.

Clearly, we expect that the criteria for judging image compression algorithms may not provide a good proxy to an accurate judgement of what information is considered relevant in a biological vision system. In particular, the existence of selective attention suggests that different aspects of image information are transmitted at different times depending on the behavioral goals and circumstances [12]. That is, a biological organism can change the relevance criteria dynamically on demand while for still image compression algorithms it is rather necessary that this assessment is made once and forever in a fixed and static fashion.

These issues are outside the scope of this paper. Instead we follow the common path in the past to use the mean squared reconstruction error for the pixel intensities. This is the measure of choice for high-rate still image compression [42]. In particular, it is common to report on the performance of a code by determining its rate–distortion curve which specifies the required information

rate for a given reconstruction error (and vice versa) [40]. Consequently, we will ask for a given information rate, how do the image representations compare with respect to the reconstruction error. As result, we will obtain a so-called rate–distortion curve which displays the average reconstruction error as a function of the information rate or vice versa. The second method used in [24,35] is an estimate of a single point on this curve for a particular fixed value of the reconstruction error.

The estimation of the rate–distortion curve is clearly the most difficult task among the three criteria. The framework of transform coding [39], which is extensively used in still image compression, makes several simplifying assumptions that allow one to obtain a clear picture. The encoding task is divided into two steps: First, the image patches $\mathbf{x}$ are linearly transformed into $\mathbf{y} = W\mathbf{x}$. Then the coefficients y_j are quantized independently of each other. Using this framework, we can ask whether the use of an ICA image transformation leads to a smaller reconstruction error after coefficient quantization than PCA or any other transform.

As for quantizing the coefficients, we resort to the framework of variable rate entropy coding [43]. In particular, we apply uniform quantization, which is close to optimal for high-rate compression [39,44]. For uniform quantization, it is only required to specify the bin width of the coefficients. There is also the possibility to use a different number of quantization levels for the different coefficients. The question of how to set these numbers is known as the 'bit allocation problem' because the amount of bits needed to encode one coefficient will depend monotonically on the number of quantization levels. The number of quantization levels can be adjusted in two different but equivalent ways: One possibility is to use a different bin width for each individual coefficient. Alternatively, it is also possible to use the same bin width for all coefficients and multiply all coefficients with an appropriate scale factor before quantization. The larger the variance of an individual coefficient, the more bits will be allocated to represent it.

Here, we will employ the latter approach, for which the bit allocation problem becomes an inherent part of the transformation: Any bit allocation scheme can be obtained via post-multiplication with a diagonal matrix. Thus, in contrast to the objective function of ICA, the rate–distortion criterion is not invariant against post-multiplication with a diagonal matrix. For ICA and PCA, we will determine the rate–distortion curve for both, normalized output variances ("white ICA" and "white PCA") and normalized basis functions ("normalized ICA" and "orthonormal PCA"), respectively.

Decorrelation Transforms

The particular shape of the ICA basis functions is obtained by minimization of the multi-information over all invertible linear transforms $\mathbf{y} = W\mathbf{x}$. In contrast, the removal of second-order correlations alone generally does not yield localized, oriented, and bandpass image basis functions. ICA additionally removes higher-order correlations which are generated by linear mixing. In order to assess the importance of this type of higher-order correlations for redundancy reduction and coding efficiency we will compare ICA to other decorrelating image bases.

Let $C = \mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}]$ be the covariance matrix of the data and $C = UDU^{\mathrm{T}}$ its eigen-decomposition. Then, any linear second-order decorrelation transform can be written as

$$W = D_2 \cdot V \cdot D^{-1/2} \cdot U^{\mathrm{T}} \tag{5}$$

where D and U are defined as above, V is an arbitrary orthogonal matrix and D_2 is an arbitrary diagonal matrix. It is easily verified that $Y = WX$ has diagonal covariance for all choices of V and D_2, i.e., all second-order correlations vanish. This means that any particular choice of V and D_2 determines a specific decorrelation transform. Based on this observation we introduce a number of linear transformations for later reference. All matrices are square and are chosen to be of determinant λ^m, where m is the number of columns (or rows) of W (i.e., $\lambda = \sqrt[m]{\prod \lambda_i}$ is the geometrical mean of the eigenvalues $\lambda_i, i = 1, \ldots, m$).

Orthogonal principal component analysis (oPCA). If the variances of the principle components (i.e., the diagonal elements of D) are all different, PCA is the only metric-preserving decorrelation transform and is heavily used in digital image coding. It corresponds to choosing $V = I_m$ as the identity matrix and $D_2 = \lambda D^{1/2}$, such that $W_{\mathrm{oPCA}} = \lambda U^{\mathrm{T}}$.

White principal component analysis (wPCA). Equalizing the output variances in the PCA representation sets the stage for the derivation of further decorrelation transforms different from PCA. In order to assess the effect of variance equalization for coding efficiency, we also include this "white PCA" representation into our analysis: Choose $V = I_m$ as for orthonormal PCA and then set $D_2 = \mu I_m$ with $\mu = \lambda \sqrt[m]{\det(D^{1/2})}$ such that $W_{\mathrm{wPCA}} = \mu D^{-1/2} U^{\mathrm{T}}$.

Symmetric whitening (SYM). Among the non-orthogonal decorrelation transforms, symmetric whitening stays as close to the input representation as possible (in Frobenius norm) [45]. In terms of early vision this may be seen as an implementation of a wiring length minimization principle. Remarkably, the basis functions of symmetric whitening resemble the center-surround shape of retinal ganglion cell receptive fields when applied to the pixel representation of natural images [17]. The symmetric whitening transform is obtained by setting $V = U$ and $D_2 = \mu I_m$ such that $W_{\mathrm{SYM}} = \mu\, UD^{-1/2} U^{\mathrm{T}}$.

Random whitening (RND). As a baseline which neither exploits a special structure with respect to the input representation nor makes use of higher-order correlations we also consider a completely random transformation. To obtain a random orthogonal matrix we first draw a random matrix G from a Gaussian matrix-variate distribution and then we set $V_{\mathrm{RND}} = \left(GG^{\mathrm{T}}\right)^{-1/2} G$. With $D_2 = \mu I_m$ we obtain $W_{\mathrm{RND}} = \mu V_{\mathrm{RND}} D^{-1/2} U^{\mathrm{T}}$.

White independent component analysis (wICA). Finally, ICA is the transformation which has been suggested to explain the orientation selectivity of V1 simple cells [8,23]. Set $V = V_{\mathrm{ICA}}$ for which the multi-information $I[Y]$ takes a minimum. With $D_2 = \mu I_m$ we obtain $W_{\mathrm{wICA}} = \mu V_{\mathrm{ICA}} D^{-1/2} U^{\mathrm{T}}$.

Normalized independent component analysis (nICA). Normalized independent component analysis (nICA) differs from white ICA (W_{wICA}) only by a different choice of the second diagonal matrix D_2. Instead of having equal variance in each coefficient, we now choose D_2 such that the corresponding basis vector of each coefficient has the same length in pixel space. It is easy to see that our first two criteria, the multi-information and the negative log-likelihood, are invariant under changes in D_2. It makes a difference for the rate–distortion curves as in our setup the variance (or, more precisely, the standard deviation) determines the bit allocation. Practically, W_{nICA} can be determined by using W_{wICA} as follows: First, we compute the matrix inverse $A := W_{\mathrm{wICA}}^{-1}$ and determine the Euclidean norm $a_1, \ldots, a_m$ of the column vectors of A. With $D_a = diag(a_1, \ldots, a_m)$, we then obtain $W_{\mathrm{nICA}} = \frac{1}{\sqrt[m]{\det(D_a)}} D_a W_{\mathrm{wICA}}$.

ICA Algorithm

If the true joint probability distribution is known, the minimization of the multi-information over all linear transformations can be formulated without any assumptions about the shape of the distribution. In practice, the multi-information has to be

estimated from a finite amount of data which requires to make assumptions about the underlying density.

There are many different ICA algorithms which differ in the assumptions made and also in the optimization technique employed. The choice of the particular ICA algorithm used here was guided by a set of requirements that arise from the specific problem setting. Although a wide variety of ICA algorithms has been published, none of them fits exactly all of our requirements.

We would like to use an ICA algorithm, which gives the ICA image basis the best chance for the comparison with other image representations. For the comparison of the multi-information reduction, we are using the OPT estimator introduced in [28] which has been found to give the most reliable results. This estimator employs a parametric estimate of the coefficient distributions based on the exponential power family which is known to provide an excellent fit to the coefficient distributions of natural images [28,46]. Our ICA algorithm should make the same assumptions about the data as we make for the final comparison of the multi-information reduction. Therefore, we are also using the exponential power family model for the marginal densities during the minimization of the multi-information. In addition, we want to have an ICA basis which is indistinguishable from the other image representations with respect to the second-order statistics. Therefore, we are using a pre-whitened ICA algorithm, whose search space is restricted to the subgroup of orthogonal matrices $SO(n)$. One of the most efficient ICA methods in the public domain specialized to pre-whitened ICA is FastICA [47]. We use this fixed-point algorithm as an initialization. Subsequently, the solution is further refined by performing a gradient ascent over the manifold of orthogonal matrices on the likelihood of the data, when each marginal is modelled by a the exponential power distribution as in the case of the OPT estimator.

In order to optimize the objective function over the subspace of orthogonal matrices, we adapted the algorithms for Stiefel manifolds proposed by Edelman et al. [48] to the simpler case of orthogonal groups and combined it with the line-search routine dbrent from [49] to achieve a rather straightforward gradient descent algorithm. For the initialization with FastICA, we use the Gaussian non-linearity, the symmetric approach and a tolerance level of 10^{-5}.

Spherically Symmetric Model

A well known result by Maxwell [50] states that the only factorial distribution invariant against arbitrary orthogonal transformations is the isotropic Gaussian distribution. Natural images exhibit marginals which are significantly more peaked than Gaussian. Nevertheless, their distribution does share the spherical symmetry with the Gaussian as already found by [51] for gabor filter pairs and lately exploited by [31] for nonlinear image representations. Therefore, it makes sense to compare the performance of the ICA model with a spherically symmetric model of the whitened data $\mathbf{y}_{\mathrm{w}} = W_{\mathrm{RND}}\mathbf{x}$. Note that any spherically symmetric model is still invariant under orthogonal transformations while only the Gaussian additionally exhibits marginal independence.

While the radial distribution of a Gaussian (i.e., the distribution over the lengths of the random vectors) is a χ-distribution, whose shape and scale parameter is determined by the number of dimensions and the variance, respectively, the spherical symmetric model may be seen as a generalization of the Gaussian, for which the radial distribution $p(r)$ with $r := \|\mathbf{y}\|_2$ can be of arbitrary shape. The density of the spherically symmetric distribution (SSD) is defined as $p_{\mathbf{Y}}(\mathbf{y}) = p_r(r)/S_n(r)$, where $S_n(r) = r^{n-1}2\pi^{n/2}/\Gamma(n/2)$ is the surface area of a sphere in $\mathbb{R}^n$ with radius r. For simplicity we will model the radial distribution with a member of the Gamma family

$$p(r) = \frac{r^{u-1}\exp\left(-\frac{r}{s}\right)}{s^u\Gamma(u)},\ r \geq 0 \tag{6}$$

with shape parameter u and scale parameter s, which can be easily matched to the mean and variance of the empirical distribution via $s = \widehat{\mathbf{Var}}[r] \big/ \hat{\mathbb{E}}[r]$ and $u = \hat{\mathbb{E}}[r]^2 \big/ \widehat{\mathbf{Var}}[r]$.

Dataset

The difference in the performance between ICA and other linear transformations clearly depends on the data. For *gray-scale* images we observed in our previous study [28] that the difference in the multi-information between ICA and any other decorrelation transform is consistently smaller than 5%. In particular, we controlled for the use of different pictures and for the effect of different pre-processing steps.

Here, we resort to the dataset used in a previous study [25,26], which among all previous studies reported the largest advantage of ICA compared to PCA. This *color* image dataset is based on the Bristol Hyperspectral Images Database [52] that contains multi-spectral recordings of natural scenes taken in the surroundings of Bristol, UK and in the greenhouses of Bristol Botanical Gardens. The authors of [26] kindly provided to us a pre-processed version of the image data where spectral radiance vectors were already converted into LMS values. During subsequent processing the reflectance standard was cut out and images were converted to log intensities [26].

All images come at a resolution of 256×256 pixels. From each image circa 5000 patches of size 7×7 pixels were drawn at random locations (circa 40000 patches in total). For chromatic images with three color channels (LMS) each patch is reshaped as a $7\times7\times3 = 147$-dimensional vector. To estimate the contribution of color information, a comparison with monochromatic images was performed where gray-value intensities were computed as $I = \log\left(\frac{1}{3}(L+M+S)\right)$ and exactly the same patches were used for analysis. In the latter case, the dimensionality of a data sample is thus reduced to 49 dimensions. All experiments are carried out over ten different training and test sets sampled independently from the original images.

Our motivation to chose 7×7 patches is to keep the same setting as in [26] for the sake of comparability. As this patch size is rather small, we performed the same analysis for patch sizes of 15×15 as well. All results in the paper refer to the case of 7×7 image patches. The results for 15×15 can be found in the supplementary material (Text S1).

The statistics of the average illumination in the image patches, the DC component, differs significantly from image to image. Therefore, we first separated the DC component from the patches before further transforming them. In order to leave the entropy of the data unaffected, we used an orthogonal transformation. The projector P_{remDC} is computed such that the first (for each color channel) component of $P_{remDC}\mathbf{x}$ corresponds to the DC component(s) of that patch. One such a possible choice is the matrix

$$p = \begin{bmatrix} 1 & 0 & 0 & \cdots \\ 1 & 1 & 0 & \cdots \\ 1 & 0 & \ddots & \cdots \\ \vdots & & & 1 \end{bmatrix}^{\mathrm{T}}$$

However, this is not an orthogonal transformation. Therefore, we decompose P into $P=QR$ where R is upper triangular and Q is an orthogonal transform. Since $P=QR$, the first column of Q must be a multiple of the vector with all coefficients equal to one (due to the upper triangluarity of R). Therefore, the first component of $Q^{\mathrm{T}}\mathbf{x}$ is a multiple of the DC component. Since Q is an orthonomal transform, using all but the first row of Q^{T} for P_{remDC} projects out the DC component. In the case of color images P_{remDC} becomes a block-diagonal matrix with Q^{T} as diagonal elements for each channel.

By removing the DC component in that manner, all linear transformations are applied in $n-1$ dimensions, if n denotes the number of pixels in the original image patch. In this case the marginal entropy of the DC-components has to be included in the computation of the multi-information in order to ensure a valid comparison with the original pixel basis. We use the same estimators as in [28] to estimate the marginal entropy of DC-component.

Results

Filter Shapes

As in previous studies [8,23] the filters derived with ICA exhibited orientation selective tuning properties similar to those observed for V1 simple cells (see Figure 1). For illustration, we also show the basis functions learned with PCA and RND in Figure 1. The basis functions A are obtained by inverting the filter matrix W (including the DC component). The result is displayed in the upper panel (Figure 1A–C). Following common practice, we also visualize the basis functions after symmetric whitening (Figure 1D–F).

The basis functions of both PCA and ICA exhibit color opponent coding but the basis functions of ICA are additionally localized and orientation selective. The basis functions of the random decorrelation transform does not exhibit any regular structure besides the fact that they are bandpass. The following quantitative comparisons will show, however, that the distinct shape of the ICA basis functions does not yield a clear advantage for redundancy reduction and coding efficiency.

Multi-Information

The multi-information is the original objective function that is minimized by ICA over all possible linear decorrelation transforms. Figure 2 shows the reduction in multi-information achieved with different decorrelation transforms including ICA for chromatic and gray value images, respectively. For each representation, the results are reported in bits per component, i.e., as marginal entropies averaged over all dimensions:

$$\langle h\rangle = \frac{1}{n}\sum_{k=1}^{n} h[p_k(y_k)] \tag{7}$$

Table 1 shows the corresponding values for the transformations RND, SYM, PCA and ICA. For both chromatic images and gray-value intensities, the lowest and highest reduction is achieved with RND or ICA, respectively. However, the additional gain in the multi-information reduction achieved with ICA on top of RND constitutes only 3.20% for chromatic images and 2.39% for achromatic in comparison with the total reduction relative to the pixel basis (PIX). This means that only a small fraction of redundancy reduction can actually be accounted to the removal of higher-order redundancies with ICA.

One may argue that the relatively small patch size of 7×7 pixel may be responsible for the small advantage of ICA as all decorrelation functions already getting the benefit of localization. In order to address the question how the patch size affects the linear redundancy reduction, we repeated the same analysis on a whole range of different patch sizes. Figure 3 shows the multi-information reduction with respect to the pixel representation (PIX) achieved by the transformations RND and ICA. The achievable reduction quickly saturates with increasing patch size such that its value for 7×7 image patches is already at about 90% of its asymptote. In particular, one can see that the relative advantage of ICA over other transformations is still small (~3%) also for large patch sizes. All Tables and Figures for patch size 15×15 can be found in the additional material (Text S1).

Average Log-Loss

Since redundancy reduction can also be interpreted as a special form of density estimation we also look at the average log-loss which quantifies how well the underlying density model of the different transformations is matched to the statistics of the data. Table 2 shows the average log-loss (ALL) and Table 3 the differential log-likelihood (DLL) in bits per component. For the average log-loss, ICA achieved an ALL of 1.78 bits per component for chromatic images and 1.85 bits per component for achromatic images. Compared to the ALL in the RND representation of 1.9 bits and 1.94 bits, respectively, the gain achieved by ICA is again small. Additionally, the ALL values were very close to the differential entropies, resulting in small DLL values. This confirms that the exponential power distribution fits the shape of the individual marginal coefficient distributions well. Therefore, we can safely conclude that the advantage of ICA is small not only in terms of redundancy reduction as measured by the multi-information, but also in the sense of density estimation.

Comparison to a Spherical Symmetric Model. The fact that ICA fits the distribution of natural images only marginally better than a random decorrelation transform implies that the generative model underlying ICA does not apply to natural images. In order to assess the importance of the actual filter shape, we fitted a spherically symmetric model to the filter responses. The likelihood of such a model is invariant under post-multiplication of an orthogonal matrix, i.e., the actual shape of the filter. Therefore, a good fit of such a model provides strong evidence against a critical role of certain filter shapes.

As shown in Table 2, the ALL of the SSD model is 1.67 bits per component for chromatic images and 1.65 bits per component for achromatic images. This is significantly smaller than the ALL of ICA indicating that it fits the distribution of natural images much better than ICA does. This result is particularly striking if one compares the number of parameters fitted in the ICA model compared to the SSD case: After whitening, the optimization in ICA is done over the manifold of orthogonal matrices which has $m(m-1)/2$ free parameters (where m denotes the number of dimensions without the DC components). The additional optimization of the shape parameters for the exponential power family fitted to each individual component adds another m parameters. For the case of 7×7 color image patches we thus have $\frac{144\cdot 145}{2}=10440$ parameters. In stark contrast, there are only two free parameters in the SSD model with a radial Gamma distribution, the shape parameter u and the scale parameter s. Nevertheless, for chromatic images the gain of the SSD model relative to random whitening is almost twice as large as that of ICA and even three and a half times as large for achromatic images.

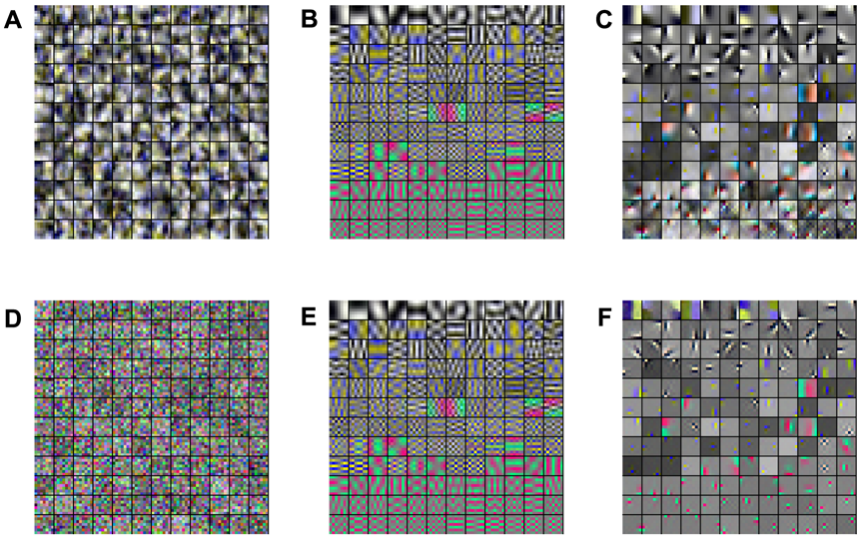


Figure 1. Examples for Receptive Fields of Various Image Transforms. Basis functions of a random decorrelation transform (RND), principal component analysis (PCA) and independent component analysis (ICA) in pixel space (A–C) and whitened space (E–F). The image representation in whitened space is obtained by left multiplication with the matrix square root of the inverse covariance matrix $C^{-1/2}$.
doi:10.1371/journal.pcbi.1000336.g001

Since the SSD model is completely independent of the choice of the orthogonal transformation after whitening, its superior performance compared with ICA provides a very strong argument against the hypothesis that orientation selectivity plays a critical role for redundancy reduction. In addition, it is also corroborates earlier arguments that has been given to show that the statistics of natural images does not conform to the generative model underlying ICA [51,53].

Besides the better fit of the data by the SSD model, there is also a more direct way of demonstrating the dependencies of the ICA coefficients: If $Y_{\mathrm{wICA}} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$ is data in the wICA representation, then the independence assumption of ICA can be simulated by applying independent random permutations to the rows of Y_{wICA}. Certainly, such a shuffling procedure does not alter the histograms of the individual coefficients but it is suited to destroy potential statistical dependencies among the coefficients. Subsequently, we can transform the shuffled data Y_{sICA} back to the RND basis $Y_{\mathrm{sRND}} = W_{\mathrm{RND}} W_{\mathrm{wICA}}^{-1} Y_{\mathrm{sICA}}$. If the ICA coefficients were independent, the shuffling procedure would not alter the joint statistics, and hence, one should find no difference in the multi-information between Y_{sRND} and Y_{RND}. But infact, we observe a large discrepancy between the two (Figure 4). The

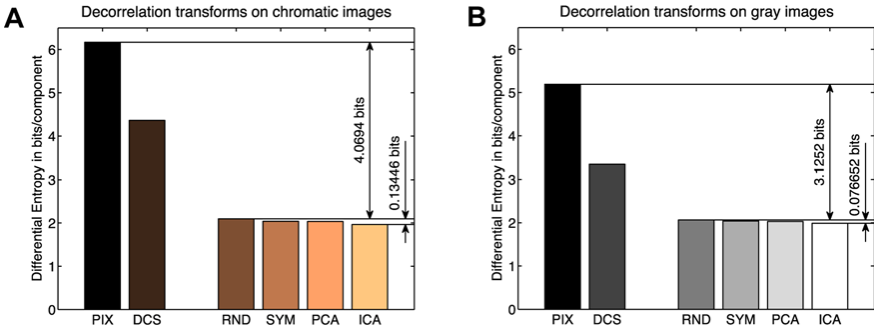


Figure 2. Multi-Information Reduction per Dimension. Average differential entropy $\langle h \rangle$ for the pixel basis (PIX), after separation of the DC component (DCS), and after application of the different decorrelation transforms. The difference between PIX and RND corresponds to the redundancy reduction that is achieved with a random second-order decorrelation transform. The small difference between RND and ICA is the maximal amount of higher-order redundancy reduction that can be achieved by ICA. Diagram (A) shows the results for chromatic images and diagram (B) for gray value images. For both types of images, only a marginal amount can be accounted to the reduction of higher order dependencies.
doi:10.1371/journal.pcbi.1000336.g002

Table 1. Comparision of the Multi-Information Reduction for Chromatic and Achromatic Images.

Absolute Difference		Relative Difference		
	Color	Gray	Color	Gray
RND-PIX	−4.0694±0.0043	−3.1252±0.0043		
SYM-RND	−0.0593±0.0004	−0.0259±0.0006	$\frac{SYM-RND}{SYM-PIX}$ 1.44±0.01	0.82±0.02
PCA-RND	−0.0627±0.0008	−0.0353±0.0011	$\frac{PCA-RND}{PCA-PIX}$ 1.52±0.02	1.12±0.03
ICA-RND	−0.1345±0.0008	−0.0767±0.0008	$\frac{ICA-RND}{ICA-PIX}$ 3.20±0.02	2.39±0.02

Differences in the multi-information reduction between various decorrelation transforms (SYM, PCA, ICA) relative to a random decorrelation transform (RND) compared to the multi-information reduction achieved with the random decorrelation transform relative to the original pixel basis (RND-PIX). The *absolute* multi-information reduction is given in bits/component on the left hand side. The right hand side shows how much more the special decorrelation transforms SYM, PCA and ICA can reduce the multi-information *relative* to the random (RND) one.

doi:10.1371/journal.pcbi.1000336.t001

Table 2. Average Log-Loss (ALL) for Chromatic and Achromatic Images.

	Color	Gray
	ALL	ALL
RND	1.9486±0.0035	1.9414±0.0044
SYM-RND	−0.0881±0.0004	−0.0402±0.0005
PCA-RND	−0.0751±0.0009	−0.0391±0.0011
ICA-RND	−0.1637±0.0007	−0.0880±0.0007
SSD-RND	−0.2761±0.0025	−0.2868±0.0032

The first row shows the average log-loss (ALL, in bits/component) of the density model determined by the linear transformation RND. The value was obtained by averaging over 10 separately sampled training and test sets of size 40.000 and 50.000, respectively. The following rows show the difference of the ALL of the models SYM, PCA, ICA and of the spherically symmetric density (SSD) to the ALL of the RND model. The smaller average log-loss of the SSD model compared to the ICA model fundamentally contradicts the assumptions underlying the ICA model.

doi:10.1371/journal.pcbi.1000336.t002

distributions of the sRND coefficients were very close to Gaussians and the average marginal entropy of sRND yielded $\langle h_{\mathrm{sRND}} - h_{\mathrm{Gauss}} \rangle \approx -0.001$ bits in contrast to $\langle h_{\mathrm{RND}} - h_{\mathrm{Gauss}} \rangle \approx -0.1$ bits. In other words, the finding that for natural images the marginals of a random decorrelation transform have Laplacian shape ($\alpha \approx 1$) stands in clear contradiction to the generative model underlying ICA. If the ICA model was valid, one would expect that the sum over the ICA coefficients would yield Gaussian marginals due to the central limit theorem. In conclusion, we have very strong evidence that the ICA coefficients are not independent in case of natural images.

Rate-Distortion Curves

There are different ways to account for the limited precision that is imposed by neural noise and firing rate limitations. As mentioned above the advantage with respect to a plain information maximization criterion can equivalently be measured by the multi-information criterion considered above [37,54]. In order to additionally account for the question which representation optimally encodes the *relevant* image information, we also present rate distortion curves which show the minimal reconstruction error as a function of the information rate.

We compare the rate–distortion curves of wICA, nICA, wPCA and oPCA (see Figure 5). Despite the fact that ICA is optimal in terms of redundancy reduction (see Table 2), oPCA performs optimal with respect to the rate-distortion trade-off. wPCA in turn performes worst and remarkably similar to wICA. Since wPCA and wICA differ only by an orthogonal transformation, both representations are bound to the same metric. oPCA is the only transformation which has the same metric as the pixel representation according to which the reconstruction error is determined. By normalizing the length of the ICA basis vectors in the pixel space, the metric of nICA becomes more similar to the pixel basis and the performance with respect to the rate-distortion trade-off

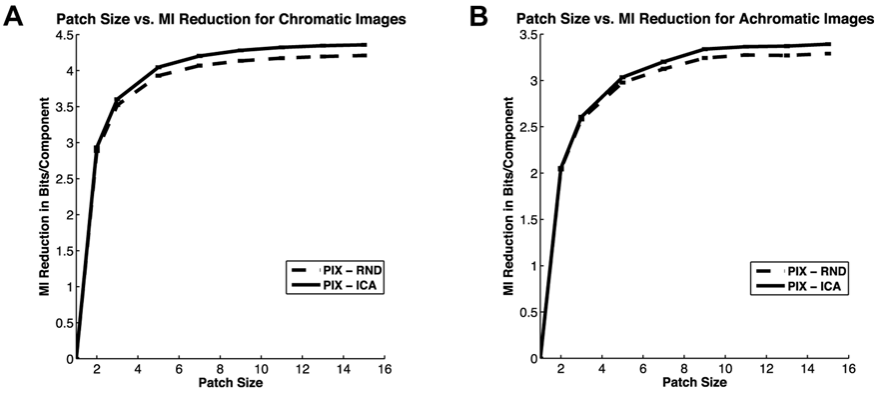


Figure 3. Redundancy Reduction as a Function of Patch Size. The graph shows the multi-information reduction achieved by the transformations RND and ICA for chromatic (A) and achromatic images (B). The gain quickly saturates with increasing patch size such that its value for 7×7 image patches is already at about 90% of its asymptote. This demonstrates that the advantage of ICA over other transformations does not increase with increasing patch size.

doi:10.1371/journal.pcbi.1000336.g003

Table 3. Differential Log-Likelihood (DLL) for Chromatic and Achromatic Images.

	Color		Gray	
	DLL	$\langle\alpha\rangle$	**DLL**	$\langle\alpha\rangle$
RND	−0.0113±0.0007	1.0413±0.0026	−0.0057±0.0006	1.0132±0.0046
SYM	−0.0388±0.0009	0.8961±0.0021	−0.0195±0.0009	0.9486±0.0040
PCA	−0.0224±0.0007	0.9145±0.0024	−0.0087±0.0007	0.9425±0.0025
ICA	−0.0378±0.0009	0.7687±0.0017	−0.0154±0.0011	0.8434±0.0025

The small DLL values suggest, that the exponential power distribution fits the shape of the individual coefficient distributions well. In addition, we also report the average exponent $\langle\alpha\rangle$ of the exponential power family fit to the individual coefficient distributions ($\alpha=1$ corresponds to a Laplacian shape).
doi:10.1371/journal.pcbi.1000336.t003

improved considerably. Nevertheless, for a fixed reconstrucion error the discrete entropy after quantization in the oPCA basis is up to 1 bit/component *smaller* than for the corresponding nICA-basis.

In order to understand this result more precisely, we analyzed how the quantization of the coefficients affects the two variables of the rate–distortion function, *discrete entropy* and *reconstruction error*.

Figure 6 shows an illustrative example in order to make the following analysis more intuitive. The example demonstrates that the quality of a transform code not only depends on the redundancy of the coefficients but also on the shape of the partition cells induced by the quantization. In particular, when the cells are small (i.e., the entropy rate is high), then the reconstruction error mainly depends on having cell shapes that minimize the average distance to the center of the cell. Linear transform codes can only produce partitions into parallelepipeds (Figure 6B). The best parallelepipeds are cubes (Figure 6A). This is why PCA yields the (close to) optimal trade-off between minimizing the redundancy *and* the distortion, as it is the only orthogonal transform that yields uncorrelated coefficients. For a more comprehensive introduction to transform coding we refer the reader to the excellent review by Goyal [39].

Discrete entropy. Given a uniform binning of width δ the discrete entropy H_δ of a probability density $p(x)$ is defined as

$$H_\delta = -\sum_i p_i \log p_i \quad \text{with} \quad p_i = \int_{B_i} p(x)dx, \tag{8}$$

where B_i denotes the interval defined by the i-th bin. For small bin-sizes $\delta\to 0$, there is a close relationship between *discrete* and *differential* entropy: Because of the mean value theorem we can approximate $p_i \approx p(\xi_i)\delta$ with $\xi\in B_i$, and hence

$$\begin{aligned} H_\delta &\approx -\sum_i p(\xi_i)\delta \log[p(\xi_i)\delta] \\ &= \underbrace{-\sum_i \delta \; p(\xi_i)\log p(\xi_i)}_{\substack{\delta\to 0 \\ \to -\int p(x)\log p(x)\,dx}} \underbrace{-\log\delta \sum_i p(\xi_i)\delta}_{\substack{\delta\to 0 \\ \to 1}}. \end{aligned}$$

Thus, we have the relationship $H_\delta \approx h - \log\delta$ for sufficiently small δ (i.e., high-rate quantization). In other words, H_δ asymptotically grows linearly with $(-\log\delta)$. Therefore, we can fit a linear function to the asymptotic branch of the function $H_\delta = H_\delta(-\log\delta)$ which is plotted in Figure 7A (more precisely we are plotting the average over all dimensions). If we take the ordinate intercept of the linear approximation, we obtain a nonparametric estimate of the differential entropy which can be compared to the entropy estimates reported above (Those estimates were determined with the OPT estimator). Equivalently, one can consider the function $h_\delta(-\log\delta) := H_\delta - (-\log\delta)$ which gives a better visualization of the error of the linear approximation (Figure 7, left, dashed line). For $h_\delta(-\log\delta)$ the differential entropy is obtained in the limit $h = \lim_{(-\log\delta)\to\infty} h_\delta = \lim_{\delta\to 0} h_\delta$.

This analysis shows that differences in differential entropy in fact translate into differences in discrete entropy after uniform quantization with sufficiently small bins. Accordingly, the minimization of the multi-information as proposed by the redundancy reduction hypothesis does in fact also minimize the discrete entropy of a uniformly quantized code. In particular, if we look at the discrete entropy of the four different transforms, oPCA, wPCA, wICA, nICA (Figure 7B), we find that asymptotically the two PCA transforms require slightly more entropy than the two ICA transforms, and there is no difference anymore between

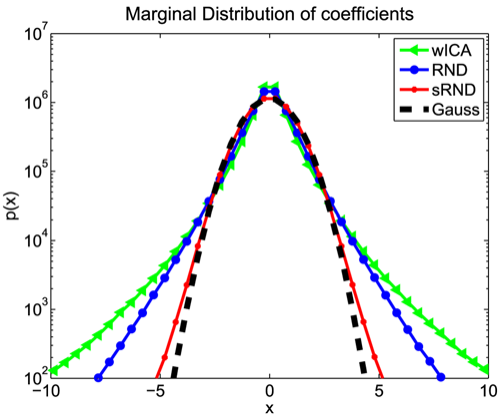


Figure 4. The Distribution of Natural Images does not Conform with the Generative Model of ICA. In order to test for statistical dependencies among the coefficients Y_{wICA} of whithened ICA for single data samples, the coefficients were shuffled among the data points along each dimension. Subsequently, we transform the resulting data matrix Y_{sICA} into $Y_{\text{sRND}} = W_{RND} W_{wICA}^{-1} Y_{\text{sICA}}$. This corresponds to a change of basis from the ICA to the random decorrelation basis (RND). The plot shows the log-histogram over the coefficients over all dimensions. If the assumptions underlying ICA were correct, there would be no difference between the histogram of Y_{sRND} and Y_{RND}.
doi:10.1371/journal.pcbi.1000336.g004

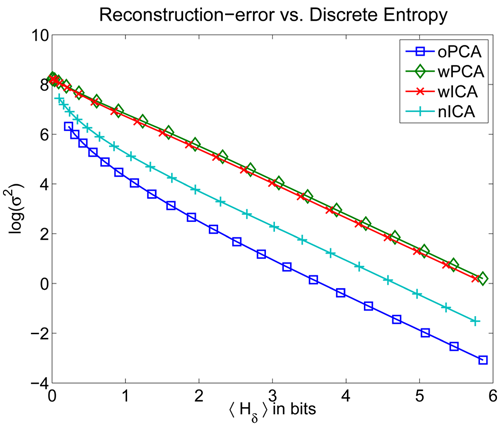


Figure 5. Rate-distortion Curves. Rate-distortion curve for PCA and ICA when equalizing the output variances (wPCA and wICA) and when equalizing the norm of the corresponding image bases in pixel space (oPCA and nICA). The plot shows the discrete entropy H_δ in bits (averaged over all dimensions) against the log of the squared reconstruction error σ^2. oPCA outperforms all other transforms in terms of the rate-distortion trade-off. wPCA in turn performes worst and remarkably similar to wICA. Since wPCA and wICA differ only by an orthogonal transformation, both representations are bound to the same metric. oPCA is the only transformation which has the same metric as the pixel representation according to which the reconstruction error is determined. By normalizing the length of the ICA basis vectors in the pixel space, the metric of nICA becomes more similar to the pixel basis and the performance with respect to the rate-distortion trade-off can be seen to improve considerably.
doi:10.1371/journal.pcbi.1000336.g005

oPCA and wPCA or wICA and nICA. This close relationship between discrete and differential entropy for high-rate quantization, however, is not sufficient to determine the coding performance evaluated by the rate–distortion curve. The latter requires to compare also the reconstruction error for the given quantization.

Reconstruction error. The reconstruction error is defined as the mean squared distance in the pixel basis between the original image and the image obtained by reconstruction from the quantized coefficients of the considered transformation. For the reconstruction, we simply use the inverse of the considered transformation, which is optimal in the limit of high-rate quantization.

When looking at the reconstruction error as a function of the bin width (Figure 8) we can observe much more pronounced differences between the different transformations than it was the case for the entropy. As a consequence, the differences in the reconstruction error turn out to be much more important for the rate-distortion trade-off than the differences in the entropy. Only the two transformations with exactly the same metric, wPCA and wICA, exhibit no difference in the reconstruction error. This suggests that minimization of the multi-information is strictly related to efficient coding if and only if the transformation with respect to the pixel basis is orthogonal. As we have seen that the potential effect of higher-order redundancy reduction is rather small, we expect that the PCA transform constitutes a close approximation to the minimizer of the multi-information among all orthogonal transforms because PCA is the only orthogonal transform which removes all second-order correlations.

Discussion

The structural organization of orientation selectivity in the primary visual cortex has been associated with self-organization since the early seventies [55], and much progress has been made to narrow down the range of possible models compatible with the empirical findings (e.g., [56–58]). The link to visual information processing, however, still remains elusive [59–61].

More abstract unsupervised learning models which obtain orientation selective filters using sparse coding [8] or ICA [23] try to address this link between image processing and the self-organization of neural structure. In particular, these models not only seek to reproduce the orientation tuning properties of V1 simple cells but they additionally address the question of how the simple cell responses collectively can instantiate a representation for arbitrary images. Furthermore, these image representations are learned from an information theoretic principle assuming that the learned filters exhibit advantageous coding properties.

The goal of this study is to quantitatively test this assumption in the simple linear transform coding framework. To this end, we investigated three criteria, the multi-information—i.e., the objective function of ICA—the average log-loss, and rate-distortion curves. There are a number of previous studies which also aimed at quantifying how large the advantage of the orientation selective ICA filters is relative to second-order decorrelation transformations. In particular, four papers [24–26,28], are most closely related to this study as all of them compare the average log-loss of different transformations. However, they did not provide a coherent answer to the question how large the advantage of ICA is compared to other decorrelation transforms.

Lewicki and Olshausen [24] found that their learned bases show a 15–20% improvement over traditional bases. However, their result cannot be used to compare second-order and higher-order redundancy reduction because the entire analysis is based on a dataset in which all images have been preprocessed with a bandpass filter as in olshausen:1996. Since bandpass filtering already removes a substantial fraction of second-order correlations in natural images, their study is likely to systematically underestimate the total amount of second-order correlations in natural images.

Lee et al. [25,26] reported an advantage of over 100% percent for ICA in the case of color images and a more moderate but substantial gain of about 20% for gray-value images. In order to avoid possible differences due to the choice of data set we here used exactly the same data as in [25,26]. Very consistently, we find only a small advantage for ICA of less than five percent for both multi-information and the average log-loss. In particular, we are not able to reproduce the very large difference between color and gray-value images that they reported. Unfortunately, we cannot pinpoint where the differences in the numbers ultimately come from because it is not clear which estimation procedure was used in [25,26].

The estimators used for the measurements in the present study have been shown previously to give correct results on artificial data [28] and we provide our code online for verification. Furthermore, Weiss and Freeman showed for an undirected probabilistic image model that whitening already yields 98% of the total performance [62]. Finally, the superior performance of the simple SSD model with only two free parameters provides a very strong explanation for why the gain achieved with ICA is so small relative to a random decorrelation transform: Since a spherically symmetric model is invariant under orthogonal transformations and provides a better fit to the data, the actual shape of the filter does not seem to be critical. It also shows that the fundamental assumption

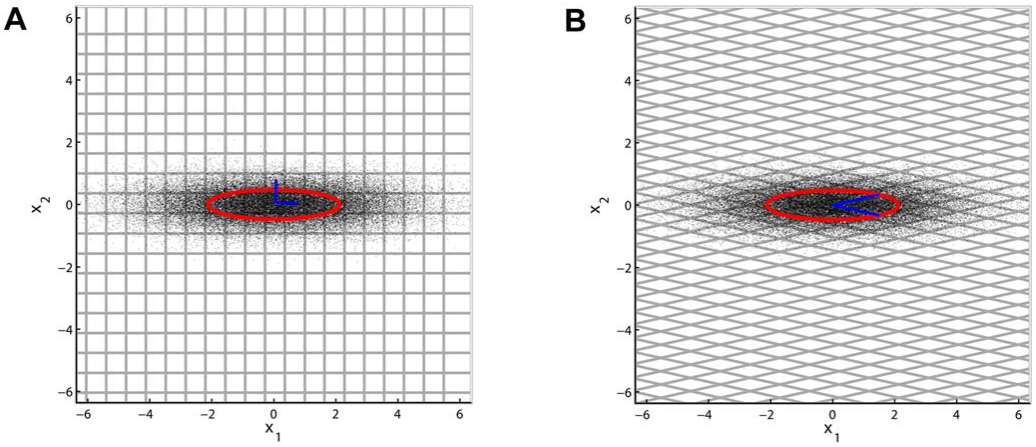


Figure 6. The Partition Cell Shape is Crucial for the Quantization Error. The quality of a source code depends on both the shapes of the partition cells and on how the sizes of the cells vary with respect to the source density. When the cells are small (i.e., the entropy rate is high), then, the quality mainly depends on having cell shapes that minimize the average distance to the center of the cell. For a given volume, a body in Euclidean space that minimizes the average distance to the center is a sphere. The best packings (including the hexagonal case) cannot be achieved with linear transform codes. Transform codes can only produce partitions into parallelepipeds, as shown here for two dimensions. The best parallelepipeds are cubes which are only obtained in the case of orthogonal transformations. Therefore PCA yields the (close to) optimal trade-off between minimizing the redundancy *and* the distortion as it is the only *orthogonal* decorrelation transform (see [39] for more details). The figure shows 50.000 samples from a bivariate Gaussian random variable. Plot (A) depicts a uniform binning (bin width $\Delta = 0.01$, only some bin borders are shown) induced by the only orthogonal basis for which the coefficients x_1 and x_2 are decorrelated. Plot (B) shows uniform binning in a decorrelated, but not orthogonal basis (indicated by the blue lines). Both cases have been chosen such that the multi-information between the coefficients is identical and the same entropy rate was used to encode the signal. However, due to the shape of the bins in plot (B) the total quadratic error increases from 0.4169 to 0.9866. The code for this example can be also downloaded from http://www.kyb.tuebingen.mpg.de/bethge/code/QICA/.
doi:10.1371/journal.pcbi.1000336.g006

underlying ICA—the data are well described by a linear generative model with independent sources—is not justified in the case of natural images.

From all these results, we can safely conclude that the actual gain of ICA compared to PCA is smaller than 5% for both gray level images and color images.

Is Smaller Than 5% Really Small?

A valid question to ask is whether comparing the amount of higher-order correlations to the amount of second-order correlations is the right thing to do. Even if the amount of higher-order correlations may be small in comparison to the amount of second-order correlations, we still know that higher-order correlations can

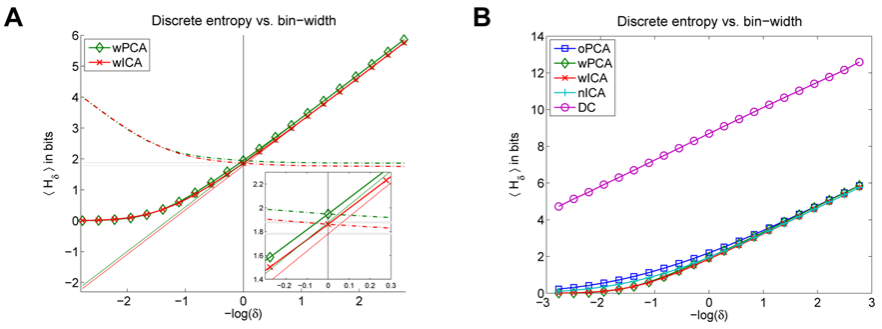


Figure 7. Discrete vs. Differential Entropy. (A) Relationship between discrete and differential entropy. Discrete entropy $\langle H_\delta \rangle$ averaged over all channels as a function of the negative log bin width. The straight lines constitute the linear approximation to the asymptotic branch of the function. Their interception with the y-axis are visualized by the gray shaded, horizontal lines. The dashed lines represent $\langle h_\delta \rangle$ which converge to the gray shaded lines for $\delta \to 0$. (B) There are only small differences in the average discrete entropy for oPCA, wPCA, wICA, nICA as a function of the negative log bin width. Since the discrete entropy of the DC component is the same for all transforms, it is not included in that average but plotted separately instead.
doi:10.1371/journal.pcbi.1000336.g007

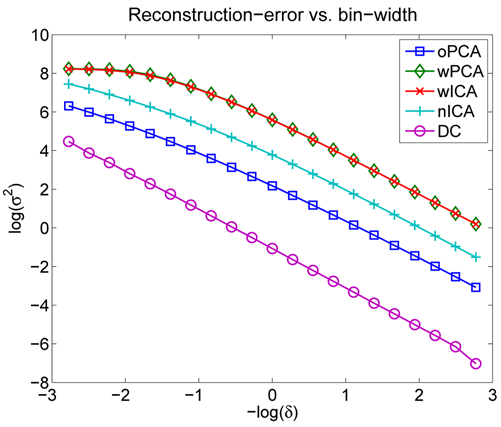


Figure 8. Reconstruction Error vs. Bin Width of Discrete Entropy. Reconstruction error σ^2 as a function of the bin width δ, shown on a logarithmic scale. The differences between the different transforms are relatively large. Only the two transformations with exactly the same metric, wPCA and wICA, exhibit no difference in the reconstruction error.
doi:10.1371/journal.pcbi.1000336.g008

be a critical signature of the content of an image. For example, textures are very useful to demonstrate how changes in higher-order correlations can change the perceptual meaning of an image.

Our results on the rate-distortion trade-off can be taken as an indication that the fraction of higher-order correlations captured by ICA is perceptually less relevant. This interpretation is further corroborated by a psychophysical comparison of the *perceptual* redundancy of the ICA and the PCA basis [63]. Another confirmation of this interpretation can be obtained if we use the learned image representations as generative models. Perceptually image patches sampled from the ICA model do not look more similar to natural image patches than those sampled from the random decorrelation basis (Figure 9). Currently, we are running psychophysical experiments which also show quantitatively that there is no significant difference between the ICA model and the PCA model if the subjects have to discriminate between textures that are generated by these models.

In summary, we were not able thus far to come up with a meaningful interpretation for which the improvement of ICA would be recognized as being large. On the basis of the present study it seems rather unlikely that such a measure can be found for linear ICA. Instead, we believe that more sophisticated, nonlinear image models are necessary to demonstrate a clear advantage of orientation selectivity.

What about Nonparametric Approaches?

The focus on linear redundancy reduction models in this study is motivated by the goal to first establish a solid and reproducible result for the simplest possible case before moving on to more involved nonlinear transformations. Nevertheless, it is important to discuss what we can expect if the restriction to linear transformations is dropped. From a nonparametric analysis [27], Petrov and Zhaoping concluded that higher-order correlations in general contribute only very little to the redundancy in natural images and, hence, are probably not the main cause for the receptive field properties in V1. The empirical support for this claim, however, is limited by the fact that their comparison is based on mutual information estimates within a very small neighborhood of five pixels only. This is problematic as it is known that many kinds of higher-order correlations in natural images become apparent only in much higher-dimensional statistics [64]. Furthermore, their estimate of the amount of second-order correlations is not invariant against pointwise nonlinear transformations of the pixel intensities.

In a more recent non-parametric study, Chandler and Field arrived at a very different result regarding the relative contribution of second-order and higher-order dependencies [29]. They use nearest-neighbor based methods to estimate the joint entropy of natural images in comparison to "spectrum-equalized" noise and white noise, where "spectrum-equalized" noise denotes Gaussian noise with exactly the same spectrum as that of natural images. As shown in Figure 18 of [29] they find a smaller difference between spectrum-equalized noise and white noise than between natural images and spectrum-equalized noise. Hence, from their finding, it seems that the amount of higher-order correlations in natural images is even larger than the amount of second-order

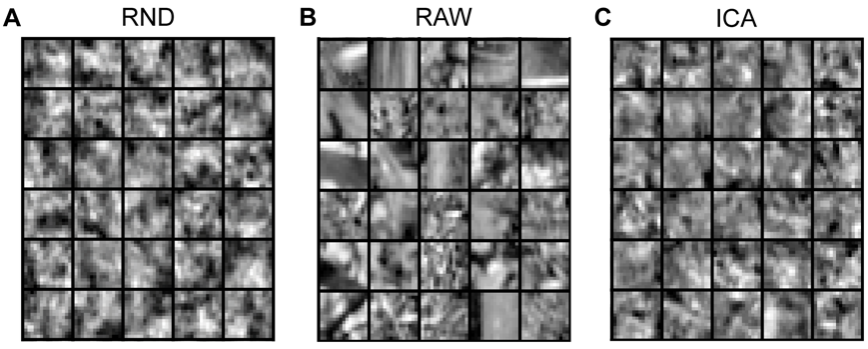


Figure 9. Comparison of Patches Sampled From Different Image Models. The figure demonstrates that the perceptual similarity between samples from the ICA image model (C) and samples from natural images (B) is not significantly increased relative to the perceptual similarity between samples from the RND image model (A) and (B).
doi:10.1371/journal.pcbi.1000336.g009

correlations. Also this result has to be taken with care: Reliable non-parametric estimates in high-dimensions are difficult to obtain even if one resorts to nearest-neighbor based methods, and the estimate of the amount of second-order correlations in [29] is not invariant against pointwise nonlinear transformations of the pixel intensities, too.

In summary, the present nonparametric studies do not give a unique answer regarding the total amount of higher-order correlations in natural images. Since estimating the absolute amount of multi-information is an extremely difficult task in high dimensions, the differences in the results can easily originate from the different assumptions and approximations made in these studies. Consequently, it remains an open question how large the true total redundancy of natural images is. In any case, it is clear that there are many higher-order redundancies in natural images that play a crucial role for visual perception. No matter how large these redundancies are in comparison to the second-order correlations, we need to develop better image models that have the right structure to capture these regularities.

What about Nonlinear Image Models?

Apart from the non-parametric approaches, a large number of nonlinear image models has been proposed over the years which are capable to capture significantly more statistical regularities of natural images than linear ICA can do (e.g., [62,65–72]). In fact, Olshausen and Field [8] already used a more general model than linear ICA when they originally derived the orientation selective filters from higher-order redundancy reduction. In contrast to plain ICA, they used an *overcomplete* generative model which assumes more source signals than pixel dimensions. In addition, the sources are modeled as latent variables like in a factor analysis model. That is the data is assumed to be generated according to $\mathbf{x}=A\mathbf{s}+\xi$ where A denotes the overcomplete dictionary, $\mathbf{s}$ is distributed according to a sparse factorial distribution, and ξ is a Gaussian random variable. The early quantitative study by Lewicki and Olshausen [24] could not demonstrate an advantage of overcomplete coding in terms of the rate-distortion trade-off and also the more recent work by Seeger [70] seems to confirm this conclusion. The addition of a Gaussian random variable ξ to $A\mathbf{s}$, however is likely to be advantageous as it may help to interpolate betweem the plain ICA model on the one hand and the spherically symmetric model on the other hand. A comparison of the average log-loss between this model and plain ICA has not been done yet but we can expect that this model can achieve a similar or even better match to the natural image statistics as the spherically symmetric model.

The spherical symmetric model can also be modeled by a redundancy reduction transformation which changes the radial component such that the output distribution is sought to match a Gaussian distribution [31]. Hence, the redundancy reduction of this model is very similar to the average log-loss of the spherically symmetric distribution. From a biological vision point of view, this type of model is particularly interesting as it allows one to draw a link to divisive normalization, a prominent contrast gain control mechanism observed for virtually all neurons in the early visual system. Our own ongoing work [30] shows that this idea can be generalized to a larger class of L_p-spherically symmetric distributions [67]. In this way, it is possible to find an optimal interpolation between ICA and the spherically symmetric case [73]. That is, one can combine orientation selectivity with divisive normalization in a joint model. Our preliminary results suggests that optimal divisive normalization together with orientation selectivity allows for about 10% improvement while divisive normalization alone (i.e., the spherical symmetric model) is only 2% worse [30].

Concluding Remarks

Taken together, the effect of orientation selectivity on redundancy reduction is very limited within the common linear filter bank model of V1 simple cells. In contrast to Zhaoping and coworkers, we do not claim that higher-order redundancy minimization is unlikely to be the main constraint in shaping the cortical receptive fields [22,27]. Our conclusion is that although there are significant higher-order correlations in natural images, orientation selective filtering turns out to be not very effective for capturing these. Nevertheless, we do expect that visual representations in the brain aim to model those higher-order correlations, because they are perceptually relevant. Therefore, we think it is important to further explore which type of nonlinear transformations would be suitable to capture more pronounced higher-order correlations. The objective functions studied in this paper are related to factorial coding, density estimation and minimization of the pixel mean square reconstruction error. Of course, there are also other alternatives that are interesting, too. For example, Zhaoping proposed that one possible goal of V1 is to explicitly represent bottom-up saliency in its neural responses for visual attentional selection [12]. As a further alternative, we are currently trying to extend the efficient coding framework to deal with other loss functions. Obviously, the goal of the visual system is not to preserve the pixel representation of the visual input. Instead, seeing serves the purpose to make successful predictions about behaviorally relevant aspects of the environment [74]. Since 3D shape inference is necessary to almost any naturally relevant task, it seems particularly interesting to explore the role of orientation selectivity in the context of 3D shape inference [75]. For a quantitative account of this problem one can seek to minimize the reconstruction error for the 3D shape rather than for its 2D image. Certainly, this task is much more involved than image reconstruction. Nevertheless, we need to think more about how to tackle the problem of visual inference within the framework of unsupervised learning in order to unravel the principles of neural processing in the brain that are ultimately responsible for our ability to see.

Supporting Information

Text S1 In the article we chose a patch size of 7×7 in order to enhance the comparability to previous work. The supplementary material contains all results (figures and tables) for patch size 15×15.

Found at: doi:10.1371/journal.pcbi.1000336.s001 (2.82 MB PDF)

Acknowledgments

We would like to thank Philipp Berens, Roland Fleming, Jakob Macke and Bruno Olshausen for fruitful discussions and helpful comments on the manuscript.

Author Contributions

Conceived and designed the experiments: MB. Performed the experiments: JE FS. Analyzed the data: JE FS. Contributed reagents/materials/analysis tools: JE FS MB. Wrote the paper: JE FS MB.

References

1. Attneave F (1954) Informational aspects of visual perception. Psychol Rev 61: 183–193.
2. Barlow H (1959) Sensory mechanisms, the reduction of redundancy, and intelligence. In: The Mechanisation of Thought Processes. London: Her Majesty's Stationery Office. pp 535–539.
3. Linsker R (1988) Self-organization in a perceptual network. Computer 21: 105–117.
4. Atick J (1992) Could information theory provide an ecological theory of sensory processing? Network 3: 213–251.
5. Barlow H (1989) Unsupervised learning. Neural Comput 1: 295–311.
6. Watanabe S (1981) Pattern recognition as a quest for minimum entropy. Pattern Recognit 13: 381–387.
7. Földiák (1990) Forming sparse representations by local anti-hebbian learning. Biol Cybern 64: 165–170.
8. Olshausen B, Field D (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381: 560–561.
9. Földiak P (1991) Learning invariance from transformation sequences. Neural Comput 3: 194–200.
10. Bialek W, Nemenman I, Tishby N (2001) Predictability, complexity, and learning. Neural Comput 13: 2409–2463.
11. Becker S, Hinton GE (1992) Self-organizing neural network that discovers surfaces in random-dot stereograms. Nature 355: 161–163.
12. Zhaoping L (2006) Theoretical understanding of the early visual processes by data compression and data selection. Network 17: 301–334.
13. Friedman JH, Stuetzle W, Schroeder A (1984) Projection pursuit density estimation. J Am Stat Assoc 19: 599–608.
14. Simoncelli E, Olshausen B (2001) Natural image statistics and neural representation. Annu Rev Neurosci 24: 1193–1216.
15. Buchsbaum G, Gottschalk A (1983) Trichromacy, opponent colours coding and optimum colour information transmission in the retina. Proc R Soc Lond B Biol Sci 220: 89–113.
16. Ruderman DL, Cronin TW, Chiao C (1998) Statistics of cone responses to natural images: implications for visual coding. J Opt Soc Am A 15: 2036–2045.
17. Atick J, Redlich A (1992) What does the retina know about natural scenes. Neural Comput 4: 196–210.
18. van Hateren J (1993) Spatiotemporal contrast sensitivity of early vision. Vision Res 33: 257–267.
19. Dong DW, Atick JJ (1995) Statistics of natural time-varying images. Network 6: 345–358.
20. Dan Y, Atick JJ, Reid RC (1996) Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory. J Neurosci 16: 3351–3362.
21. Hancock PJB, Baddeley RJ, Smith LS (1992) The principal components of natural images. Network 3: 61–70.
22. Li Z, Atick JJ (1994) Toward a theory of the striate cortex. Neural Comput 6: 127–146.
23. Bell A, Sejnowski T (1997) The "independent components" of natural scenes are edge filters. Vision Res 37: 3327–3338.
24. Lewicki M, Olshausen B (1999) Probabilistic framework for the adaptation and comparison of image codes. J Opt Soc Am A 16: 1587–1601.
25. Wachtler T, Lee TW, Sejnowski TJ (2001) Chromatic structure of natural scenes. J Opt Soc Am A 18: 65–77.
26. Lee TW, Wachtler T, Sejnowski TJ (2002) Color opponency is an efficient representation of spectral properties in natural scenes. Vision Res 42: 2095–2103.
27. Petrov Y, Zhaoping L (2003) Local correlations, information redundancy, and the sufficient pixel depth in natural images. J Opt Soc Am A 20: 56–66.
28. Bethge M (2006) Factorial coding of natural images: How effective are linear model in removing higher-order dependencies? J Opt Soc Am A 23: 1253–1268.
29. Chandler DM, Field DJ (2007) Estimates of the information content and dimensionality of natural scenes from proximity distributions. J Opt Soc Am A 24: 922–941.
30. Sinz FH, Bethge M (2008) How much can orientation selectivity and contrast gain control reduce the redundancies in natural images. Technical Report 169. Tübingen, Germany: Max Planck Institute for Biological Cybernetics.
31. Lyu S, Simoncelli EP (2008) Nonlinear image representation using divisive normalization. IEEE Conf Comput Vis Pattern Recognit 2008: 1–8.
32. Perez A (1977) ε-admissible simplification of the dependence structure of a set of random variables. Kybernetika 13: 439–444.
33. Cover T, Thomas J (1991) Elements of Information Theory. New York: J. Wiley & Sons.
34. Bernardo JM (1979) Expected information as expected utility. Ann Stat 7: 686–690.
35. Lewicki M, Sejnowski T (2000) Learning overcomplete representations. Neural Comput 12: 337–365.
36. Hulle MMV (2005) Mixture density modeling, kullback-leibler divergence, and differential log-likelihood. Signal Processing 85: 951–963.
37. Nadal JP, Parga N (1994) Nonlinear neurons in the low-noise limit: a factorial code maximizes information transfer. Network 5: 565–581.
38. Bell A, Sejnowski T (1995) An information maximisation approach to blind separation and blind deconvolution. Neural Comput 7: 1129–1159.
39. Goyal V (2001) Theoretical foundations of transform coding. IEEE Signal Process Mag 18: 9–21.
40. Gray R (1990) Entropy and Information Theory. New York: Springer.
41. Barlow H (2001) The exploitation of regularities in the environment by the brain. Behav Brain Sci 24: 602–607.
42. Wang Z, Bovic A, Lu L (2002) Why is image quality assessment so difficult? Proc IEEE Int Conf Acoust Speech Signal Process ICASSP 4: 3313–3316.
43. Gray R, Neuhoff D (1998) Quantization. IEEE Trans Inf Theory 44: 2325–2383.
44. Gish H, Pierce JN (1968) Asymptotically efficient quantizing. IEEE Trans Inf Theory 14: 676–683.
45. Fan K, Hoffman AJ (1955) Some metric inequalities in the space of matrices. Proc Am Math Soc 6: 111–116.
46. Srivastava A, Lee A, Simoncelli E, Zhu S (2003) On advances in statistical modeling of natural images. J Math Imaging Vis 18: 17–33.
47. Hyvärinen A, Karhunen J, Oja E (2001) Independent Component Analysis. New York: John Wiley & Sons.
48. Edelman A, Arias TA, Smith ST (1999) The geometry of algorithms with orthogonality constraints. SIAM J Matrix Anal Appl 20: 303–353.
49. Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) Numerical Recipes in C: The Art of Scientific Computing. New York: Cambridge University Press.
50. Maxwell JC (1855) Experiments on colour as perceived by the eye, with remarks on colour-blindness. Transactions of the Royal Society of Edinburgh XXI 2: 275–298.
51. Zetzsche C, Krieger G, Wegmann B (1999) The atoms of vision: Cartesian or polar? J Opt Soc Am A 16: 1554–1565.
52. Brelstaff GJ, Parraga A, Troscianko T, Carr D (1995) Hyperspectral camera system: acquisition and analysis. In: Proceedings of SPIE. Lurie BJ, Pearson JJ, Zilioli E, eds. Volume 2587, pp. 150–159. The database can be downloaded from: http://psy223.psy.bris.ac.uk/hyper/.
53. Baddeley R (1996) An efficient code in v1. Nature 381: 560–561.
54. Nadal JP, Brunel N, Parga N (1998) Nonlinear feedforward networks with stochastic outputs: infomax implies redundancy reduction. Network 9: 207–217.
55. Malsburg (1973) Self-organization of orientation sensitive cells in the striate cortex. Biol Cybern 14: 85–100.
56. Kaschube M, Schnabel M, Loewel S, Coppola D, White LE, et al. (2006) Universal pinwheel statistics in the visual cortex. In: Neuroscience Meeting Planner, 545.9/T11. Atlanta, Georgia: Society for Neuroscience.
57. Wolf F (2005) Symmetry, multistability, and long-range interactions in brain development. Phys Rev Lett 95: 208701.
58. Wimbauer, Wenisch, Miller, van Hemmen (1997) Development of spatiotemporal receptive fields of simple cells: I. model formulation. Biol Cybern 77: 453–461.
59. Horton JC, Adams DL (2005) The cortical column: a structure without a function. Philos Trans R Soc Lond B Biol Sci 360: 837–862.
60. Olshausen BA, Field DJ (2005) How close are we to understanding v1? Neural Comput 17: 1665–1699.
61. Masland RH, Martin PR (2007) The unsolved mystery of vision. Curr Biol 17: R577–R582.
62. Weiss Y, Freeman W (2007) What makes a good model of natural images? IEEE Conf Comput Vis Pattern Recognit 2007: 1–8.
63. Bethge M, Wiecki TV, Wichmann FA (2007) The independent components of natural images are perceptually dependent. In: Proceedings of SPIE. Volume 6492. pp A1–A12.
64. Bethge M, Berens P (2008) Near-maximum entropy models for binary neural representations of natural images. In: 21th Neural Information Processing Systems Conference. Platt JC, Koller D, Singer Y, Roweis S, eds. Cambridge, Massachusetts: MIT Press. pp 97–104.
65. Wainwright M, Simoncelli E (2000) Scale mixtures of Gaussians and the statistics of natural images. In: Advances in Neural Information Processing Systems (NIPS*99) 12. Solla S, Leen T, Müller KR, eds. Cambridge, Massachusetts: MIT Press. pp 855–861.
66. Karklin Y, Lewicki MS (2005) A hierarchical Bayesian model for learning nonlinear statistical regularities in nonstationary natural signals. Neural Comput 17: 397–423.
67. Hyvärinen A, Köster U (2007) Complex cell pooling and the statistics of natural images. Network 18: 81–100.
68. Osindero S, Hinton G (2008) Modeling image patches with a directed hierarchy of Markov random fields. In: Advances in Neural Information Processing Systems 20. Platt J, Koller D, Singer Y, Roweis S, eds. Cambridge, Massachusetts: MIT Press. pp 1121–1128.
69. Garrigues P, Olshausen B (2008) Learning horizontal connections in a sparse coding model of natural images. In: Platt J, Koller D, Singer Y, Roweis S, eds. Advances in Neural Information Processing Systems 20. Cambridge, MA: MIT Press. pp 505–512.
70. Seeger MW (2008) Bayesian inference and optimal design for the sparse linear model. J Mach Learn Res 9: 759–813.
71. Guerrero-Colón JA, Simoncelli EP, Portilla J (2008) Image denoising using mixtures of Gaussian scale mixtures. Proc Int Conf Image Proc 15: 565–568.

72. Hammond DK, Simoncelli EP (2008) Image modeling and denoising with orientation-adapted Gaussian scale mixtures. IEEE Trans Image Process 17: 2089–2101.
73. Sinz FH, Gerwinn S, Bethge M (2009) Characterization of the p-generalized normal distribution. J Multivar Anal 100: 817–820.
74. Helmholtz H (1878) The facts of perception. In: Selected Writings of Hermann Helmholtz. Kahl R, ed. Middletown, Connecticut: Wesleyan University Press.
75. Fleming RW, Torralba A, Adelson EH (2004) Specular reflections and the perception of shape. J Vis 4: 798–820.

4.2 Natural Image Coding in V1: How Much Use Is Orientation Selectivity: Supplementary Material

Additional Material

August 14, 2008

Figures and Tables for Patch Size 15×15

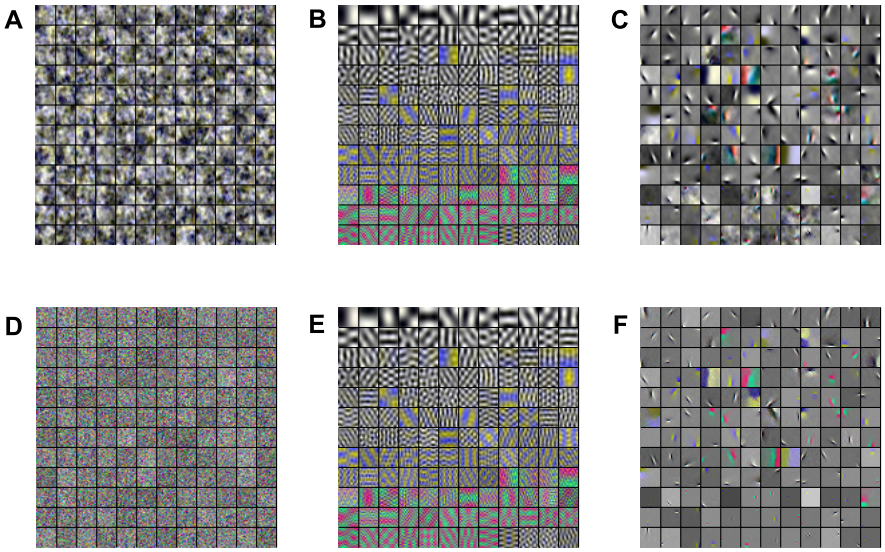


Figure 1: **Examples for Receptive Fields of Various Image Transforms** Basis functions of a random decorrelation transform (**RND**), principal component analysis (**PCA**) and independent component analysis (**ICA**) in pixel space (**A**-**C**) and whitened space (**E**-**F**). The image representation in whitened space is obtained by left multiplication with the matrix square root of the inverse covariance matrix r$C^{-1/2}$. This figure can only give a rough idea of the shape of the basis functions. For a detailed inspection of the basis functions we refer the reader to our web page `http://www.kyb.mpg.de/bethge/code/QICA/` where we provide all the data and code used in this paper.

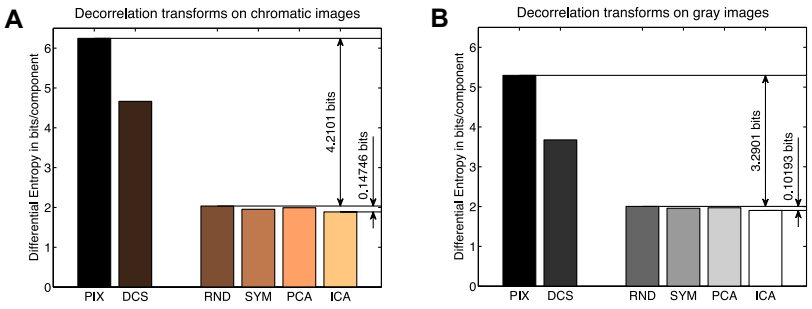


Figure 2: **Multi-Information Reduction per Dimension** Average differential entropy $\langle h \rangle$ for the pixel basis (PIX), after separation of the DC component (DCS), and after application of the different decorrelation transforms. The diagram shows the results for chromatic images (**A**) and the diagram for gray value images (**B**). For both types of images, only a marginal amount can be accounted to the reduction of higher order dependencies.

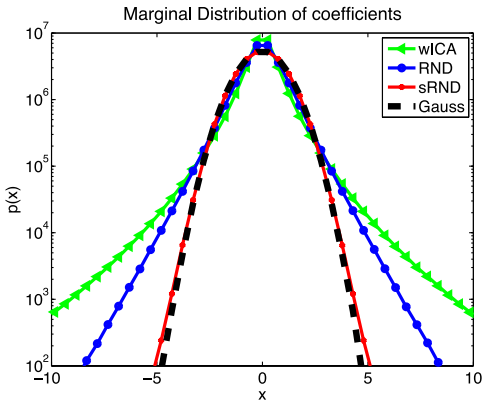


Figure 3: **The Distribution of Natural Images does not Conform with the Generative Model of ICA** In order to test for statistical dependencies among the coefficients Y_{wICA} of whithened ICA for single data samples, the coefficients were shuffled among the data points along each dimension. Subsequently, we transform the resulting data matrix Y_{sICA} into $Y_{\mathrm{sRND}} = W_{RND}W_{wICA}^{-1}Y_{\mathrm{sICA}}$. This corresponds to a change of basis from the ICA to the random decorrelation basis (RND). The plot shows the log-histogram over the coefficients over all dimensions. If the assumptions underlying ICA were correct, there would be no difference between the histogram of Y_{sRND} and Y_{RND}.

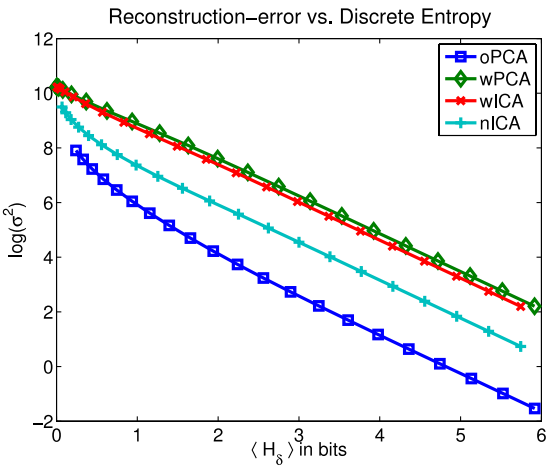


Figure 4: **Rate-distortion Curves** Rate-distortion curve for PCA and ICA when equalizing the output variances (wPCA and wICA) and when equalizing the norm of the corresponding image bases in pixel space (oPCA and nICA). The plot shows the discrete entropy H_δ in bits (averaged over all dimensions) against the log of the squared reconstruction error σ^2. oPCA outperforms all other transforms in terms of coding efficiency. wPCA in turn performed the worst and remarkably similar to wICA. Since wPCA and wICA differ only by an orthogonal transformation, both representations are bound to the same metric. oPCA is the only transformation which has the same metric as the pixel representation according to which the reconstruction error is determined. By normalizing the length of the ICA basis vectors in the pixel space, the metric of nICA becomes more similar to the pixel basis and the performance with respect to coding efficiency can be seen to improved considerably.

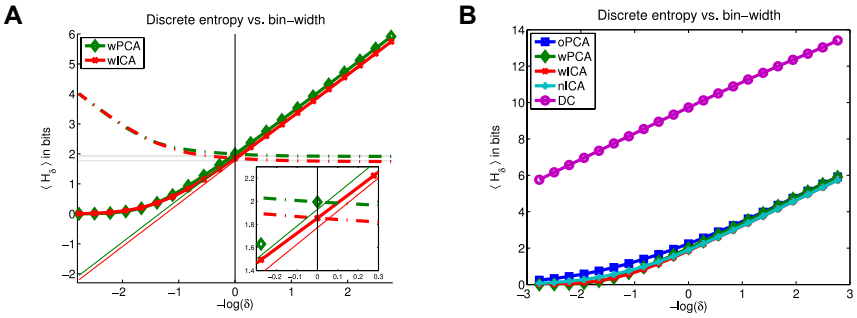


Figure 5: **Discrete vs. Differential Entropy** **A.** Relationship between discrete and differential entropy. Discrete entropy $\langle H_\delta \rangle$ averaged over all channels as a function of the negative log-bin-width. The straight lines constitute the linear approximation to the asymptotic branch of the function. Their interception with the y-axis are visualized by the gray shaded, horizontal lines. The dashed lines represent $\langle h_\delta \rangle$ which converge to the gray shaded lines for $\delta \to 0$. **B.** There are only small differences in the average discrete entropy for oPCA, wPCA, wICA, nICA as a function of the negative log-bin-width. Since the discrete entropy of the DC component is the same for all transforms, it is not included in that average but plotted separately instead.

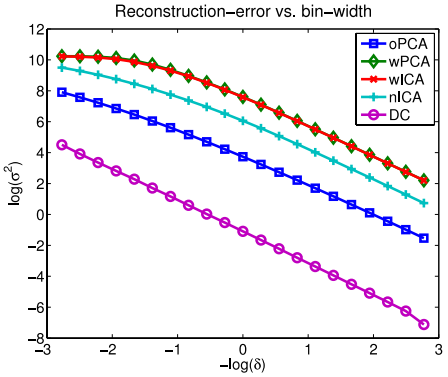


Figure 6: **Reconstruction Error vs. Bin Width of Discrete Entropy** Reconstruction error σ^2 as a function of the bin-width δ, shown on a logarithmic scale. The differences between the different transforms are relatively large. Only the two transformations with exactly the same metric, wPCA and wICA, exhibit no difference in the reconstruction error.

Absolute Difference			**Relative Difference**		
	Color	Gray		Color	Gray
RND-PIX	-4.2101 ± 0.0020	-3.2901 ± 0.0019			
SYM-RND	-4.2915 ± 0.0023	-3.3360 ± 0.0022	$\frac{\text{SYM-RND}}{\text{SYM-PIX}}$	1.90 ± 0.01	1.37 ± 0.01
PCA-RND	-4.2534 ± 0.0022	-3.3239 ± 0.0022	$\frac{\text{PCA-RND}}{\text{PCA-PIX}}$	1.02 ± 0.01	1.01 ± 0.01
ICA-RND	-4.3575 ± 0.0024	-3.3921 ± 0.0026	$\frac{\text{ICA-RND}}{\text{ICA-PIX}}$	3.38 ± 0.02	3.01 ± 0.02

Table 1: **Comparision of the Multi-Information Reduction for Chromatic and Achromatic Images** Differences in the multi-information reduction between various decorrelation transforms (SYM, PCA, ICA) relative to a random decorrelation transform (RND) compared to the multi-information reduction achieved with the random decorrelation transform relative to the original pixel basis (RND-PIX). The *absolute* multi-information reduction is given in bits/component on the left hand side. How much more the special decorrelation transforms SYM, PCA and ICA can reduce the multi-information *relative* to the random (RND) one is given in percent on the right hand side.

	Color		**Gray**	
A	ALL		ALL	
RND	1.9925 ± 0.0041		1.9685 ± 0.0038	
SYM-RND	-0.1203 ± 0.0007		-0.0682 ± 0.0005	
PCA-RND	-0.0511 ± 0.0004		-0.0364 ± 0.0005	
ICA-RND	-0.1829 ± 0.0009		-0.1191 ± 0.0010	
SSD-RND	-0.2461 ± 0.0022		-0.2742 ± 0.0030	
B	DLL	$\langle\alpha\rangle$	DLL	$\langle\alpha\rangle$
RND	-0.0086 ± 0.0002	1.1273 ± 0.0039	-0.0060 ± 0.0004	1.0811 ± 0.0034
SYM	-0.0472 ± 0.0005	0.9034 ± 0.0027	-0.0282 ± 0.0006	0.9535 ± 0.0032
PCA	-0.0162 ± 0.0003	1.0229 ± 0.0033	-0.0085 ± 0.0005	1.0100 ± 0.0031
ICA	-0.0434 ± 0.0004	0.7540 ± 0.0019	-0.0227 ± 0.0007	0.8237 ± 0.0025

Table 2: **Comparision of the Average Log-Loss (ALL) and the Differential Log-Likelihood (DLL) Chromatic and Achromatic Images** **A.** The first row shows the average log-loss (ALL, in bits/component) of the density model determined by the linear transformation RND. The value was obtained by averaging over 10 separately sampled training and test sets of size 40.000 and 50.000, respectively. The following rows shows the difference of the ALL of the models SYM, PCA, ICA and the spherically symmetric density (SSD) to the ALL determined by linear transformation RND. The large value for RND−ICA fundamentally contradicts the assumptions underlying the ICA model. **B.** The small DLL values suggest, that the exponential power distribution fits the shape of the individual coefficient distributions well. In addition, we also report the average exponent $\langle\alpha\rangle$ of the exponential power family fit to the individual coefficient distributions ($\alpha = 1$ corresponds to a Laplacian shape).

Multi Information for Small Patches

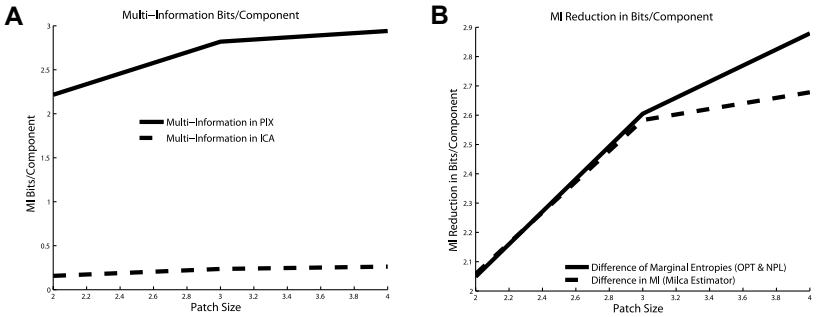


Figure 7: **Multi-Information for Small Patches** **A** Multi-information for patch sizes 2×2, 3×3 and 4×4 in the representations PIX and ICA. **B** Multi-information reduction as estimated by the multi-information from the left plot and by the differences in the marginal entropies. For patch size 4×4, the estimations start to disagree. Since the multi-information is much harder to estimate than the marginal entropies, we conclude that from patch size 4×4 on, the multi-information estimates are not reliable anymore.

	2×2	3×3	4×4
PIX	2.2157	2.8193	2.9405
ICA	0.1573	0.2358	0.2622

Table 3: **Multi-Information for Small Patch Sizes** The table shows the multi-information in the representations PIX and ICA in bits/pixel as computed with the estimator from the MILCA package by Kraskov.

4.3 The Conjoint Effect of Divisive Normalization and Orientation Selectivity on Redundancy Reduction: Original Article

The Conjoint Effect of Divisive Normalization and Orientation Selectivity on Redundancy Reduction in Natural Images

Fabian Sinz
MPI for Biological Cybernetics
72076 Tübingen, Germany
fabee@tuebingen.mpg.de

Matthias Bethge
MPI for Biological Cybernetics
72076 Tübingen, Germany
mbethge@tuebingen.mpg.de

Abstract

Bandpass filtering, orientation selectivity, and contrast gain control are prominent features of sensory coding at the level of V1 simple cells. While the effect of bandpass filtering and orientation selectivity can be assessed within a linear model, contrast gain control is an inherently nonlinear computation. Here we employ the class of L_p elliptically contoured distributions to investigate the extent to which the two features—orientation selectivity and contrast gain control—are suited to model the statistics of natural images. Within this framework we find that contrast gain control can play a significant role for the removal of redundancies in natural images. Orientation selectivity, in contrast, has only a very limited potential for redundancy reduction.

1 Introduction

It is a long standing hypothesis that sensory systems are adapted to the statistics of their inputs. These natural signals are by no means random, but exhibit plenty of regularities. Motivated by information theoretic principles, Attneave and Barlow suggested that one important purpose of this adaptation in sensory coding is to model and reduce the redundancies [4; 3] by transforming the signal into a statistically independent representation.

The problem of redundancy reduction can be split into two parts: (i) finding a good statistical model of the natural signals and (ii) a way to map them into a factorial representation. The first part is relevant not only to the study of biological systems, but also to technical applications such as compression and denoising. The second part offers a way to link neural response properties to computational principles, since neural representations of natural signals must be advantageous in terms of redundancy reduction if the hypothesis were true. Both aspects have been extensively studied for natural images [2; 5; 8; 19; 20; 21; 24]. In particular, it has been shown that applying Independent Component Analysis (ICA) to natural images consistently and robustly yields filters that are localized, oriented and show bandpass characteristics [19; 5]. Since those features are also ascribed to the receptive fields of neurons in the primary visual cortex (V1), it has been suggested that the receptive fields of V1 neurons are shaped to form a minimally redundant representation of natural images [5; 19].

From a redundancy reduction point of view, ICA offers a small but significant advantage over other linear representations [6]. In terms of density estimation, however, it is a poor model for natural images since already a simple non-factorial spherically symmetric model yields a much better fit to the data [10].

Recently, Lyu et al. proposed a method that converts any spherically symmetric distribution into a (factorial) Gaussian (or Normal distribution) by using a non-linear transformation of the norm of

the image patches [17]. This yields a non-linear redundancy reduction mechanism, which exploits the superiority of the spherically symmetric model over ICA. Interestingly, the non-linearity of this Radial Gaussianization method closely resembles another feature of the early visual system, known as contrast gain control [13] or divisive normalization [20]. However, since spherically symmetric models are invariant under orthogonal transformations, they are agnostic to the particular choice of basis in the whitened space. Thus, there is no role for the shape of the filters in this model.

Combining the observations from the two models of natural images, we can draw two conclusions: On the one hand, ICA is not a good model for natural images, because a simple spherically symmetric model yields a much better fit [10]. On the other hand, the spherically symmetric model in Radial Gaussianization cannot capture that ICA filters do yield a higher redundancy reduction than other linear transformations. This leaves us with the questions whether we can understand the emergence of oriented filters in a more general redundancy reduction framework, which also includes a mechanism for contrast gain control.

In this work we address this question by using the more general class of L_p-spherically symmetric models [23; 12; 15]. These models are quite similar to spherically symmetric models, but do depend on the particular shape of the linear filters. Just like spherically symmetric models can be non-linearly transformed into isotropic Gaussians, L_p-spherically symmetric models can be mapped into a unique class of factorial distributions, called p-generalized Normal distributions [11]. Thus, we are able to quantify the influence of orientation selective filters and contrast gain control on the redundancy reduction of natural images in a joint model.

2 Models and Methods

2.1 Decorrelation and Filters

All probabilistic models in this paper are defined on whitened natural images. Let $\boldsymbol{C}$ be the covariance matrix of the pixel intensities for an ensemble $\boldsymbol{x}_1, ..., \boldsymbol{x}_m$ of image patches, then $\boldsymbol{C}^{-\frac{1}{2}}$ constitutes the symmetric whitening transform. Note that all vectors $\boldsymbol{y} = \boldsymbol{V}\boldsymbol{C}^{-\frac{1}{2}}\boldsymbol{x}$, with $\boldsymbol{V}$ being an orthogonal matrix, have unit covariance. $\boldsymbol{V}\boldsymbol{C}^{-\frac{1}{2}}$ yield the linear filters that are applied to the raw image patches before feeding them in the probabilistic models described below. Since any decorrelation transform can be written as $\boldsymbol{V}\boldsymbol{C}^{-\frac{1}{2}}$, the choice of $\boldsymbol{V}$ determines the shape of the linear filters. In our experiments, we use three different kinds of $\boldsymbol{V}$:

SYM The simplest choice is $\boldsymbol{V}_{\text{SYM}} = \boldsymbol{I}$, i. e. $\boldsymbol{y} = \boldsymbol{C}^{-\frac{1}{2}}\boldsymbol{x}$ contains the coefficients in the symmetric whitening basis. From a biological perspective, this case is interesting as the filters resemble receptive fields of retinal ganglion cells with center-surround properties.

ICA The filters $\boldsymbol{V}_{\text{ICA}}$ of ICA are determined by maximizing the non-Gaussanity of the marginal distributions. For natural image patches, ICA is known to yield orientation selective filters in resemblance to V1 simple cells. While other orientation selective bases are possible, the filters defined by $\boldsymbol{V}_{\text{ICA}}$ correspond to the optimal choice for redundancy reduction under the restriction to linear models.

HAD The coefficients in the basis $\boldsymbol{V}_{\text{HAD}} = \frac{1}{\sqrt{m}}\boldsymbol{H}\boldsymbol{V}_{\text{ICA}}$, with $\boldsymbol{H}$ denoting an arbitrary Hadamard matrix, correspond to a sum over the different ICA coefficients, each possibly having a flipped sign. Hadamard matrices are defined by the two properties $H_{ij} = \pm 1$ and $\boldsymbol{H}\boldsymbol{H}^\top = m\boldsymbol{I}$. This case can be seen as the opposite extreme to the case of ICA. Instead of running an independent search for the most Gaussian marginals, the central limit theorem is used to produce the most Gaussian components by using the Hadamard transformation to mix all ICA coefficients with equal weight resorting to the independence assumption underlying ICA.

2.2 L_p-spherically Symmetric Distributions

The contour lines of spherically symmetric distributions have constant Euclidean norm. Similarly, the contour lines of L_p-spherically symmetric distributions have constant p-norm[1] $||\boldsymbol{y}||_p :=$

[1]Note that $||\boldsymbol{y}||_p$ is only a norm in the strict sense if $p \geq 1$. However, since the following considerations also hold for $0 < p < 1$, we will employ the term "p-norm" and the notation "$||\boldsymbol{y}||_p$" for notational convenience.

$\sqrt[p]{\sum_{i=1}^{n} |y_i|^p}$ The set of vectors with constant p-norm $\mathbb{S}_p^{n-1}(r) := \{\boldsymbol{y} \in \mathbb{R}^n : \ ||\boldsymbol{y}||_p = r, \ p > 0, \ r > 0\}$ is called p-sphere of radius r. Different examples of p-spheres are shown along the coordinate axis of Figure 1. For $p \neq 2$ the distribution is not invariant under arbitrary orthogonal transformations, which means that the choice of the basis $\boldsymbol{V}$ can make a difference in the likelihood of the data.

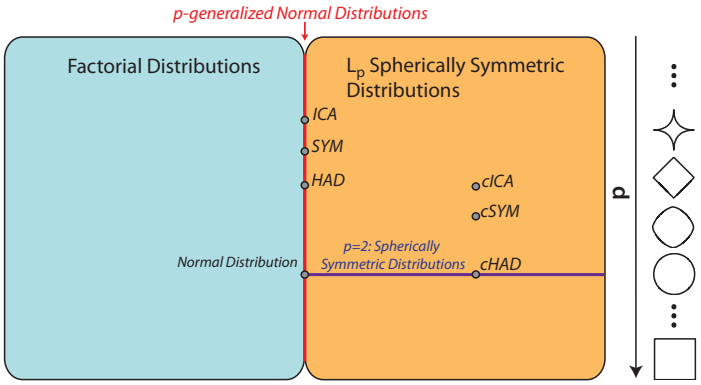


Figure 1: The spherically symmetric distributions are a subset of the L_p-spherical symmetric distributions. The right shapes indicate the iso-density lines for the different distributions. The Gaussian is the only L_2-spherically symmetric distribution with independent marginals. Like the Gaussian distribution, all p-generalized Normal distributions have independent marginals. *ICA, SYM, ...* denote the models used in the experiments below.

A multivariate random variable Y is called L_p-spherically symmetric distributed if it can be written as a product $Y = RU$, where U is uniformly distributed on $\mathbb{S}_p^{n-1}(1)$ and R is a univariate non-negative random variable with an arbitrary distribution [23; 12]. Intuitively, R corresponds to the radial component, i. e. the length $||\boldsymbol{y}||_p$ measured with the p-norm. U describes the directional components in a polar-like coordinate system (see Extra Material). It can be shown that this definition is equivalent to the density $\varrho(\boldsymbol{y})$ of Y having the form $\varrho(\boldsymbol{y}) = f(||\boldsymbol{y}||_p^p)$ [12]. This immediately suggests two ways of constructing an L_p-spherically symmetric distribution. Most obviously, one can specify a density $\varrho(\boldsymbol{y})$ that has the form $\varrho(\boldsymbol{y}) = f(||\boldsymbol{y}||_p^p)$. An example is the p-generalized Normal distribution (gN) [11]

$$\varrho(\boldsymbol{y}) \quad = \quad \frac{p^n}{\Gamma^n\left(\frac{1}{p}\right)(2\sigma^2)^{\frac{n}{p}}2^n} \exp\left(-\frac{\sum_{i=1}^{n}|y_i|^p}{2\sigma^2}\right) = f(||\boldsymbol{y}||_p^p). \tag{1}$$

Analogous to the Gaussian being the only factorial spherically symmetric distribution [1], this distribution is the only L_p-spherically symmetric distribution with independent marginals [22]. For the p-generalized Normal, the marginals are members of the exponential power family.

In our experiments, we will use the p-generalized Normal to model linear marginal independence by fitting it to the coefficients of the various bases in whitened space. Since this distribution is sensitive to the particular filter shapes for $p \neq 2$, we can assess how well the distribution of the linearly transformed image patches is matched by a factorial model.

An alternative way of constructing an L_p-spherically symmetric distribution is to specify the radial distribution ϱ_r. One example, which will be used later, is obtained by choosing a mixture of Log-Normal distributions (RMixLogN). In Cartesian coordinates, this yields the density

$$\varrho(\boldsymbol{y}) \quad = \quad \frac{p^{n-1}\Gamma\left(\frac{n}{p}\right)}{2^n\Gamma^n\left(\frac{1}{p}\right)} \sum_{k=1}^{K} \frac{\eta_k}{||\boldsymbol{y}||_p^n \sigma_k \sqrt{2\pi}} \exp\left(-\frac{(\log||\boldsymbol{y}||_p - \mu_k)^2}{2\sigma_k^2}\right). \tag{2}$$

An immediate consequence of any L_p-spherically symmetric distribution being specified by its radial density is the possibility to change between any two of those distributions by transforming the radial component with $(\mathcal{F}_2^{-1} \circ \mathcal{F}_1)(||\boldsymbol{y}||_p)$, where $\mathcal{F}_1$ and $\mathcal{F}_2$ are cumulative distribution functions (cdf) of the source and the target density, respectively. In particular, for a fixed p, any L_p-spherically symmetric distribution can be transformed into a factorial one by the transform

$$\boldsymbol{z} = g(\boldsymbol{y}) \cdot \boldsymbol{y} = \frac{(\mathcal{F}_2^{-1} \circ \mathcal{F}_1)(||\boldsymbol{y}||_p)}{||\boldsymbol{y}||_p} \boldsymbol{y}.$$

This transform closely resembles contrast gain control models for primary visual cortex [13; 20], which use a different gain function having the form $\tilde{g}(\boldsymbol{y}) = \frac{1}{c+r}$ with $r = ||\boldsymbol{y}||_2^2$ [17].

We will use the distribution of equation (2) to describe the joint model consisting of a linear filtering step followed by a contrast gain control mechanism. Once, the linear filter responses in whitened space are fitted with this distribution, we non-linearly transform it into a the factorial p-generalized Normal by the transformation $g(\boldsymbol{y}) \cdot \boldsymbol{y} = (\mathcal{F}_{\mathsf{gN}}^{-1} \circ \mathcal{F}_{\mathsf{RMixLogN}})(||\boldsymbol{y}||_p)/||\boldsymbol{y}||_p \cdot \boldsymbol{y}$.

Finally, note that because a L_p-spherically symmetric distribution is specified by its univariate radial distribution, fitting it to data boils down to estimating the univariate density for R, which can be done efficiently and robustly.

3 Experiments and Results

3.1 Dataset

We use the dataset from the Bristol Hyperspectral Images Database [7], which was already used in previous studies [25; 16]. All images had a resolution of 256×256 pixels and were converted to gray level by averaging over the channels. From each image circa 5000 patches of size 15×15 pixels were drawn at random locations for training (circa 40000 patches in total) as well as circa 6250 patches per image for testing (circa 50000 patches in total). In total, we sampled ten pairs of training and test sets in that way. All results below are averaged over those. Before computing the linear filters, the DC component was projected out with an orthogonal transformation using a QR decomposition. Afterwards, the data was rescaled in order to make whitening a volume conserving transformation (a transformation with determinant one) since those transformations leave the entropy unchanged.

3.2 Evaluation Measure

In all our experiments, we used the Average Log Loss (ALL) to assess the quality of the fit and the redundancy reduction achieved. The $\mathrm{ALL} = \frac{1}{n}\mathbb{E}_\varrho[-\log_2 \hat{\varrho}(\boldsymbol{y})] \approx \frac{1}{mn}\sum_{k=1}^{m} -\log_2 \hat{\varrho}(\boldsymbol{y})$ is the negative mean log-likelihood of the model distribution under the true distribution. If the model distribution matches the true one, the ALL equals the entropy. Otherwise, the difference between the ALL and the entropy of the true distribution is exactly the Kullback-Leiber divergence between the two. The difference between the ALLs of two models equals the reduction in multi-information (see Extra Material) and can therefore be used to quantify the amount of redundancy reduction.

3.3 Experiments

We fitted the L_p-spherically symmetric distributions from equations (1) and (2) to the image patches in the bases HAD, SYM, and ICA by a maximum likelihood fit on the radial component. For the mixture of Log-Normal distributions, we used EM for a mixture of Gaussians on the logarithm of the p-norm of the image patches.

For each model, we computed the maximum likelihood estimate of the model parameters and determined the best value for p according to the ALL in bits per component on a training set. The final ALL was computed on a separate test set.

For ICA, we performed a gradient descent over the orthogonal group on the log-likelihood of a product of independent exponential power distributions, where we used the result of the FastICA algorithm by Hyvärinen et al. as initial starting point [14]. All transforms were computed separately for each training set.

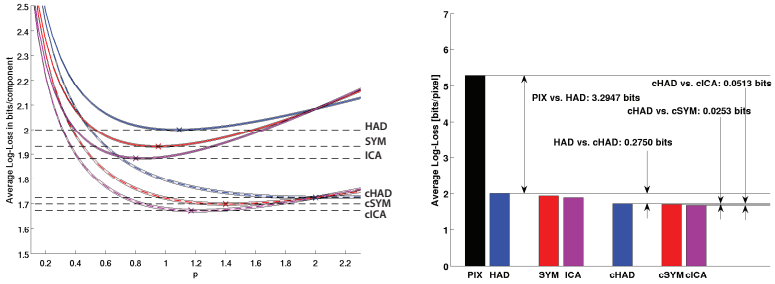


Figure 2: ALL in bits per component as a function of p. The linewidth corresponds to the standard deviation over ten pairs of training and test sets. *Left:* ALL for the bases HAD, SYM and ICA under the p-generalized Normal (HAD, SYM, ICA) and the factorial L_p-spherically symmetric model with the radial component modeled by a mixture of Log-Normal distributions (cHAD, cSYM, cICA). *Right:* Bar plot for the different ALL indicated by horizontal lines in the left plot.

In order to compare the redundancy reduction of the different transforms with respect to the pixel basis (PIX), we computed a non-parametric estimate of the marginal entropies of the patches before the DC component was projected out [6]. Since the estimation is not bound to a particular parametric model, we used the mean of the marginal entropies as an estimate of the average log-loss in the pixel representation.

3.4 Results

Figure 2 and Table 1 show the ALL for the bases HAD, SYM, and ICA as a function of p. The upper curve bundle represents the factorial p-generalized Normal model, the lower bundle the non-factorial model with the radial component modeled by a mixture of Log-Normal distributions with five mixtures. The ALL for the factorial models always exceeds the ALL for the non-factorial models. At $p = 2$, all curves intersect, because all models are invariant under a change of basis for that value. Note that the smaller ALL of the non-factorial model cannot be attributed to the mixture of Log-Normal distributions having more degrees of freedom. As mentioned in the introduction, the p-generalized Normal is the only factorial L_p-spherically symmetric distribution [22]. Therefore, marginal independence is such a rigid assumption that the output scale is the only degree of freedom left.

From the left plot in Figure 2, we can assess the influence of the different filter shapes and contrast gain control on the redundancy reduction of natural images. We used the best ALL of the HAD basis under the p-generalized Normal as a baseline for a whitening transformation without contrast gain control (HAD). Analogously, we used the best ALL of the HAD basis under the non-factorial model as a baseline for a pure contrast gain control model (cHAD). We compared these values to the best ALL obtained by using the SYM and the ICA basis under both models. Because the filters of SYM and ICA resemble receptive field properties of retinal ganglion cells and V1 simple cells, respectively, we can assess their possible influence on the redundancy reduction with and without contrast gain control. The factorial model corresponds to the case without contrast gain control (SYM and ICA). Since we have shown that the non-factorial model can be transformed into a factorial one by a p-norm based divisive normalization operation, these scores correspond to the cases with contrast gain control (cSYM and cICA). The different cases are depicted by the horizontal lines in Figure 2.

As already reported in other works, plain orientation selectivity adds only very little to the redundancy reduction achieved by decorrelation and is less effective than the baseline contrast gain control model [10; 6; 17]. If both orientation selectivity and contrast gain control are combined (cICA) it is possible to achieve about 9% extra redundancy reduction in addition to baseline whitening

Absolute Difference [Bits/Comp.]		Relative Difference [% wrt. cICA]
HAD - PIX	-3.2947 ± 0.0018	91.0016 ± 0.0832
SYM- PIX	-3.3638 ± 0.0022	92.9087 ± 0.0782
ICA - PIX	-3.4110 ± 0.0024	94.2135 ± 0.0747
cHAD - PIX	-3.5692 ± 0.0045	98.5839 ± 0.0134
cSYM - PIX	-3.5945 ± 0.0047	99.2815 ± 0.0098
cICA - PIX	-3.6205 ± 0.0049	100.0000 ± 0.0000

Table 1: Difference in ALL for gray value images with standard deviation over ten training and test set pairs. The column on the left displays the absolute difference to the PIX representation. The column on the right shows the relative difference with respect to the largest reduction achieved by ICA with non-factorial model.

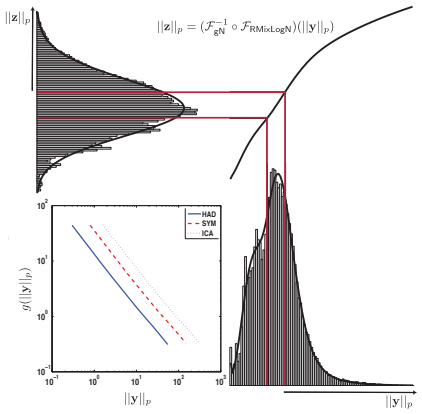


Figure 3: The curve in the upper right corner depicts the transformation $||\boldsymbol{z}||_p = (\mathcal{F}_{\mathsf{gN}}^{-1} \circ \mathcal{F}_{\mathsf{RMixLogN}})(||\boldsymbol{y}||_p)$ of the radial component in the ICA basis for gray scale images. The resulting radial distribution over $||\boldsymbol{z}||_p$ corresponds to the radial distribution of the p-generalized Normal. The inset shows the gain function $g(||\boldsymbol{y}||_p) = \frac{\mathcal{F}_{\mathsf{RMixLogN}}(||\boldsymbol{y}||_p)}{||\boldsymbol{y}||_p}$ in log-log coordinates. The scale parameter of the p-generalized normal was chosen such that the marginal had unit variance.

(HAD). By setting the other models in relation to the best joint model (cICA:$= 100\%$), we are able to tell apart the relative contributions of bandpass filtering (HAD$= 91\%$), particular filter shapes (SYM$= 93\%$, ICA$= 94\%$), contrast gain control (cHAD$= 98.6\%$) as well as combined models (cSYM$= 99\%$, cICA $:= 100\%$) to redundancy reduction (see Table 1). Thus, orientation selectivity (ICA) contributes less to the overall redundancy reduction than any model with contrast gain control (cHAD, cSYM, cICA). Additionally, the relative difference between the joint model (cICA) and plain contrast gain control (cHAD) is only about 1.4%. For cSYM it is even less, about 0.7%. The difference in redundancy reduction between center-surround filters and orientation selective filters becomes even smaller in combination with contrast gain control (1.3% for ICA vs. SYM, 0.7% for cICA vs. cSYM). However, it is still significant (t-test, $p = 5.5217 \cdot 10^{-9}$).

When examining the gain functions $g(||\boldsymbol{y}||_p) = \frac{(\mathcal{F}_{\mathsf{gN}}^{-1} \circ \mathcal{F}_{\mathsf{RMixLogN}})(||\boldsymbol{y}||_p)}{||\boldsymbol{y}||_p}$ resulting from the transformation of the radial components, we find that they approximately exhibit the form $g(||\boldsymbol{y}||_p) = \frac{c}{||\boldsymbol{y}||_p^{\kappa}}$. The inset in Figure 3 shows the gain control function $g(||\boldsymbol{y}||_p)$ in a log-log plot. While standard contrast gain control models assume $p = 2$ and $\kappa = 2$, we find that κ between 0.90 and 0.93 to be optimal for redundancy reduction. p depends on the shape of the linear filters and ranges from approximately 1.2 to 2. In addition, existing contrast gain models assume the form $g(||\boldsymbol{y}||_2) = \frac{1}{\sigma + ||y||_2^2}$, while we find that σ must be approximately zero.

In the results above, the ICA filters always achieve the lowest ALL under both p-spherically symmetric models. For examining whether these filters really represent the best choice, we also optimized the filter shapes under the model of equation (2) via maximum likelihood estimation on the orthogonal group in whitened space [9; 18]. Figure 4 shows the filter shapes for ICA and the ones obtained from the optimization, where we used either the ICA solution or a random orthogonal matrix as starting point. Qualitatively, the filters look exactly the same. The ALL also changed just

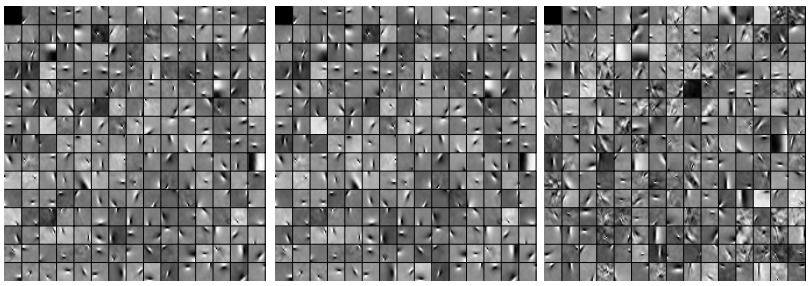


Figure 4: Filters optimized for ICA (*left*) and for the p-spherically symmetric model with radial mixture of Log-Normal distributions starting from the ICA solution (*middle*) and from a random basis (*right*). The first filter corresponds to the DC component, the others to the filter shapes under the respective model. Qualitatively the filter shapes are very similar. The ALL for the ICA basis under the mixture of Log-Normal model is 1.6748 ± 0.0058 bits/component (*left*), the ALL with the optimized filters is 1.6716 ± 0.0056 (*middle*) and 1.6841 ± 0.0068 (*right*).

marginally from 1.6748 ± 0.0058 to 1.6716 ± 0.0056 or 1.6841 ± 0.0068, respectively. Thus, the ICA filters are a stable and optimal solution under the model with contrast gain control, too.

4 Summary

In this report, we studied the conjoint effect of contrast gain control and orientation selectivity on redundancy reduction for natural images. In particular, we showed how the L_p-spherically distribution can be used to tune a nonlinearity of contrast gain control to remove higher-order redundancies in natural images.

The idea of using an L_p-spherically symmetric model for natural images has already been brought up by Hyvärinen and Köster in the context of Independent Subspace Analysis [15]. However, they do not use the L_p-distribution for contrast gain control, but apply a global contrast gain control filter on the images before fitting their model. They also use a less flexible L_p-distribution since their goal is to fit an ISA model to natural images and not to carry out a quantitative comparison as we did.

In our work, we find that the gain control function turns out to follow a power law, which parallels the classical model of contrast gain control. In addition, we find that edge filters also emerge in the non-linear model which includes contrast gain control. The relevance of orientation selectivity for redundancy reduction, however, is further reduced. In the linear framework (possibly endowed with a point-wise nonlinearity for each neuron) the contribution of orientation selectivity to redundancy reduction has been shown to be smaller than 5% relative to whitening (i. e. bandpass filtering) alone [6; 10]. Here, we found that the contribution of orientation selectivity is even smaller than two percent relative to whitening plus gain control. Thus, this quantitative model comparison provides further evidence that orientation selectivity is not critical for redundancy reduction, while contrast gain control may play a more important role.

Acknowledgements

The authors would like to thank Reshad Hosseini, Sebastian Gerwinn and Philipp Berens for fruitful discussions. This work is supported by the German Ministry of Education, Science, Research and Technology through the Bernstein award to MB (BMBF; FKZ: 01GQ0601), a scholarship of the German National Academic Foundation to FS, and the Max Planck Society.

References

[1] S. F. Arnold and J. Lynch. On Ali's characterization of the spherical normal distribution. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(1):49–51, 1982.

[2] J. J. Atick. Could information theory provide an ecological theory of sensory processing? *Network*, 3:213–251, 1992.

[3] F. Attneave. Informational aspects of visual perception. *Psychological Review*, 61:183–193, 1954.

[4] H. B. Barlow. Sensory mechanisms, the reduction of redundancy, and intelligence. In *The Mechanisation of Thought Processes*, pages 535–539, London: Her Majesty's Stationery Office, 1959.

[5] A. J. Bell and T. J. Sejnowski. The "independent components" of natural scenes are edge filters. *Vision Res.*, 37(23):3327–38, 1997.

[6] M. Bethge. Factorial coding of natural images: How effective are linear model in removing higher-order dependencies? *J. Opt. Soc. Am. A*, 23(6):1253–1268, June 2006.

[7] G. J. Brelstaff, A. Parraga, T. Troscianko, and D. Carr. Hyperspectral camera system: acquisition and analysis. In B. J. Lurie, J. J. Pearson, and E. Zilioli, editors, *Proceedings of SPIE*, volume 2587, pages 150–159, 1995. The database can be downloaded from: `http://psy223.psy.bris.ac.uk/hyper/`.

[8] G. Buchsbaum and A. Gottschalk. Trichromacy, opponent colours coding and optimum colour information transmission in the retina. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 220:89–113, November 1983.

[9] A. Edelman, T. A. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, 20(2):303–353, 1999.

[10] J. Eichhorn, F. Sinz, and M. Bethge. Simple cell coding of natural images in V1: How much use is orientation selectivity? (arxiv:0810.2872v1). 2008.

[11] I. R. Goodman and S. Kotz. Mutltivariate θ-generalized normal distributions. *Journal of Multivariate Analysis*, 3:204–219, 1973.

[12] A. K. Gupta and D. Song. l_p-norm spherical distribution. *Journal of Statistical Planning and Inference*, 60:241–260, 1997.

[13] D. J. Heeger. Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, 9:181–198, 1992.

[14] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, 2001.

[15] A. Hyvärinen and U. Köster. Complex cell pooling and the statistics of natural images. *Network*, 18:81–100, 2007.

[16] T.-W. Lee, T. Wachtler, and T. J. Sejnowski. Color opponency is an efficient representation of spectral properties in natural scenes. *Vision Res*, 42(17):2095–2103, Aug 2002.

[17] S. Lyu and E. P. Simoncelli. Nonlinear image representation using divisive normalization. In *Proc. Computer Vision and Pattern Recognition*, June 2008. To Appear.

[18] J. H. Manton. Optimization algorithms exploiting unitary constraints. *IEEE Transactions on Signal Processing*, 50:635 – 650, 2002.

[19] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, June 1996.

[20] O. Schwartz and E. P. Simoncelli. Natural signal statistics and sensory gain control. *Nature Neuroscience*, 4(8):819–825, August 2001.

[21] E. P. Simoncelli and O. Schwartz. Modeling surround suppression in V1 neurons with a statistically-derived normalization model. In M. S. Kearns, S. A. Solla, and D. A. Cohn, editors, *Adv. Neural Information Processing Systems (NIPS*98)*, volume 11, pages 153–159, Cambridge, MA, 1999. MIT Press.

[22] F. H. Sinz, S. Gerwinn, and M. Bethge. Characterization of the p-generalized normal distribution. *Journal of Multivariate Analysis*, 07/26/ 2008.

[23] D. Song and A. K. Gupta. l_p-norm uniform distribution. *Proceedings of the American Mathematical Society*, 125:595–601, 1997.

[24] J. H. van Hateren and A. van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc R Soc Lond B Biol Sci.*, 265(1394):1724–1726, 1998.

[25] T Wachtler, T W Lee, and T J Sejnowski. Chromatic structure of natural scenes. *Journal of the Optical Society of America. A, Optics, image science, and vision*, 18:65–77, 2001. PMID: 11152005.

4.4 The Conjoint Effect of Divisive Normalization and Orientation Selectivity on Redundancy Reduction: Supplementary Material

Extra Material

1. Data Preprocessing

1.1. **Removing the DC Component with an Orthogonal Projection.** The projector P_{remDC} is computed such that the first (for each color channel) component of $P_{remDC}\mathbf{x}$ corresponds to the DC component(s) of that patch. The transpose of the matrix

$$P = \begin{pmatrix} 1 & 0 & 0 & \cdots \\ 1 & 1 & 0 & \cdots \\ 1 & 0 & \ddots & \cdots \\ \vdots & & & 1 \end{pmatrix}$$

has exactly the required property. However, it is not an orthogonal transformation. Therefore, we decompose P into $P = QR$ where R is upper triangular and Q is an orthogonal transform. Since $P = QR$, the first column of Q must be a multiple of the vector with all coefficients equal to one (due to the upper triangluarity of R). Therefore, the first component of $Q^\top \mathbf{x}$ is a multiple of the DC component. Since Q is an orthonomal transform, using all but the first row of $Q^\top$ for P_{remDC} projects out the DC component. In case of color images the same trick is applied to each channel by making P_{remDC} a block-diagonal matrix with $Q^\top$ as diagonal elements.

1.2. **Rescaling the Data to Make Whitening an Volume Conserving Transform.** Secondly, the data was scaled such that the whitening transform has determinant one, i.e. that the determinant of the globally scaled data is one. This is done by setting $\eta = \prod \lambda_i^{\frac{1}{2n}}$, where λ_i are the eigenvalues of the covariance matrix of the training data and n is their dimension. Therefore, the determinant of the covariance matrix of the data after scaling with $\frac{1}{\eta}$ is

$$\frac{1}{\eta^{2n}} \prod \lambda_i = \frac{\prod \lambda_i}{\left(\prod \lambda_i^{\frac{1}{2n}}\right)^{2n}} = 1.$$

Since the whitening transform consist of $D^{-\frac{1}{2}}U^\top$ with $UDU^\top = C$ (C is the determinant of the scaled data), the whitening must have determinant one due to

$$1 = \det(C) = \det(UDU^\top) = \det(D^{-\frac{1}{2}}U^\top)^2$$

Note, that the same scaling factor is used for the training and test set.

2. Measures Of Redundancy

Redundancies can be quantified by a comparison of coding costs. According to Shannon's channel coding theorem the entropy of a discrete random variable is an attainable lower bound on the coding cost for error-free encoding [1]. For the construction of such a code, it is necessary to know the true distribution of the random variable. If the assumed distribution $\hat{P}(k)$ used for the construction of an optimal code is different from the true distribution $P(k)$, the coding cost is given by the log-loss

$$\mathbb{E}_P[-\log(\hat{P}(k))] \quad = \quad -\sum_k P(k) \log \hat{P}(k) = H[k] + D_{KL}[P(k)||\hat{P}(k)]\,.$$

The Kullback-Leibler divergence quantifies the additional coding cost caused by using a model distribution different from the true one. As long as it is positive, the representation can be still compressed further, which means that there are still redundancies left.

For continuous random variables, the total amount of bits required for loss-less encoding is infinite. However, in analogy to the discrete case, we can use the Kullback-Leibler divergence of the true distribution to a given model distribution. The goal of redundancy reduction is to map a random variable Y to a new random variable $Z = f(Y)$ such that the distribution of Z is as close to a factorial distribution as possible. Thus we can use the Kullback-Leibler divergence of the true distribution to the product of its marginals to measure redundancy. This quantity is known as multi-information

$$I[\rho(\mathbf{z})] \quad = \quad D_{\mathrm{KL}}\left[\rho(\mathbf{z})||\prod_{j=1}^{n}\rho_j(z_j)\right] \; = \; \int \rho(\mathbf{z}) \log \frac{\rho(\mathbf{z})}{\prod_{j=1}^{n}\rho_j(z_j)} d\mathbf{z}.$$

Algorithmically, redundancy can be reduced by finding a representation $Z = f(Y)$ such that a factorial model distribution $\hat{\rho}(\mathbf{z}) = \prod_{j=1}^{n}\hat{\rho}_j(z_j)$ is as close as possible to the true distribution $\rho(\mathbf{z})$. Since the multi-information $I[\rho(\mathbf{z})]$ is hard to estimate, one looks at the difference between the multi-informations of Y and $Z = f(Y)$, i.e. the quantity

$$\begin{aligned} \Delta I &= I[\rho(\mathbf{z})] - I[\varrho(\mathbf{y})] \\ &= D_{\mathrm{KL}}\left[\rho(\mathbf{z})||\prod_{j=1}^{n}\hat{\rho}_j(z_j)\right] - D_{\mathrm{KL}}\left[\varrho(\mathbf{y})||\prod_{j=1}^{n}\hat{\varrho}_j(y_j)\right] \; , \end{aligned}$$

where $\prod_{j=1}^{n}\hat{\varrho}_j(y_j)$ is a factorial model distribution for the representation Y. The following calculation shows that evaluating the redundancy reduction achieved with a mapping $\mathbf{z} = f(\mathbf{y})$ is equivalent to evaluating the difference between the log-loss of two particular model distributions.

Before doing the actual calculation, it is useful to define the different distributions involved and state some interrelations between them:

(1) $\rho(\mathbf{z})$ and $\varrho(\mathbf{y})$ are the true distributions of the random variables Y and $Z = f(Y)$. They are related by

$$\begin{aligned} \rho(\mathbf{z})d\mathbf{z} &= \rho(f(\mathbf{y})) \cdot \left|\det \frac{\partial z_i}{\partial y_j}\right| d\mathbf{y} = \varrho(\mathbf{y})d\mathbf{y} \\ \varrho(\mathbf{y})d\mathbf{y} &= \varrho(f^{-1}(\mathbf{z})) \cdot \left|\det \frac{\partial y_i}{\partial z_j}\right| d\mathbf{z} = \rho(\mathbf{z})d\mathbf{z} \; , \end{aligned}$$

where $\frac{\partial z_i}{\partial y_j}$ denotes the Jacobian for f and $\frac{\partial y_i}{\partial z_j}$ the Jacobian of f^{-1}. Note that $\left|\det \frac{\partial z_i}{\partial y_j}\right| = \left|\det \frac{\partial y_i}{\partial z_j}\right|^{-1}$.

(2) $\hat{\rho}(\mathbf{z}) := \prod_{j=1}^{n}\hat{\rho}_j(z_j)$, $\hat{\varrho}_f(\mathbf{y})$ and $\prod_{j=1}^{n}\hat{\varrho}_j(y_j)$ are the model distributions. $\prod_{j=1}^{n}\hat{\varrho}_j(y_j)$ is the factorial model for the representation Y. The non-factorial model distribution $\hat{\varrho}_f(\mathbf{y})$ was chosen such that the function f maps it into a factorial distribution, i.e.

$$\begin{aligned} \prod_{j=1}^{n}\hat{\rho}_j(z_j) &\overset{\text{choice of } f}{=} \hat{\rho}(\mathbf{z}) \\ &= \hat{\rho}_f(f(\mathbf{y})) \cdot \left|\det \frac{\partial z_i}{\partial y_j}\right| \\ &= \hat{\varrho}_f(\mathbf{y}). \end{aligned}$$

Now, we can write the difference in multi-information as

$$\begin{aligned}
\Delta I &= I[\rho(\mathbf{z})] - I[\varrho(\mathbf{y})] \\
&= D_{\mathrm{KL}}\left[\rho(\mathbf{z})||\prod_{j=1}^{n}\hat{\rho}_j(z_j)\right] - D_{\mathrm{KL}}\left[\varrho(\mathbf{y})||\prod_{j=1}^{n}\hat{\varrho}_j(y_j)\right] \\
&= \mathbb{E}_{\rho}\left[\log\frac{\rho(\mathbf{z})}{\prod_{j=1}^{n}\hat{\rho}_j(z_j)}\right] - \mathbb{E}_{\varrho}\left[\log\frac{\varrho(\mathbf{y})}{\prod_{j=1}^{n}\hat{\varrho}_j(y_j)}\right] \\
&= \mathbb{E}_{\varrho}\left[\log\frac{\rho(f(\mathbf{y}))\cdot\left|\det\frac{\partial z_i}{\partial y_j}\right|}{\hat{\varrho}_f(\mathbf{y})}\right] - \mathbb{E}_{\varrho}\left[\log\frac{\varrho(\mathbf{y})}{\prod_{j=1}^{n}\hat{\varrho}_j(y_j)}\right] \\
&= \mathbb{E}_{\varrho}\left[\log\frac{\rho(f(\mathbf{y}))\cdot\left|\det\frac{\partial z_i}{\partial y_j}\right|}{\hat{\varrho}_f(\mathbf{y})} - \log\frac{\varrho(\mathbf{y})}{\prod_{j=1}^{n}\hat{\varrho}_j(y_j)}\right] \\
&= \mathbb{E}_{\varrho}\left[\log\frac{\prod_{j=1}^{n}\hat{\varrho}_j(y_j)}{\hat{\varrho}_f(\mathbf{y})}\cdot\frac{\overbrace{\rho(f(\mathbf{y}))\cdot\left|\det\frac{\partial z_i}{\partial y_j}\right|}^{=\varrho(\mathbf{y})}}{\varrho(\mathbf{y})}\right] \\
&= \mathbb{E}_{\varrho}\left[\log\frac{\prod_{j=1}^{n}\hat{\varrho}_j(y_j)}{\hat{\varrho}_f(\mathbf{y})}\right] \\
&= \mathbb{E}_{\varrho}\left[-\log\hat{\varrho}_f(\mathbf{y})\right] - \mathbb{E}_{\varrho}\left[-\log\prod_{j=1}^{n}\hat{\varrho}_j(y_j)\right].
\end{aligned}$$

Thus, if we have a model density which does not factorize with respect to $\mathbf{y}$ and we have a (possibly nonlinear) mapping $\mathbf{z} = f(\mathbf{y})$ such that the transformed model density with respect to $\mathbf{z}$ becomes factorial, we can evaluate the redundancy reduction achieved with the mapping f simply by estimating the difference in the average log-loss obtained for $\hat{\varrho}_f(\mathbf{y})$ and $\prod_{j=1}^{n}\hat{\varrho}_j(y_j)$.

In order to get a measure which is less dependent on the number of dimensions n we define the average log-loss (ALL) to be $\mathrm{ALL} = \frac{1}{n}\mathbb{E}[-\log\hat{\varrho}(\mathbf{y})]$ for any given model distribution $\hat{\varrho}(\mathbf{y})$.

In practice, the ALL can estimated by with the empirial mean

$$\frac{1}{n}\mathbb{E}_{\varrho}\left[-\log\hat{\varrho}_f(\mathbf{y})\right] \approx \frac{1}{n\cdot m}\sum_{i=1}^{m} -\log\hat{\varrho}_f(\mathbf{y}_i).$$

3. L_p-Spherically Symmetric Distributions

3.1. **Definitions, Lemmas and Theorems.** In this part, we provide the rigorous definitions, lemmas and theorems used in the paper. Most results and proofs are not new and have been collected from papers and books. Nevertheless, in many cases we adapted the original statements to our need and provided more detailed versions of the proofs. The original sources are mentioned at the respective lemmas and theorems.

Definition 1. *p-Norm*

Let $\mathbf{y} \in \mathbb{R}^n$ be an arbitrary vector. We define

$$||\mathbf{y}||_p \quad = \quad \left(\sum_{i=1}^{n} |y_i|^p \right)^{\frac{1}{p}}, \; p > 0$$

as the p-norm of $\mathbf{y}$. Note, that only for $p > 1$, $||\mathbf{y}||_p$ is a norm in the strict sense. However, we will also use the term "p-norm" even if only $0 < p$.

Definition 2. ***p*-Sphere**
The unit p-sphere $\mathbb{S}_p^{n-1}$ in n dimensions is the set of points that fulfill

$$\mathbb{S}_p^{n-1} \quad := \quad \{\mathbf{y} \in \mathbb{R}^n | \, ||\mathbf{y}||_p = 1, \, p > 0\}.$$

Lemma 3. ***Transformation in Radial and Spherical Coordinates*** [3]
Let $\mathbf{y} = (y_1, ...y_n)^\top$ $n \geq 2$ be a vector in $\mathbb{R}^n \backslash \{\mathbf{0}\}$. Consider the transformation

$$\mathbf{y} \quad \mapsto \quad (r, u_1, ..., u_{n-1}) = \left(||\mathbf{y}||_p, \frac{y_1}{||\mathbf{y}||_p}, ..., \frac{y_{n-1}}{||\mathbf{y}||_p} \right).$$

The absolute value of the determinants of the transformation on the upper and lower halfspaces

$$\begin{aligned} \mathbb{R}_+^n \quad &:= \quad \{\mathbf{y} \in \mathbb{R}^n | \, y_n \geq 0\} \\ \mathbb{R}_-^n \quad &:= \quad \{\mathbf{y} \in \mathbb{R}^n | \, y_n < 0\} \end{aligned}$$

are equal and are given by

$$|\det \mathcal{J}| \quad = \quad r^{n-1} \left(1 - \sum_{i=1}^{n-1} |u_i|^p \right)^{\frac{1-p}{p}}.$$

Proof. The proof is a more detailed version of the proof found in [3].
Let

$$\Delta_i \quad := \quad \begin{cases} 1, \, u_i \geq 0 \\ -1, \, u_i < 0. \end{cases}$$

Then we can write $|u_i| = \Delta_i u_i$. The above transformation is bijective on each of the regions $\mathbb{R}_+^n$ and $\mathbb{R}_-^n$. Let $\sigma = \text{sign}(y_n)$, then the inverse is given by

$$\begin{aligned} y_i \quad &= \quad u_i r \, , \, 1 \leq i \leq n-1 \\ y_n \quad &= \quad \sigma r \left(1 - \sum_{i=1}^{n-1} |u_i|^p \right)^{\frac{1}{p}} = \sigma r \left(1 - \sum_{i=1}^{n-1} (\Delta_i u_i)^p \right)^{\frac{1}{p}}. \end{aligned}$$

Note, that the $\sigma = \text{sign}(y_n)$ determines the halfspace in which the transformation is inverted.

First, we determine the Jacobian $\mathcal{J}$. We start with computing the derivatives

$$\begin{aligned}
\frac{\partial y_i}{\partial u_j} &= \delta_{ij} r,\ 1 \leq i,j \leq n-1 \\
\frac{\partial y_n}{\partial u_j} &= -\sigma r \left(1 - \sum_{i=1}^{n-1} |u_i|^p\right)^{\frac{1-p}{p}} \Delta_i^p u_i^{p-1},\ 1 \leq j \leq n-1 \\
\frac{\partial y_i}{\partial r} &= u_i,\ 1 \leq i \leq n-1 \\
\frac{\partial y_n}{\partial r} &= \sigma \left(1 - \sum_{i=1}^{n-1} (\Delta_i u_i)^p\right)^{\frac{1}{p}}.
\end{aligned}$$

Therefore, the Jacobian, is given by

$$\begin{aligned}
\mathcal{J} &= \begin{pmatrix}
\frac{\partial y_1}{\partial u_1} & & \frac{\partial y_1}{\partial u_{n-1}} & \frac{\partial y_1}{\partial r} \\
\vdots & \ddots & & \vdots \\
\vdots & & \ddots & \vdots \\
\frac{\partial y_n}{\partial u_1} & & \frac{\partial y_n}{\partial u_{n-1}} & \frac{\partial y_n}{\partial r}
\end{pmatrix} \\
&= \begin{pmatrix}
r & & 0 & \dots & u_1 \\
0 & & r & & u_2 \\
\vdots & & & \ddots & \vdots \\
-\sigma r \left(1 - \sum_{i=1}^{n-1} |u_i|^p\right)^{\frac{1-p}{p}} \Delta_1^p u_1^{p-1} & & \dots & \dots & \sigma \left(1 - \sum_{i=1}^{n-1} (\Delta_i u_i)^p\right)^{\frac{1}{p}}
\end{pmatrix}.
\end{aligned}$$

Before actually computing the absolute value of the determinant $|\det \mathcal{J}|$, we can factor out r from the first $n-1$ columns. Furthermore, we can factor out σ from the last row. Since we take the absolute value of $\det \mathcal{J}$ and $\sigma = \{-1, 1\}$, we can remove it completely afterwards. Now we can use Laplace's formula to expand the determinant along the last column. With this, we get

$$\begin{aligned}
\frac{1}{r^{n-1}} |\det \mathcal{J}| &= \sum_{k=1}^{n-1} (-1)^{n+k} \cdot u_k \cdot (-1)^{n-1-k} \cdot -\Delta_k^p u_k^{p-1} \cdot \left(1 - \sum_{i=1}^{n-1} |u_i|^p\right)^{\frac{1-p}{p}} \\
&\quad + (-1)^{2n} \left(1 - \sum_{i=1}^{n-1} |u_i|^p\right)^{\frac{1}{p}} \\
&= \sum_{k=1}^{n-1} |u_k|^p \left(1 - \sum_{i=1}^{n-1} |u_i|^p\right)^{\frac{1-p}{p}} + \left(1 - \sum_{i=1}^{n-1} |u_i|^p\right)^{\frac{1}{p}} \\
&= \left(1 - \sum_{i=1}^{n-1} |u_i|^p\right)^{\frac{1-p}{p}} \left(\sum_{k=1}^{n-1} |u_k|^p + 1 - \sum_{k=1}^{n-1} |u_k|^p\right) \\
&= \left(1 - \sum_{i=1}^{n-1} |u_i|^p\right)^{\frac{1-p}{p}}.
\end{aligned}$$

Resolving the result for $|\det \mathcal{J}|$ completes the proof. □

Theorem 4. ***p-Spherical Uniform Distribution*** [3]
Let $Y = (Y_1, ..., Y_n)^\top$ be a random vector. Let the Y_i be i.i.d. distributed with p.d.f.

$$\varrho(\mathbf{y}) = \frac{p^{1-\frac{1}{p}}}{2\Gamma\left(\frac{1}{p}\right)} \exp\left(-\frac{|y|^p}{p}\right),\ y \in \mathbb{R}.$$

Let $U_i = \frac{Y_i}{||Y||_p}$ for $i = 1, ..., n$. Then $\sum_{i=1}^n |U_i|^p = 1$ and the joint p.d.f of $U_1, ..., U_{n-1}$ is

$$q_u(u_1, ..., u_{n-1}) = \frac{p^{n-1}\Gamma\left(\frac{n}{p}\right)}{2^{n-1}\Gamma^n\left(\frac{1}{p}\right)} \left(1 - \sum_{i=1}^{n-1} |u_i|^p\right)^{\frac{1-p}{p}}$$

with $-1 < u_i < 1$, $i = 1, ..., n-1$ and $\sum_{i=1}^{n-1} |u_i|^p < 1$.

Proof. The joint p.d.f. of Y is given by

$$\varrho(\mathbf{y}) = \frac{p^{n-\frac{n}{p}}}{2^n\Gamma^n\left(\frac{1}{p}\right)} \exp\left(-\frac{1}{p}\sum_{i=1}^{n} |y_i|^p\right)$$

with $y_i \in \mathbb{R}$ and $i = 1, ..., n$. Applying the transformation

$$(y_1, ..., y_n) = (r, u_1, ..., u_{n-1})$$

from Lemma 3 and taking into account that each $(u_1, ..., u_{n-1})$ corresponds to $(y_1, ..., y_n)$ and $(y_1, ..., -y_n)$ we obtain

$$q(u_1, ..., u_{n-1}, r) = 2 \cdot \frac{p^{n-\frac{n}{p}}}{2^n\Gamma^n\left(\frac{1}{p}\right)} r^{n-1} \exp\left(-\frac{r^p}{p}\right) \left(1 - \sum_{i=1}^{n-1} |u_i|^p\right)^{\frac{1-p}{p}}.$$

By integrating out r, we obtain $q_u(u_1, ..., u_n)$:

$$\int_0^\infty q(u_1, ..., u_{n-1}, r)dr = \frac{p^{n-\frac{n}{p}}}{2^{n-1}\Gamma^n\left(\frac{1}{p}\right)} \left(1 - \sum_{i=1}^{n-1} |u_i|^p\right)^{\frac{1-p}{p}} \int_0^\infty r^{n-1} \exp\left(-\frac{r^p}{p}\right) dr.$$

In order to compute the integral, we use the substitution $z = \frac{r^p}{p}$ or $r = (zp)^{\frac{1}{p}}$. This yields $dr = (zp)^{\frac{1}{p}-1}dz$ and, therefore,

$$\begin{aligned}
\int_0^\infty r^{n-1} \exp\left(-\frac{r^p}{p}\right) dr &= \int_0^\infty (zp)^{\frac{n-1}{p}} \exp(-z)(zp)^{\frac{1-p}{p}} dz \\
&= p^{\frac{n-p}{p}} \int_0^\infty z^{\frac{n}{p}-1} \exp(-z)dz \\
&= p^{\frac{n-p}{p}} \Gamma\left(\frac{n}{p}\right).
\end{aligned}$$

Hence,

$$\begin{aligned} q_u(u_1,...,u_{n-1}) &= \int_0^\infty q(u_1,...,u_{n-1},r)dr \\ &= \frac{p^{n-\frac{n}{p}}}{2^{n-1}\Gamma^n\left(\frac{1}{p}\right)}\left(1-\sum_{i=1}^{n-1}|u_i|^p\right)^{\frac{1-p}{p}} p^{\frac{n-p}{p}}\Gamma\left(\frac{n}{p}\right) \\ &= \frac{p^{n-1}\Gamma\left(\frac{n}{p}\right)}{2^{n-1}\Gamma^n\left(\frac{1}{p}\right)}\left(1-\sum_{i=1}^{n-1}|u_i|^p\right)^{\frac{1-p}{p}}. \end{aligned}$$

□

In order to see, why q_u is called uniform on $\mathbb{S}_p^{n-1}$ we must observe that q_u of $\left(1-\sum_{i=1}^{n-1}|u_i|^p\right)^{\frac{1-p}{p}}$ which is due to the coordinate transformation and $\frac{p^{n-1}\Gamma\left(\frac{n}{p}\right)}{2^{n-1}\Gamma^n\left(\frac{1}{p}\right)}$ which corresponds to twice the surface area of the p-sphere (see Lemma 5). Since each $\mathbf{u}$ corresponds to two $\mathbf{y}$ before the coordinate transform (one on the upper and one on the lower halfsphere), the density of $\mathbf{u}$ in $\mathbf{y}$-coordinates corresponds to $\frac{1}{S_p^{n-1}}$ where $S_p^{n-1}=\frac{2^n\Gamma(\frac{1}{p})^n}{p^{n-1}\Gamma(\frac{n}{p})}$ is the surface area of the unit p-sphere (see Lemma 5).

As we will see in Lemma 7, $\frac{Y}{||Y||_p}$ is independent of $||Y||_p$ and, therefore, the specific form of the density ϱ does not matter as long as it is p-spherically symmetric.

Lemma 5. ***Volume and Surface of the p-Sphere***
The volume $V_p^{n-1}(r)$ of the p-Sphere with radius r is given by

$$V_p^{n-1}(r) = \frac{r^n 2^n\Gamma(\frac{1}{p})^n}{np^{n-1}\Gamma(\frac{n}{p})}.$$

The surface $S_p^{n-1}(r)$ is given by

$$\begin{aligned} S_p^{n-1}(r) &= \frac{d}{dr}V_p^{n-1}(r) \\ &= \frac{r^{n-1}2^n\Gamma(\frac{1}{p})^n}{p^{n-1}\Gamma(\frac{n}{p})}. \end{aligned}$$

As a convention, we leave out the argument of $V_p^{n-1}(r)$ and $S_p^{n-1}(r)$ when denoting the volume or the surface of the unit p-sphere, i.e.

$$\begin{aligned} V_p^{n-1} &:= V_p^{n-1}(1) \\ S_p^{n-1} &:= S_p^{n-1}(1). \end{aligned}$$

Proof. In order to compute the volume of the p-sphere in n-dimension, we must solve the integral $\int_{\mathbb{S}_p^{n-1}} d\mathbf{u}$. Using the volume element transformation from lemma

3, we can transform the integral into

$$\begin{aligned}\int_{\mathbb{S}_p^{n-1}} d\mathbf{u} &= 2\int_0^r \int r^{n-1}\left(1-\sum_{i=1}^{n-1}|u_i|^p\right)^{\frac{1-p}{p}} dr\, d\mathbf{u} \\ &= 2\int_0^r r^{n-1}dr \cdot \int \left(1-\sum_{i=1}^{n-1}|u_i|^p\right)^{\frac{1-p}{p}} d\mathbf{u} \\ &= \frac{1}{n}r^n \cdot 2\int \left(1-\sum_{i=1}^{n-1}|u_i|^p\right)^{\frac{1-p}{p}} d\mathbf{u}.\end{aligned}$$

In theorem 4 we prove that $q(u_1,...,u_{n-1}) = \frac{p^{n-1}\Gamma\left(\frac{n}{p}\right)}{2^{n-1}\Gamma^n\left(\frac{1}{p}\right)}\left(1-\sum_{i=1}^{n-1}|u_i|^p\right)^{\frac{1-p}{p}}$ is a probability density. In particular, this means that

$$\begin{aligned}\int q(u_1,...,u_{n-1})d\mathbf{u} &= \frac{p^{n-1}\Gamma\left(\frac{n}{p}\right)}{2^{n-1}\Gamma^n\left(\frac{1}{p}\right)}\int \left(1-\sum_{i=1}^{n-1}|u_i|^p\right)^{\frac{1-p}{p}} d\mathbf{u} \\ &= 1\end{aligned}$$

which is equivalent to

$$\int \left(1-\sum_{i=1}^{n-1}|u_i|^p\right)^{\frac{1-p}{p}} d\mathbf{u} = \frac{2^{n-1}\Gamma^n\left(\frac{1}{p}\right)}{p^{n-1}\Gamma\left(\frac{n}{p}\right)}.$$

Therefore,

$$\begin{aligned}V_p^{n-1}(r) &= \int_{\mathbb{S}_p^{n-1}} d\mathbf{u} \\ &= \frac{2}{n}r^n \cdot \int \left(1-\sum_{i=1}^{n-1}|u_i|^p\right)^{\frac{1-p}{p}} d\mathbf{u} \\ &= \frac{r^n 2^n \Gamma^n\left(\frac{1}{p}\right)}{np^{n-1}\Gamma\left(\frac{n}{p}\right)}\end{aligned}$$

Differentiation of $V_p^{n-1}(r)$ with respect to r yields the result for the surface area. □

Definition 6. **L_p-Spherically Symmetric Distribution** [2] A random vector $Y = (Y_1,...,Y_n)^\top$ is said to have a L_p-spherically symmetric distribution if Y can be written as a product of two independent random variables $Y = R \cdot U$, where R is a non-negative univariate random variable with density $q_r : \mathbb{R}^+ \to \mathbb{R}^+$ and U is uniformly distributed on the unit p-sphere, i.e.

$$q_u(u_1,...,u_n) = \frac{p^{n-1}\Gamma\left(\frac{n}{p}\right)}{2^{n-1}\Gamma^n\left(\frac{1}{p}\right)}\left(1-\sum_{i=1}^{n-1}|u_i|^p\right)^{\frac{1-p}{p}}$$

(see Theorem 4).

Lemma 7. ***Probability Density Functions*** [2]

Let $Y = (Y_1, ..., Y_n)^\top$ be an n-dimensional random variable with $P\{Y = \mathbf{0}\} = 0$ and a density of the form $Y \sim \tilde{\varrho}(||\mathbf{y}||_p^p)$. Then the following three statements hold:

(1) The random variables $R = ||Y||_p$ and $U = \frac{Y}{||Y||_p}$ are independent.
(2) $U = \frac{Y}{||Y||_p}$ is uniformly distributed on the unit p-sphere $\mathbb{S}_p^{n-1}$.
(3) $R = ||Y||_p$ has a density q_r, where q_r relates to $\tilde{\varrho}$ via

$$\begin{aligned} q_r(r) &= \frac{r^{n-1} 2^n \Gamma(\frac{1}{p})^n}{p^{n-1}\Gamma(\frac{n}{p})} \tilde{\varrho}(r^p) \\ &= S_p^{n-1}(r)\tilde{\varrho}(r^p),\ r > 0. \end{aligned}$$

Proof. The proof is a more detailed version of the proof found in [2].

First we transform the density of Y with the transformation of lemma 3 and obtain the new density in spherical and radial coordinates

$$\begin{aligned} q(u_1, ..., u_{n-1}, r) &= 2\left(1 - \sum_{i=1}^{n-1} |u_i|^p\right)^{\frac{1-p}{p}} \tilde{\varrho}(r^p) r^{n-1} \\ &\quad -1 < u_i < 1,\ 1 \le i \le n-1,\ \sum_{i=1}^{n} |u_i|^p < 1. \end{aligned}$$

Since q can be written as a product of a function of r and a function of $\mathbf{u} = (u_1, ..., u_{n-1})$, U and R are independent. Thus, $||Y||_p = R$ and $U = \frac{Y}{||Y||_p}$ are independent as well.

In order to get $q_u(u_1, ..., u_{n-1})$, we must integrate out r. However, we do not know the exact form of $\tilde{\varrho}$. But since q is a probability density, we know that

$$\int_0^\infty \int q(u_1, ..., u_{n-1}, r) d\mathbf{u} dr = 1.$$

Since Y and R are independent, we can write this integral as

$$\int_0^\infty \int q(u_1, ..., u_{n-1}, r) d\mathbf{u} dr = 2\int \left(1 - \sum_{i=1}^{n-1} |u_i|^p\right)^{\frac{1-p}{p}} d\mathbf{u} \cdot \int_0^\infty \tilde{\varrho}(r^p) r^{n-1} dr.$$

From that, we can immediately derive

$$\int_0^\infty \tilde{\varrho}(r^p) r^{n-1} dr = \left(2\int \left(1 - \sum_{i=1}^{n-1} |u_i|^p\right)^{\frac{1-p}{p}} d\mathbf{u}\right)^{-1}.$$

In order to solve $\left(2\int \left(1 - \sum_{i=1}^{n-1} |u_i|^p\right)^{\frac{1-p}{p}} d\mathbf{u}\right)^{-1}$ we can use theorem 4. In this theorem, we showed that $q_u(u_1, ..., u_{n-1}) = \frac{p^{n-1}\Gamma\left(\frac{n}{p}\right)}{2^{n-1}\Gamma^n\left(\frac{1}{p}\right)} \left(1 - \sum_{i=1}^{n-1} |u_i|^p\right)^{\frac{1-p}{p}}$ is the uniform distribution on the p-unit sphere. In particular, we know that $\int q(u_1, ..., u_{n-1}) d\mathbf{u} = 1$ and, therefore,

$$\int \left(1 - \sum_{i=1}^{n-1} |u_i|^p\right)^{\frac{1-p}{p}} d\mathbf{u} = \frac{2^{n-1}\Gamma^n\left(\frac{1}{p}\right)}{p^{n-1}\Gamma\left(\frac{n}{p}\right)}.$$

Thus,

$$\begin{aligned}\int_0^\infty \tilde{\varrho}(r^p) r^{n-1} dr &= \left(2\int\left(1-\sum_{i=1}^{n-1}|u_i|^p\right)^{\frac{1-p}{p}} d\mathbf{u}\right)^{-1} \\ &= \frac{p^{n-1}\Gamma\left(\frac{n}{p}\right)}{2^n\Gamma^n\left(\frac{1}{p}\right)}\end{aligned}$$

and

$$\begin{aligned}q_u(u_1,...,u_{n-1}) &= \int_0^\infty q(u_1,...,u_{n-1},r)dr \\ &= \left(1-\sum_{i=1}^{n-1}|u_i|^p\right)^{\frac{1-p}{p}} \frac{p^{n-1}\Gamma\left(\frac{n}{p}\right)}{2^{n-1}\Gamma^n\left(\frac{1}{p}\right)}.\end{aligned}$$

This shows that Y is uniformly distributed on the unit p-sphere.

The density of R can be computed by integrating out $u_1,...,u_{n-1}$

$$\begin{aligned}q_r(r) &= \int q(u_1,...,u_{n-1},r)d\mathbf{u} \\ &= \frac{2^n\Gamma^n\left(\frac{1}{p}\right)}{p^{n-1}\Gamma\left(\frac{n}{p}\right)} r^{n-1}\tilde{\varrho}(r^p),\ r>0\end{aligned}$$

by the same argument as in 2. This completes the proof. □

The next theorem tells us that Y is L_p-spherically symmetric distributed if and only if its density has the form $\tilde{\varrho}(||\mathbf{y}||_p^p)$.

Theorem 8. ***Form of*** L_p***-Spherically Symmetric Distribution*** [2] *Let* $Y=(Y_1,...,Y_n)^\top$ *be an* n*-dimensional random variable with* $P\{Y=\mathbf{0}\}=0$*. Then, the density of* Y *has the form* $\tilde{\varrho}(||\mathbf{y}||_p^p)$*, where* $g:\mathbb{R}_+\to\mathbb{R}_+$ *is a measurable function, if and only if* $Y=RU$ *is spherically symmetric distributed, with independent* R *and* U*, where* R *has the density*

$$q_r(r) = \frac{2^n\Gamma^n\left(\frac{1}{p}\right)}{p^{n-1}\Gamma\left(\frac{n}{p}\right)} r^{n-1}g(r^p),\ r>0.$$

Proof. <u>*Sufficiency:*</u> Assume $Y=RU$ with independent R and U, where U is uniformly distributed on the p-sphere and R has the density q_r. Then the joint density is given by (see theorem 4):

$$\begin{aligned}q(r,u_1,...,u_{n-1}) &= q_r(r)\frac{p^{n-1}\Gamma\left(\frac{n}{p}\right)}{2^{n-1}\Gamma^n\left(\frac{1}{p}\right)}\left(1-\sum_{i=1}^{n-1}|u_i|^p\right)^{\frac{1-p}{p}} \\ &\quad -1<u_i<1,\ 1\le i\le n-1,\ \sum_{i=1}^{n-1}|u_i|^p<1,\ r>0.\end{aligned}$$

Now let $y_i = ru_i$ for $1 \leq i \leq n-1$ and $|y_n| = r\left(1 - \sum_{i=1}^{n-1} |u_i|^p\right)^{\frac{1}{p}}$. We can use 3 to see that the absolute value of the determinant of the Jacobian is given by

$$\left(r^{n-1}\left(1 - \sum_{i=1}^{n-1} |u_i|^p\right)^{\frac{1-p}{p}}\right)^{-1} = r^{1-n}\left(1 - \sum_{i=1}^{n-1} |u_i|^p\right)^{\frac{p-1}{p}}.$$

Therefore,

$$\begin{aligned} p(y_1, ..., y_n) &= \frac{p^{n-1}\Gamma\left(\frac{n}{p}\right)}{2^{n-1}\Gamma^n\left(\frac{1}{p}\right)} q_r(||\mathbf{y}||_p)||\mathbf{y}||_p^{1-n} \\ &= \tilde{\varrho}(||\mathbf{y}||_p^p). \end{aligned}$$

Necessity: Assume $Y \sim \tilde{\varrho}(||Y||_p^p)$. According to lemma 7 $\frac{Y}{||Y||_p}$ and Y are independent and $\frac{Y}{||Y||_p}$ is uniformly distributed on the p-sphere. Again in lemma 7 we showed that R has the density

$$q_r(r) = \frac{2^n\Gamma^n\left(\frac{1}{p}\right)}{p^{n-1}\Gamma\left(\frac{n}{p}\right)} r^{n-1}\tilde{\varrho}(r^p),\ r > 0.$$

Therefore, Y is L_p-spherically symmetric distributed if and only if $Y \sim \tilde{\varrho}(||Y||_p^p)$ for some density $\tilde{\varrho}$. □

3.2. **Distributions.**

3.2.1. *The p-Spherically Symmetric Distribution with Radial Mixture of Log-Normal Distribution.* We obtain this distribution by modeling the radial component with a mixture of log-Normal distributions

$$q_r(r) = \sum_{k=1}^{K} \frac{\eta_k}{r\sigma_k\sqrt{2\pi}} \exp\left(-\frac{(\log r - \mu_k)^2}{2\sigma_k^2}\right).$$

Here, η_k with $\sum_k \eta_k = 1$ constitute the "prior" probability of selecting one log-Normal distribution from the mixture, and μ_k and σ_k^2 denote the mean and the variance of the kth mixture. Taking into account the uniform distribution on the p-sphere, we get

$$q(\mathbf{u}, r) = \left(1 - \sum_{i=1}^{n-1} |u_i|^p\right)^{\frac{1-p}{p}} \frac{p^{n-1}\Gamma\left(\frac{n}{p}\right)}{2^{n-1}\Gamma^n\left(\frac{1}{p}\right)} \sum_{k=1}^{K} \frac{\eta_k}{r\sigma_k\sqrt{2\pi}} \exp\left(-\frac{(\log r - \mu_k)^2}{2\sigma_k^2}\right).$$

Reversing the coordinate transform, we obtain the distribution in Euclidean coordinates

$$\varrho(\mathbf{y}) = \frac{p^{n-1}\Gamma\left(\frac{n}{p}\right)}{2^n\Gamma^n\left(\frac{1}{p}\right)} \sum_{k=1}^{K} \frac{\eta_k}{||\mathbf{y}||_p^n\sigma_k\sqrt{2\pi}} \exp\left(-\frac{(\log ||\mathbf{y}||_p - \mu_k)^2}{2\sigma_k^2}\right).$$

Since $||\mathbf{y}||_p$ being log-Normal distributed means $\log ||\mathbf{y}||_p$ being Gaussian distributed, we can use the standard EM for a mixture of Gaussians on the log-domain to estimate the parameters of the mixture. This is justified because log (or exp) is a

strictly monotonic increasing (decreasing) function and the multiplicative determinant of the Jacobian does not depend on the parameters. Therefore, the maximizing parameter values for one the mixture of log-Normal distributions also maximizes the log-likelihood of the mixture of Gaussians in the log-domain.

In order to transform the radial component into the radial component of the p-generalized distribution, we will need the cumulative distribution function, which is given by

$$\begin{aligned}
\mathcal{F}(r_0) &= \int_0^{r_0} q_r(r)dr \\
&= \int_0^{r_0} \sum_{k=1}^{K} \frac{\eta_k}{r\sigma_k\sqrt{2\pi}} \exp\left(-\frac{(\log r - \mu_k)^2}{2\sigma_k^2}\right) dr \\
&= \sum_{k=1}^{K} \eta_k \int_0^{r_0} \frac{1}{r\sigma_k\sqrt{2\pi}} \exp\left(-\frac{(\log r - \mu_k)^2}{2\sigma_k^2}\right) dr \\
&= \sum_{k=1}^{K} \eta_k \mathcal{F}_k(r_0; \mu_k, \sigma_k) \ ,
\end{aligned}$$

where $\mathcal{F}_k(r_0; \mu_k, \sigma_k)$ is simply the cumulative distribution function of the log-Normal distribution with parameters μ_k and σ_k.

3.2.2. *The p-generalized Normal distribution.* The p-generalized Normal distribution is obtained by choosing Y to be a collection of n i.i.d. random variables Y_i, each distributed according to the exponential power distribution

$$\begin{aligned}
Y_i \sim p(y) &= \frac{p}{\Gamma\left(\frac{1}{p}\right)(2\sigma^2)^{\frac{1}{p}}2} \exp\left(-\frac{|y|^p}{2\sigma^2}\right) \\
Y \sim \varrho(\mathbf{y}) = \prod_{i=1}^{n} p(y_i) &= \left(\frac{p}{\Gamma\left(\frac{1}{p}\right)(2\sigma^2)^{\frac{1}{p}}2}\right)^n \exp\left(-\frac{\sum_{i=1}^{n}|y_i|^p}{2\sigma^2}\right)
\end{aligned}$$

Since $\varrho(\mathbf{y})$ has the form $\tilde{\varrho}(||\mathbf{y}||_p^p)$, it is a proper p-spherically symmetric distribution due to Theorem 8. Note, that for the case of $p = 2$, the p-generalized Normal distribution reduces to a multivariate isotropic Gaussian. In order to compute the contrast gain control function, we need to compute the radial distribution q_r of $p(\mathbf{x})$. Transforming p according to Lemma 3 yields

$$q(r, \mathbf{u}) = \frac{p^n r^{n-1}}{\Gamma^n\left(\frac{1}{p}\right)(2\sigma)^{\frac{n}{p}}2^{n-1}} \exp\left(-\frac{r^p}{2\sigma}\right)\left(1 - \sum_{i=1}^{n-1}|u_i|^p\right)^{\frac{1-p}{p}} .$$

By integrating over $\mathbf{u}$ (see lemma 5 how to carry out the integral) we get

$$q_r(r) = \frac{p\, r^{n-1}}{\Gamma\left(\frac{n}{p}\right)(2\sigma^2)^{\frac{n}{p}}} \exp\left(-\frac{r^p}{2\sigma^2}\right)$$

In order to estimate the scale parameter σ from data $X = \{r_1, ..., r_m\} = \{||\mathbf{x}_1||_p, ..., ||\mathbf{x}_m||_p\}$, we carry out the usual procedure for maximum likelihood estimation and obtain

$$\begin{aligned}
\frac{d}{d\sigma}\log q_r(r) &= \frac{d}{d\sigma}\left(-\frac{2n}{p}\log(\sigma) - \frac{r^p}{2\sigma^2}\right) \\
&= \frac{r^p p - 2n\sigma^2}{p\sigma^3} \\
\frac{d}{d\sigma}\sum_{i=1}^{m}\log q_r(r_i) &= \sum_{i=1}^{m}\frac{r_i^p p - 2n\sigma^2}{p\sigma^3} \\
&\stackrel{!}{=} 0.
\end{aligned}$$

This yields

$$\hat{\sigma} = \sqrt{\frac{p}{2mn}\sum_{i=1}^{m} r_i^p}.$$

For the transformation of the radial component, we will also need the cumulative distribution function of

$$q_r(r) = \frac{p\, r^{n-1}}{\Gamma\left(\frac{n}{p}\right)(2\sigma^2)^{\frac{n}{p}}}\exp\left(-\frac{r^p}{2\sigma^2}\right).$$

It can be computed via simple integration with the substitution $y = \frac{r^p}{2\sigma^2}$

$$\begin{aligned}
\mathcal{F}_{\mathcal{N}_p}(a) &= \int_0^a \frac{p r^{n-1}}{\Gamma\left(\frac{n}{p}\right)(2\sigma^2)^{\frac{n}{p}}}\exp\left(-\frac{r^p}{2\sigma^2}\right)dr \\
&= \frac{p}{\Gamma\left(\frac{n}{p}\right)(2\sigma^2)^{\frac{n}{p}}}\int_0^a r^{n-1}\exp\left(-\frac{r^p}{2\sigma^2}\right)dr \\
&= \frac{1}{\Gamma\left(\frac{n}{p}\right)}\int_0^{\frac{a^p}{2\sigma^2}} y^{\frac{n}{p}-1}\exp(-y)dy \\
&= \frac{\Gamma\left(\frac{n}{p}, \frac{a^p}{2\sigma^2}\right)}{\Gamma\left(\frac{n}{p}\right)},
\end{aligned}$$

where $\Gamma(z,b) = \int_0^b y^{z-1}\exp(-y)dy$ is the incomplete Γ-function.

4. Log-Likelihood of Filters under the Log-Normal Mixture Model

The log-likelihood of a basis $\mathbf{W}$ in whitened space, given a set of whitened images $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_m\}$, is given by

$$\begin{aligned}
\mathcal{L}(\mathbf{W}|\eta,\mu,\sigma) &= \sum_{i=1}^{m}\log p(\mathbf{y}_i|\eta,\mu,\sigma,\mathbf{x}_i,\mathbf{W}) \\
&= m(n-1)\log p + m\log\Gamma\left(\frac{n}{p}\right) - mn\log 2 - mn\log\Gamma\left(\frac{1}{p}\right) + \\
&\quad \sum_{i=1}^{m}\log\left(\sum_{k=1}^{K}\frac{\eta_k}{||\mathbf{W}\mathbf{x}||_p^n\sigma_k\sqrt{2\pi}}\exp\left(-\frac{(\log||\mathbf{W}\mathbf{x}_i||_p - \mu_k)^2}{2\sigma_k^2}\right)\right).
\end{aligned}$$

Taking the derivative with respect to the jth row $\mathbf{w}_j$ of $\mathbf{W}$ yields

$$
\begin{aligned}
& \frac{\partial}{\partial \mathbf{w}_j} \mathcal{L}(\mathbf{W}|\eta,\mu,\sigma) \\
= & \sum_{i=1}^{m} \frac{\partial}{\partial \mathbf{w}_j} \log \Bigg(\underbrace{\sum_{k=1}^{K} \frac{\eta_k}{||\mathbf{W}\mathbf{x}_i||_p^n \sigma_k \sqrt{2\pi}} \exp\left(-\frac{(\log ||\mathbf{W}\mathbf{x}_i||_p - \mu_k)^2}{2\sigma_k^2} \right)}_{=:\mathcal{L}_1(\mathbf{W}|\eta,\mu,\sigma,\mathbf{x}_i)} \Bigg) \\
= & \sum_{i=1}^{m} \mathcal{L}_1(\mathbf{W}|\eta,\mu,\sigma,\mathbf{x}_i)^{-1} \cdot \sum_{k=1}^{K} \frac{\eta_k}{\sigma_k\sqrt{2\pi}} \frac{\partial}{\partial \mathbf{w}_j} \left(||\mathbf{W}\mathbf{x}_i||_p^{-n} \exp\left(-\frac{(\log ||\mathbf{W}\mathbf{x}_i||_p - \mu_k)^2}{2\sigma_k^2} \right) \right) \\
= & \sum_{i=1}^{m} \mathcal{L}_1(\mathbf{W}|\eta,\mu,\sigma,\mathbf{x}_i)^{-1} \times \\
& \sum_{k=1}^{K} \frac{\eta_k}{\sigma_k\sqrt{2\pi}} ||\mathbf{W}\mathbf{x}_i||_p^{-(n+1)} \exp\left(-\frac{(\log ||\mathbf{W}\mathbf{x}_i||_p - \mu_k)^2}{2\sigma_k^2} \right) \left(-n - \frac{1}{\sigma_k^2}(\log ||\mathbf{W}\mathbf{x}_i||_p - \mu_k) \right) \frac{\partial}{\partial \mathbf{w}_j} ||\mathbf{W}\mathbf{x}_i||_p \\
= & \sum_{i=1}^{m} \mathcal{L}_1(\mathbf{W}|\eta,\mu,\sigma,\mathbf{x}_i)^{-1} ||\mathbf{W}\mathbf{x}_i||_p^{-(n+p)} \cdot \mathbf{x}_i^\top \times \\
& \sum_{k=1}^{K} \frac{\eta_k}{\sigma_k\sqrt{2\pi}} \exp\left(-\frac{(\log ||\mathbf{W}\mathbf{x}_i||_p - \mu_k)^2}{2\sigma_k^2} \right) \left(-n - \frac{1}{\sigma_k^2}(\log ||\mathbf{W}\mathbf{x}_i||_p - \mu_k) \right) \Delta_j |\mathbf{w}_j \mathbf{x}_i|^{p-1} \ ,
\end{aligned}
$$

since $\frac{\partial}{\partial \mathbf{w}_j} ||\mathbf{W}\mathbf{x}_i||_p = \frac{\partial}{\partial \mathbf{w}_j} \left(\sum_{i=1}^{n} |\mathbf{w}_i \mathbf{x}|^p \right)^{\frac{1}{p}} = ||\mathbf{W}\mathbf{x}_i||_p^{1-p} \cdot \Delta_j |\mathbf{w}_j \mathbf{x}_i|^{p-1} \cdot \mathbf{x}_i^\top$ with $\Delta_{ij} := \operatorname{sgn}(\mathbf{w}_j \mathbf{x}_i)$.

Therefore, the gradient $\frac{\partial}{\partial \mathbf{W}} \mathcal{L}(\mathbf{W}|\eta,\mu,\sigma)$ can be written as an product between two matrices $\frac{\partial}{\partial \mathbf{W}} \mathcal{L}(\mathbf{W}|\eta,\mu,\sigma) = \mathbf{A} \cdot \mathbf{B}$ with

$$
\begin{aligned}
(\mathbf{A})_{ji} &= -\Delta_{ij} |\mathbf{w}_j \mathbf{x}_i|^{p-1} \sum_{k=1}^{K} \frac{\eta_k}{\sigma_k\sqrt{2\pi}} \exp\left(-\frac{(\log ||\mathbf{W}\mathbf{x}_i||_p - \mu_k)^2}{2\sigma_k^2} \right) \left(n + \frac{1}{\sigma_k^2}(\log ||\mathbf{W}\mathbf{x}_i||_p - \mu_k) \right) \\
(\mathbf{B})_{i\ell} &= \mathcal{L}_1(\mathbf{W}|\eta,\mu,\sigma,\mathbf{x}_i)^{-1} ||\mathbf{W}\mathbf{x}_i||_p^{-(n+p)} \cdot x_{i\ell} \\
&= \left(||\mathbf{W}\mathbf{x}_i||_p^p \sum_{k=1}^{K} \frac{\eta_k}{\sigma_k\sqrt{2\pi}} \exp\left(-\frac{(\log ||\mathbf{W}\mathbf{x}_i||_p - \mu_k)^2}{2\sigma_k^2} \right) \right)^{-1} \cdot x_{i\ell}
\end{aligned}
$$

Absolute Difference [Bits/Comp.]			Relative Difference [% wrt. cICA]	
	Color	Gray	Color	Gray
HAD - PIX	-4.0778 ± 0.0039	$-3.1275 + -0.0040$	$92.0797 + -0.0581$	$90.8566 + -0.0854$
SYM - PIX	-4.1665 ± 0.0040	$-3.1697 + -0.0037$	94.0826 ± 0.0534	$92.0834 + -0.0876$
ICA - PIX	-4.2376 ± 0.0041	$-3.2146 + -0.0037$	95.6872 ± 0.0489	$93.3870 + -0.0823$
cHAD - PIX	-4.3516 ± 0.0055	-3.4149 ± 0.0058	98.2622 ± 0.0086	99.2077 ± 0.0103
cSYM - PIX	-4.3819 ± 0.0056	-3.4242 ± 0.0058	98.9454 ± 0.0098	99.4770 ± 0.0099
cICA - PIX	-4.4286 ± 0.0057	-3.4422 ± 0.0059	100.0000 ± 0.0000	100.0000 ± 0.0000

TABLE 1. Difference in ALL for gray value and color images with standard devation over ten training and test set pairs. For computational efficiency the patch size has been chosen 7×7. The columns on the left display the absolute difference to the PIX representation. The columns on the right show the percentual difference with respect to the largest reduction achieved by ICA with non-factorial model.

5. ALL Scores For Color and Gray Value Images

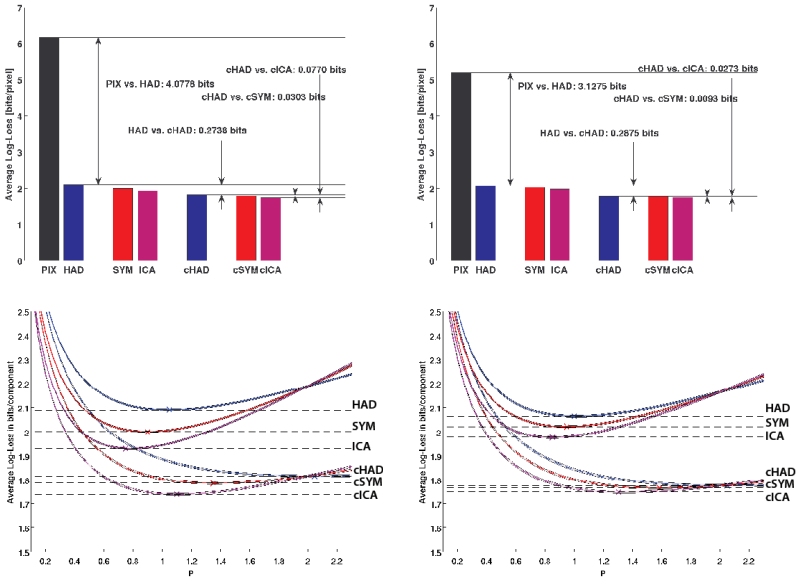


FIGURE 5.1. ALL in Bits per component as a function of p for achromatic (*right*) and chromatic (left) images. For computational efficiency both plots have been computed on patches of size 7×7. The slightly brighter envelope depicts the standard deviation over ten pairs of training and test sets. For further details see the respective figure in the paper.

References

[1] T.M. Cover and J.A. Thomas. *Elements of information theory*. J. Wiley & Sons, New York, 1991.

[2] A. K. Gupta and D. Song. l_p-norm spherical distribution. *Journal of Statistical Planning and Inference*, 60:241–260, 1997.

[3] D. Song and A. K. Gupta. l_p-norm uniform distribution. *Proceedings of the American Mathematical Society*, 125:595–601, 1997.

4.5 Characterization of the p-generalized normal distribution: Original Article

Journal of Multivariate Analysis 100 (2009) 817–820

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Note(s)

Characterization of the p-generalized normal distribution

Fabian Sinz*, Sebastian Gerwinn, Matthias Bethge

Max Planck Institute for Biological Cybernetics, Spemannstraße 41, 72076 Tübingen, Germany

ARTICLE INFO

Article history:
Received 16 April 2008
Available online 26 July 2008

AMS subject classification:
62E10

Keywords:
Characterization
Generalized normal distribution
Exponential power distribution
L_p-spherically symmetric distributions

ABSTRACT

It is a well known fact that invariance under the orthogonal group and marginal independence uniquely characterizes the isotropic normal distribution. Here, a similar characterization is provided for the more general class of differentiable bounded L_p-spherically symmetric distributions: Every factorial distribution in this class is necessarily p-generalized normal.

1. Introduction

Kac's characterization of the normal distribution [5] states that the isotropic Gaussian is the only distribution in the intersection of the class of factorial distributions and the class of spherically symmetric distributions. A natural extension to the latter are the L_p-spherically symmetric distributions [6,4]. A random variable X is L_p-spherically symmetric distributed if it can be written as a product of two independent random variables R and U, where R is a univariate non-negative random variable with an arbitrary distribution and U is uniformly distributed on the set $\mathcal{S}_p^{n-1} := \{\mathbf{x} \in \mathbb{R}^n : \sum_{i=1}^n |x_i|^p = 1\}$. Equivalently, X is L_p-spherically distributed if its density has the form $g\left(\sum_{i=1}^n |x_i|^p\right)$.

This class of distributions is of great practical interest: It offers more flexibility than the spherically symmetric model, but is still easy to fit to data since it only requires estimating the univariate radial distribution. An interesting subclass is the p-generalized Normal distribution [3]

$$g\left(\sum_{i=1}^n |x_i|^p\right) = \frac{p^n}{\left(2\Gamma\left(\frac{1}{p}\right)(2\sigma^2)^{\frac{1}{p}}\right)^n} \mathrm{e}^{-\frac{\sum_{i=1}^n |x_i|^p}{2\sigma^2}},$$

which contains the Normal distribution as a special case for $p = 2$.

Note, that the p-generalized Normal distribution is factorial with marginals from the exponential power family [1]. In that sense, the p-generalized Normal distribution is the analog of a Gaussian for L_p-spherically symmetric distributions. Surprisingly, to the best of our knowledge, we could not find any reference that characterizes the p-generalized Normal distribution as the only marginally independent L_p-spherically symmetric distribution. Here, we provide this characterization for the class of differentiable and bounded L_p-spherically symmetric densities.

* Corresponding author.
E-mail addresses: fabee@tuebingen.mpg.de (F. Sinz), sgerwinn@tuebingen.mpg.de (S. Gerwinn), mbethge@tuebingen.mpg.de (M. Bethge).
URLs: http://www.kyb.tuebingen.mpg.de/~fabee (F. Sinz), http://www.kyb.tuebingen.mpg.de/~sgerwinn (S. Gerwinn), http://www.kyb.tuebingen.mpg.de/~mbethge (M. Bethge).

0047-259X/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2008.07.006

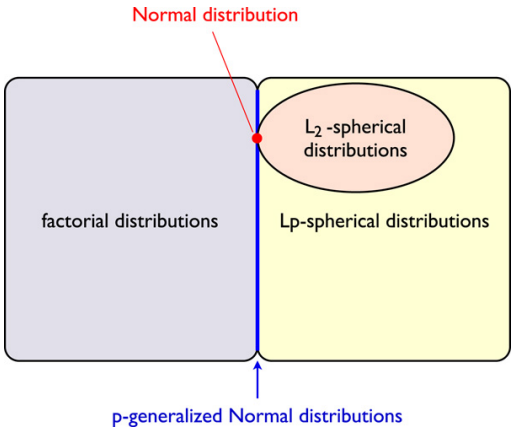


Fig. 1. Properties of the p-generalized Normal distribution. The Gaussian is the only L_2-spherically symmetric distribution with independent marginals. Like the Gaussian, all p-generalized Normal distributions have independent marginals and the property of spherical symmetry is a special case of the L_p-spherical symmetry in this class. We prove that the p-generalized Normal distributions are the only distributions which combine these two properties simultaneously.

2. Characterization

Theorem 1. *Let $g : \mathbb{R}^n \to \mathbb{R}_+$ be an differentiable multivariate L_p-spherically symmetric density. If g has the following properties:*

(1) $g \in C^1(\mathbb{R}^n)$
(2) *g and $\frac{\partial}{\partial x_i} g$ are bounded for all $i = 1, \ldots, n$*

then marginal independence, i.e. $g\left(\sum_{i=1}^n |x_i|^p\right) = \prod_{k=1}^n h(|x_k|^p)$, implies that g is p-generalized Normal, i.e.

$$h(|x_k|) = \frac{p}{2\Gamma\left(\frac{1}{p}\right)(2\sigma^2)^{\frac{1}{p}}} \exp\left(-\frac{|x_k|^p}{2\sigma^2}\right).$$

Proof. Let g be factorial, i.e. $g\left(\sum_{i=1}^n |x_i|^p\right) = \prod_{k=1}^n h_k(|x_k|^p)$, and let P be a permutation matrix. Since g is L_p-spherically symmetric, g is invariant under permutation of the basis elements, i.e. $g\left(\sum_{i=1}^n |x_i|^p\right) = g\left(\sum_{i=1}^n |y_i|^p\right)$ with $\mathbf{y} = P\mathbf{x}$. Choose a $\mathbf{v} \in \mathbb{R}^n$ for some $a \in \mathbb{R}$ with $v_j = a \cdot \delta_{ij}$ and P such that $P\mathbf{v} = \mathbf{w}$ with $w_j = a \cdot \delta_{kj}$. Thus,

$$\begin{aligned}
&g\left(\sum_{i=1}^n |v_i|^p\right) = g\left(\sum_{i=1}^n |w_i|^p\right) \\
&\Rightarrow h_i(|a|^p) \prod_{\substack{\ell=1 \\ \ell\neq i}}^n h_\ell(0) = h_k(|a|^p) \prod_{\substack{\ell=1 \\ \ell\neq k}}^n h_\ell(0) \\
&\Rightarrow h_i(a^p) = h_k(a^p) \cdot c \quad \forall a \in \mathbb{R}_+ \text{ with } c = \frac{h_i(0)}{h_k(0)}.
\end{aligned}$$

Since all h_i integrate to one, c must be one as well.

Note that none of the $h_i(0)$ can be zero because they can be written as

$$h_i(0) = \int_{\mathbb{R}^{n-1}} g\left(\sum_{k\neq i} |x_k|^p\right) \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_{i-1} \mathrm{d}x_{i+1} \ldots \mathrm{d}x_n$$

and g a non-negative function which does not vanish everywhere. Therefore, all marginals h must have the same form, that is $g\left(\sum_{i=1}^n |x_i|^p\right) = \prod_{k=1}^n h(|x_k|^p)$.

With the particular choice of $\mathbf{v}$ it follows $g\left(\sum_{i=1}^n |v_i|^p\right) = g\left(|a|^p\right) = h(|a|^p) \cdot h(0)^{n-1}$ or just $g(u) = h(u) \cdot h(0)^{n-1}$ by substitution $u := |a|^p$. Now, choosing $(a, b, 0, \ldots, 0)^\top \in \mathbb{R}^n$ we can write

$$g(|a|^p + |b|^p) = h(|a|^p) h(|b|^p) h(0)^{n-2}$$

$$
\begin{aligned}
&= g(|a|^p)h(0)^{1-n} \cdot g(|b|^p)h(0)^{1-n} \cdot h(0)^{n-2} \\
&= g(|a|^p)g(|b|^p)h(0)^{-n} \\
&= g(|a|^p)g(|b|^p)/g(0)
\end{aligned}
$$

or just $g(u+\epsilon) = g(u)g(\epsilon)/g(0)$ for all $u, \epsilon \in \mathbb{R}_+$.

Thus, we obtain

$$
g\,(u+\varepsilon) - g\,(u) = \frac{g\,(u)}{g(0)} \cdot (g\,(\varepsilon) - g(0))
$$

and it follows immediately

$$
g'(u) = \frac{g(u)}{g(0)} g'(0)
$$

Solving this differential equation uniquely yields the functional form

$$
\begin{aligned}
g\,(u) &= g(0) \exp\left(\frac{g'(0)}{g(0)} \cdot u\right) \\
&= \exp(c_1 a + c_0).
\end{aligned}
$$

Choosing a value for c_1 corresponds to setting the scale of the distribution. Taking into account that g must integrate to one determines c_0 and yields that h is in the exponential power family. Thus, g is p-generalized Normal. □

3. Discussion

The theorem presented in this paper provides an important theoretical insight showing that the intersection between the space of L_p-spherical distributions and the space of factorial distributions is a low-dimensional manifold known as the family of p-generalized Normal distributions. In particular, the previous characterization of the isotropic Gaussian as the only spherically symmetric factorial distribution can now be understood as the special case of the more general theorem when $p = 2$ (see Fig. 1). Consequently, the range of potential applications is now extended from the special case of isotropic distributions to arbitrary L_p-spherical distributions.

An immediate consequence of the theorem concerns density estimation on empirical data. Assuming marginal independence and L_p-spherical symmetry not only implies that the marginals must be exponential power distributions, but also decreases the degrees of freedom to the mean and the scale parameter of the p-generalized Normal. This shows that marginal independence is a very restrictive assumption in the class of L_p-spherical symmetric distributions which turns the infinite dimensional estimation problem of the radial distribution into a one-dimensional one.

Other consequences and applications arise from the fact that each L_p-spherically symmetric distributed random variable X has a stochastic representation $\mathsf{X} = \mathsf{RU}$. By changing the radial component with the transform $\mathcal{F}_2^{-1} \circ \mathcal{F}_1 : \mathbb{R}_+ \to \mathbb{R}_+$, where $\mathcal{F}_1$ and $\mathcal{F}_2$ are the cumulative distribution functions of the source and the target radial distribution, respectively, one can change the distribution of X within the class of L_p-spherically symmetric distributions for a particular fixed p.

From our theorem we know that there is a unique factorial distribution (up to a scale parameter) each L_p-spherically symmetric distribution can be mapped into by choosing $\mathcal{F}_2$ to be the c.d.f. of the p-generalized Normal distribution

$$
\mathcal{F}_2(r) = \mathcal{F}_p(r) = \frac{\Gamma\left(\frac{n}{p}, \frac{r^p}{2\sigma^2}\right)}{\Gamma\left(\frac{n}{p}\right)},
$$

with $\Gamma(z, a)$ denoting the incomplete Γ-function.

Conversely, one can also use this relationship for efficient sampling from arbitrary L_p-spherically symmetric distributions. The idea is to first sample from a p-generalized Normal distribution and subsequently transform the radial component by setting $\mathcal{F}_1 = \mathcal{F}_p$ and setting $\mathcal{F}_2$ equal to the c.d.f. of the radial component R of the target distribution. That is, each random vector x sampled from the p-generalized Normal is transformed by $x \mapsto \frac{(\mathcal{F}_2^{-1} \circ \mathcal{F}_p)(r)}{r} x$ with $r = \left(\sum_{i=1}^n |x_i|^p\right)^{\frac{1}{p}}$ which can be computed very fast. Furthermore, sampling from the p-generalized Normal is easy as one can sample from the univariate marginal distributions independently. Our theorem implies that the exponential power distribution is the only admissible marginal distribution with which such a sampling scheme is possible.

Finally, our theorem is also useful for constructing an independence test for L_p-spherically symmetric distributed random variables. For a given set of samples $x_1, \ldots, x_m \in \mathbb{R}^n$, the radial distribution

$$
q_r(r) = \frac{p\, r^{n-1}}{\Gamma\left(\frac{n}{p}\right) (2\sigma^2)^{\frac{n}{p}}} \mathrm{e}^{-\frac{r^p}{2\sigma^2}}
$$

of the p-generalized Normal is fitted to the radial components $r_k = \left(\sum_{i=1}^{n} |x_{ki}|^p\right)^{\frac{1}{p}}, k = 1, \ldots, m$ of the data points. Afterwards, a goodness of fit test (e.g. Kolmogorov–Smirnov) can be used to test whether the x_k come from a factorial L_p-spherically symmetric distribution. Since the p-generalized Normal is the only L_p-spherically symmetric distribution with independent marginals, the test should succeed if the marginals are independent and fail if they are not. Such an independence test can be of particular interest in the context of Independent Component Analysis [2] in order to verify whether the data actually comply with the independence assumption underlying this method.

References

[1] G.E.P. Box, G.C. Tiao, Bayesian Inference in Statistical Analysis, John Wiley & Sons Inc., 1992.
[2] P. Comon, Independent component analysis, a new concept? Signal Processing 36 (3) (1994) 287–314.
[3] I. Goodman, S. Kotz, Multivariate θ-generalized normal distributions, Journal of Multivariate Analysis 3 (1973) 204–219.
[4] A. Gupta, D. Song, l_p-norm spherical distribution, Journal of Statistical Planning and Inference 60 (1997) 241–260.
[5] M. Kac, On a characterization of the normal distribution, American Journal of Mathematics 61 (3) (1939) 726–728.
[6] J. Osiewalski, M. Steel, Robust Bayesian inference in l_q-spherical models, Biometrika 80 (1993) 456–460.

4.6 Hierarchical Modeling of Local Image Features through L_p-Nested Symmetric Distributions: Original Article

Hierarchical Modeling of Local Image Features through L_p-Nested Symmetric Distributions

Fabian Sinz
Max Planck Institute for Biological Cybernetics
Spemannstraße 41
72076 Tübingen, Germany
fabee@tuebingen.mpg.de

Eero P. Simoncelli
Center for Neural Science, and Courant Institute
of Mathematical Sciences, New York University
New York, NY 10003
eero.simoncelli@nyu.edu

Matthias Bethge
Max Planck Institute for Biological Cybernetics
Spemannstraße 41
72076 Tübingen, Germany
mbethge@tuebingen.mpg.de

Abstract

We introduce a new family of distributions, called *L_p-nested symmetric distributions*, whose densities are expressed in terms of a hierarchical cascade of L_p-norms. This class generalizes the family of spherically and L_p-spherically symmetric distributions which have recently been successfully used for natural image modeling. Similar to those distributions it allows for a nonlinear mechanism to reduce the dependencies between its variables. With suitable choices of the parameters and norms, this family includes the Independent Subspace Analysis (ISA) model as a special case, which has been proposed as a means of deriving filters that mimic complex cells found in mammalian primary visual cortex. L_p-nested distributions are relatively easy to estimate and allow us to explore the variety of models between ISA and the L_p-spherically symmetric models. By fitting the generalized L_p-nested model to 8×8 image patches, we show that the subspaces obtained from ISA are in fact more dependent than the individual filter coefficients within a subspace. When first applying contrast gain control as preprocessing, however, there are no dependencies left that could be exploited by ISA. This suggests that complex cell modeling can only be useful for redundancy reduction in larger image patches.

1 Introduction

Finding a precise statistical characterization of natural images is an endeavor that has concerned research for more than fifty years now and is still an open problem. A thorough understanding of natural image statistics is desirable from an engineering as well as a biological point of view. It forms the basis not only for the design of more advanced image processing algorithms and compression schemes, but also for a better comprehension of the operations performed by the early visual

system and how they relate to the properties of the natural stimuli that are driving it. From both perspectives, redundancy reducing algorithms such as Principal Component Analysis (PCA), Independent Component Analysis (ICA), Independent Subspace Analysis (ISA) and Radial Factorization [11; 21] have received considerable interest since they yield image representations that are favorable for compression and image processing and at the same time resemble properties of the early visual system. In particular, ICA and ISA yield localized, oriented bandpass filters which are reminiscent of receptive fields of simple and complex cells in primary visual cortex [4; 16; 10]. Together with the Redundancy Reduction Hypothesis by Barlow and Attneave [3; 1], those observations have given rise to the idea that these filters represent an important aspect of natural images which is exploited by the early visual system.

Several result, however, show that the density model of ICA is too restricted to provide a good model for natural images patches. Firstly, several authors have demonstrated that filter responses of ICA filters on natural images are not statistically independent [20; 23; 6]. Secondly, after whitening, the optimum of ICA in terms of statistical independence is very shallow or, in other words, all whitening filters yield almost the same redundancy reduction [5; 2]. A possible explanation for that finding is that, after whitening, densities of local image features are approximately spherical [24; 23; 12; 6]. This implies that those densities cannot be made independent by ICA because (i) all whitening filters differ only by an orthogonal transformation, (ii) spherical densities are invariant under orthogonal transformations, and (iii) the only spherical and factorial distribution is the Gaussian. Once local image features become more distant from each other, the contour lines of the density deviates from spherical and become more star-shaped. In order to capture this star-shaped contour lines one can use the more general L_p-spherically symmetric distributions which are characterized by densities of the form $\rho(\boldsymbol{y}) = g(\|\boldsymbol{y}\|_p)$ with $\|\boldsymbol{y}\|_p = (\sum |y_i|^p)^{1/p}$ and $p > 0$ [9; 10; 21].

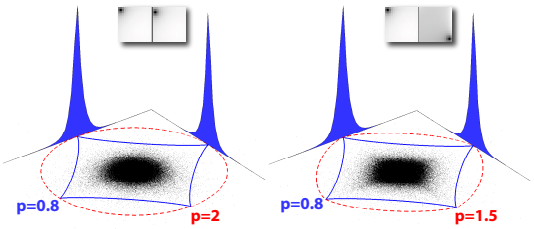


Figure 1: Scatter plots and marginal histograms of neighboring (*left*) and distant (*right*) symmetric whitening filters which are shown at the top. The dashed Contours indicate the unit sphere for the optimal p of the best fitting non-factorial (*dashed line*) and factorial (*solid line*) L_p-spherically symmetric distribution, respectively. While close filters exhibit $p = 2$ (spherically symmetric distribution), the value of p decreases for more distant filters.

As illustrated in Figure 1, the relationship between local bandpass filter responses undergoes a gradual transition from L_2-spherical for nearby to star-shaped (L_p-spherical with $p < 2$) for more distant features [12; 21]. Ultimately, we would expect extremely distant features to become independent, having a factorial density with $p \approx 0.8$. When using a single L_p-spherically symmetric model for the joint distribution of nearby and more distant features, a single value of p can only represent a compromise for the whole variety of iso-probability contours. This raises the question whether a combination of local spherical models, as opposed to a single L_p-spherical model, yields a better characterization of the statistics of natural image patches. Possible ways to join several local models are Independent Subspace Analysis (ISA) [10], which uses a factorial combination of locally L_p-spherical densities, or Markov Random Fields (MRFs) [18; 13]. Since MRFs have the drawback of being implicit density models and computationally very expensive for inference, we will focus on ISA and our model. In principle, ISA could choose its subspaces such that nearby features are grouped into a joint subspace which can then be well described by a spherical symmetric model ($p = 2$) while more distant pixels, living in different subspaces, are assumed to be independent. In fact, previous studies have found ISA to perform better than ICA for image patches as small as 8×8 and to yield an optimal $p \approx 2$ for the local density models [10]. On the other hand, the ISA model assumes a binary partition into either a L_p-spherical or a factorial distribution which does not seem to be fully justified considering the gradual transition described above.

Here, we propose a new family of hierarchical models by replacing the L_p-norms in the L_p-spherical models by L_p-nested functions, which consist of a cascade of nested L_p-norms and therefore allow for different values of p for different groups of filters. While this family includes the L_p-spherical family and ISA models, it also includes densities that avoid the hard partition into either factorial or L_p-spherical. At the same time, parameter estimation for these models can still be similarly efficient and robust as for L_p-spherically symmetric models. We find that this family (i) fits the data significantly better than ISA and (ii) generates interesting filters which are grouped in a sensible way within the hierarchy. We also find that, although the difference in performance between L_p-spherical and L_p-nested models is significant, it is small on 8×8 patches, suggesting that within this limited spatial range, the iso-probability contours of the joint density can still be reasonably approximated by a single L_p-norm. Preliminary results on 16×16 patches exhibit a more pronounced difference between the L_p-nested and the L_p-spherically symmetric distribution, suggesting that the change in p becomes more important for modelling densities over a larger spatial range.

2 Models

L_p-Nested Symmetric Distributions Consider the function

$$f(\boldsymbol{y}) = \left(\left(\sum_{i=1}^{n_1} |y_i|^{p_1} \right)^{\frac{p_\emptyset}{p_1}} + \ldots + \left(\sum_{i=n_1+\ldots+n_{\ell-1}+1}^{n} |y_i|^{p_\ell} \right)^{\frac{p_\emptyset}{p_\ell}} \right)^{\frac{1}{p_\emptyset}} \tag{1}$$
$$= \left\| \left(\|\boldsymbol{y}_{1:n_1}\|_{p_1}, \ldots, \|\boldsymbol{y}_{n-n_\ell+1:n}\|_{p_\ell} \right)^\top \right\|_{p_\emptyset}.$$

We call this type of functions L_p-*nested* and the resulting class of distributions L_p-*nested symmetric*. L_p-nested symmetric distributions are a special case of the ν-spherical distributions which have a density characterized by the form $\rho(\boldsymbol{y}) = g(\nu(\boldsymbol{y}))$ where $\nu : \mathbb{R}^n \to \mathbb{R}$ is a positively homogeneous function of degree one, i.e. it fulfills $\nu(a\boldsymbol{y}) = a\nu(\boldsymbol{y})$ for any $a \in \mathbb{R}_+$ and $\boldsymbol{y} \in \mathbb{R}^n$ [7]. L_p-nested functions are obviously positively homogeneous. Of course, L_p-nested functions of L_p-nested functions are again L_p-nested. Therefore, an L_p-nested function f in its general form can be visualized by a tree in which each inner node corresponds to an L_p-norm while the leaves stand for the coefficients of the vector $\boldsymbol{y}$.

Because of the positive homogeneity it is possible to normalize a vector $\boldsymbol{y}$ with respect to ν and obtain a coordinate respresentation $x = r \cdot \boldsymbol{u}$ where $r = \nu(\boldsymbol{y})$ and $\boldsymbol{u} = \boldsymbol{y}/\nu(\boldsymbol{y})$. This implies that the random variable Y has the stochastic representation $Y \doteq RU$ with independent U and R [7] which makes it a generalization of the Gaussian Scale Mixture model [23]. It can be shown that for a given ν, U always has the same distribution while the distribution $\varrho(r)$ of R determines the specific $\rho(\boldsymbol{y})$ [7]. For a general ν, it is difficult to determine the distribution of U since the partition function involves the surface area of the ν-unit sphere which is not analytically tractable in most cases. Here, we show that L_p-nested functions allow for an analytical expression of the partition function. Therefore, the corresponding distributions constitute a flexible yet tractable subclass of ν-spherical distributions.

In the remaining paper we adopt the following notational convention: We use multi-indices to index single nodes of the tree. This means that $I = \emptyset$ denotes the root node, $I = (\emptyset, i) = i$ denotes its i^{th} child, $I = (i, j)$ the j^{th} child of i and so on. The function values at individual inner nodes I are denoted by f_I, the vector of function values of the children of an inner node I by $\mathbf{f}_{I,1:\ell_I} = (\mathsf{f}_{I,1}, \ldots, \mathsf{f}_{I,\ell_I})^\top$. By definition, parents and children are related via $\mathsf{f}_I = \|\mathbf{f}_{I,1:\ell_I}\|_{p_I}$. The number of children of a particular node I is denoted by ℓ_I.

L_p-nested symmetric distributions are a very general class of densities. For instance, since every L_p-norm $\|\cdot\|_p$ is an L_p-nested function, L_p-nested distributions includes the family of L_p-spherically symmetric distributions including (for $p = 2$) the family of spherically symmetric distributions. When e.g. setting $f = \|\cdot\|_2$ or $f = (\|\cdot\|_2^p)^{1/p}$, and choosing the radial distribution ϱ appropriately, one can recover the Gaussian $\rho(\boldsymbol{y}) = Z^{-1} \exp\left(-\|\boldsymbol{y}\|_2^2\right)$ or the generalized spherical Gaussian $\rho(\boldsymbol{y}) = Z^{-1} \exp\left(-\|\boldsymbol{y}\|_2^p\right)$, respectively. On the other hand, when choosing the L_p-nested function f as in equation (1) and ϱ to be the radial distribution of a p-generalized Normal distribution $\varrho(r) =$

$Z^{-1}r^{n-1}\exp\left(-r^{p_\emptyset}/s\right)$ [8; 22], the inner nodes $\mathbf{f}_{1:\ell_\emptyset}$ become independent and we can recover an ISA model. Note, however, that not all ISA models are also L_p-nested since L_p-nested symmetry requires the radial distribution to be that of a p-generalized Normal.

In general, for a given radial distribution ϱ on the L_p-nested radius $f(\boldsymbol{y})$, an L_p-nested symmetric distribution has the form

$$\rho(\boldsymbol{y}) = \frac{1}{\mathcal{S}_f(f(\boldsymbol{y}))} \cdot \varrho(f(\boldsymbol{y})) = \frac{1}{\mathcal{S}_f(1) \cdot f^{n-1}(\boldsymbol{y})} \cdot \varrho(f(\boldsymbol{y})) \tag{2}$$

where $\mathcal{S}_f(f(\boldsymbol{y})) = \mathcal{S}_f(1) \cdot f^{n-1}(\boldsymbol{y})$ is the surface area of the L_p-nested sphere with the radius $f(\boldsymbol{y})$. This means that the partition function of a general L_p-nested symmetric distribution is the partition function of the radial distribution normalized by the surface area of the L_p-nested sphere with radius $f(\boldsymbol{y})$. For a given f and a radius $\mathsf{f}_\emptyset = f(\boldsymbol{y})$ this surface area is given by the equation

$$\mathcal{S}_f(\mathsf{f}_\emptyset) = \mathsf{f}_\emptyset^{n-1} 2^n \prod_{I \in \mathcal{I}} \frac{1}{p_I^{\ell_I - 1}} \prod_{k=1}^{\ell_I - 1} B\left[\frac{\sum_{i=1}^{k} n_{I,k}}{p_I}, \frac{n_{I,k+1}}{p_I}\right] = \mathsf{f}_\emptyset^{n-1} 2^n \prod_{I \in \mathcal{I}} \frac{\prod_{k=1}^{\ell_I} \Gamma\left[\frac{n_{I,k}}{p_I}\right]}{p_I^{\ell_I - 1} \Gamma\left[\frac{n_I}{p_I}\right]}$$

where $\mathcal{I}$ denotes the set of all multi-indices of inner nodes, n_I the number of leaves of the subtree under I and $B[a,b]$ the beta function. Therefore, if the partition function of the radial distribution can be computed easily, so can the partition function of the multivariate L_p-nested distribution.

Since the only part of equation (2) that includes free parameters is the radial distribution ϱ, maximum likelihood estimation of those parameters $\boldsymbol{\vartheta}$ can be carried out on the univariate distribution ϱ only, because

$$\operatorname{argmax}_{\boldsymbol{\vartheta}} \log \rho(\boldsymbol{y}|\boldsymbol{\vartheta}) \overset{(2)}{=} \operatorname{argmax}_{\boldsymbol{\vartheta}} \left(-\log \mathcal{S}_f(f(\boldsymbol{y})) + \log \varrho(f(\boldsymbol{y})|\boldsymbol{\vartheta})\right) = \operatorname{argmax}_{\boldsymbol{\vartheta}} \log \varrho(f(\boldsymbol{y})|\boldsymbol{\vartheta}).$$

This means that parameter estimation can be done efficiently and robustly on the values of the L_p-nested function.

Since, for a given f, an L_p-nested distribution is fully specified by a radial distribution, changing the radial distribution also changes the L_p-nested distribution. This suggests an image decomposition constructed from a cascade of nonlinear, gain-control-like mappings reducing the dependence between the filter coefficients. Similar to Radial Gaussianization or L_p-Radial Factorization algorithms [12; 21], the radial distribution $\varrho_\emptyset$ of the root node is mapped into the radial distribution of a p-generalized Normal via histogram equalization, thereby making its children exponential power distributed and statistically independent [22]. This procedure is then repeated recursively for each of the children until the leaves of the tree are reached.

Below, we estimate the multi-information (MI) between the filters or subtrees at different levels of the hierarchy. In order to do that robustly, we need to know the joint distribution of their values. In particular, we are interested in the joint distribution of the children $\mathbf{f}_{I,1:\ell_I}$ of a node I (e.g. layer 2 in Figure 2). Just from the form of an L_p-nested function one might guess that those children are L_p-spherically symmetric distributed. However, this is not the case. For example, the children $\mathbf{f}_{1:\ell_\emptyset}$ of the root node (assuming that none of them is a leaf) follow the distribution

$$\rho(\mathbf{f}_{1:\ell_\emptyset}) = \frac{\varrho_\emptyset(\|\mathbf{f}_{1:\ell_\emptyset}\|_{p_\emptyset})}{S_{\|\cdot\|_{p_\emptyset}}(\|\mathbf{f}_{1:\ell_\emptyset}\|_{p_\emptyset})} \prod_{i=1}^{\ell_\emptyset} \mathsf{f}_i^{n_i - 1}. \tag{3}$$

This implies that $\mathbf{f}_{1:\ell_\emptyset}$ can be represented as a product of two independent random variables $\boldsymbol{u} = \mathbf{f}_{1:\ell_\emptyset}/\|\mathbf{f}_{1:\ell_\emptyset}\|_{p_\emptyset} \in \mathbb{R}_+^{\ell_\emptyset}$ and $r = \|\mathbf{f}_{1:\ell_\emptyset}\|_{p_\emptyset} \in \mathbb{R}_+$ with $r \sim \varrho_\emptyset$ and $\left(u_1^{p_\emptyset}, ..., u_{\ell_\emptyset}^{p_\emptyset}\right) \sim \operatorname{Dir}\left[n_1/p_\emptyset, ..., n_{\ell_\emptyset}/p_\emptyset\right]$ following a Dirichlet distribution (see Additional Material). We call this distribution a *Dirichlet Scale Mixture (DSM)*. A similar form can be shown for the joint distribution of leaves and inner nodes (summarizing the whole subtree below them). Unfortunately, only the children $\mathbf{f}_{1:\ell_\emptyset}$ of the root node are really DSM distributed. We were not able to analytically calculate the marginal distribution of an arbitrary node's children $\mathbf{f}_{I,1:\ell_I}$, but we suspect it to have a similar form. For that reason we fit DSMs to those children $\mathbf{f}_{I,1:\ell_\emptyset}$ in the experiments below and use the estimated model to assess the dependencies between them. We also use it for measuring the dependencies between the subspaces of ISA.

Fitting DSMs via maximum likelihood can be carried out similarly to estimating L_p-nested distributions: Since the radial variables $\boldsymbol{u}$ and r are independent, the Dirichlet and the radial distribution can be estimated on the normalized data points $\{\boldsymbol{u}_i\}_{i=1}^m$ and their respective norms $\{r_i\}_{i=1}^m$ independently.

L_p-Spherically Symmetric Distributions and Independent Subspace Analysis The family of L_p-spherically symmetric distributions are a special case of L_p-nested distributions for which $f(\boldsymbol{y}) = \|\boldsymbol{y}\|_p$ [9]. We use the ISA model by [10] where the filter responses $\boldsymbol{y}$ are modelled by a factorial combination of L_p-spherically symmetric distributions sitting on each subspace

$$\rho(\boldsymbol{y}) = \prod_{k=1}^{K} \rho_k(\|\boldsymbol{y}_{I_k}\|_{p_k}).$$

3 Experiments

Given an image patch $\boldsymbol{x}$, all models used in this paper define densities over filter responses $\boldsymbol{y} = W\boldsymbol{x}$ of linear filters. This means, that all models have the form $\rho(\boldsymbol{y}) = |\det W| \cdot \rho(W\boldsymbol{x})$. The $(n-1)\times n$ matrix W has the form $W = QSP$ where $P \in \mathbb{R}^{(n-1)\times n}$ has mutually orthogonal rows and projects onto the orthogonal complement of the DC-filter (filter with equal coefficients), $S \in \mathbb{R}^{(n-1)\times(n-1)}$ is a whitening matrix and $Q \in SO_{n-1}$ is an orthogonal matrix determining the final filter shapes of W. When we speak of optimizing the filters according to a model, we mean optimizing Q over SO_{n-1}. The reason for projecting out the DC component is, that it can behave quite differently depending on the dataset. Therefore, it is usually removed and modelled separately. Since the DC component is the same for all models and would only add a constant offset to the measures we use in our experiments, we ignore it in the experiments below.

Data We use ten pairs of independently sampled training and test sets of 8×8 (16×16) patches from the van Hateren dataset, each containing $100,000$ ($500,000$) examples. Hyvärinen and Köster [10] report that ISA already finds several subspaces for 8×8 image patches. We perform all experiments with two different types of preprocessing: either we only whiten the data (WO-data), or we whiten it and apply an additional contrast gain control step (CGC-data), for which we use the radial factorization method described in [12; 21] with $p = 2$ in the symmetric whitening basis.

We use the same whitening procedure as in [21; 6]: Each dataset is centered on the mean over examples and dimensions and rescaled such that whitening becomes volume conserving. Similarly, we use the same orthogonal matrix to project out the DC-component of each patch (matrix P above). On the remaining $n-1$ dimensions, we perform symmetric whitening (SYM) with $S = C^{-\frac{1}{2}}$ where C denotes the covariance matrix of the DC-corrected data $C = \operatorname{cov}[PX]$.

Evaluation Measures We use the *Average Log Loss* per component (ALL) for assessing the quality of the different models, which we estimate by taking the empirical average over a large ensemble of test points $ALL = -\frac{1}{n-1}\langle \log \rho(\boldsymbol{y})\rangle_Y \approx -\frac{1}{m(n-1)}\sum_{i=1}^m \log \rho(\boldsymbol{y}_i)$. The ALL equals the entropy if the model distribution equals the true distribution and is larger otherwise. For the CGC-data, we adjust the ALL by the log-determinant of the CGC transformation [11]. In contrast to [10] this allows us to quantitively compare models across the two different types of preprocessing (WO and CGC), which was not possible in [10].

In order to measure the dependence between different random variables, we use the *multi-information* per component (MI) $\frac{1}{n-1}\left(\sum_{i=1}^d H[Y_i] - H[Y]\right)$ which is the difference between the sum of the marginal entropies and the joint entropy. The MI is a positive quantity which is zero if and only if the joint distribution is factorial. We estimate the marginal entropies by a jackknifed MLE entropy estimator [17] (corrected for the log of the bin width in order to estimate the differential entropy) where we adjust the bin width of the histograms suggested by Scott [19]. Instead of the joint entropy, we use the ALL of an appropriate model distribution. Since the ALL is theoretically always larger than the true joint entropy (ignoring estimation errors) using the ALL instead of the joint entropy should underestimate the true MI, which is still sufficient for our purpose.

Parameter Estimation For all models (ISA, DSM, L_p-spherical and L_p-nested), we estimate the parameters $\boldsymbol{\vartheta}$ for the radial distribution as described above in Section 2. For a given filter matrix

W the values of the exponents p are estimated by minimizing the ALL at the ML estimates $\hat{\boldsymbol{\vartheta}}$ over $\boldsymbol{p} = (p_1, ..., p_q)^\top$. For the L_p-nested distributions, we use the Nelder-Mead [15] method for the optimization over $\boldsymbol{p} = (p_1, ..., p_q)^\top$ and for the L_p-spherically symmetric distributions we use Golden Search over the single p. For the ISA model, we carry out a Golden Search over p for each subspace independently. For the L_p-spherical and the single models on the ISA subspaces, we use a search range of $p \in [0.1, 2.1]$ on p. For estimating the Dirichlet Scale Mixtures, we use the `fastfit` package by Tom Minka to estimate the parameters of the Dirichlet distribution. The radial distribution is estimated independently as described above.

When fitting the filters W to the different models (ISA, L_p-spherical and L_p-nested), we use a gradient ascent on the log-likelihood over the orthogonal group by alternating between optimizing the parameters $\mathbf{p}$ and ϑ and optimizing for W. For the gradient ascent, we compute the standard Euclidean gradient with respect to $W \in \mathbb{R}^{(n-1)\times(n-1)}$ and project it back onto the tangent space of SO_{n-1}. Using the gradient ∇W obtained in that manner, we perform a line search with respect to t using the backprojections of $W + t \cdot \nabla W$ onto SO_{n-1}. This method is a simplified version of the one proposed by [14].

Experiments with Independent Subspace Analysis and L_p-Spherically Symmetric Distributions We optimized filters for ISA models with $K = 2, 4, 8, 16$ subspaces embracing $32, 16, 8, 4$ components (one subspace always had one dimension less due to the removal of the DC component), and for an L_p-spherically symmetric model. When optimizing for W we use a radial Γ-distribution for the L_p-spherically symmetric models and a radial χ^p distribution ($\|\boldsymbol{y}_{I_k}\|_{p_k}^{p_k}$ is Γ-distributed) for the models on the single single subspaces of ISA, which is closer to the one used by [10]. After optimization, we make a final optimization for $\boldsymbol{p}$ and $\boldsymbol{\vartheta}$ using a mixture of log normal distributions ($\log \mathcal{N}$) with $K = 6$ mixture components on the radial distribution(s).

L_p-Nested Symmetric Distributions As for the L_p-spherically symmetric models, we use a radial Γ-distribution for the optimization of W and a mixture of $\log \mathcal{N}$ distributions for the final fit. We use two different kind of tree structures for our experiments with L_p-nested symmetric distributions. In the *deep tree* (DT) structure we first group 2×2 blocks of four neighboring SYM filters. Afterwards, we group those blocks again in a quadtree manner until we reached the root node (see Figure 2**A**). The second tree structure (PND_k) was motivated by ISA. Here, we simply group the filter within each subspace and joined them at the root node afterwards (see Figure 2**B**). In order to speed up parameter estimation, each layer of the tree shared the same value of p.

Multi-Information Measurements For the ISA models, we estimated the MI between the filter responses within each subspace and between the L_p-radii $\|\boldsymbol{y}_{I_k}\|_{p_k}, 1 \leq k \leq K$. In the former case we used the ALL of an L_p-spherically symmetric distribution with especially optimized p and $\boldsymbol{\vartheta}$, in the latter a DSM with optimized radial and Dirichlet distribution as a surrogate for the joint entropy. For the L_p-nested distribution, we estimate the MI between the children $\mathbf{f}_{I,1:\ell_I}$ of all inner nodes I. In case the children are leaves, we use the ALL of an L_p-spherically symmetric distribution as surrogate for the joint entropy, in case the children are inner nodes themselves, we use the ALL of an DSM. The red arrows in Figure 2**A** exemplarily depict the entities between which the MI was estimated.

4 Results and Discussion

Figure (2) shows the optimized filters for the DT and the PND_{16} tree structure (we included the filters optimized on the first of ten datasets for all tree structures in the Additional Material). For both tree structures, the filters on the lowest level are grouped according to spatial frequency and orientation, whereas the variation in orientation is larger for the PND_{16} tree structure and some filters are unoriented. The next layer of inner nodes, which is only present in the DT tree structure, roughly joins spatial location, although each of those inner nodes has one child whose leaves are global filters.

When looking at the various values of p at the inner nodes, we can observe that nodes which are higher up in the tree usually exhibit a smaller value of p. Surprisingly, as can be seen in Figure 3 **B** and **C**, a smaller value of p does not correspond to a larger independence between the subtrees, which are even more correlated because almost every subtree contains global filters. The small value of p is caused by the fact that the DSM (the distribution of the subtree values) has to account for this correlation which it can only do by decreasing the value of p (see Figure 3 and the DSM in

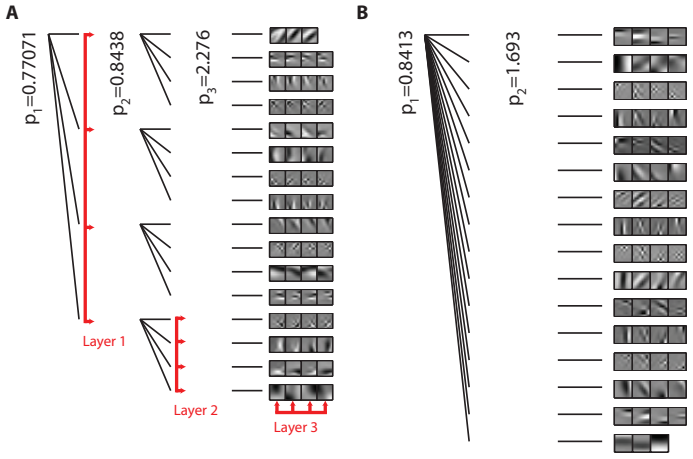


Figure 2: **Examples for the tree structures of L_p-nested distributions used in the experiments:** (**A**) shows the DT structure with the corresponding optimized values. The red arrows display examples of groups of filters or inner nodes, respectively, for which we estimated the MI. (**B**) shows the PND_{16} tree structure with the corresponding values of p at the inner nodes and the optimized filters.

the Additional Material). Note that this finding is exactly opposite to the assumptions in the ISA model which can usually not generate such a behavior (Figure 3**A**) as it models the two subtrees to be independent. This is likely to be one reason for the higher ALL of the ISA models (see Table 1).

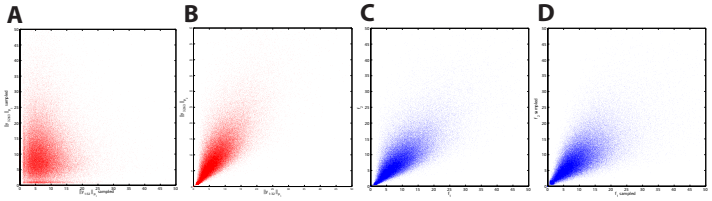


Figure 3: Independence of subspaces for WO-data not justfied: (**A**) Subspace radii sampled from ISA_2, (**B**) subspace radii of natural image patches in the ISA_2 basis, (**C**) subtree values of the PND_2 in the PND_2 basis, and (**D**) samples from the PND_2 model. While the ISA_2 model spreads out the radii almost over the whole positive quadrant due to the independence assumption the samples from the L_p-nested subtrees are more concentrated around the diagonal like the true data. The L_p-nested model can achieve this behavior since (i) it does not assume a radial distribution that leads to independent radii on the subtrees and (ii) the subtree values f_1 and f_2 are $\text{DSM}[n_1/p_\emptyset, n_2/p_\emptyset,]$ distributed. By changing the value of $p_\emptyset$, the DSM model can put more mass towards the diagonal, which produces the "beam-like" behavior shown in the plot.

Table 1 shows the ALL and the MI measurements for all models. Except for the ISA models on WO-data, all performances are similar, whereas the L_p-nested models usually achieve the lowest ALL independent of the particular tree structure used. For the WO-data, the L_p-spherical and the ISA_2 model come close to the performance of the L_p-nested models. For the other ISA models on WO-data the ALL gets worse with increasing number of subspaces (bold font number in Table 1). This reflects the effect described above: Contrary to the assumptions of the ISA model, the responses of the different subspaces become in fact more correlated than the single filter responses. This can also be seen in the MI measurements discussed below.

When looking at the ALL for CGC data, on the other hand, ISA suddenly becomes competitive. This importance of CGC for ISA has already been noted in [10]. The small differences between all the models in the CGC case shows that the contour change of the joint density for 8×8 patches is too small to allow for a large advantage of the L_p-nested model, because contrast gain control (CGC)

directly corresponds to modeling the distribution with an L_p-spherically symmetric distribution [21]. Preliminary results on 16×16 data (1.39 ± 0.003 for the L_p-nested and 1.45 ± 0.003 for the L_p-spherical model on WO-data), however, show a more pronounced improvement with for the L_p-nested model, indicating that a single p does not suffice anymore to capture all dependencies when going to larger patch sizes.

When looking at the MI measurements between the filters/subtrees at different levels of the hierarchy in the L_p-nested, L_p-spherically symmetric and ISA models, we can observe that for the WO-data, the MI actually increases when going from lower to higher layers. This means that the MI between the direct filter responses (layer 3 for DT and layer 2 for all others) is in fact lower than the MI between the subspace radii or the inner nodes of the L_p-nested tree (layer 1-2 for DT, layer 1 for all others). The highest MI is achieved between the children of the root node for the DT tree structure (DT layer 1). As explained above this observation contradicts the assumptions of the ISA model and probably causes it worse performance on the WO-data.

For the CGC-data, on the other hand, the MI has been substantially decreased by CGC over all levels of the hierarchy. Furthermore, the single filter responses inside a particular subspace or subtree are now more dependent than the subtrees or subspaces themselves. This suggests that the competitive performance of ISA is not due to the model but only due to the fact that CGC made the data already independent. In order to double check this result, we fitted an ICA model to the CGC-data [21] and found an ALL of 1.41 ± 0.004 which is very close to the performance of ISA and the L_p-nested distributions (which would not be the case for WO-data [21]).

Taken together, the ALL and the MI measurements suggest that ISA is not the best way to join multiple local models into a single joint model. The basic assumption of the ISA model for natural images is that filter coefficients can either be dependent within a subspace or must be independent between different subspaces. However, the increasing ALL for an increasing number of subspaces and the fact that the MI between subspaces is actually higher than within the subspaces, demonstrates that this hard partition is not justified when the data is only whitened.

Family	**L_p-nested**				
Model	Deep Tree	PND_2	PND_4	PND_8	PND_{16}
ALL	1.39 ± 0.004	1.39 ± 0.004	1.39 ± 0.004	1.40 ± 0.004	1.39 ± 0.004
ALL CGC	1.39 ± 0.005	1.40 ± 0.004	1.40 ± 0.005	1.40 ± 0.004	1.39 ± 0.004
MI Layer 1	0.84 ± 0.019	0.48 ± 0.008	0.7 ± 0.002	0.75 ± 0.003	0.61 ± 0.0036
MI Layer 1 CGC	0.0 ± 0.004	0.10 ± 0.002	0.02 ± 0.003	0.0 ± 0.009	0.0 ± 0.01
MI Layer 2	0.42 ± 0.021	0.35 ± 0.017	0.33 ± 0.017	0.28 ± 0.019	0.25 ± 0.025
MI Layer 2 CGC	0.002 ± 0.005	0.01 ± 0.0008	0.01 ± 0.004	0.01 ± 0.006	0.02 ± 0.008
MI Layer 3	0.28 ± 0.036	-	-	-	-
MI Layer 3 GCG	0.04 ± 0.005	-	-	-	-
Family	**L_p-spherical**	**ISA**			
Model	-	ISA_2	ISA_4	ISA_8	ISA_{16}
ALL	1.41 ± 0.004	$\mathbf{1.40 \pm 0.005}$	$\mathbf{1.43 \pm 0.006}$	$\mathbf{1.46 \pm 0.006}$	$\mathbf{1.55 \pm 0.006}$
ALL CGC	1.41 ± 0.004	1.41 ± 0.008	1.39 ± 0.007	1.40 ± 0.005	1.41 ± 0.007
MI Layer 1	0.34 ± 0.004	0.47 ± 0.01	0.69 ± 0.012	0.7 ± 0.018	0.63 ± 0.0039
MI Layer 1 CGC	0.00 ± 0.005	0.00 ± 0.09	0.00 ± 0.06	0.00 ± 0.04	0.00 ± 0.02
MI Layer 2	-	0.36 ± 0.017	0.33 ± 0.019	0.31 ± 0.032	0.24 ± 0.024
MI Layer 2 CGC	-	0.004 ± 0.003	0.03 ± 0.012	0.02 ± 0.018	0.0006 ± 0.013

Table 1: **ALL and MI for all models:** The upper part shows the results for the L_p-nested models, the lower part show the results for the L_p-spherical and the ISA models. The ALL for the L_p-nested models is almost equal for all tree structures and a bit lower compared to the L_p-spherical and the ISA models. For the whitened only data, the ALL increases significantly with the number of subspaces (**bold** font). For the CGC data, most models perform similarly well. When looking at the MI, we can see that higher layers for whitened only data are in fact more dependent than lower ones. For CGC data, the MI has dropped substantially over all layers due to CGC. In that case, the lower layers are more independent.

In summary, our results show that L_p-nested symmetric distributions yield a good performance on natural image patches, although the advantage over L_p-spherically symmetric distributions is fairly small, suggesting that the distribution within these small patches (8×8) is captured reasonably well by a single L_p-norm. Furthermore, our results demonstrate that—at least for 8×8 patches—the assumptions of ISA are too rigid for WO-data and are trivially fulfilled for the CGC-data, since CGC already removed most of the dependencies. We are currently working to extend this study to larger patches, which we expect will reveal a more significant advantage for L_p-nested models.

References

[1] F. Attneave. Informational aspects of visual perception. *Psychological Review*, 61:183–193, 1954.

[2] R. Baddeley. Searching for filters with "interesting" output distributions: an uninteresting direction to explore? *Network: Computation in Neural Systems*, 7(2):409–421, 1996.

[3] H. B. Barlow. *Sensory mechanisms, the reduction of redundancy, and intelligence*. 1959.

[4] Anthony J. Bell and Terrence J. Sejnowski. An Information-Maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, November 1995.

[5] Matthias Bethge. Factorial coding of natural images: how effective are linear models in removing higher-order dependencies? *Journal of the Optical Society of America A*, 23(6):1253–1268, June 2006.

[6] Jan Eichhorn, Fabian Sinz, and Matthias Bethge. Natural image coding in v1: How much use is orientation selectivity? *PLoS Comput Biol*, 5(4):e1000336, April 2009.

[7] Carmen Fernandez, Jacek Osiewalski, and Mark F. J. Steel. Modeling and inference with ν-spherical distributions. *Journal of the American Statistical Association*, 90(432):1331–1340, Dec 1995.

[8] Irwin R. Goodman and Samuel Kotz. Multivariate θ-generalized normal distributions. *Journal of Multivariate Analysis*, 3(2):204–219, Jun 1973.

[9] A. K. Gupta and D. Song. l_p-norm spherical distribution. *Journal of Statistical Planning and Inference*, 60:241–260, 1997.

[10] A. Hyvarinen and U. Koster. Complex cell pooling and the statistics of natural images. *Network: Computation in Neural Systems*, 18(2):81–100, 2007.

[11] S Lyu and E P Simoncelli. Nonlinear extraction of 'independent components' of natural images using radial Gaussianization. *Neural Computation*, 21(6):1485–1519, June 2009.

[12] S Lyu and E P Simoncelli. Reducing statistical dependencies in natural signals using radial Gaussianization. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Adv. Neural Information Processing Systems 21*, volume 21, pages 1009–1016, Cambridge, MA, May 2009. MIT Press.

[13] Siwei Lyu and E.P. Simoncelli. Modeling multiscale subbands of photographic images with fields of gaussian scale mixtures. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(4):693–706, 2009.

[14] J. H. Manton. Optimization algorithms exploiting unitary constraints. *IEEE Transactions on Signal Processing*, 50:635 – 650, 2002.

[15] J. A. Nelder and R. Mead. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313, Jan 1965.

[16] Bruno A. Olshausen and David J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, June 1996.

[17] Liam Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253, Jun 2003.

[18] S. Roth and M.J. Black. Fields of experts: a framework for learning image priors. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 860–867 vol. 2, 2005.

[19] David W. Scott. On optimal and data-based histograms. *Biometrika*, 66(3):605–610, Dec 1979.

[20] E.P. Simoncelli. Statistical models for images: compression, restoration and synthesis. In *Signals, Systems & Computers, 1997. Conference Record of the Thirty-First Asilomar Conference on*, volume 1, pages 673–678 vol.1, 1997.

[21] F. Sinz and M. Bethge. The conjoint effect of divisive normalization and orientation selectivity on redundancy reduction. In *Neural Information Processing Systems 2008*, 2009.

[22] F. H. Sinz, S. Gerwinn, and M. Bethge. Characterization of the p-generalized normal distribution. *Journal of Multivariate Analysis*, 100(5):817–820, 05 2009.

[23] M. J. Wainwright and E. P. Simoncelli. Scale mixtures of gaussians and the statistics of natural images. In *Advances in neural information processing systems*, volume 12, pages 855–861, 2000.

[24] Christoph Zetzsche, Gerhard Krieger, and Bernhard Wegmann. The atoms of vision: Cartesian or polar? *Journal of the Optical Society of America A*, 16(7):1554–1565, Jul 1999.

4.7 Hierarchical Modeling of Local Image Features through L_p-Nested Symmetric Distributions: Supplementary Material

L_p-NESTED SYMMETRIC DISTRIBUTIONS

FABIAN SINZ, EERO SIMONCELLI, MATTHIAS BETHGE

1. Introduction

A important part in statistical analysis of data is to find a class of models that is flexible and rich enough to model the regularities in the data, but at the same time exhibits enough symmetry and structure itself to still be computationally and analytically tractable. One special way of introducing such a symmetry is to fix the general form of the isodensity contour lines. This approach was taken by [2] who modelled the contour lines by the level sets of a positively homogeneous function of degree one. Unfortunately, in the general case it is hard to derive the normalization constant for an arbitrary such function. For a special kind of ν-spherical distributions, the L_p-spherically symmetric distributions [5; 3] this problem becomes tractable by restricting the contour lines to L_p-spheres, but at the prize of introducing permutation symmetry. The L_p-spherically symmetric distribution itself generalize the class of L_2-spherically symmetric distributions which exhibit rotational symmetry [4; 1]. In some cases permutation or even rotational symmetry might be an appropriate assumption for the data. However, in other cases such symmetries might actually make the model miss important structure present in the data.

Here, we present a generalization of the class of L_p-spherically symmetric distribution within the class of ν-spherical distributions. Instead of using a single L_p-norm to define the contour of the density, we use nested L_p-norms where the coefficients, the L_p-norm is computed over, can be L_p-norms themselves—with possibly different p. This preserves positive homogeneity and replaces permutational invariance with invariance under reflection at the coordinate axes. Due to the nested structure, we call this new class of distributions *L_p-nested symmetric distributions*. As we demonstrate below, this construction still bears enough structure to define polar-like coordinates similar to those of [6; 3] and thereby to compute the normalization constant of the distribution given an arbitrary univariate distribution on the function values. By that construction, we can leverage most important properties of the L_p-spherically symmetric distributions to the L_p-nested distributions.

The remaining part of the paper is structured as follows: In section 2 we introduce some helpful nomenclature and define L_p-nested functions. In section 3 we define coordinates in the spirit of [3] and derive the Jacobian of the determinant. In section 4 we introduce the uniform distribution on the L_p-nested unit sphere which allows us to leverage some of the results of [3] to L_p-nested symmetric distributions in section 5. In section 6 we derive a sampling scheme for L_p-nested symmetric distributions. We conclude by presenting a potential application for the class of L_p-nested symmetric distributions.

Date: October 30, 2009.

2. Nomenclature and Definitions

Definition 2.1 (L_p-nested functions)**.** We call a function $f : \mathbb{R}^n \to \mathbb{R}_0^+$ *L_p-nested* if f fulfills the following recursive definition:

(i) The function $f : \mathbb{R}^n \to \mathbb{R}$ is the L_p-norm of its ℓ children $(f_1(\mathbf{x}_1), ..., f_\ell(\mathbf{x}_\ell))^\top$:

$$f(\mathbf{x}) = ||(f_1(\mathbf{x}_1), ..., f_\ell(\mathbf{x}_\ell))^\top||_p,$$

where the $\mathbf{x}_j \in \mathbb{R}^{n_j}$ are a partition of the vector $\mathbf{x}$ into ℓ parts.

(ii) The children f_i are either L_p-nested functions themselves or compute the absolute value of a single coefficient x_i, i.e. $f_j(\mathbf{x}_j) = |x_i|$ if and only if $\mathbf{x}_j = x_i \in \mathbb{R}$.

This gives rise to a tree structure of f which is depicted in Figure 1. Note, that every L_p-nested function is positively homogeneous by construction. In order to present results for arbitrary L_p-nested functions, we start by introducing some helpful notation.

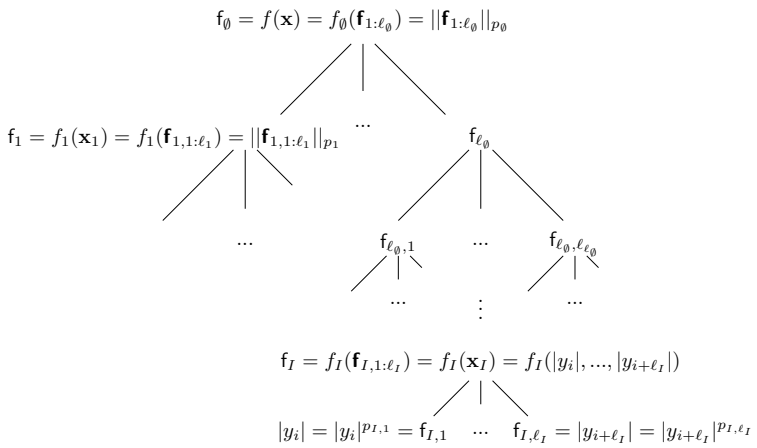


Figure 1. **Tree structure associated with an L_p-nested function f:** Every parent node I gets its value f_I by computing the L_{p_I}-norm of the values of its children $\mathbf{f}_{I,1:\ell_I}$. The leafs of the tree correspond to the (absolute values) of the coefficients in the vector $\mathbf{x}$. The values of the p at the leaf nodes are set to the value $p = 1$ by definition, e.g. $p_{I,1} = ... = p_{I,\ell_I} = 1$ in the diagram.

Definition 2.2 (Notation and Conventions for L_p-nested functions)**.** We use the following notational conventions:

(i) We use multi-indices to denote the different nodes of the tree corresponding to an L_p-nested function f. The function f itself corresponds to the root node and is denoted by $f_\emptyset$. The functions corresponding to its children are denoted by $f_1, ..., f_{\ell_\emptyset}$. The children of the i^{th} child are denoted by $f_{i,1}, ..., f_{i,\ell_i}$. In this manner, an index is added for each layer of the tree.

(ii) We always use the letter "ℓ" to denote the total amount of children of a node.

(iii) For notational convenience, we assign a p to each of the leaf nodes (i.e. the absolute values $|x_i|$) but fix their values to $p = 1$ by definition.

(iv) For the sake of compact notation, we denote a list of indices with a single multi-index $I = i_1, ..., i_\ell$. The range of the single indices and the length of the multi-index should be clear from the context. Multi-indices are always denoted by upper-case letters. A concatenation I, k of a multi-index I with another index k corresponds to adding k to the index list, i.e. $I, k = i_1, ..., i_m, k$. We use the convention that $I, \emptyset = I$.

(v) Those coefficients of the vector $\mathbf{x}$ that correspond to leafs of the subtree under a node with the index I are denoted by $\mathbf{x}_I$. The number of leafs in a subtree under a node I is denoted by n_I. If I denotes a leaf then $n_I = 1$.

(vi) The L_p-nested function associated with the subtree under a node I is denoted by

$$f_I(\mathbf{x}_I) = ||(f_{I,1}(\mathbf{x}_{I,1}), ..., f_{I,\ell_I}(\mathbf{x}_{I,\ell_I}))^\top||_{p_I}.$$

We use sans-serif font to denote the function value $\mathsf{f}_I = f_I(\mathbf{x}_I)$ of a subtree I. In many cases we use f_I and $f_I(\mathbf{x}_I)$ interchangeably. Whether f_I is to be considered as a function of its children or merely the value of the node I should always be clear from the context.

A vector with the function values of the children of I is denoted with bold sans-serif font and the following index-list notation:

$$\begin{aligned} f_I(\mathbf{x}_I) &= ||(f_{I,1}(\mathbf{x}_{I,1}), ..., f_{I,\ell_I}(\mathbf{x}_{I,\ell_I}))^\top||_{p_I} \\ &= ||(\mathsf{f}_{I,1}, ..., \mathsf{f}_{I,\ell_I})^\top||_{p_I} \\ &= ||\mathbf{f}_{I,1:\ell_I}||_{p_I} \end{aligned}$$

(vii) The function computing the value of the ℓ^{th} —and therefore by convention last—child of a node I when fixing the value f_I of that node, is denoted by

$$\begin{aligned} g_{I,\ell_I}(\mathsf{f}_I, \mathsf{f}_{I,1}, ..., \mathsf{f}_{I,\ell_I-1}) &= \left(\mathsf{f}_I^{p_I} - \sum_{k=1}^{\ell_I-1} \mathsf{f}_{I,k}^{p_I}\right)^{\frac{1}{p_I}} \\ &= g_{I,\ell_I}(\mathbf{f}_{I,\emptyset:\ell_I-1}) \\ &= \mathsf{g}_{I,\ell_I}. \end{aligned}$$

Notice the small but important difference that the value f_I depends only on the values of its children $\mathsf{f}_{I,1}, ..., \mathsf{f}_{I,\ell_I}$, while the value g_{I,ℓ_I} depends on the value of its neighbors $\mathsf{f}_{I,1}, ..., \mathsf{f}_{I,\ell_I-1}$ and its parent $\mathsf{f}_I = \mathsf{f}_{I,\emptyset}$.

(viii) Vectors in $\mathbb{R}^n$ that lie on the L_p-nested unit sphere, i.e. that fulfill $f(\mathbf{u}) = 1$ are denoted by the letter $\mathbf{u}$.

Vectors $\tilde{\mathbf{u}} \in \mathbb{R}^{\ell_I}$ that lie on the L_{p_I} unit sphere associated with the inner node I, i.e. that fulfill $\mathbf{f}_{I:1:\ell_I} = \mathsf{f}_I \tilde{\mathbf{u}}$ are denoted by the letter $\tilde{\mathbf{u}}$. Note that the coordinates $\mathbf{u}$ and $\tilde{\boldsymbol{u}}$ are different: $f_I(\tilde{\mathbf{u}}) = 1$ while $f_I(\mathbf{u}_I) \leq 1$.

When defining polar-like coordinates in section 3 only all but the last coefficients of $\mathbf{u}$ or $\tilde{\boldsymbol{u}}$ are needed, since the last can be computed from the remaining ones. We often still denote this shorter vectors by $\mathbf{u}$ or $\tilde{\boldsymbol{u}}$. The actual dimensionality should be clear from the context.

Let us demonstrate the above definitions with a simple example.

Example 2.1. Consider the L_p-nested function

$$\begin{aligned} f(\mathbf{x}) &= \left(\left(|x_1|^{p_1} + |x_2|^{p_1} \right)^{\frac{p_\emptyset}{p_1}} + |x_3|^{p_\emptyset} \right)^{\frac{1}{p_\emptyset}} \\ &= \left(\left(\left(\mathsf{f}_{1,1}^{p_{1,1}}\right)^{\frac{p_1}{p_{1,1}}} + \left(\mathsf{f}_{1,2}^{p_{1,2}}\right)^{\frac{p_1}{p_{1,2}}} \right)^{\frac{p_\emptyset}{p_1}} + \left(\mathsf{f}_2^{p_2}\right)^{\frac{p_\emptyset}{p_2}} \right)^{\frac{1}{p_\emptyset}} \\ &= \left(f_1 \left(\mathbf{f}_{1,1:2} \right)^{p_\emptyset} + f_2 \left(\mathsf{f}_{2,1} \right)^{p_\emptyset} \right)^{\frac{1}{p_\emptyset}} \\ &= f_\emptyset \left(\mathbf{f}_{1:2} \right) \end{aligned}$$

with $\ell_\emptyset = 2$, $\ell_1 = 2$ and $p_{1,1} = p_{1,2} = p_2 = 1$ by definition. Resolving $f(x_1, x_2, x_3) = a$ for $|x_3|$ yields the functions g

$$\begin{aligned} |x_3| &= \mathsf{g}_2 \\ &= g_2 \left(\mathsf{f}_\emptyset, \mathsf{f}_1 \right) \\ &= \left(\mathsf{f}_\emptyset^{p_\emptyset} - \mathsf{f}_1^{p_\emptyset} \right)^{\frac{1}{p_\emptyset}} \\ &= \left(a^{p_\emptyset} - f_1 \left(\mathbf{f}_{1,1:2} \right)^{p_\emptyset} \right)^{\frac{1}{p_\emptyset}} \\ &= \left(a^{p_\emptyset} - \left(|x_1|^{p_1} + |x_2|^{p_1} \right)^{\frac{p_\emptyset}{p_1}} \right)^{\frac{1}{p_\emptyset}} \end{aligned}$$

3. L_p-nested Coordinate Transformation and the Determinant of its Jacobian

The most important consequence of the positive homogeneity of f is that it can be used to normalized vectors and, by that property, to generalize the polar-like coordinates using L_p-norms of [3].

Definition 3.1 (Polar-like Coordinates)**.** We define the following polar-like coordinates for a vector $\mathbf{x} \in \mathbb{R}^n$:

$$\begin{aligned} u_i &= \frac{x_i}{f(\mathbf{x})} \text{ for } i = 1, ..., n-1 \\ r &= f(\mathbf{x}). \end{aligned}$$

The inverse coordinate transformation is given by

$$\begin{aligned} x_i &= r u_i \text{ for } i = 1, ..., n-1 \\ x_n &= r \Delta_n u_n \end{aligned}$$

where we define $\Delta_n = \operatorname{sgn} x_n$ and u_n to be the value of the leaf corresponding to $|x_n|$ when setting $\mathsf{f}_\emptyset = 1$.

The definition of the coordinates is basically equivalent to that of [3] with the difference that the L_p-norm is replaced by an L_p-nested function. Just as in the case of L_p-spherical coordinates, it will turn out that the Jacobian of the coordinate

transformation does not depend on the value of Δ_n. This is basically a consequence of the invariance under reflection at the coordinate axes.

The remaining part of this section will be devoted to computing the determinant of the Jacobian. We start by stating the general form of the determinant in terms of the partial derivatives $\frac{\partial u_n}{\partial u_k}$, u_k and r. Afterwards we demonstrate that those partial derivatives have a special form and that most of them cancel in the Laplace expansion of the determinant.

Lemma 3.1 (Determinant of the Jacobian)**.** *Let r and $\mathbf{u}$ be defined as in Definition (3.1). The general form of the determinant of the Jacobian $\mathcal{J}$ of the inverse coordinate transformation is given by*

$$|\det \mathcal{J}| = r^{n-1} \left(-\sum_{k=1}^{n-1} \frac{\partial u_n}{\partial u_k} \cdot u_k + u_n \right). \tag{1}$$

Proof. The partial derivatives of the inverse coordinate transformation are given by:

$$\begin{aligned}
\frac{\partial}{\partial u_k} y_i &= \delta_{ik} r \text{ for } 1 \leq i, k \leq n-1 \\
\frac{\partial}{\partial u_k} y_n &= \Delta_n r \frac{\partial u_n}{\partial u_k} \text{ for } 1 \leq k \leq n-1 \\
\frac{\partial}{\partial r} y_i &= u_i \text{ for } 1 \leq i \leq n-1 \\
\frac{\partial}{\partial r} y_n &= \Delta_n u_n.
\end{aligned}$$

Therefore, the structure of the Jacobian is given by

$$\mathcal{J} = \begin{pmatrix} r & \dots & 0 & u_1 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & r & u_{n-1} \\ \Delta_n r \frac{\partial u_n}{\partial u_1} & \dots & \Delta_n r \frac{\partial u_n}{\partial u_{n-1}} & \Delta_n u_n \end{pmatrix}.$$

Since we are only interested in the absolute value of the determinant and since $\Delta_n \in \{-1, 1\}$, we can factor out Δ_n and drop it. Furthermore, we can factor out r from the first $n-1$ columns which yields

$$|\det \mathcal{J}| = r^{n-1} \left| \det \begin{pmatrix} 1 & \dots & 0 & u_1 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & 1 & u_{n-1} \\ \frac{\partial u_n}{\partial u_1} & \dots & \frac{\partial u_n}{\partial u_{n-1}} & u_n \end{pmatrix} \right|.$$

Now we can use Laplace formula to expand the determinant with respect to the last column. For that purpose, let $\mathcal{J}_i$ denote the matrix which is obtained by deleting

the last column and the ith row from $\mathcal{J}$. This matrix has the following structure

$$\mathcal{J}_i = \begin{pmatrix} 1 & & & 0 & & & \\ & \ddots & & 0 & & & \\ & & 1 & 0 & & & \\ & & & \vdots & 1 & & \\ & & & 0 & & \ddots & \\ & & & 0 & & & 1 \\ \frac{\partial u_n}{\partial u_1} & & & \frac{\partial u_n}{\partial u_i} & & & \frac{\partial u_n}{\partial u_{n-1}} \end{pmatrix}.$$

We can transform $\mathcal{J}_i$ into a lower triangular matrix by moving the column with all zeros and $\frac{\partial u_n}{\partial u_i}$ bottom entry to the rightmost column of $\mathcal{J}_i$. Each swapping of two columns introduces a factor of -1. In the end, we can compute the value of $\det \mathcal{J}_i$ by simply taking the product of the diagonal entries and obtain $\det \mathcal{J}_i = (-1)^{n-1-i} \frac{\partial u_n}{\partial u_i}$. This yields

$$\begin{aligned} |\det \mathcal{J}| &= r^{n-1} \left(\sum_{k=1}^{n} (-1)^{n+k} u_k \det \mathcal{J}_k \right) \\ &= r^{n-1} \left(\sum_{k=1}^{n-1} (-1)^{n+k} u_k \det \mathcal{J}_k + (-1)^{2n} \frac{\partial y_n}{\partial r} \right) \\ &= r^{n-1} \left(\sum_{k=1}^{n-1} (-1)^{n+k} u_k (-1)^{n-1-k} \frac{\partial u_n}{\partial u_k} + u_n \right) \\ &= r^{n-1} \left(- \sum_{k=1}^{n-1} u_k \frac{\partial u_n}{\partial u_k} + u_n \right). \end{aligned}$$

□

For a given L_p-nested function f, the terms r, u_k and $\frac{\partial u_n}{\partial u_k}$ needed to compute the determinant with equation (1) can be computed easily. However, as already mentioned, most constituents of those terms cancel each other as the following example demonstrates. We urge the reader to follow the next example as it contains the important ideas for the general case below.

Example 3.1. Consider the function from the previous example

$$f(\mathbf{y}) = \left((|x_1|^{p_1} + |x_2|^{p_1})^{\frac{p_\emptyset}{p_1}} + |x_3|^{p_\emptyset} \right)^{\frac{1}{p_\emptyset}}.$$

Setting $\mathbf{u} = \frac{\mathbf{x}}{f(\mathbf{x})}$ and solving for u_3 yields

$$\begin{aligned} & f(\mathbf{u}) = 1 \\ \Leftrightarrow \quad & u_3 = \left(1 - (|u_1|^{p_1} + |u_2|^{p_1})^{\frac{p_\emptyset}{p_1}} \right)^{\frac{1}{p_\emptyset}} \end{aligned}$$

Now, let G_2 and F_1 denote

$$\begin{aligned} \mathsf{G}_2 &= \left(1 - (|u_1|^{p_1} + |u_2|^{p_1})^{\frac{p_\emptyset}{p_1}} \right)^{\frac{1-p_\emptyset}{p_\emptyset}} \\ \mathsf{F}_1 &= (|u_1|^{p_1} + |u_2|^{p_1})^{\frac{p_\emptyset - p_1}{p_1}}. \end{aligned}$$

Essentially, G_2 and F_1 are terms that evolve from the application from the chain rule when computing the partial derivative. G_2 originates from using the chain rule upwards in the tree and F_1 from using it downwards. The indices correspond the multi-indices of the respective nodes. Computing the derivative yields

$$\begin{aligned}
\frac{\partial u_3}{\partial u_k} &= \frac{\partial}{\partial u_k}\left(1-\left(|u_1|^{p_1}+|u_2|^{p_1}\right)^{\frac{p_\emptyset}{p_1}}\right)^{\frac{1}{p_\emptyset}} \\
&= \frac{1}{p_\emptyset}\mathsf{G}_2\cdot -\frac{\partial}{\partial u_k}\left(|u_1|^{p_1}+|u_2|^{p_1}\right)^{\frac{p_\emptyset}{p_1}} \\
&= \frac{1}{p_\emptyset}\frac{p_\emptyset}{p_1}\mathsf{G}_2\cdot -\mathsf{F}_1\frac{\partial}{\partial u_k}|u_k|^{p_1} \\
&= -\mathsf{G}_2\mathsf{F}_1\Delta_k {u_k}^{p_1-1}.
\end{aligned}$$

By inserting the results in equation (1) we obtain

$$\begin{aligned}
\frac{1}{r^2}|\mathcal{J}| &= -\sum_{k=1}^{2}\frac{\partial u_n}{\partial u_k}\cdot u_k + u_3 \\
&= \sum_{k=1}^{2}\mathsf{G}_2\mathsf{F}_1|u_k|^{p_1} + u_3 \\
&= \mathsf{G}_2\left(\sum_{k=1}^{2}\mathsf{F}_1|u_k|^{p_1} + \mathsf{G}_2^{-1}\left(1-\left(|u_1|^{p_1}+|u_2|^{p_1}\right)^{\frac{p_\emptyset}{p_1}}\right)^{\frac{1}{p_\emptyset}}\right)
\end{aligned}$$

$$\begin{aligned}
&= \mathsf{G}_2\left(\sum_{k=1}^{2}\mathsf{F}_1|u_k|^{p_1} + \left(1-\left(|u_1|^{p_1}+|u_2|^{p_1}\right)^{\frac{p_\emptyset}{p_1}}\right)^{-\frac{1-p_\emptyset}{p_\emptyset}}\left(1-\left(|u_1|^{p_1}+|u_2|^{p_1}\right)^{\frac{p_\emptyset}{p_1}}\right)^{\frac{1}{p_\emptyset}}\right) \\
&= \mathsf{G}_2\left(\sum_{k=1}^{2}\mathsf{F}_1|u_k|^{p_1} + 1-\left(|u_1|^{p_1}+|u_2|^{p_1}\right)^{\frac{p_\emptyset}{p_1}}\right) \\
&= \mathsf{G}_2\left(\mathsf{F}_1\sum_{k=1}^{2}|u_k|^{p_1} + 1-\mathsf{F}_1\mathsf{F}_1^{-1}\left(|u_1|^{p_1}+|u_2|^{p_1}\right)^{\frac{p_\emptyset}{p_1}}\right) \\
&= \mathsf{G}_2\left(\mathsf{F}_1\sum_{k=1}^{2}|u_k|^{p_1} + 1-\mathsf{F}_1\left(|u_1|^{p_1}+|u_2|^{p_1}\right)^{-\frac{p_\emptyset-p_1}{p_1}}\left(|u_1|^{p_1}+|u_2|^{p_1}\right)^{\frac{p_\emptyset}{p_1}}\right) \\
&= \mathsf{G}_2\left(\mathsf{F}_1\sum_{k=1}^{2}|u_k|^{p_1} + 1-\mathsf{F}_1\sum_{k=1}^{2}|u_k|^{p_1}\right) \\
&= \mathsf{G}_2.
\end{aligned}$$

In the example above, the terms from using the chain rule downwards in the tree canceled while the terms from using the chain rule upwards remained. It will turn out that this is true in general. Before we state the general equation we introduce a short notation for the terms that cancel and for those that remain.

Definition 3.2. Let $I = i_1, ..., i_{r-1}$. In the following, we denote

$$\begin{aligned} \mathsf{G}_{I,\ell_I} &= \mathsf{g}_{I,\ell_I}^{p_{I,\ell_I}-p_I} \\ &= \left(\mathsf{g}_I^{p_I} - \sum_{j=1}^{\ell-1} \mathsf{f}_{I,j}^{p_I}\right)^{\frac{p_{I,\ell_I}-p_I}{p_I}} \end{aligned} \tag{2}$$

and

$$\begin{aligned} \mathsf{F}_{I,i_r} &= \mathsf{f}_{I,i_r}^{p_I-p_{I,i_r}} \\ &= \left(\sum_{k=1}^{\ell} \mathsf{f}_{I,i_r,k}^{p_{I,i_r}}\right)^{\frac{p_I-p_{I,i_r}}{p_{I,i_r}}} . \end{aligned}$$

Note that the term F_{I,i_r} is a function of its children while G_{I,i_r} is a function of the parent node and all but the last children.

Before going on, let us quickly outline the essential mechanism when taking the chain rule $\frac{u_n}{u_q}$. Imagine the tree corresponding to f. By definition u_n is the rightmost leaf of the tree. Let L, ℓ_L be the multi-index of u_n. The calculation of $\frac{\partial u_n}{\partial u_q}$ will obviously involve heavy usage of the chain rule. As in the example, the chain rule starts at the leaf u_n ascends in the the tree until it reaches the lowest node whose subtree contains both, u_n and u_q. At this point, it starts descending the tree until it reaches the leaf u_q. Depending on whether the chain rule ascends or descends, two different forms of derivatives occur: At $u_n = \mathsf{g}_{L,\ell_L}$ the chain rule will start ascending by taking the derivative of the term

$$\mathsf{g}_{L,\ell_L} \quad = \quad \left(\mathsf{g}_L^{p_L} - \sum_{k=1}^{\ell_L-1} \mathsf{f}_{L,k}^{p_L}\right)^{\frac{1}{p_L}}$$

which will produce a G-term and move the chain rule one step up in the tree.

If the parent of u_n is already the lowest node whose subtree contains u_q a u_n, then u_q is hidden somewhere in the f-terms and the g-term is independent of u_q. However, if this node is still higher in the tree, then the situation is reversed, i.e. the f-terms are independent of u_q which is hidden in the g-term. When going on, the chain rule will produce a G-term when ascending the tree and an F-term when descending. The situation is depicted in Figure 2. The next lemma states a few helpful properties of the F- and G-terms.

Lemma 3.2. *Let $I = i_1, ..., i_{r-1}$ and f_{I,i_r} be any node of the tree associated with an L_p-nested function f. Then the following recursions hold for the derivatives of $\mathsf{g}_{I,i_r}^{p_{I,i_r}}$ and $\mathsf{f}_{I,i_r}^{p_I}$ w.r.t u_q: If u_q is not in the subtree under the node I, i_r, i.e. $u_k \notin \mathsf{f}_{I,i_r}$, then (remember that $p_{I,i_r} = 1$ for leaf nodes by notational convention):*

$$\frac{\partial}{\partial u_q} \mathsf{f}_{I,i_r}^{p_I} = 0$$

and

$$\frac{\partial}{\partial u_q} \mathsf{g}_{I,i_r}^{p_{I,i_r}} = \frac{p_{I,i_r}}{p_I} \mathsf{G}_{I,i_r} \cdot \begin{cases} \frac{\partial}{\partial u_q} \mathsf{g}_I^{p_I} & \text{if } u_q \in \mathsf{g}_I \\ -\frac{\partial}{\partial u_q} \mathsf{f}_{I,j}^{p_I} & \text{if } u_q \in \mathsf{f}_{I,j} \end{cases}$$

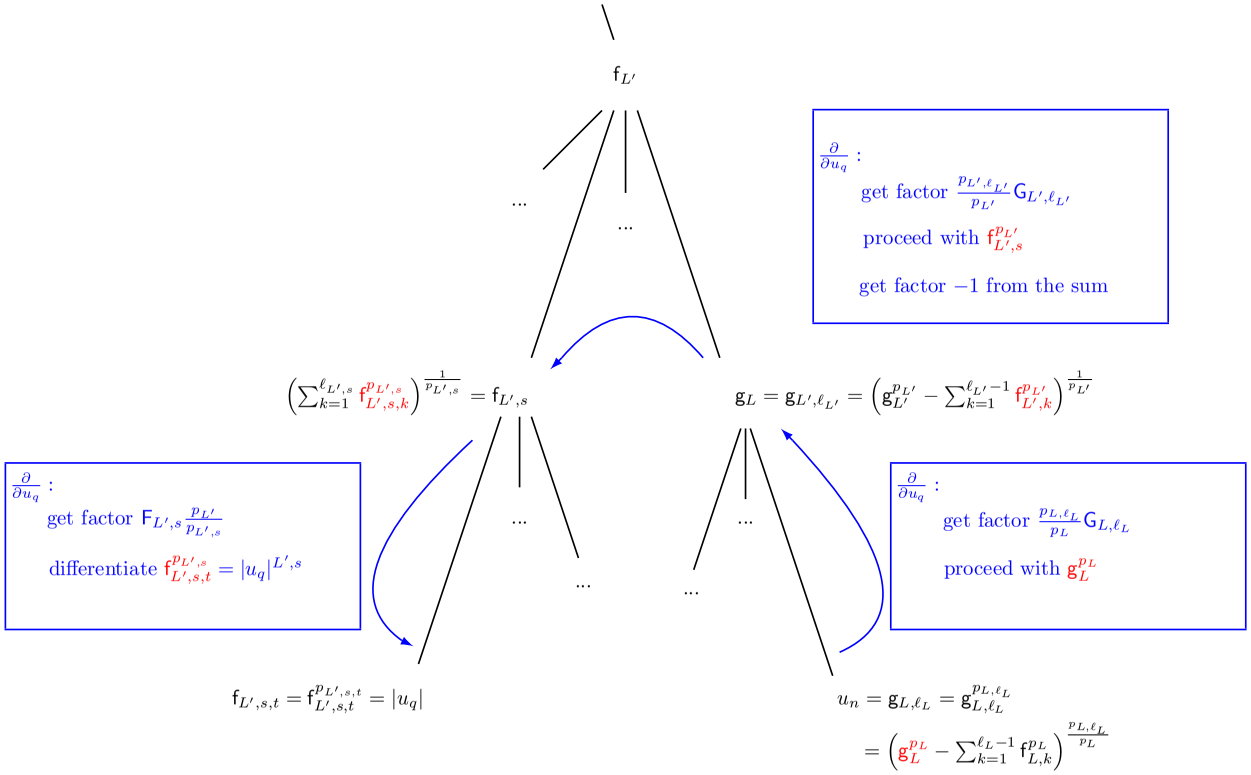


FIGURE 2. **Example scheme for the steps of the chain rule when calculating $\frac{\partial u_n}{\partial u_q}$:** Note that, $L = L', \ell_{L'}$, that the $p_{L,\ell_L}, p_{L,s,t}$ at the leafs are equal to one by definition and that succeeding ratios of the p cancel each other.

for $u_q \in \mathsf{f}_{I,j}$ *and* $u_q \notin \mathsf{f}_{I,k}$ *for* $k \neq j$. *Otherwise*

$$\frac{\partial}{\partial u_q}\mathsf{g}_{I,i_r}^{p_{I,i_r}} = 0$$

and

$$\frac{\partial}{\partial u_q}\mathsf{f}_{I,i_r}^{p_I} = \frac{p_I}{p_{I,i_r}}\mathsf{F}_{I,i_r}\frac{\partial}{\partial u_q}\mathsf{f}_{I,i_r,s}^{p_{I,i_r}}$$

for $u_q \in \mathsf{f}_{I,i_r,s}$ *and* $u_q \notin \mathsf{f}_{I,i_r,k}$ *for* $k \neq s$.

Proof. Both first equations are obvious, since only those nodes have a non-zero derivative for which the subtree actually depends on u_q. The second equations can be seen by computation

$$\begin{aligned}
\frac{\partial}{\partial u_q}\mathsf{g}_{I,i_r}^{p_{I,i_r}} &= p_{I,i_r}\mathsf{g}_{I,i_r}^{p_{I,i_r}-1}\frac{\partial}{\partial u_q}\mathsf{G}_{I,i_r} \\
&= p_{I,i_r}\mathsf{g}_{I,i_r}^{p_{I,i_r}-1}\frac{\partial}{\partial u_q}\left(\mathsf{g}_I^{p_I} - \sum_{j=1}^{\ell_I-1}\mathsf{f}_{I,j}^{p_I}\right)^{\frac{1}{p_I}} \\
&= \frac{p_{I,i_r}}{p_I}\mathsf{g}_{I,i_r}^{p_{I,i_r}-1}\mathsf{g}_{I,i_r}^{1-p_I}\frac{\partial}{\partial u_q}\left(\mathsf{g}_I^{p_I} - \sum_{j=1}^{\ell_I-1}\mathsf{f}_{I,j}^{p_I}\right) \\
&= \frac{p_{I,i_r}}{p_I}\mathsf{G}_{I,i_r}\cdot\begin{cases}\frac{\partial}{\partial u_q}\mathsf{g}_I^{p_I} & \text{if } u_q \in \mathsf{g}_I \\ -\frac{\partial}{\partial u_q}\mathsf{f}_{I,j}^{p_I} & \text{if } u_q \in \mathsf{f}_{I,j}\end{cases}
\end{aligned}$$

Similarly

$$\begin{aligned}
\frac{\partial}{\partial u_q}\mathsf{f}_{I,i_r}^{p_I} &= p_I\mathsf{f}_{I,i_r}^{p_I-1}\frac{\partial}{\partial u_q}\mathsf{f}_{I,i_r} \\
&= p_I\mathsf{f}_{I,i_r}^{p_I-1}\frac{\partial}{\partial u_q}\left(\sum_{k=1}^{\ell_{I,i_r}}\mathsf{f}_{I,i_r,k}^{p_{I,i_r}}\right)^{\frac{1}{p_{I,i_r}}} \\
&= \frac{p_I}{p_{I,i_r}}\mathsf{f}_{I,i_r}^{p_I-1}\mathsf{f}_{I,i_r}^{1-p_{I,i_r}}\frac{\partial}{\partial u_q}\mathsf{f}_{I,i_r,s}^{p_{I,i_r}} \\
&= \frac{p_I}{p_{I,i_r}}\mathsf{F}_{I,i_r}\frac{\partial}{\partial u_q}\mathsf{f}_{I,i_r,s}^{p_{I,i_r}}
\end{aligned}$$

for $u_k \in \mathsf{f}_{I,i_r,s}$. □

The next lemma states the form of the derivative $\frac{\partial u_n}{\partial u_q}$ in terms of the G- and F-terms.

Lemma 3.3. *Let* $|u_q| = \mathsf{f}_{\ell_1,\ldots,\ell_r,i_1,\ldots,i_t}$, $|u_n| = \mathsf{f}_{\ell_1,\ldots,\ell_d}$ *with* $r < d$ *and, therefore, the shortest path from* u_n *to* u_q *be* $(\ell_1,\ldots,\ell_d),(\ell_1,\ldots,\ell_{d-1}),\ldots,(\ell_1,\ldots,\ell_r),(\ell_1,\ldots,\ell_r,i_1),\ldots,(\ell_1,\ldots,\ell_r,i_1,\ldots,i_t)$. *The derivative of* u_n *w.r.t.* u_q *is given by*

$$\frac{\partial}{\partial u_q}u_n = -\mathsf{G}_{\ell_1,\ldots,\ell_d}\cdot\ldots\cdot\mathsf{G}_{\ell_1,\ldots,\ell_{r+1}}\cdot\mathsf{F}_{\ell_1,\ldots,\ell_r,i_1}\cdot\mathsf{F}_{\ell_1,\ldots,\ell_r,i_1,\ldots,i_{t-1}}\cdot\Delta_q u_q^{p_{\ell_1,\ldots,\ell_r,i_1,\ldots,i_{t-1}}-1}$$

with $\Delta_q = sgn\, u_q$ *and* $|u_q|^p = (\Delta_q u_q)^p$. *In particular*

$$u_q \frac{\partial}{\partial u_q} u_n = -\mathsf{G}_{\ell_1,\ldots,\ell_d} \cdot \ldots \cdot \mathsf{G}_{\ell_1,\ldots,\ell_{r+1}} \cdot \mathsf{F}_{\ell_1,\ldots,\ell_r,i_1} \cdot \mathsf{F}_{\ell_1,\ldots,\ell_r,i_1,\ldots,i_{t-1}} \cdot |u_q|^{p_{\ell_1,\ldots,\ell_r,i_1,\ldots,i_{t-1}}}.$$

Proof. Successive application of Lemma (3.2). □

Before finally deriving the expression for the determinant, we state two more helpful equations.

Lemma 3.4. *Let* $I = i_1, \ldots, i_{r-1}$, *then*

$$\mathsf{G}_{I,i_r}^{-1} \mathsf{g}_{I,i_r}^{p_{I,i_r}} = \mathsf{g}_{I,i_r}^{p_I} \tag{3}$$

$$= \mathsf{g}_I^{p_I} - \sum_{k=1}^{\ell_I - 1} \mathsf{F}_{I,k} \mathsf{f}_{I,k}^{p_{I,k}} \tag{4}$$

and

$$\mathsf{f}_{I,i_r}^{p_{I,i_r}} = \sum_{k=1}^{\ell_{I,i_r}} \mathsf{F}_{I,i_r,k} \mathsf{f}_{I,i_r,k}^{p_{I,i_r,k}} \tag{5}$$

Proof. First, we prove the equalities (3) and (4):

$$\begin{aligned}
\mathsf{G}_{I,i_r}^{-1} \mathsf{g}_{I,i_r}^{p_{I,i_r}} &= \mathsf{g}_{I,i_r}^{-(p_{I,i_r} - p_I)} \mathsf{g}_{I,i_r}^{p_{I,i_r}} \\
&= \mathsf{g}_{I,i_r}^{p_I} \text{ q.e.d. (3)} \\
&= \left(\mathsf{g}_I^{p_I} - \sum_{k=1}^{\ell_I - 1} \mathsf{f}_{I,k}^{p_I} \right)^{\frac{p_I}{p_I}} \\
&= \mathsf{g}_I^{p_I} - \sum_{k=1}^{\ell_I - 1} \mathsf{f}_{I,k}^{p_I - p_{I,k}} \mathsf{f}_{I,k}^{p_{I,k}} \\
&= \mathsf{g}_I^{p_I} - \sum_{k=1}^{\ell_I - 1} \mathsf{F}_{I,k} \mathsf{f}_{I,k}^{p_{I,k}} \text{ q.e.d. (4).}
\end{aligned}$$

In a similar fashion, equality (5) can be proven by substituting definitions and introducing one in the exponent.

□

Proposition 3.1 (Determinant of the Jacobian)**.** *Let* $\mathcal{L}$ *be the set of multi-indices of the path from the leaf* u_n *to the root node (excluding the root node). The determinant of the Jacobian for an* L_p*-nested function is given by*

$$\det |\mathcal{J}| = r^{n-1} \prod_{L \in \mathcal{L}} \mathsf{G}_L.$$

Proof. Let $L = \ell_1, \ldots \ell_{d-1}$ be the multi-index of the parent of u_n. We compute $\frac{1}{r^{n-1}} |\det \mathcal{J}|$ and obtain the result by solving for $|\det \mathcal{J}|$. As shown in Lemma (3.1) $\frac{1}{r^{n-1}} |\det \mathcal{J}|$ has the form

$$\frac{1}{r^{n-1}} |\det \mathcal{J}| = -\sum_{k=1}^{n-1} \frac{\partial u_n}{\partial u_k} \cdot u_k + u_n.$$

By definition $u_n = \mathsf{g}_{L,\ell_d} = \mathsf{g}_{L,\ell_d}^{p_{L,\ell_d}}$. Now, assume that $u_r, ..., u_{n-1}$ are children of f_L, i.e. $u_k = \mathsf{f}_{L,I,i_t}$ for some $I, i_t = i_1, ..., i_t$ and $r \leq k < n$. Remember, that by Lemma (3.3) the terms $u_q \frac{\partial}{\partial u_q} u_n$ for $r \leq q < n$ have the form

$$u_q \frac{\partial}{\partial u_q} u_n = -\mathsf{G}_{L,\ell_d} \cdot \mathsf{F}_{L,i_1} \cdot ... \cdot \mathsf{F}_{L,I} \cdot |u_q|^{p_{\ell_1,...,\ell_{d-1},i_1,...,i_{t-1}}}.$$

Now, we can expand the determinant as follows

$$\begin{aligned}
&-\sum_{k=1}^{n-1} \frac{\partial u_n}{\partial u_k} \cdot u_k + \mathsf{g}_{L,\ell_d}^{p_{L,\ell_d}} \\
&= -\sum_{k=1}^{r-1} \frac{\partial u_n}{\partial u_k} \cdot u_k - \sum_{k=r}^{n-1} \frac{\partial u_n}{\partial u_k} \cdot u_k + \mathsf{g}_{L,\ell_d}^{p_{L,\ell_d}} \\
&= -\sum_{k=1}^{r-1} \frac{\partial u_n}{\partial u_k} \cdot u_k + \mathsf{G}_{L,\ell_d} \left(-\sum_{k=r}^{n-1} \mathsf{G}_{L,\ell_d}^{-1} \frac{\partial u_n}{\partial u_k} \cdot u_k + \mathsf{G}_{L,\ell_d}^{-1} \mathsf{g}_{L,\ell_d}^{p_{L,\ell_d}} \right) \\
&= -\sum_{k=1}^{r-1} \frac{\partial u_n}{\partial u_k} \cdot u_k + \mathsf{G}_{L,\ell_d} \left(-\sum_{k=r}^{n-1} \mathsf{G}_{L,\ell_d}^{-1} \frac{\partial u_n}{\partial u_k} \cdot u_k + \mathsf{g}_L^{p_L} - \sum_{k=1}^{\ell_d - 1} \mathsf{F}_{L,k} \mathsf{f}_{L,k}^{p_{L,k}} \right)
\end{aligned}$$

by equality (4) of Lemma (3.4). Note that all terms $\mathsf{G}_{L,\ell_d}^{-1} \frac{\partial u_n}{\partial u_k} \cdot u_k$ for $r \leq k < n$ now have the form

$$\mathsf{G}_{L,\ell_d}^{-1} u_k \frac{\partial}{\partial u_k} u_n = -\mathsf{F}_{L,i_1} \cdot ... \cdot \mathsf{F}_{L,I} \cdot |u_q|^{p_{\ell_1,...,\ell_{d-1},i_1,...,i_{t-1}}}$$

since we constructed them to be neighbors of u_n. However, with equation (5) of Lemma (3.4), we can further expand the sum $\sum_{k=1}^{\ell_d - 1} \mathsf{F}_{L,k} \mathsf{f}_{L,k}^{p_{L,k}}$ down to the leafs $u_r, ..., u_{n-1}$. When doing so we end up with the same factors $\mathsf{F}_{L,i_1} \cdot ... \cdot \mathsf{F}_{L,I} \cdot |u_q|^{p_{\ell_1,...,\ell_{d-1},i_1,...,i_{t-1}}}$ as in the derivatives $\mathsf{G}_{L,\ell_d}^{-1} u_q \frac{\partial}{\partial u_q} u_n$. This means exactly that

$$-\sum_{k=r}^{n-1} \mathsf{G}_{L,\ell_d}^{-1} \frac{\partial u_n}{\partial u_k} \cdot u_k = \sum_{k=1}^{\ell_d - 1} \mathsf{F}_{L,k} \mathsf{f}_{L,k}^{p_{L,k}}$$

and, therefore,

$$\begin{aligned}
&= -\sum_{k=1}^{r-1} \frac{\partial u_n}{\partial u_k} \cdot u_k + \mathsf{G}_{L,\ell_d} \left(-\sum_{k=r}^{n-1} \mathsf{G}_{L,\ell_d}^{-1} \frac{\partial u_n}{\partial u_k} \cdot u_k + \mathsf{g}_L^{p_L} - \sum_{k=1}^{\ell_d - 1} \mathsf{F}_{L,k} \mathsf{f}_{L,k}^{p_{L,k}} \right) \\
&= -\sum_{k=1}^{r-1} \frac{\partial u_n}{\partial u_k} \cdot u_k + \mathsf{G}_{L,\ell_d} \left(\sum_{k=1}^{\ell_d - 1} \mathsf{F}_{L,k} \mathsf{f}_{L,k}^{p_{L,k}} + \mathsf{g}_L^{p_L} - \sum_{k=1}^{\ell_d - 1} \mathsf{F}_{L,k} \mathsf{f}_{L,k}^{p_{L,k}} \right) \\
&= -\sum_{k=1}^{r-1} \frac{\partial u_n}{\partial u_k} \cdot u_k + \mathsf{G}_{L,\ell_d} \mathsf{g}_L^{p_L}.
\end{aligned}$$

By factoring out G_{L,ℓ_d} from the equation, the terms $\frac{\partial u_n}{\partial u_k} \cdot u_k$ loose the G_{L,ℓ_d} in front and we get basically the same equation as before, only that the new leaf (the new "u_n") is $\mathsf{g}_L^{p_L}$ and we got rid of all the children of f_L. By repeating that procedure up to the root node, we successively factor out all $\mathsf{G}_{L'}$ for $L' \in \mathcal{L}$ until

all terms of the sum vanish and we are only left with $\mathsf{f}_\emptyset = 1$. Therefore, the determinant is

$$\frac{1}{r^{n-1}} |\det \mathcal{J}| = \prod_{L \in \mathcal{L}} \mathsf{G}_L$$

which completes the proof. □

4. L_p-nested Uniform Distribution

In analogy to [6] we define a uniform distribution on the L_P-nested sphere. Naturally, the density of this distribution is the inverse of the surface area of the L_p-nested unit sphere. In this section we first compute the surface of the L_p-nested sphere and then define the L_p-nested uniform distribution in terms of the polar-like coordinates from the section before. Before we start, we start by computing the surface and the volume of an arbitrary L_p-nested sphere.

Proposition 4.1 (Volumen and Surface of the L_p-nested Sphere)**.** *Let f be an L_p-nested function and let $\mathcal{I}$ be the set of all multi-indices denoting the inner nodes of the tree structure associated with f. Let n_I denote the number of leafs contained in the subtree under the node I (if I is a leaf already, $n_I = 1$). The volumen $\mathcal{V}_f(R)$ and the surface $\mathcal{S}_f(R)$ of the L_p-nested sphere with radius R is given by*

$$\mathcal{V}_f(R) = \frac{R^n 2^n}{n} \prod_{I \in \mathcal{I}} \frac{1}{p_I^{\ell_I - 1}} \prod_{k=1}^{\ell_I - 1} B\left[\frac{\sum_{i=1}^{k} n_{I,k}}{p_I}, \frac{n_{I,k+1}}{p_I} \right] \tag{6}$$

$$= \frac{R^n 2^n}{n} \prod_{I \in \mathcal{I}} \frac{\prod_{k=1}^{\ell_I} \Gamma\left[\frac{n_{I,k}}{p_I}\right]}{p_I^{\ell_I - 1} \Gamma\left[\frac{n_I}{p_I}\right]} \tag{7}$$

$$\mathcal{S}_f(R) = R^{n-1} 2^n \prod_{I \in \mathcal{I}} \frac{1}{p_I^{\ell_I - 1}} \prod_{k=1}^{\ell_I - 1} B\left[\frac{\sum_{i=1}^{k} n_{I,k}}{p_I}, \frac{n_{I,k+1}}{p_I} \right] \tag{8}$$

$$= R^{n-1} 2^n \prod_{I \in \mathcal{I}} \frac{\prod_{k=1}^{\ell_I} \Gamma\left[\frac{n_{I,k}}{p_I}\right]}{p_I^{\ell_I - 1} \Gamma\left[\frac{n_I}{p_I}\right]} \tag{9}$$

Proof. We obtain the volumen by computing the integral $\int_{f(\mathbf{x}) \leq R} d\mathbf{x}$. Differentiation with respect to R yields the surface area. For symmetry reasons we can compute the volume only on the positive quadrant $\mathbb{R}^n_+$ and multiply the result with 2^n later to obtain the full volumen and surface area. The strategy for computing the volumen is as follows. We start off with inner nodes I that are parents of leafs only. The value f_I of such a node is simply the L_{p_I} norm of its children. Therefore, we can convert the integral over the children of I with the transformation of [3]. This maps the leafs $\mathbf{f}_{I,1:\ell_I}$ into f_I and "angular" variables $\tilde{\boldsymbol{u}}_{\ell_I - 1}$. Since integral borders of the original integral depend only on the value of f_I and not on $\tilde{\boldsymbol{u}}$, we can separate the variables $\tilde{\boldsymbol{u}}$ from the radial variables f_I and integrate the variables $\tilde{\boldsymbol{u}}_{\ell_I - 1}$ separately. The integration over $\tilde{\boldsymbol{u}}_{\ell_I - 1}$ yields a certain factor, while the variable f_I effectively becomes a new leaf.

Now suppose I is the parent of leafs only. W.l.o.g. let the ℓ_I leafs correspond to the last ℓ_I coefficients of $\mathbf{x}$. Let $\mathbf{x} \in \mathbb{R}^n_+$. Carrying out the first transformation and

integration yields

$$\int_{f(\mathbf{x})\leq R} d\mathbf{x} = \int_{f(\mathbf{x}_{1:n-\ell_I},\mathsf{f}_I)\leq R} \int_{\tilde{\boldsymbol{u}}_{\ell_I-1}\in\mathcal{V}_+^{\ell_I-1}} \mathsf{f}_I^{\ell_I-1} \left(1-\sum_{i=1}^{\ell_I-1}\tilde{u}_i^{p_I}\right)^{\frac{1-p_I}{p_I}} d\mathsf{f}_I d\tilde{\boldsymbol{u}}_{\ell_I-1} d\mathbf{x}_{1:n-\ell_I}$$

$$= \int_{f(\mathbf{x}_{1:n-\ell_I},\mathsf{f}_I)\leq R} \mathsf{f}_I^{n_I-1} d\mathsf{f}_I d\mathbf{x}_{1:n-\ell_I} \times \int_{\tilde{\boldsymbol{u}}_{\ell_I-1}\in\mathcal{V}_+^{\ell_I-1}} \left(1-\sum_{i=1}^{\ell_I-1}\tilde{u}_i^{p_I}\right)^{\frac{n_{I,\ell_I}-p_I}{p_I}} d\tilde{\boldsymbol{u}}_{\ell_I-1}.$$

For solving the second integral we make the pointwise transformation $s_i = \tilde{u}_i^{p_I}$ and obtain

$$\begin{aligned}\int_{\tilde{\boldsymbol{u}}_{\ell_I-1}\in\mathcal{V}_+^{\ell_I-1}} \left(1-\sum_{i=1}^{\ell_I-1}\tilde{u}_i^{p_I}\right)^{\frac{n_{I,\ell_I}-p_I}{p_I}} d\tilde{\boldsymbol{u}}_{\ell_I-1} &= \frac{1}{p_I^{\ell_I-1}} \int_{\sum s_i\leq 1} \left(1-\sum_{i=1}^{\ell_I-1} s_i\right)^{\frac{n_{I,\ell_I}}{p_I}-1} \prod_{i=1}^{\ell_I-1} s_i^{\frac{1}{p_I}-1} d\boldsymbol{s}_{\ell_I-1} \\ &= \frac{1}{p_I^{\ell_I-1}} \prod_{k=1}^{\ell_I-1} B\left[\frac{\sum_{i=1}^{k} n_{I,k}}{p_I}, \frac{n_{I,k+1}}{p_I}\right] \\ &= \frac{1}{p_I^{\ell_I-1}} \prod_{k=1}^{\ell_I-1} B\left[\frac{k}{p_I}, \frac{1}{p_I}\right]\end{aligned}$$

by using the fact that the transformed integral has the form of an unnormalized Dirichlet distribution and, therefore, the value of the integral must equal its normalization constant.

Now, we go on with solving the integral

$$\int_{f(\mathbf{x}_{1:n-\ell_I},\mathsf{f}_I)\leq R} \mathsf{f}_I^{n_I-1} d\mathsf{f}_I d\mathbf{x}_{1:n-\ell_I}. \tag{10}$$

We carry this out in exactly the same manner as we solved the previous integral. We only need to make sure that we only contract nodes that have only leafs as children (remember that radii of contracted nodes become leafs) and we need to find a formula how the factors $\mathsf{f}_I^{n_I-1}$ propagate through the tree.

For the latter, we first state the formula and then prove it via induction. For notational convenience let $\hat{\mathbf{x}}$ denote the remaining coefficients of $\mathbf{x}$, $\hat{\mathbf{f}}$ the vector of leafs resulting from contraction and $\mathcal{J}$ the set of multi-indices corresponding to the contracted leafs. The integral which is left to solve after integrating over all $\tilde{\boldsymbol{u}}$ is given by (remember that n_J denotes real leafs, i.e. the ones corresponding to coefficients of $\mathbf{x}$):

$$\int_{f(\hat{\mathbf{x}},\hat{\mathbf{f}})\leq R} \prod_{J\in\mathcal{J}} \mathsf{f}_J^{n_J-1} d\hat{\mathbf{f}} d\hat{\mathbf{x}}.$$

We already proved the first induction step by computing equation (10). For computing the general induction step suppose I is an inner node whose children are leafs or contracted leafs. Let $\mathcal{J}'$ be the set of contracted leafs under I and $\hat{\mathcal{J}} = \mathcal{J}\backslash\mathcal{J}'$. Furthermore, let $\tilde{\mathbf{f}}$ and $\tilde{\mathbf{x}}$ be the leafs belonging to the set $\hat{\mathcal{J}}$. For notational convenience, we will denote all children of I with $\mathsf{f}_{I,k}$ no matter whether they are real leafs y_i or result from a previous contraction. Transforming the children of I into

radial coordinates by [3] yields

$$\begin{aligned}
\int_{f(\hat{\mathbf{x}},\hat{\mathbf{f}})\le R} \prod_{J\in\mathcal{J}} \mathsf{f}_J^{n_J-1} d\hat{\mathbf{f}} d\hat{\mathbf{x}} &= \int_{f(\hat{\mathbf{x}},\hat{\mathbf{f}})\le R} \left(\prod_{\hat{\jmath}\in\hat{\mathcal{J}}} \mathsf{f}_{\hat{\jmath}}^{n_{\hat{\jmath}}-1}\right) \cdot \left(\prod_{J'\in\mathcal{J}'} \mathsf{f}_{J'}^{n_{J'}-1}\right) d\hat{\mathbf{f}} d\hat{\mathbf{x}} \\
&= \int_{f(\tilde{\mathbf{x}},\tilde{\mathbf{f}},\mathsf{f}_I)\le R} \int_{\tilde{\boldsymbol{u}}_{\ell_I-1}\in\mathcal{V}_+^{\ell_I-1}} \left(\left(1-\sum_{i=1}^{\ell_I-1}\tilde{u}_i^{p_I}\right)^{\frac{1-p_I}{p_I}} \mathsf{f}_I^{\ell_I-1}\right) \cdot \left(\prod_{\hat{\jmath}\in\hat{\mathcal{J}}} \mathsf{f}_{\hat{\jmath}}^{n_{\hat{\jmath}}-1}\right) \\
&\times \left(\left(\mathsf{f}_I\left(1-\sum_{i=1}^{\ell_I-1}\tilde{u}_i^{p_I}\right)\right)^{\frac{n_{\ell_I}-1}{p_I}} \prod_{k=1}^{\ell_I-1} (\mathsf{f}_I\tilde{u}_k)^{n_k-1}\right) d\tilde{\mathbf{x}} d\tilde{\mathbf{f}} d\mathsf{f}_I d\tilde{\boldsymbol{u}}_{\ell_I-1} \\
&= \int_{f(\tilde{\mathbf{x}},\tilde{\mathbf{f}},\mathsf{f}_I)\le R} \int_{\tilde{\boldsymbol{u}}_{\ell_I-1}\in\mathcal{V}_+^{\ell_I-1}} \left(\prod_{\hat{\jmath}\in\hat{\mathcal{J}}} \mathsf{f}_{\hat{\jmath}}^{n_{\hat{\jmath}}-1}\right) \\
&\times \left(\mathsf{f}_I^{\ell_I-1+\sum_{i=1}^{\ell_I}(n_i-1)} \left(1-\sum_{i=1}^{\ell_I-1}\tilde{u}_i^{p_I}\right)^{\frac{n_{\ell_I}-p_I}{p_I}} \prod_{k=1}^{\ell_I-1}\tilde{u}_k^{n_k-1}\right) d\tilde{\mathbf{x}} d\tilde{\mathbf{f}} d\mathsf{f}_I d\tilde{\boldsymbol{u}}_{\ell_I-1} \\
&= \int_{f(\tilde{\mathbf{x}},\tilde{\mathbf{f}},\mathsf{f}_I)\le R} \left(\prod_{\hat{\jmath}\in\hat{\mathcal{J}}} \mathsf{f}_{\hat{\jmath}}^{n_{\hat{\jmath}}-1}\right) \mathsf{f}_I^{n_I-1} d\tilde{\mathbf{x}} d\tilde{\mathbf{f}} d\mathsf{f}_I \\
&\times \int_{\tilde{\boldsymbol{u}}_{\ell_I-1}\in\mathcal{V}_+^{\ell_I-1}} \left(1-\sum_{i=1}^{\ell_I-1}\tilde{u}_i^{p_I}\right)^{\frac{n_{\ell_I}-p_I}{p_I}} \prod_{k=1}^{\ell_I-1}\tilde{u}_k^{n_k-1} d\tilde{\boldsymbol{u}}_{\ell_I-1}.
\end{aligned}$$

Again, by transforming it into a Dirichlet distribution, the latter integral has the solution

$$\int_{\tilde{\boldsymbol{u}}_{\ell_I-1}\in\mathcal{V}_+^{\ell_I-1}} \left(1-\sum_{i=1}^{\ell_I-1}\tilde{u}_i^{p_I}\right)^{\frac{n_{\ell_I}-p_I}{p_I}} \prod_{k=1}^{\ell_I-1}\tilde{u}_k^{n_k-1} d\tilde{\boldsymbol{u}}_{\ell_I-1} = \prod_{k=1}^{\ell_I-1} B\left[\frac{\sum_{i=1}^{k} n_{I,k}}{p_I}, \frac{n_{I,k+1}}{p_I}\right]$$

while the remaining former integral has the form

$$\int_{f(\tilde{\mathbf{x}},\tilde{\mathbf{f}},\mathsf{f}_I)\le R} \left(\prod_{\hat{\jmath}\in\hat{\mathcal{J}}} \mathsf{f}_{\hat{\jmath}}^{n_{\hat{\jmath}}-1}\right) \mathsf{f}_I^{n_I-1} d\tilde{\mathbf{x}} d\tilde{\mathbf{f}} d\mathsf{f}_I = \int_{f(\hat{\mathbf{x}},\hat{\mathbf{f}})\le R} \prod_{J\in\mathcal{J}} \mathsf{f}_J^{n_J-1} d\hat{\mathbf{f}} d\hat{\mathbf{x}}$$

as claimed.

By carrying out the integration up to the root node the remaining integral becomes

$$\int_{\mathsf{f}_\emptyset\le R} \mathsf{f}_\emptyset^{n-1} d\mathsf{f}_\emptyset = \int_0^R \mathsf{f}_\emptyset^{n-1} d\mathsf{f}_\emptyset = \frac{R^n}{n}.$$

Collecting the factors from integration over the $\tilde{\boldsymbol{u}}$ proves the equations (6) and (8). Using $B[a,b] = \frac{\Gamma[a]\Gamma[b]}{\Gamma[a+b]}$ yields equations (7) and (9). □

In order to clarify the proof we explicitly carry out the integration for our first example.

Example 4.1. Again, let the L_p-nested function be given by

$$f(\mathbf{x}) = \left((|x_1|^{p_1} + |x_2|^{p_1})^{\frac{p_\emptyset}{p_1}} + |x_3|^{p_\emptyset} \right)^{\frac{1}{p_\emptyset}}.$$

Let $\mathbf{x} \in \mathbb{R}^3_+$. Carrying out the steps from the proof above yields

$$\begin{aligned}
\int_{f(\mathbf{x})\leq R} d\mathbf{x} &= \int_{f(\mathsf{f}_1,x_3)\leq R} \int_0^1 (1-\tilde{u}^{p_1})^{\frac{1-p_1}{p_1}} \mathsf{f}_1^{\ell_1-1} d\tilde{u} d\mathsf{f}_1 dx_3 \\
&= \int_{f(\mathsf{f}_1,x_3)\leq R} \mathsf{f}_1^{\ell_1-1} d\mathsf{f}_1 dx_3 \times \int_0^1 (1-\tilde{u}^{p_1})^{\frac{1-p_1}{p_1}} d\tilde{u} \\
&= \int_{f(\mathsf{f}_1,x_3)\leq R} \mathsf{f}_1^{\ell_1-1} d\mathsf{f}_1 dx_3 \times \frac{1}{p_1} B\left[\frac{1}{p_1}, \frac{1}{p_1}\right].
\end{aligned}$$

Solving the first integral yields

$$\begin{aligned}
\int_{f(\mathsf{f}_1,x_3)\leq R} \mathsf{f}_1^{\ell_1-1} d\mathsf{f}_1 &= \int_{\mathsf{f}_\emptyset \leq R} \int_0^1 \mathsf{f}_\emptyset^{\ell_\emptyset - 1} (\mathsf{f}_\emptyset \tilde{u}^{p_0})^{\ell_1-1} (1-\tilde{u}^{p_0})^{\frac{1-p_\emptyset}{p_\emptyset}} d\tilde{u} d\mathsf{f}_\emptyset \\
&= \int_{\mathsf{f}_\emptyset \leq R} \int_0^1 \mathsf{f}_\emptyset^{\ell_\emptyset + \ell_1 - 2} \tilde{u}^{\ell_1-1} (1-\tilde{u}^{p_0})^{\frac{1-p_\emptyset}{p_\emptyset}} d\tilde{u} d\mathsf{f}_\emptyset \\
&= \int_{\mathsf{f}_\emptyset \leq R} \mathsf{f}_\emptyset^2 d\mathsf{f}_\emptyset \times \int_0^1 \tilde{u} (1-\tilde{u}^{p_0})^{\frac{1-p_\emptyset}{p_\emptyset}} d\tilde{u} \\
&= \frac{R^3}{3} \cdot \frac{1}{p_\emptyset} B\left[\frac{2}{p_\emptyset}, \frac{1}{p_\emptyset}\right].
\end{aligned}$$

Collecting all factors yields

$$\int_{f(\mathbf{x})\leq R} d\mathbf{x} = \frac{R^3}{3} \cdot \frac{1}{p_\emptyset} \frac{1}{p_1} B\left[\frac{2}{p_\emptyset}, \frac{1}{p_\emptyset}\right] B\left[\frac{1}{p_1}, \frac{1}{p_1}\right].$$

Extending the domain such that $\mathbf{x} \in \mathbb{R}^3$, simply introduces a factor 2^3. The surface is obtained by differentiating with respect to R. This yields the final equations

$$\begin{aligned}
\mathcal{V}_f(R) &= \frac{R^3 2^3}{3} \cdot \frac{1}{p_\emptyset} \frac{1}{p_1} B\left[\frac{2}{p_\emptyset}, \frac{1}{p_\emptyset}\right] B\left[\frac{1}{p_1}, \frac{1}{p_1}\right] \\
\mathcal{S}_f(R) &= R^2 2^3 \cdot \frac{1}{p_\emptyset} \frac{1}{p_1} B\left[\frac{2}{p_\emptyset}, \frac{1}{p_\emptyset}\right] B\left[\frac{1}{p_1}, \frac{1}{p_1}\right]
\end{aligned}$$

Proposition 4.2 (L_p-nested Uniform Distribution)**.** *Let f be an L_p-nested function. Let $\mathcal{L}$ be set set of multi-indices on the path from the root node to the leaf corresponding to y_n and let $\tilde{L}$ be the multi-index of x_n. The uniform distribution on the L_p-nested unit sphere, i.e. the set $\{\mathbf{x} \in \mathbb{R}^n | f(\mathbf{x}) = 1\}$ is given by*

$$\rho(\mathbf{u}) = \left(\frac{1}{2^{n-1}} \prod_{I \in \mathcal{I}} p_I^{\ell_I - 1} \prod_{k=1}^{\ell_I - 1} B\left[\frac{\sum_{i=1}^k n_{I,k}}{p_I}, \frac{n_{I,k+1}}{p_I} \right]^{-1} \right) \cdot \prod_{L \in \mathcal{L}} \mathsf{G}_L$$

where the support of $p(\mathbf{u})$ is given by

$$\text{supp } \rho = \left\{ \mathbf{u} \in \mathbb{R}^{n-1} | f(\mathbf{u}, g_{\tilde{L}}(\mathbf{u})) = 1 \right\}$$

Proof. Since the L_p-nested sphere is a compact set, the density of the uniform distribution is simply one over the surface area of the unit L_p-nested sphere. The surface $\mathcal{S}_f(1)$ is given by Proposition 4.1. Transforming $\frac{1}{\mathcal{S}_f(1)}$ into the coordinates

of Definition 3.1 introduces the determinant of the Jacobian from Proposition 3.1 and an additional factor of 2 since the $\mathbf{u} \in \mathbb{R}^{n-1}$ have to account for both half-shells of the L_p-nested unit sphere. This yields the expression above. □

Example 4.2 (L_p-spherically symmetric uniform distribution). We consider L_p-norm as a special case of an L_p-nested function

$$f(\mathbf{x}) = ||\mathbf{x}||_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}.$$

The corresponding tree has only one single inner node, which is the root node. Using Proposition 4.1, the surface area is given by

$$\begin{aligned}
\mathcal{S}_{||\cdot||_p} &= 2^n \frac{1}{p_\emptyset^{\ell_\emptyset - 1}} \prod_{k=1}^{\ell_\emptyset - 1} B\left[\frac{\sum_{i=1}^{k} n_k}{p_\emptyset}, \frac{n_{k+1}}{p_\emptyset} \right] \\
&= 2^n \frac{1}{p^{n-1}} \prod_{k=1}^{n-1} B\left[\frac{k}{p}, \frac{1}{p} \right] \\
&= 2^n \frac{1}{p^{n-1}} \prod_{k=1}^{n-1} \frac{\Gamma\left[\frac{k}{p}\right] \Gamma\left[\frac{1}{p}\right]}{\Gamma\left[\frac{k+1}{p}\right]} \\
&= \frac{2^n \Gamma^n\left[\frac{1}{p}\right]}{p^{n-1} \Gamma\left[\frac{n}{p}\right]}.
\end{aligned}$$

The factor G_n is given by $\left(1 - \sum_{i=1}^{n-1} |u_i|^p\right)^{\frac{1-p}{p}}$, which together with the factor 2 yields the uniform distribution on the L_p-sphere as defined in [6]

$$p(\mathbf{u}) = \frac{p^{n-1} \Gamma\left[\frac{n}{p}\right]}{2^{n-1} \Gamma^n\left[\frac{1}{p}\right]} \left(1 - \sum_{i=1}^{n-1} |u_i|^p\right)^{\frac{1-p}{p}}.$$

5. L_p-Nested Symmetric Distributions

Definition 5.1 (L_p-Nested Symmetric Distribution)**.** A n-dimensional random vector $\boldsymbol{X}$ is called *L_p-nested symmetrically distributed with respect to f* if f is an L_p-nested function, $\boldsymbol{X} = R\boldsymbol{U}$ for two independent random variables R and $\boldsymbol{U}$, where R is a non-negative univariate random variable and $\boldsymbol{U}$ is a n-dimensional random variable uniformly distributed on the L_p-nested unit sphere corresponding to f, i.e. $f(\boldsymbol{U}) = 1$ and $U_1, .., U_{n-1}$ follow the distribution of Proposition 4.2.

This definition of L_p-nested symmetric distribution is a straightforward generalization of Gupta and Song's definition of L_p-spherically symmetric distributions. By exactly the same reasoning as their's [3] the definition implies that $f(X) \doteq R$ and $\frac{\boldsymbol{X}}{f(\boldsymbol{X})} \doteq \boldsymbol{U}$ and, therefore, that $f(\boldsymbol{X})$ and $\frac{\boldsymbol{X}}{f(\boldsymbol{X})}$ are independent. This also means that being able to sample from any L_p-nested symmetric distribution makes it possible to sample from any other L_p-nested symmetric distribution as long as the radial distribution of it is known. One simply has to normalize the samples $\boldsymbol{X}$ from the first distribution to obtain an instance of a uniformly distributed random variable on the L_p-unit sphere, sample a new radius and scale the normalized

sample with it. Based on that idea, we derive a sampling scheme for L_p-nested distributions in section 6.

Another consequence resulting from the definition of L_p-nested symmetric distributions is the following proposition, which is almost equivalent to Lemma 2.1 and Theorem 2.1 in [3] which themselves are a special case of the results in [2].

Proposition 5.1. *Each L_p-nested symmetric density on $\mathbb{R}^n$ (with zero probability mass at zero) has the form $\tilde{\rho}(\boldsymbol{X}) = \rho(f(\boldsymbol{X}))$ and gives rise to a univariate (radial) density ϱ on $\mathbb{R}_+$. On the other hand, each univariate density ρ on $\mathbb{R}_+$ gives rise to a L_p-nested symmetric distribution on $\mathbb{R}^n$. The relation between the two densities is given by*

$$\begin{aligned}\varrho(r) &= \mathcal{S}_f(1) r^{n-1} \rho(r) \\ &= \mathcal{S}_f(r) \rho(r)\end{aligned}$$

and

$$\begin{aligned}\rho(\mathbf{x}) &= \frac{1}{\mathcal{S}_f(1) \cdot f^{n-1}(\mathbf{x})} \varrho(f(\mathbf{x})) \\ &= \frac{1}{\mathcal{S}_f(f(\mathbf{x}))} \varrho(f(\mathbf{x})).\end{aligned}$$

This shows again, that L_p-nested symmetric distributions are parameterized over univariate radial distributions. The maximum likelihood estimation of the parameters of L_p-nested symmetric distributions therefore becomes very easy since $\operatorname{argmax}_\vartheta \log \rho(\boldsymbol{X}|\vartheta) = \operatorname{argmax}_\vartheta \log \varrho(f(\boldsymbol{X})|\vartheta)$ which means that parameter estimation can be carried out over a univariate instead of an n-dimensional multivariate distribution, which is more robust and computationally efficient.

By the form of a general L_p-nested function and the corresponding symmetric distribution, one might suspect, that the children of the root node, i.e. the $\mathbf{f}_{1:\ell_\emptyset}$ are $L_{p_\emptyset}$-spherically symmetric distributed. This is actually not the case as the next proposition shows.

Proposition 5.2. *Let f be an L_p-nested function. Suppose we remove complete subtrees (not single branches) from the tree associated with f. Let $\hat{\mathbf{x}} \in \mathbb{R}^m$ denote a subset of the coefficients of $\mathbf{x} \in \mathbb{R}^n$ that are still part of that smaller tree and let $\hat{\mathbf{f}}$ denote the vector of inner nodes that became new leafs. The joint distribution of $\hat{\mathbf{x}}$ and $\hat{\mathbf{f}}$ is given by.*

$$\rho(\hat{\mathbf{x}}, \hat{\mathbf{f}}) = \frac{\varrho(f(\hat{\mathbf{x}}, \hat{\mathbf{f}}))}{S_f(f(\hat{\mathbf{x}}, \hat{\mathbf{f}}))} \prod_{J \in \mathcal{J}} \mathsf{f}_J^{n_J - 1}$$

where J is the set of multi-indices for the elements of $\hat{\mathbf{f}}$ and n_J is the number of leafs (in the original tree) in the subtree under the node J.

Proof.

$$\begin{aligned}\rho(\mathbf{x}) &= \frac{\varrho(f(\mathbf{x}))}{\mathcal{S}_f(f(\mathbf{x}))} \\ &= \frac{\varrho(f(\mathbf{x}_{1:n-\ell_I}, \mathsf{f}_I, \tilde{\boldsymbol{u}}_{\ell_I - 1}, \Delta_n))}{\mathcal{S}_f(f(\mathbf{x}))} \cdot \mathsf{f}_I^{\ell_I - 1} \left(1 - \sum_{i=1}^{\ell_I - 1} |\tilde{u}_i|^{p_I}\right)^{\frac{1-p_I}{p_I}}\end{aligned}$$

where $\Delta_n = \text{sign}(x_n)$. Note that f is invariant to the actual value of Δ_n. However, when integrating it out, it yields a factor of 2. Integrating out $\tilde{\boldsymbol{u}}_{\ell_I - 1}$ and Δ_n now yields

$$\begin{aligned}\rho(\mathbf{x}_{1:n-\ell_I}, \mathsf{f}_I) &= \frac{\varrho(f(\mathbf{x}_{1:n-\ell_I}, \mathsf{f}_I))}{\mathcal{S}_f(f(\mathbf{x}))} \cdot \mathsf{f}_I^{\ell_I - 1} \frac{2^{\ell_I} \Gamma^{\ell_I} \left[\frac{1}{p_I}\right]}{p_I^{\ell_I - 1} \Gamma \left[\frac{\ell_I}{p_I}\right]} \\ &= \frac{\varrho(f(\mathbf{x}_{1:n-\ell_I}, \mathsf{f}_I))}{\mathcal{S}_f(f(\mathbf{x}_{1:n-\ell_I}, \mathsf{f}_I))} \cdot \mathsf{f}_I^{\ell_I - 1}\end{aligned}$$

Now, we can go on an integrate out more subtrees. For that purpose, let $\hat{\mathbf{x}}$ denote the remaining coefficients of $\mathbf{x}$, $\hat{\mathbf{f}}$ the vector of leafs resulting from the kind of contraction just shown for f_I and $\mathcal{J}$ the set of multi-indices corresponding to the "new leafs", i.e the node f_I after contraction. We obtain the following equation

$$\rho(\hat{\mathbf{x}}, \hat{\mathbf{f}}) = \frac{\varrho(f(\hat{\mathbf{x}}, \hat{\mathbf{f}}))}{S_f(f(\hat{\mathbf{x}}, \hat{\mathbf{f}}))} \prod_{J \in \mathcal{J}} \mathsf{f}_J^{n_J - 1}.$$

where n_J denotes the number of leafs in the subtree under the node J. The proof is basically the same as the one for proposition (4.1).

□

Corollary 5.1. *The children of the root node* $\mathbf{f}_{1:\ell_\emptyset} = (\mathsf{f}_1, ..., \mathsf{f}_{\ell_\emptyset})^\top$ *follow the distribution*

$$\rho(\mathbf{f}_{1:\ell_\emptyset}) = \frac{p_\emptyset^{\ell_\emptyset - 1} \Gamma \left[\frac{n}{p_\emptyset}\right]}{f^{n-1}(\mathsf{f}_1, ..., \mathsf{f}_{\ell_\emptyset}) 2^m \prod_{k=1}^{\ell_\emptyset} \Gamma \left[\frac{n_k}{p_\emptyset}\right]} \varrho\left(f(\mathsf{f}_1, ..., \mathsf{f}_{\ell_\emptyset})\right) \prod_{i=1}^{\ell_\emptyset} \mathsf{f}_i^{n_i - 1}$$

where $m \leq \ell_\emptyset$ *is the number of leafs directly attached to the root node. In particular,* $\mathbf{f}_{1:\ell_\emptyset}$ *can be written as the product RU, where* R *is the* L_p*-nested radius and the single* $|U_i|^{p_\emptyset}$ *are Dirichlet distributed, i.e.* $(|U_1|^{p_\emptyset}, ..., |U_{\ell_\emptyset}|^{p_\emptyset}) \sim Dir\left[\frac{n_1}{p_\emptyset}, ... \frac{n_{\ell_\emptyset}}{p_\emptyset}\right]$.

Proof. The joint distribution is simply the application of Proposition (5.2). Note that $f(\mathsf{f}_1, ..., \mathsf{f}_{\ell_\emptyset}) = ||\mathbf{f}_{1:\ell_\emptyset}||_{p_\emptyset}$. Applying the pointwise transformation $s_i = |u_i|^{p_\emptyset}$ yields $(|U_1|^{p_\emptyset}, ..., |U_{\ell_\emptyset - 1}|^{p_\emptyset}) \sim \text{Dir}\left[\frac{n_1}{p_\emptyset}, ... \frac{n_{\ell_\emptyset}}{p_\emptyset}\right]$ (see also [6]).

□

6. Sampling from L_p-Nested Symmetric Distributions

In this section, we derive a sampling scheme for L_p-nested symmetric distributions. Since the radial and the uniform component are independent, normalizing a the sample from any L_p-nested distribution to f-length one yields samples from the uniform distribution on the L_p-unit sphere. By multiplying those uniform samples with new samples from another radial distribution, one obtains samples from another L_p-nested distribution. Therefore, for each L_p-nested function f one needs to find only a single L_p-nested distribution one is able to sample from. Sampling from all other L_p-nested distributions with respect to f then comes for free due to the trick just described. Gupta and Song [3] sample from the L_p-generalized Normal distribution since it has independent marginals which makes it easy to sample

from it. Due to the tree structure of L_p-nested distributions, this is not possible in general. Instead we choose to sample from the uniform distribution inside the L_P-nested unit ball.

From Proposition (4.1) we already know the normalization constant. Therefore, the distribution has the form $\rho(\mathbf{x}) = \frac{1}{\mathcal{V}_f(1)}$. In order to sample from that distribution, we will first only consider the uniform distribution in the positive quadrant of the unit L_p-nested ball which has the form $\rho(\mathbf{x}) = \frac{2^n}{\mathcal{V}_f(1)}$. Samples from the uniform distributions in the whole ball can be obtained by multiplying each coordinate of a sample with independent samples from the uniform distribution in $\{-1, 1\}$.

Again, from the proof of Proposition (4.1), we are now able to derive the sampling scheme. The idea of the proof is to successively transform the inner nodes of the tree associated with f into L_p-radial coordinates as defined by [6]. This yields a series of independent integrals over expressions like

$$\int_{\tilde{\boldsymbol{u}}_{\ell_I-1} \in \mathcal{V}_+^{\ell_I-1}} \left(1 - \sum_{i=1}^{\ell_I-1} \tilde{u}_i^{p_I}\right)^{\frac{n_{\ell_I}-p_I}{p_I}} \prod_{k=1}^{\ell_I-1} \tilde{u}_k^{n_k-1} d\tilde{\boldsymbol{u}}_{\ell_I-1}$$

and a final integral over the radius $\mathsf{f}_\emptyset$ which always is

$$\int_0^1 \mathsf{f}_\emptyset^{n-1} d\mathsf{f}_\emptyset.$$

Since all variables together integrate to one, $\rho(\mathbf{x})$ is still a density on those variables. Because we can integrate the independently, the final radial variable $\mathsf{f}_\emptyset$ and the uniform variables are independent. Now, it is easy to see that $\mathsf{f}_\emptyset$ can be drawn from a β-distribution and the single u^{p_I} can be drawn from a Dirichlet distribution. By reversing the transformations we obtain samples from the uniform distribution inside the unit L_p-nested ball. Normalizing those samples yields uniformly distributed points on the L_p-nested unit sphere which can be transformed into samples from any L_p-nested distribution by multiplying with the appropriate radial samples.

This provides us with the following sampling scheme:

(1) Sample $\mathsf{f}_\emptyset$ from a beta distribution $\beta[n, 1]$.
(2) For each inner node I of the tree associated with f sample $\boldsymbol{s}_I$ from a Dirichlet distribution $\mathrm{Dir}\left[\frac{n_{I,1}}{p_I}, ..., \frac{n_{I,\ell_I}}{p_I}\right]$ where $n_{I,k}$ are the number of leafs in the subtree under node I, k. Obtain uniform coordinates on the L_p-sphere by $s_k \mapsto s_k^{\frac{1}{p_I}} = \tilde{u}_k$.
(3) Apply the reverse transformation to map the $\tilde{\boldsymbol{u}}$ and $\mathsf{f}_\emptyset$ into Cartesian coordinates $\mathbf{x}$.
(4) Normalize $\mathbf{x}$ to get a uniform sample from the sphere $\boldsymbol{z} = \frac{\mathbf{x}}{f(\mathbf{x})}$.
(5) Sample a new radius $\tilde{\mathsf{f}}_\emptyset$ from the radial distribution of the target L_p-nested distribution $\rho_\emptyset$ and obtain the sample via $\tilde{\mathbf{x}} = \tilde{\mathsf{f}}_\emptyset \cdot \boldsymbol{z}$.

References

[1] K. T. Fang, S. Kotz, and K. W. Ng. *Symmetric multivariate and related distributions*. Chapman and Hall New York, 1990. 1

[2] Carmen Fernandez, Jacek Osiewalski, and Mark F. J. Steel. Modeling and inference with ν-spherical distributions. *Journal of the American Statistical Association*, 90(432):1331–1340, Dec 1995. 1, 18

[3] A.K. Gupta and D. Song. l_p-norm spherical distribution. *Journal of Statistical Planning and Inference*, 60:241–260, 1997. 1, 4, 13, 15, 17, 18, 19

[4] Douglas Kelker. Distribution theory of spherical distributions and a location-scale parameter generalization. *Sankhya: The Indian Journal of Statistics, Series A*, 32(4):419–430, Dec 1970. 1

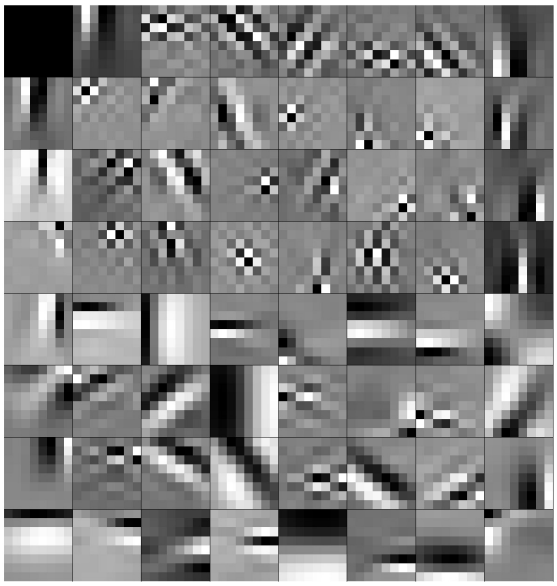
[5] Jacek Osiewalski and Mark F. J. Steel. Robust bayesian inference in l_q-spherical models. *Biometrika*, 80(2):456–460, Jun 1993. 1

[6] D. Song and A.K. Gupta. l_p-norm uniform distribution. *Proceedings of the American Mathematical Society*, 125:595–601, 1997. 1, 13, 17, 19, 20

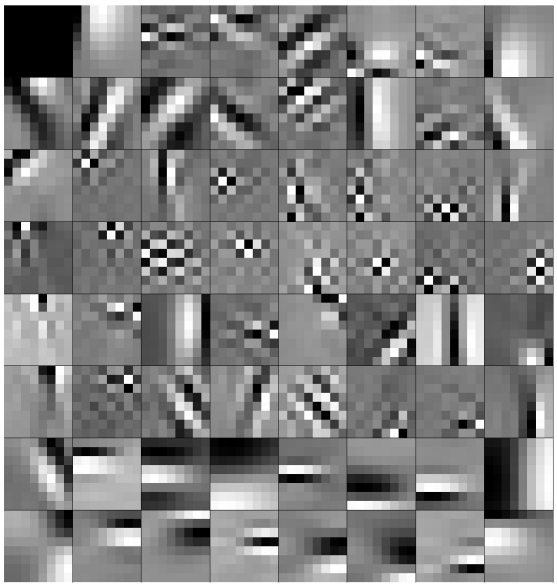
1. Optimal Filters For All Different Models

Independent Subspace Models

Independent Subspace ISA for 2 Subspaces without CGC.

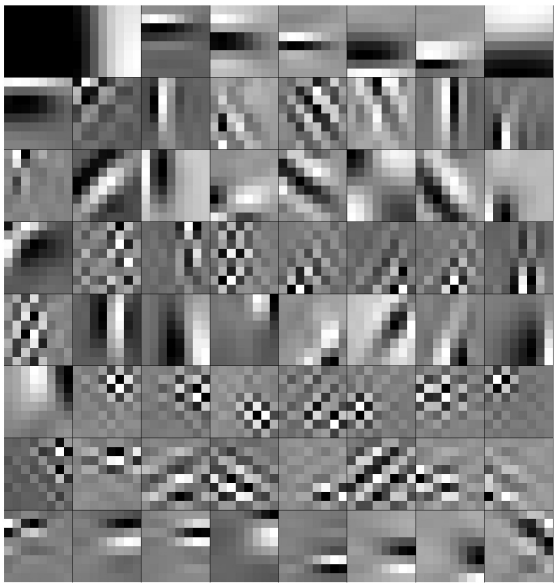


Independent Subspace ISA for 4 Subspaces without CGC.

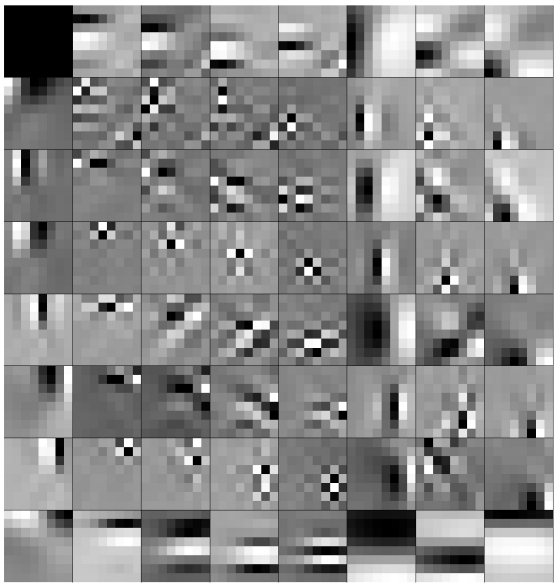


3

Independent Subspace ISA for 8 Subspaces without CGC.

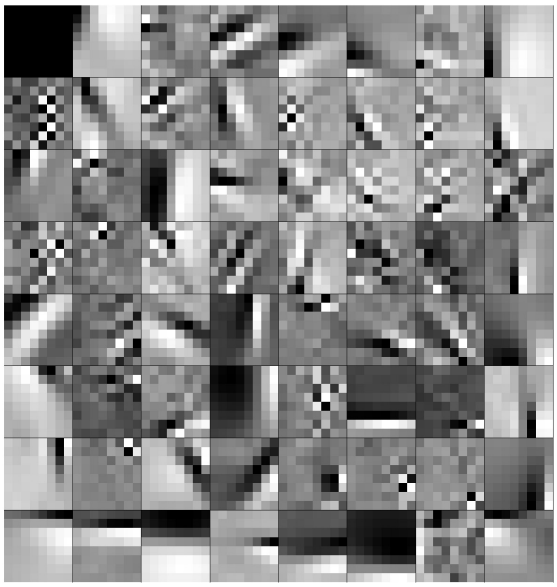


Independent Subspace ISA for 16 Subspaces without CGC.

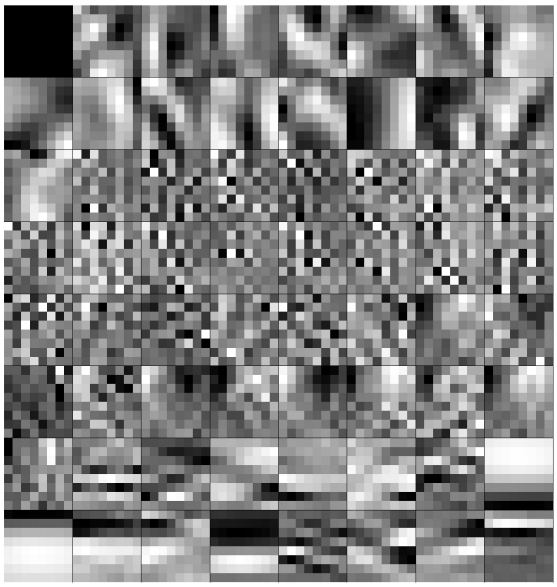


5

Independent Subspace ISA for 2 Subspaces with CGC.

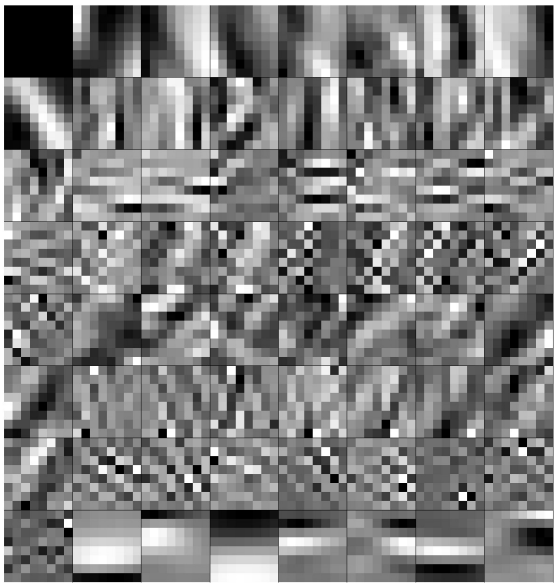


Independent Subspace ISA for 4 Subspaces with CGC.



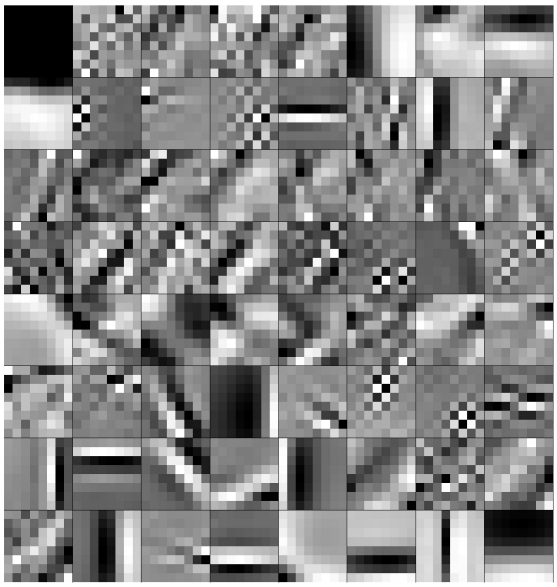
7

Independent Subspace ISA for 8 Subspaces with CGC.



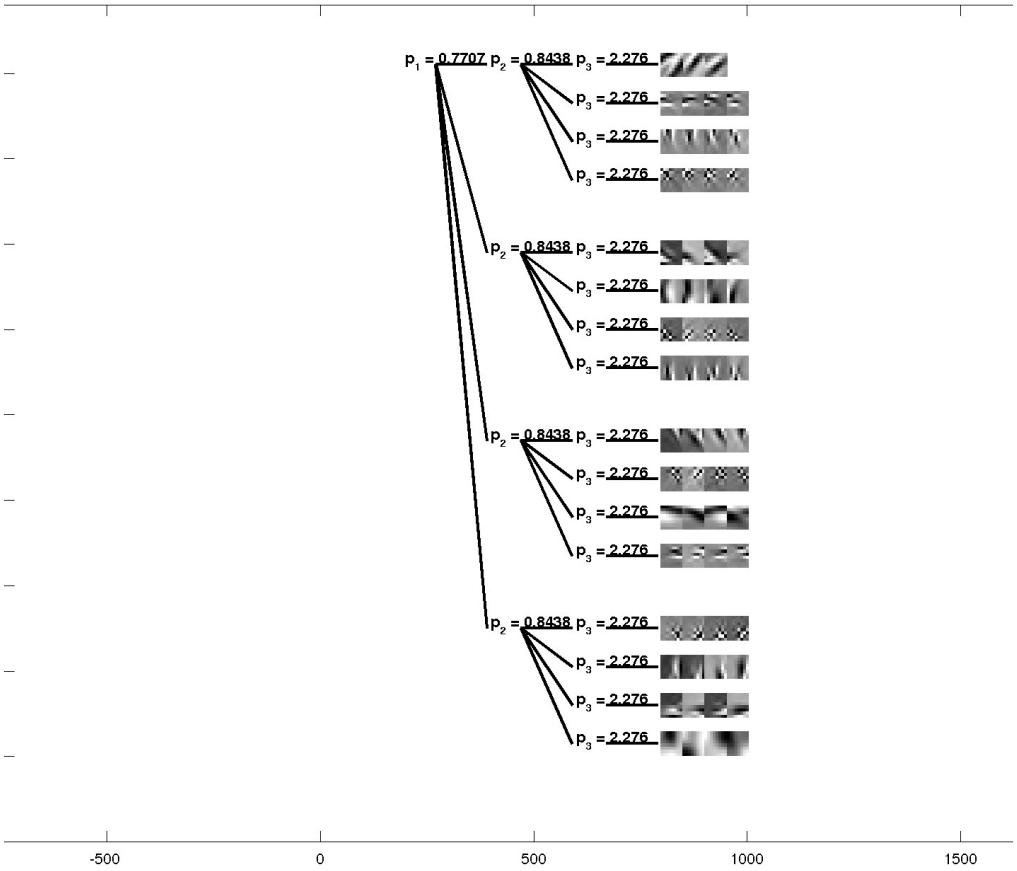
8

Independent Subspace ISA for 16 Subspaces with CGC.



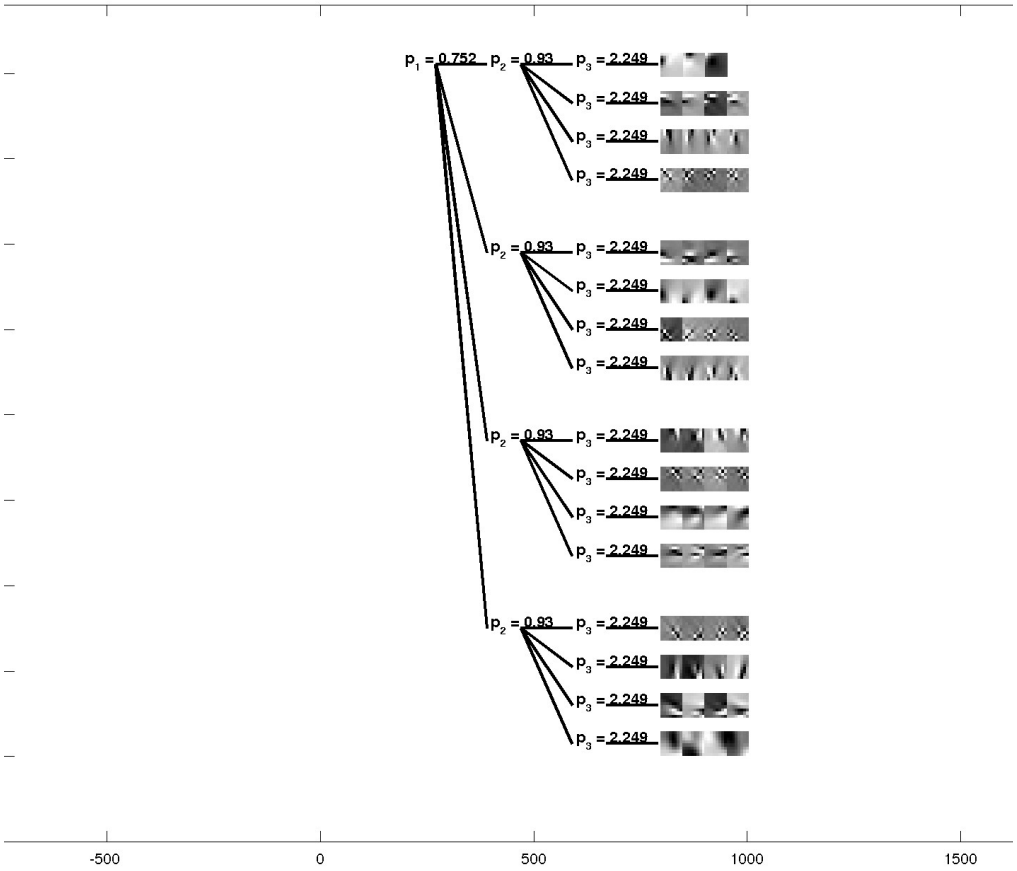
9

L_p-nested model with DT tree structure without CGC.



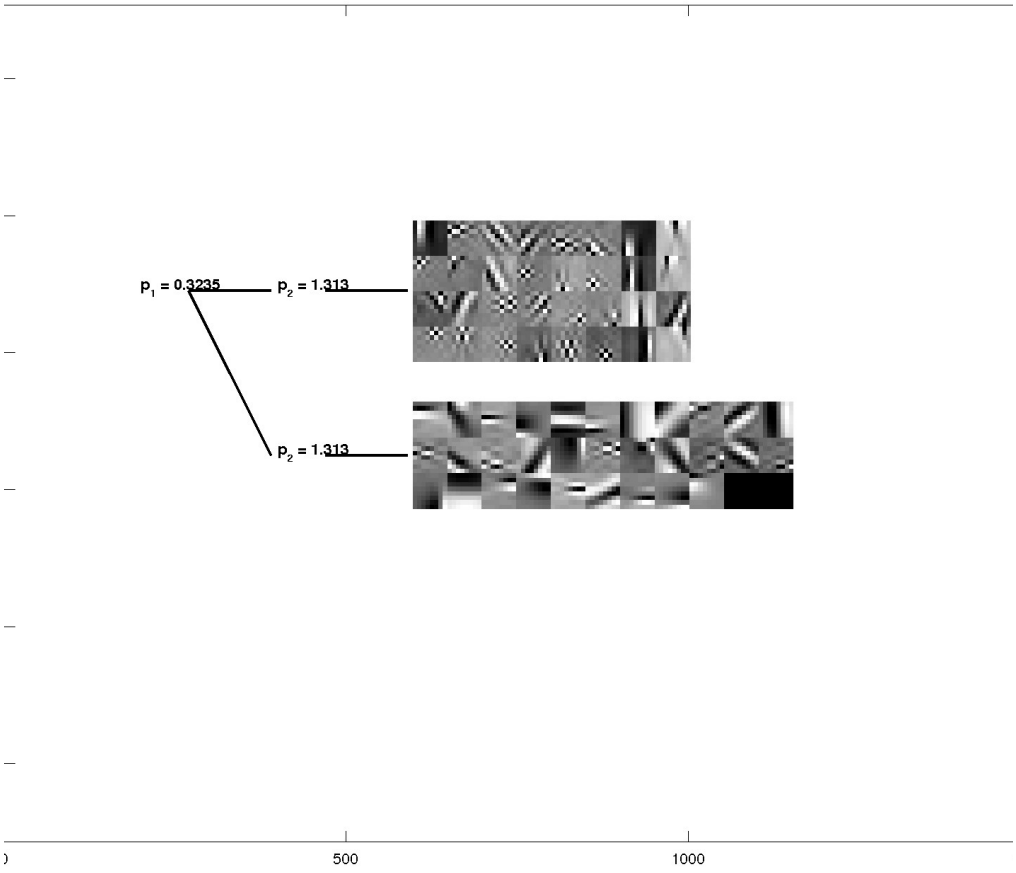
10

L_p-nested model with DT tree structure with CGC.

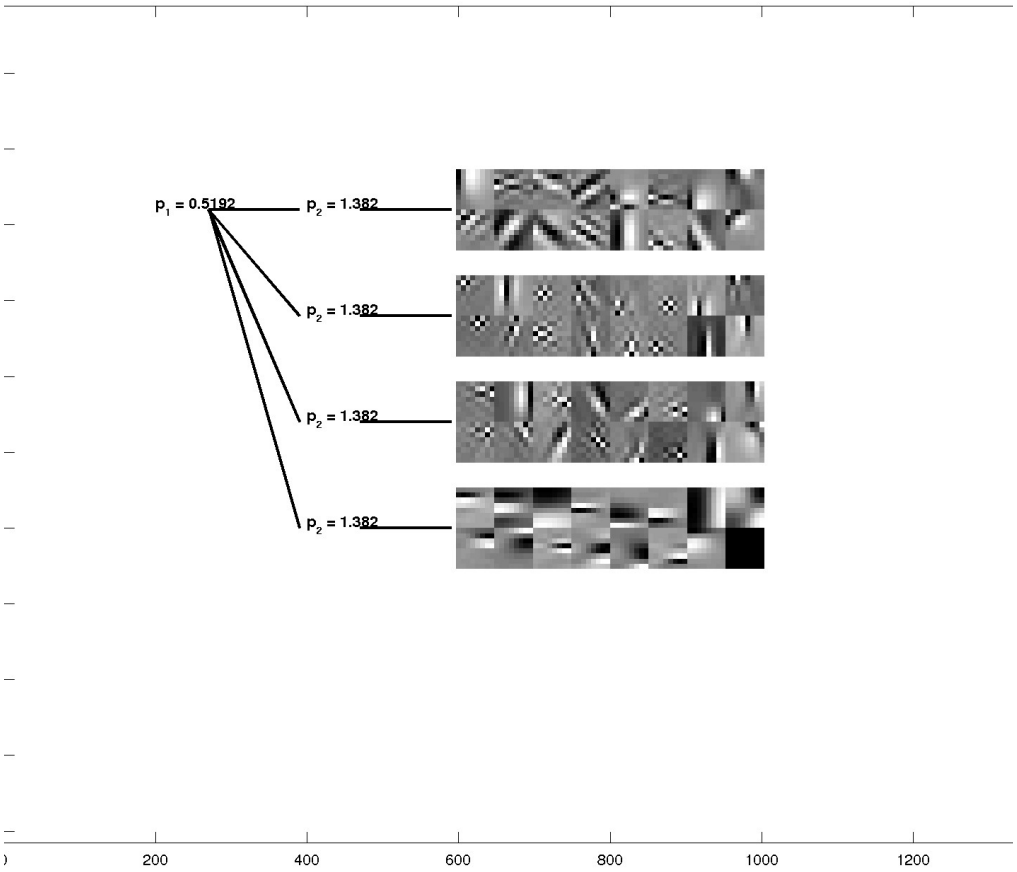


11

L_p-nested model with PND_2 tree structure without CGC.

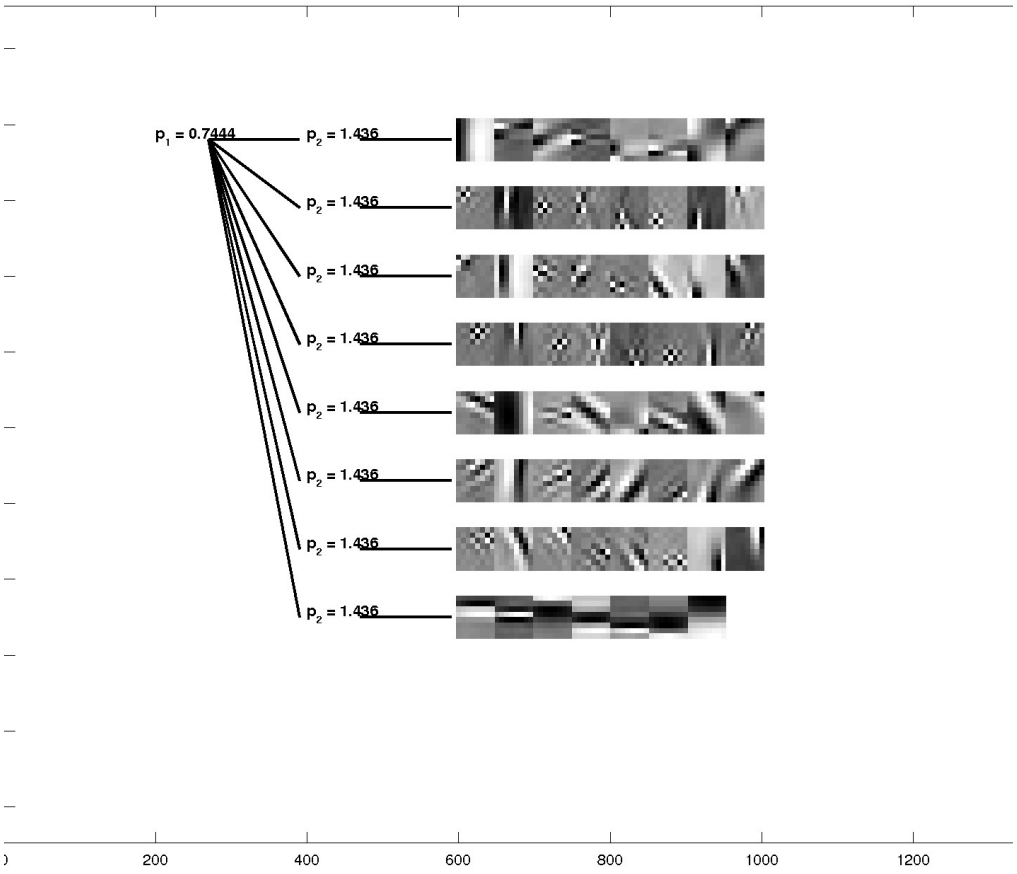


L_p-nested model with PND_4 tree structure without CGC.

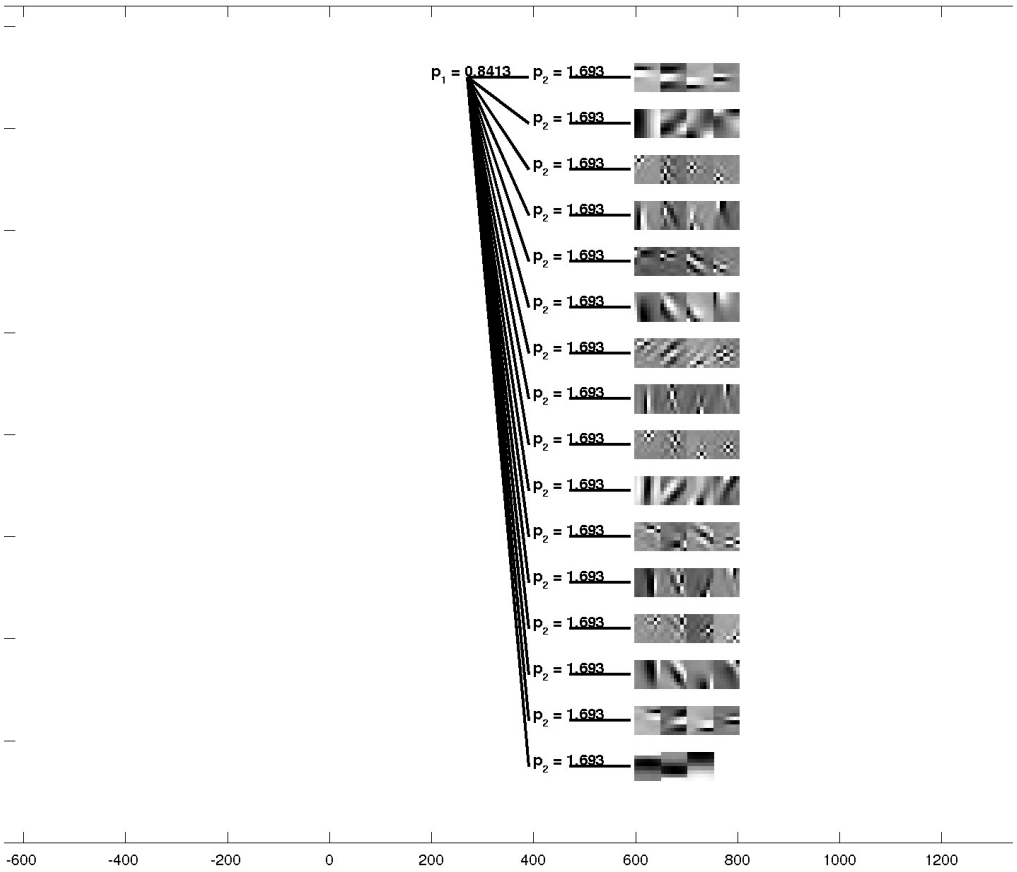


13

L_p-nested model with PND_8 tree structure without CGC.

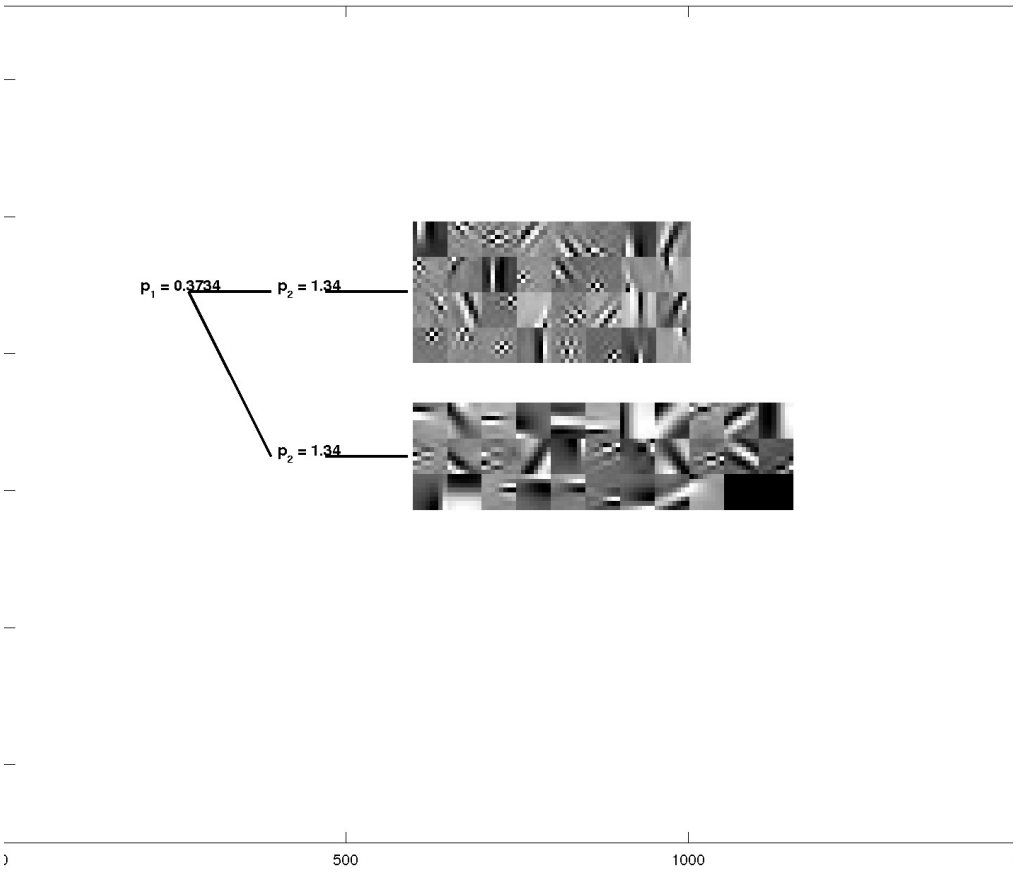


L_p-nested model with PND_{16} tree structure without CGC.

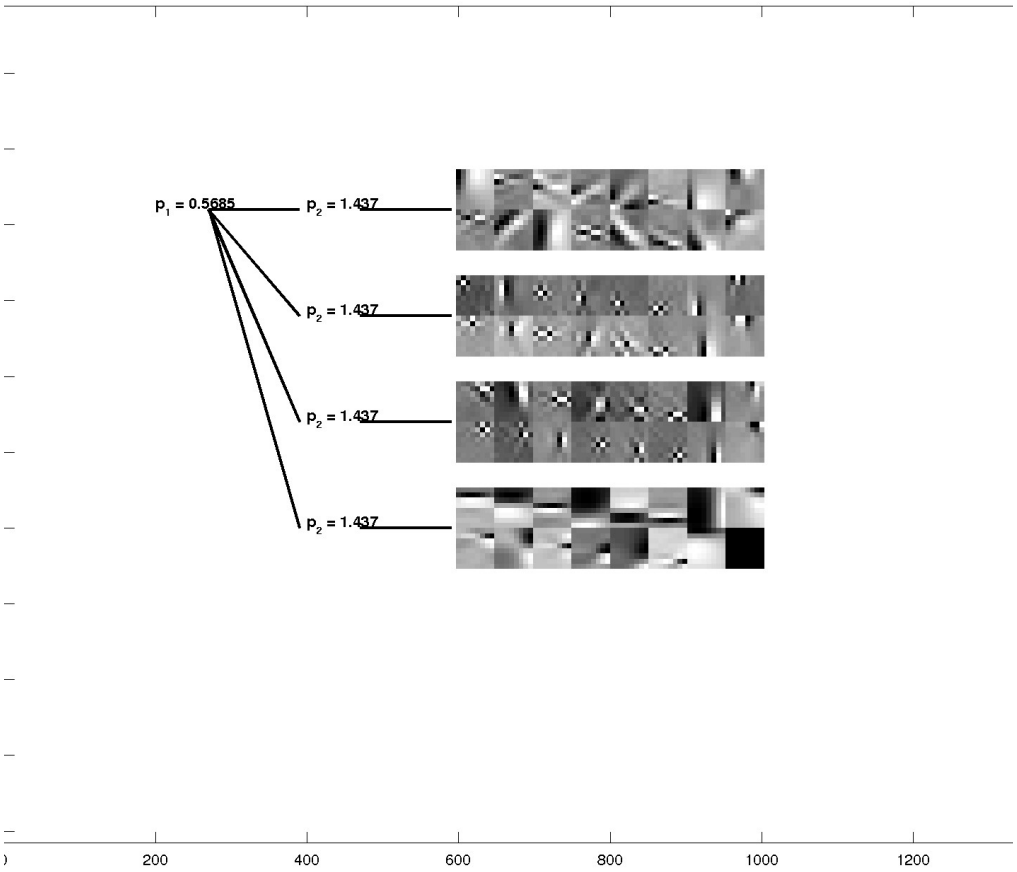


15

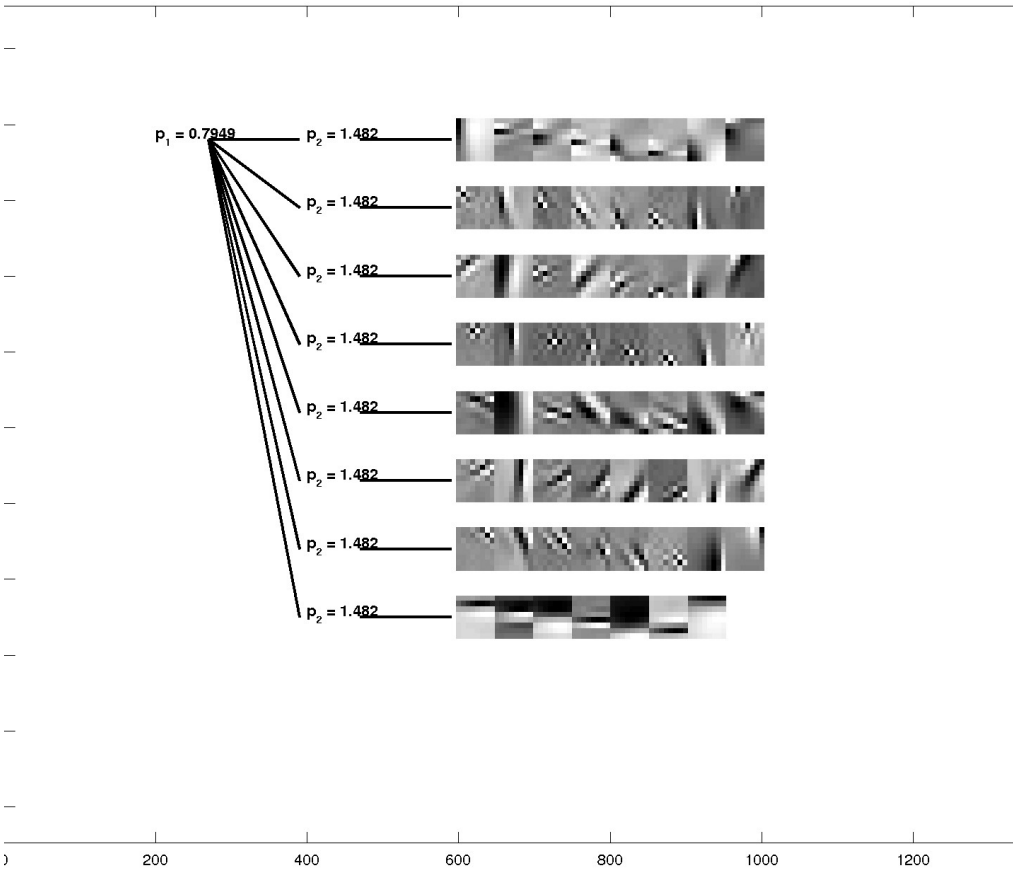
L_p-nested model with PND_2 tree structure with CGC.



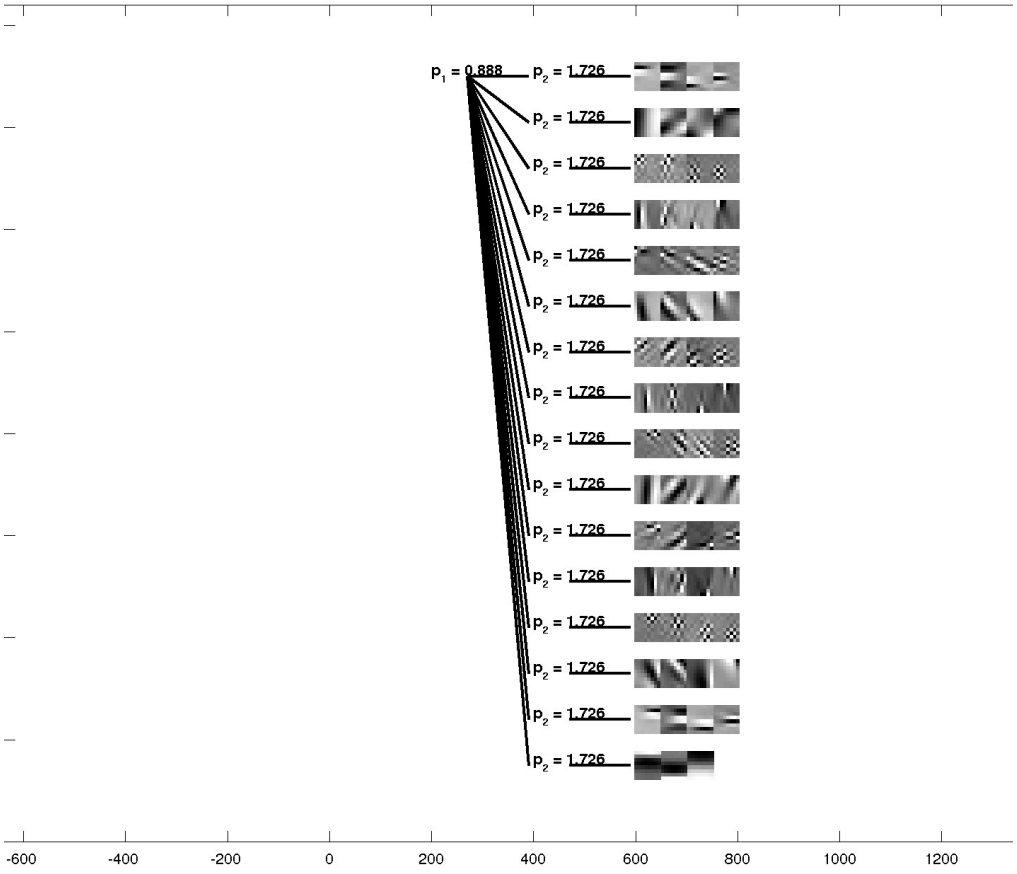
L_p-nested model with PND_4 tree structure with CGC.



L_p-nested model with PND_8 tree structure with CGC.



L_p-nested model with PND_{16} tree structure with CGC.



4.8 L_p-Nested Symmetric Distributions: Original Article

Journal of Machine Learning Research 11 (2010) 3409-3451 Submitted 12/09; Revised 8/10; Published 12/10

L_p-Nested Symmetric Distributions

Fabian Sinz FABEE@TUEBINGEN.MPG.DE
Matthias Bethge MBETHGE@TUEBINGEN.MPG.DE
Werner Reichardt Center for Integrative Neuroscience
Bernstein Center for Computational Neuroscience
Max Planck Institute for Biological Cybernetics
Spemannstraße 41, 72076 Tübingen, Germany

Editor: Aapo Hyvärinen

Abstract

In this paper, we introduce a new family of probability densities called L_p-nested symmetric distributions. The common property, shared by all members of the new class, is the same functional form $\rho(\boldsymbol{x}) = \tilde{\rho}(f(\boldsymbol{x}))$, where f is a nested cascade of L_p-norms $\|\boldsymbol{x}\|_p = (\sum |x_i|^p)^{1/p}$. L_p-nested symmetric distributions thereby are a special case of ν-spherical distributions for which f is only required to be positively homogeneous of degree one. While both, ν-spherical and L_p-nested symmetric distributions, contain many widely used families of probability models such as the Gaussian, spherically and elliptically symmetric distributions, L_p-spherically symmetric distributions, and certain types of independent component analysis (ICA) and independent subspace analysis (ISA) models, ν-spherical distributions are usually computationally intractable. Here we demonstrate that L_p-nested symmetric distributions are still computationally feasible by deriving an analytic expression for its normalization constant, gradients for maximum likelihood estimation, analytic expressions for certain types of marginals, as well as an exact and efficient sampling algorithm. We discuss the tight links of L_p-nested symmetric distributions to well known machine learning methods such as ICA, ISA and mixed norm regularizers, and introduce the nested radial factorization algorithm (NRF), which is a form of non-linear ICA that transforms any linearly mixed, non-factorial L_p-nested symmetric source into statistically independent signals. As a corollary, we also introduce the uniform distribution on the L_p-nested unit sphere.

Keywords: parametric density model, symmetric distribution, ν-spherical distributions, non-linear independent component analysis, independent subspace analysis, robust Bayesian inference, mixed norm density model, uniform distributions on mixed norm spheres, nested radial factorization

1. Introduction

High-dimensional data analysis virtually always starts with the measurement of first and second-order moments that are sufficient to fit a multivariate Gaussian distribution, the maximum entropy distribution under these constraints. Natural data, however, often exhibit significant deviations from a Gaussian distribution. In order to model these higher-order correlations, it is necessary to have more flexible distributions available. Therefore, it is an important challenge to find generalizations of the Gaussian distribution which are more flexible but still computationally and analytically tractable. In particular, density models with an explicit normalization constant are desirable because they make direct model comparison possible by comparing the likelihood of held out test

samples for different models. Additionally, such models often allow for a direct optimization of the likelihood.

One way of imposing structure on probability distributions is to fix the general form of the iso-density contour lines. This approach was taken by Fernandez et al. (1995). They modeled the contour lines by the level sets of a positively homogeneous function of degree one, that is functions ν that fulfill $\nu(a \cdot \boldsymbol{x}) = a \cdot \nu(\boldsymbol{x})$ for $\boldsymbol{x} \in \mathbb{R}^n$ and $a \in \mathbb{R}_0^+$. The resulting class of ν-spherical distributions have the general form $\rho(\boldsymbol{x}) = \tilde{\rho}(\nu(\boldsymbol{x}))$ for an appropriate $\tilde{\rho}$ which causes $\rho(\boldsymbol{x})$ to integrate to one. Since the only access of ρ to $\boldsymbol{x}$ is via ν one can show that, for a fixed ν, those distributions are generated by a univariate radial distribution. In other words, ν-spherically distributed random variables can be represented as a product of two independent random variables: one positive radial variable and another variable which is uniform on the 1-level set of ν. This property makes this class of distributions easy to fit to data since the maximum likelihood procedure can be carried out on the univariate radial distribution instead of the joint density. Unfortunately, deriving the normalization constant for the joint distribution in the general case is intractable because it depends on the surface area of those level sets which can usually not be computed analytically.

Known tractable subclasses of ν-spherical distributions are the Gaussian, elliptically contoured, and L_p-spherical distributions. The Gaussian is a special case of elliptically contoured distributions. After centering and whitening $\boldsymbol{x} := C^{-1/2}(\boldsymbol{s} - E[\boldsymbol{s}])$ a Gaussian distribution is spherically symmetric and the squared L_2-norm $||\boldsymbol{x}||_2^2 = x_1^2 + \cdots + x_n^2$ of the samples follow a χ^2-distribution (that is, the radial distribution is a χ-distribution). Elliptically contoured distributions other than the Gaussian are obtained by using a radial distribution different from the χ-distribution (Kelker, 1970; Fang et al., 1990).

The extension from L_2- to L_p-spherically symmetric distributions is based on replacing the L_2-norm by the L_p-norm

$$\nu(\boldsymbol{x}) = \|\boldsymbol{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}, \; p > 0$$

in the density definition. That is, the density of L_p-spherically symmetric distributions can always be written in the form $\rho(\boldsymbol{x}) = \tilde{\rho}(||\boldsymbol{x}||_p)$. Those distributions have been studied by Osiewalski and Steel (1993) and Gupta and Song (1997). We will adopt the naming convention of Gupta and Song (1997) and call $\|\boldsymbol{x}\|_p$ an *L_p-norm* even though the triangle inequality only holds for $p \geq 1$. L_p-spherically symmetric distributions with $p \neq 2$ are no longer invariant with respect to rotations (transformations from $SO(n)$). Instead, they are only invariant under permutations of the coordinate axes. In some cases, it may not be too restrictive to assume permutation or even rotational symmetry for the data. In other cases, such symmetry assumptions might not be justified and cause the model to miss important regularities.

Here, we present a generalization of the class of L_p-spherically symmetric distributions within the class of ν-spherical distributions that makes weaker assumptions about the symmetries in the data but still is analytically tractable. Instead of using a single L_p-norm to define the contour of the density, we use a nested cascade of L_p-norms where an L_p-norm is computed over groups of L_p-norms over groups of L_p-norms ..., each of which having a possibly different p. Due to this nested structure we call this new class of distributions *L_p-nested symmetric distributions*. The nested combination of L_p-norms preserves positive homogeneity but does not require permutation invariance anymore. While L_p-nested symmetric distributions are still invariant under reflections of the coordinate axes, permutation symmetry only holds within the subspaces of the L_p-norms at the bottom of

the cascade. As demonstrated in Sinz et al. (2009b), one possible application domain of L_p-nested symmetric distributions is natural image patches. In the current paper, we would like to present a formal treatment of this class of distributions. Readers interested in the application of these distributions to natural images should refer to Sinz et al. (2009b).

We demonstrate below that the construction of the nested L_p-norm cascade still bears enough structure to compute the Jacobian of polar-like coordinates similar to those of Song and Gupta (1997), and Gupta and Song (1997). With this Jacobian at hand it is possible to compute the univariate radial distribution for an arbitrary L_p-nested symmetric density and to define the uniform distribution on the L_p-nested unit sphere $\mathbb{L}_\nu = \{\boldsymbol{x} \in \mathbb{R}^n | \nu(\boldsymbol{x}) = 1\}$. Furthermore, we compute the surface area of the L_p-nested unit sphere and, therefore, the general normalization constant for L_p-nested symmetric distributions. By deriving these general relations for the class of L_p-nested symmetric distributions we have determined a new class of tractable ν-spherical distributions which is so far the only one containing the Gaussian, elliptically contoured, and L_p-spherical distributions as special cases.

L_p-spherically symmetric distributions have been used in various contexts in statistics and machine learning. Many results carry over to L_p-nested symmetric distributions which allow a wider application range. Osiewalski and Steel (1993) showed that the posterior on the location of a L_p-spherically symmetric distributions together with an improper Jeffrey's prior on the scale does not depend on the particular type of L_p-spherically symmetric distribution used. Below, we show that this results carries over to L_p-nested symmetric distributions. This means that we can robustly determine the location parameter by Bayesian inference for a very large class of distributions.

A large class of machine learning algorithms can be written as an optimization problem on the sum of a regularizer and a loss function. For certain regularizers and loss functions, like the sparse L_1 regularizer and the mean squared loss, the optimization problem can be seen as the maximum a posteriori (MAP) estimate of a stochastic model in which the prior and the likelihood are the negative exponentiated regularizer and loss terms. Since $\rho(\boldsymbol{x}) \propto \exp(-||\boldsymbol{x}||_p^p)$ is an L_p-spherically symmetric model, regularizers which can be written in terms of a norm have a tight link to L_p-spherically symmetric distributions. In an analogous way, L_p-nested symmetric distributions exhibit a tight link to mixed-norm regularizers which have recently gained increasing interest in the machine learning community (see, e.g., Zhao et al., 2008; Yuan and Lin, 2006; Kowalski et al., 2008). L_p-nested symmetric distributions can be used for a Bayesian treatment of mixed-norm regularized algorithms. Furthermore, they can be used to understand the prior assumptions made by such regularizers. Below we discuss an implicit dependence assumption between the regularized variables that follows from the theory of L_p-nested symmetric distributions.

Finally, the only factorial L_p-spherically symmetric distribution (Sinz et al., 2009a), the p-generalized Normal distribution, has been used as an ICA model in which the marginals follow an exponential power distribution. This class of ICA is particularly suited for natural signals like images and sounds (Lee and Lewicki, 2000; Zhang et al., 2004; Lewicki, 2002). Interestingly, L_p-spherically symmetric distributions other than the p-generalized Normal give rise to a non-linear ICA algorithm called radial Gaussianization for $p = 2$ (Lyu and Simoncelli, 2009) or radial factorization for arbitrary p (Sinz and Bethge, 2009). As discussed below, L_p-nested symmetric distributions are a natural extension of the linear L_p-spherically symmetric ICA algorithm to ISA, and give rise to a more general non-linear ICA algorithm in the spirit of radial factorization.

The remaining part of the paper is structured as follows: in Section 2 we define polar-like coordinates for L_p-nested symmetrically distributed random variables and present an analytical expression

for the determinant of the Jacobian for this coordinate transformation. Using this expression, we define the uniform distribution on the L_p-nested unit sphere and the class of L_p-nested symmetric distributions for an arbitrary L_p-nested function in Section 3. In Section 4 we derive an analytical form of L_p-nested symmetric distributions when marginalizing out lower levels of the L_p-nested cascade and demonstrate that marginals of L_p-nested symmetric distributions are not necessarily L_p-nested symmetric. Additionally, we demonstrate that the only factorial L_p-nested symmetric distribution is necessarily L_p-spherically symmetric and discuss the implications of this result for mixed norm regularizers. In Section 5 we propose an algorithm for fitting arbitrary L_p-nested symmetric models. We derive a sampling scheme for arbitrary L_p-nested symmetric distributions in Section 6. In Section 7 we generalize a result by Osiewalski and Steel (1993) on robust Bayesian inference on the location parameter to L_p-nested symmetric distributions. In Section 8 we discuss the relationship of L_p-nested symmetric distributions to ICA and ISA, and their possible role as priors on hidden variables in over-complete linear models. Finally, we derive a non-linear ICA algorithm for linearly mixed non-factorial L_p-nested symmetric sources in Section 9 which we call nested radial factorization (NRF).

2. L_p-nested Functions, Coordinate Transformation and Jacobian

Consider the function

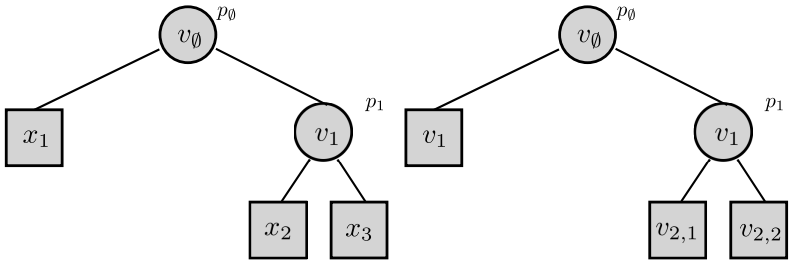
$$f(\boldsymbol{x}) = \left(|x_1|^{p_\emptyset} + (|x_2|^{p_1} + |x_3|^{p_1})^{\frac{p_\emptyset}{p_1}}\right)^{\frac{1}{p_\emptyset}} \tag{1}$$

with $p_\emptyset, p_1 \in \mathbb{R}^+$. This function is obviously a cascade of two L_p-norms and is thus positively homogeneous of degree one. Figure 1(a) shows this function visualized as a tree. Naturally, any tree like the ones in Figure 1 corresponds to a function of the kind of Equation (1). In general, the n leaves of the tree correspond to the n coefficients of the vector $\boldsymbol{x} \in \mathbb{R}^n$ and each inner node computes the L_p-norm of its children using its specific p. We call the class of functions which is generated in this way L_p-*nested* and the corresponding distributions, which are symmetric or invariant with respect to it, L_p-*nested symmetric distributions*.

L_p-nested functions are much more flexible in creating different shapes of level sets than single L_p-norms. Those level sets become the iso-density contours in the family of L_p-nested symmetric distributions. Figure 2 shows a variety of contours generated by the simplest non-trivial L_p-nested function shown in Equation (1). The shapes show the unit spheres for all possible combinations of $p_\emptyset, p_1 \in \{0.5, 1, 2, 10\}$. On the diagonal, $p_\emptyset$ and p_1 are equal and therefore constitute L_p-norms. The corresponding distributions are members of the L_p-spherically symmetric class.

To make general statements about general L_p-nested functions, we introduce a notation that is suitable for the tree structure of L_p-nested functions. As we will heavily use that notation in the remainder of the paper, we would like to emphasize the importance of the following paragraphs. We will illustrate the notation with an example below. Additionally, Figure 1 and Table 1 can be used for reference.

We use multi-indices to denote the different nodes of the tree corresponding to an L_p-nested function f. The function $f = f_\emptyset$ itself computes the value $v_\emptyset$ at the root node (see Figure 1). Those values are denoted by variables v. The functions corresponding to its children are denoted by $f_1, ..., f_{\ell_\emptyset}$, that is, $f(\cdot) = f_\emptyset(\cdot) = \|(f_1(\cdot), ..., f_{\ell_\emptyset}(\cdot))\|_{p_\emptyset}$. We always use the letter "ℓ" indexed by the node's multi-index to denote the total number of direct children of that node. The functions of



(a) Equation (1) as tree. (b) Equation (1) as tree in multi-index notation.

Figure 1: Equation (1) visualized as a tree with two different naming conventions. Figure (a) shows the tree where the nodes are labeled with the coefficients of $\boldsymbol{x} \in \mathbb{R}^n$. Figure (b) shows the same tree in multi-index notation where the multi-index of a node describes the path from the root node to that node in the tree. The leaves $v_1, v_{2,1}$ and $v_{2,2}$ still correspond to x_1, x_2 and x_3, respectively, but have been renamed to the multi-index notation used in this article.

$f(\cdot) = f_\emptyset(\cdot)$	L_p-nested function
$I = i_1, ..., i_m$	Multi-index denoting a node in the tree: The single indices describe the path from the root node to the respective node I.
$\boldsymbol{x}_I$	All entries in $\boldsymbol{x}$ that correspond to the leaves in the subtree under the node I
$\boldsymbol{x}_{\widehat{I}}$	All entries in $\boldsymbol{x}$ that are not leaves in the subtree under the node I
$f_I(\cdot)$	L_p-nested function corresponding to the subtree under the node I
$v_\emptyset$	Function value at the root node
v_I	Function value at an arbitrary node with multi-index I
ℓ_I	The number of direct children of a node I
n_I	The number of leaves in the subtree under the node I
$\boldsymbol{v}_{I,1:\ell_I}$	Vector with the function values at the direct children of a node I

Table 1: Summary of the notation used for L_p-nested functions in this article.

the children of the i^{th} child of the root node are denoted by $f_{i,1}, ..., f_{i,\ell_i}$ and so on. In this manner, an index is added for denoting the children of a particular node in the tree and each multi-index denotes the path to the respective node in the tree. For the sake of compact notation, we use upper case letters to denote a single multi-index $I = i_1, ..., i_\ell$. The range of the single indices and the length of the multi-index should be clear from the context. A concatenation I, k of a multi-index I with a single index k corresponds to adding k to the index tuple, that is, $I, k = i_1, ..., i_m, k$. We use the

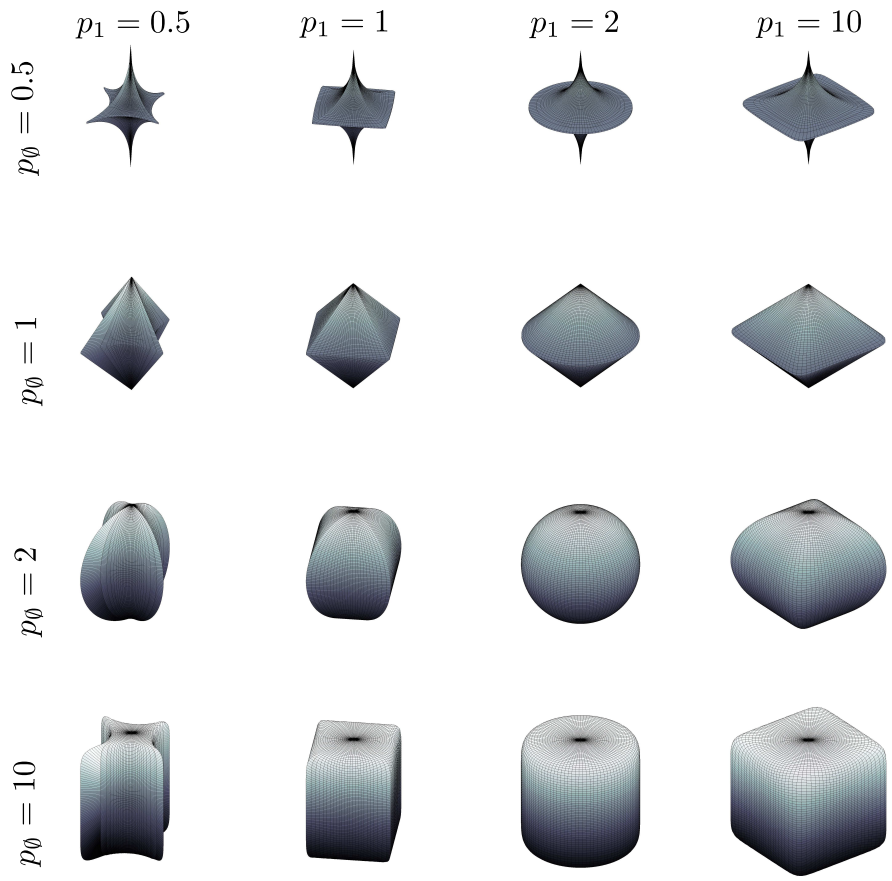


Figure 2: Variety of contours created by the L_p-nested function of Equation (1) for all combinations of $p_\emptyset, p_1 \in \{0.5, 1, 2, 10\}$.

convention that $I, \emptyset = I$. Those coefficients of the vector $\boldsymbol{x}$ that correspond to leaves of the subtree under a node with the index I are denoted by $\boldsymbol{x}_I$. The complement of those coefficients, that is, the ones that are not in the subtree under the node I, are denoted by $\boldsymbol{x}_{\hat{I}}$. The number of leaves in a subtree under a node I is denoted by n_I. If I denotes a leaf then $n_I = 1$.

The L_p-nested function associated with the subtree under a node I is denoted by

$$f_I(\boldsymbol{x}_I) = ||(f_{I,1}(\boldsymbol{x}_{I,1}), ..., f_{I,\ell_I}(\boldsymbol{x}_{I,\ell_I}))^\top||_{p_I}.$$

Just like for the root node, we use the variable v_I to denote the function value $v_I = f_I(\boldsymbol{x}_I)$ of a subtree I. A vector with the function values of the children of I is denoted with bold font $\boldsymbol{v}_{I,1:\ell_I}$ where the colon indicates that we mean the vector of the function values of the ℓ_I children of node I:

$$\begin{aligned} f_I(\boldsymbol{x}_I) &= ||(f_{I,1}(\boldsymbol{x}_{I,1}),...,f_{I,\ell_I}(\boldsymbol{x}_{I,\ell_I}))^\top||_{p_I} \\ &= ||(v_{I,1},...,v_{I,\ell_I})^\top||_{p_I} = ||\boldsymbol{v}_{I,1:\ell_I}||_{p_I}. \end{aligned}$$

Note that we can assign an arbitrary p to leaf nodes since ps for single variables always cancel. For that reason we can choose an arbitrary p for convenience and fix its value to $p = 1$. Figure 1(b) shows the multi-index notation for our example of Equation (1).

To illustrate the notation: Let $I = i_1,...,i_d$ be the multi-index of a node in the tree. $i_1,...,i_d$ describes the path to that node, that is, the respective node is the i_d^{th} child of the i_{d-1}^{th} child of the i_{d-2}^{th} child of the ... of the i_1^{th} child of the root node. Assume that the leaves in the subtree below the node I cover the vector entries $x_2,...,x_{10}$. Then $\boldsymbol{x}_I = (x_2,...,x_{10})$, $\boldsymbol{x}_{\hat{I}} = (x_1,x_{11},x_{12},...)$, and $n_I = 9$. Assume that node I has $\ell_I = 2$ children. Those would be denoted by $I,1$ and $I,2$. The function realized by node I would be denoted by f_I and only acts on $\boldsymbol{x}_I$. The value of the function would be $f_I(\boldsymbol{x}_I) = v_I$ and the vector containing the values of the children of I would be $\boldsymbol{v}_{I,1:2} = (v_{I,1},v_{I,2})^\top = (f_{I,1}(\boldsymbol{x}_{I,1}), f_{I,2}(\boldsymbol{x}_{I,2}))^\top$.

We now introduce a coordinate representation specially tailored to L_p-nested symmetrically distributed variables: One of the most important consequences of the positive homogeneity of f is that it can be used to "normalize" vectors and, by that property, create a polar like coordinate representation of a vector $\boldsymbol{x}$. Such polar-like coordinates generalize the coordinate representation for L_p-norms by Gupta and Song (1997).

Definition 1 (Polar-like Coordinates) *We define the following polar-like coordinates for a vector $\boldsymbol{x} \in \mathbb{R}^n$:*

$$\begin{aligned} u_i &= \frac{x_i}{f(\boldsymbol{x})} \text{ for } i = 1,...,n-1, \\ r &= f(\boldsymbol{x}). \end{aligned}$$

The inverse coordinate transformation is given by

$$\begin{aligned} x_i &= r u_i \text{ for } i = 1,...,n-1, \\ x_n &= r \Delta_n u_n \end{aligned}$$

where $\Delta_n = \operatorname{sgn} x_n$ and $u_n = \frac{|x_n|}{f(\boldsymbol{x})}$.

Note that u_n is not part of the coordinate representation since normalization with $1/f(\boldsymbol{x})$ decreases the degrees of freedom $\boldsymbol{u}$ by one, that is, u_n can always be computed from all other u_i by solving $f(\boldsymbol{u}) = f(\boldsymbol{x}/f(\boldsymbol{x})) = 1$ for u_n. We use the term u_n only for notational simplicity. With a slight abuse of notation, we will use $\boldsymbol{u}$ to denote the normalized vector $\boldsymbol{x}/f(\boldsymbol{x})$ or only its first $n-1$ components. The exact meaning should always be clear from the context.

The definition of the coordinates is exactly the same as the one by Gupta and Song (1997) with the only difference that the L_p-norm is replaced by an L_p-nested function. Just as in the case of L_p-spherical coordinates, it will turn out that the determinant of the Jacobian of the coordinate

transformation does not depend on the value of Δ_n and can be computed analytically. The determinant is essential for deriving the uniform distribution on the unit L_p-nested sphere $\mathbb{L}_f$, that is, the 1-level set of f. Apart from that, it can be used to compute the radial distribution for a given L_p-nested symmetric distribution. We start by stating the general form of the determinant in terms of the partial derivatives $\frac{\partial u_n}{\partial u_k}$, u_k and r. Afterwards we demonstrate that those partial derivatives have a special form and that most of them cancel in Laplace's expansion of the determinant.

Lemma 2 (Determinant of the Jacobian) *Let r and $\boldsymbol{u}$ be defined as in Definition 1. The general form of the determinant of the Jacobian $\mathcal{J} = \left(\frac{\partial x_i}{\partial y_j}\right)_{ij}$ of the inverse coordinate transformation for $y_1 = r$ and $y_i = u_{i-1}$ for $i = 2,\ldots,n$, is given by*

$$|\det \mathcal{J}| = r^{n-1}\left(-\sum_{k=1}^{n-1}\frac{\partial u_n}{\partial u_k}\cdot u_k + u_n\right). \tag{2}$$

Proof The proof can be found in the Appendix A. ■

The problematic parts in Equation (2) are the terms $\frac{\partial u_n}{\partial u_k}$, which obviously involve extensive usage of the chain rule. Fortunately, most of them cancel when inserting them back into Equation (2), leaving a comparably simple formula. The remaining part of this section is devoted to computing those terms and demonstrating how they vanish in the formula for the determinant. Before we state the general case we would like to demonstrate the basic mechanism through a simple example. We urge the reader to follow this example as it illustrates all important ideas about the coordinate transformation and its Jacobian.

Example 1 *Consider an L_p-nested function very similar to our introductory example of Equation (1):*

$$f(\boldsymbol{x}) = \left(\left(|x_1|^{p_1} + |x_2|^{p_1}\right)^{\frac{p_0}{p_1}} + |x_3|^{p_0}\right)^{\frac{1}{p_0}}.$$

Setting $\boldsymbol{u} = \frac{\boldsymbol{x}}{f(\boldsymbol{x})}$ and solving for u_3 yields

$$f(\boldsymbol{u}) = 1 \Leftrightarrow\ u_3 = \left(1 - \left(|u_1|^{p_1} + |u_2|^{p_1}\right)^{\frac{p_0}{p_1}}\right)^{\frac{1}{p_0}}. \tag{3}$$

We would like to emphasize again, that u_3 is actually not part of the coordinate representation and only used for notational simplicity. By construction, u_3 is always positive. This is no restriction since Lemma 2 shows that the determinant of the Jacobian does not depend on its sign. However, when computing the volume and the surface area of the L_p-nested unit sphere, it will become important since it introduces a factor of 2 to account for the fact that u_3 (or u_n in general) can in principle also attain negative values.

Now, consider

$$G_2(\boldsymbol{u}_{\hat{2}}) = g_2(\boldsymbol{u}_{\hat{2}})^{1-p_0} = \left(1 - \left(|u_1|^{p_1} + |u_2|^{p_1}\right)^{\frac{p_0}{p_1}}\right)^{\frac{1-p_0}{p_0}},$$
$$F_1(\boldsymbol{u}_1) = f_1(\boldsymbol{u}_1)^{p_0-p_1} = \left(|u_1|^{p_1} + |u_2|^{p_1}\right)^{\frac{p_0-p_1}{p_1}},$$

where the subindices of $\boldsymbol{u}$, f, g, G *and* F *have to be read as multi-indices. The function* g_I *computes the value of the node* I *from all other leaves that are not part of the subtree under* I *by fixing the value of the root node to one.*

$G_2(\boldsymbol{u}_{\widehat{2}})$ *and* $F_1(\boldsymbol{u}_1)$ *are terms that arise from applying the chain rule when computing the partial derivatives* $\frac{\partial u_3}{\partial u_k}$. *Taking those partial derivatives can be thought of as peeling off layer by layer of Equation* (3) *via the chain rule. By doing so, we "move" on a path between* u_3 *and* u_k. *Each application of the chain rule corresponds to one step up or down in the tree. First, we move upwards in the tree, starting from* u_3. *This produces the G-terms. In this example, there is only one step upwards, but in general, there can be several, depending on the depth of* u_n *in the tree. Each step up will produce one G-term. At some point, we will move downwards in the tree to reach* u_k. *This will produce the F-terms. While there are as many G-terms as upward steps, there is one term less when moving downwards. Therefore, in this example, there is one term* $G_2(\boldsymbol{u}_{\widehat{2}})$ *which originates from using the chain rule upwards in the tree and one term* $F_1(\boldsymbol{u}_1)$ *from using it downwards. The indices correspond to the multi-indices of the respective nodes.*

Computing the derivative yields

$$\frac{\partial u_3}{\partial u_k} = -G_2(\boldsymbol{u}_{\widehat{2}})F_1(\boldsymbol{u}_1)\Delta_k|u_k|^{p_1-1}.$$

By inserting the results in Equation (2) we obtain

$$\begin{aligned}
\frac{1}{r^2}|\mathcal{J}| &= \sum_{k=1}^{2} G_2(\boldsymbol{u}_{\widehat{2}})F_1(\boldsymbol{u}_1)|u_k|^{p_1} + u_3 \\
&= G_2(\boldsymbol{u}_{\widehat{2}})\left(F_1(\boldsymbol{u}_1)\sum_{k=1}^{2}|u_k|^{p_1} + 1 - F_1(\boldsymbol{u}_1)F_1(\boldsymbol{u}_1)^{-1}\left(|u_1|^{p_1}+|u_2|^{p_1}\right)^{\frac{p_0}{p_1}}\right) \\
&= G_2(\boldsymbol{u}_{\widehat{2}})\left(F_1(\boldsymbol{u}_1)\sum_{k=1}^{2}|u_k|^{p_1} + 1 - F_1(\boldsymbol{u}_1)\sum_{k=1}^{2}|u_k|^{p_1}\right) \\
&= G_2(\boldsymbol{u}_{\widehat{2}}).
\end{aligned}$$

The example suggests that the terms from using the chain rule downwards in the tree cancel while the terms from using the chain rule upwards remain. The following proposition states that this is true in general.

Proposition 3 (Determinant of the Jacobian) *Let* $\mathcal{L}$ *be the set of multi-indices of the path from the leaf* u_n *to the root node (excluding the root node) and let the terms* $G_{I,\ell_I}(\boldsymbol{u}_{\widehat{I,\ell_I}})$ *recursively be defined as*

$$G_{I,\ell_I}(\boldsymbol{u}_{\widehat{I,\ell_I}}) = g_{I,\ell_I}(\boldsymbol{u}_{\widehat{I,\ell_I}})^{p_{I,\ell_I}-p_I} = \left(g_I(\boldsymbol{u}_{\widehat{I}})^{p_I} - \sum_{j=1}^{\ell-1} f_{I,j}(\boldsymbol{u}_{I,j})^{p_I}\right)^{\frac{p_{I,\ell_I}-p_I}{p_I}}$$

where each of the functions g_{I,ℓ_I} *computes the value of the* ℓ^{th} *child of a node* I *as a function of its neighbors* $(I,1)$, ..., (I,ℓ_I-1) *and its parent* I *while fixing the value of the root node to one. This is equivalent to computing the value of the node* I *from all coefficients* $\boldsymbol{u}_{\widehat{I}}$ *that are not leaves in the subtree under* I. *Then, the determinant of the Jacobian for an* L_p*-nested function is given by*

$$|\det \mathcal{J}| = r^{n-1}\prod_{L\in\mathcal{L}} G_L(\boldsymbol{u}_{\widehat{L}}).$$

Proof The proof can be found in the Appendix A. ■

Let us illustrate the determinant with two examples:

Example 2 *Consider a normal L_p-norm*

$$f(\boldsymbol{x}) = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}$$

which is obviously also an L_p-nested function. Resolving the equation for the last coordinate of the normalized vector $\boldsymbol{u}$ yields $g_n(\boldsymbol{u}_{\hat{n}}) = u_n = \left(1 - \sum_{i=1}^{n-1} |u_i|^p\right)^{\frac{1}{p}}$. Thus, the term $G_n(\boldsymbol{u}_{\hat{n}})$ is given by $\left(1 - \sum_{i=1}^{n-1} |u_i|^p\right)^{\frac{1-p}{p}}$ which yields a determinant of $|\det \mathcal{J}| = r^{n-1} \left(1 - \sum_{i=1}^{n-1} |u_i|^p\right)^{\frac{1-p}{p}}$. This is exactly the one derived by Gupta and Song (1997).

Example 3 *Consider the introductory example*

$$f(\boldsymbol{x}) = \left(|x_1|^{p_0} + (|x_2|^{p_1} + |x_3|^{p_1})^{\frac{p_0}{p_1}} \right)^{\frac{1}{p_0}}.$$

Normalizing and resolving for the last coordinate yields

$$u_3 = \left((1 - |u_1|^{p_0})^{\frac{p_1}{p_0}} - |u_2|^{p_1} \right)^{\frac{1}{p_1}}$$

and the terms $G_2(\boldsymbol{u}_{\widehat{2}})$ and $G_{2,2}(\boldsymbol{u}_{\widehat{2,2}})$ of the determinant $|\det \mathcal{J}| = r^2 G_2(\boldsymbol{u}_{\widehat{2}}) G_{2,2}(\boldsymbol{u}_{\widehat{2,2}})$ are given by

$$G_2(\boldsymbol{u}_{\widehat{2}}) = (1 - |u_1|^{p_0})^{\frac{p_1 - p_0}{p_0}},$$
$$G_{2,2}(\boldsymbol{u}_{\widehat{2,2}}) = \left((1 - |u_1|^{p_0})^{\frac{p_1}{p_0}} - |u_2|^{p_1} \right)^{\frac{1-p_1}{p_1}}.$$

Note the difference to Example 1 where x_3 was at depth one in the tree while x_3 is at depth two in the current case. For that reason, the determinant of the Jacobian in Example 1 involved only one G-term while it has two G-terms here.

3. L_p-Nested Symmetric and L_p-Nested Uniform Distribution

In this section, we define the L_p-nested symmetric and the L_p-nested uniform distribution and derive their partition functions. In particular, we derive the surface area of an arbitrary L_p-nested unit sphere $\mathbb{L}_f = \{\boldsymbol{x} \in \mathbb{R}^n \mid f(\boldsymbol{x}) = 1\}$ corresponding to an L_p-nested function f. By Equation (5) of Fernandez et al. (1995) every ν-spherical and hence any L_p-nested symmetric density has the form

$$\rho(\boldsymbol{x}) = \frac{\phi(f(\boldsymbol{x}))}{f(\boldsymbol{x})^{n-1} \mathcal{S}_f(1)}, \tag{4}$$

where $\mathcal{S}_f$ is the surface area of $\mathbb{L}_f$ and ϕ is a density on $\mathbb{R}^+$. Thus, we need to compute the surface area of an arbitrary L_p-nested unit sphere to obtain the partition function of Equation (4).

Proposition 4 (Volume and Surface of the L_p-nested Sphere) *Let f be an L_p-nested function and let I be the set of all multi-indices denoting the inner nodes of the tree structure associated with f. The volume $\mathcal{V}_f(R)$ and the surface $\mathcal{S}_f(R)$ of the L_p-nested sphere with radius R are given by*

$$\mathcal{V}_f(R) = \frac{R^n 2^n}{n} \prod_{I \in I} \left(\frac{1}{p_I^{\ell_I - 1}} \prod_{k=1}^{\ell_I - 1} B\left[\frac{\sum_{i=1}^{k} n_{I,k}}{p_I}, \frac{n_{I,k+1}}{p_I} \right] \right) \tag{5}$$

$$= \frac{R^n 2^n}{n} \prod_{I \in I} \frac{\prod_{k=1}^{\ell_I} \Gamma\left[\frac{n_{I,k}}{p_I}\right]}{p_I^{\ell_I - 1} \Gamma\left[\frac{n_I}{p_I}\right]}, \tag{6}$$

$$\mathcal{S}_f(R) = R^{n-1} 2^n \prod_{I \in I} \left(\frac{1}{p_I^{\ell_I - 1}} \prod_{k=1}^{\ell_I - 1} B\left[\frac{\sum_{i=1}^{k} n_{I,k}}{p_I}, \frac{n_{I,k+1}}{p_I} \right] \right) \tag{7}$$

$$= R^{n-1} 2^n \prod_{I \in I} \frac{\prod_{k=1}^{\ell_I} \Gamma\left[\frac{n_{I,k}}{p_I}\right]}{p_I^{\ell_I - 1} \Gamma\left[\frac{n_I}{p_I}\right]} \tag{8}$$

where $B[a,b] = \frac{\Gamma[a]\Gamma[b]}{\Gamma[a+b]}$ denotes the β*-function.*

Proof The proof can be found in the Appendix B. ■

Inserting the surface area in Equation 4, we obtain the general form of an L_p-nested symmetric distribution for any given radial density ϕ.

Corollary 5 (L_p-nested Symmetric Distribution) *Let f be an L_p-nested function and ϕ a density on $\mathbb{R}^+$. The corresponding L_p-nested symmetric distribution is given by*

$$\rho(\boldsymbol{x}) = \frac{\phi(f(\boldsymbol{x}))}{f(\boldsymbol{x})^{n-1} \mathcal{S}_f(1)}$$

$$= \frac{\phi(f(\boldsymbol{x}))}{2^n f(\boldsymbol{x})^{n-1}} \prod_{I \in I} \left(p_I^{\ell_I - 1} \prod_{k=1}^{\ell_I - 1} B\left[\frac{\sum_{i=1}^{k} n_{I,k}}{p_I}, \frac{n_{I,k+1}}{p_I} \right]^{-1} \right). \tag{9}$$

The results of Fernandez et al. (1995) imply that for any ν-spherically symmetric distribution, the radial part is independent of the directional part, that is, r is independent of $\boldsymbol{u}$. The distribution of $\boldsymbol{u}$ is entirely determined by the choice of ν, or by the L_p-nested function f in our case. The distribution of r is determined by the radial density ϕ. Together, an L_p-nested symmetric distribution is determined by both, the L_p-nested function f and the choice of ϕ. From Equation (9), we can see that its density function must be the inverse of the surface area of $\mathbb{L}_f$ times the radial density when transforming (4) into the coordinates of Definition 1 and separating r and $\boldsymbol{u}$ (the factor $f(\boldsymbol{x})^{n-1} = r$ cancels due to the determinant of the Jacobian). For that reason we call the distribution of $\boldsymbol{u}$ *uniform on the L_p-sphere* $\mathbb{L}_f$ in analogy to Song and Gupta (1997). Next, we state its form in terms of the coordinates $\boldsymbol{u}$.

Proposition 6 (L_p-nested Uniform Distribution) *Let f be an L_p-nested function. Let $\mathcal{L}$ be the set of multi-indices on the path from the root node to the leaf corresponding to x_n. The uniform*

distribution on the L_p-nested unit sphere, that is, the set $\mathbb{L}_f = \{\boldsymbol{x} \in \mathbb{R}^n | f(\boldsymbol{x}) = 1\}$ is given by the following density over $u_1, ..., u_{n-1}$

$$\rho(u_1, , ..., u_{n-1}) = \frac{\prod_{L \in \mathcal{L}} G_L(\boldsymbol{u}_{\widehat{L}})}{2^{n-1}} \prod_{I \in I} \left(p_I^{\ell_I - 1} \prod_{k=1}^{\ell_I - 1} B \left[\frac{\sum_{i=1}^{k} n_{I,k}}{p_I}, \frac{n_{I,k+1}}{p_I} \right]^{-1} \right).$$

Proof Since the L_p-nested sphere is a measurable and compact set, the density of the uniform distribution is simply one over the surface area of the L_p-nested unit sphere. The surface $\mathcal{S}_f(1)$ is given by Proposition 4. Transforming $\frac{1}{\mathcal{S}_f(1)}$ into the coordinates of Definition 1 introduces the determinant of the Jacobian from Proposition 3 and an additional factor of 2 since the $(u_1, ..., u_{n-1}) \in \mathbb{R}^{n-1}$ have to account for both half-shells of the L_p-nested unit sphere, that is, to account for the fact that u_n could have been be positive or negative. This yields the expression above. ■

Example 4 *Let us again demonstrate the proposition at the special case where f is an L_p-norm $f(\boldsymbol{x}) = ||\boldsymbol{x}||_p = (\sum_{i=1}^{n} |x_i|^p)^{\frac{1}{p}}$. Using Proposition 4, the surface area is given by*

$$\mathcal{S}_{||\cdot||_p} = 2^n \frac{1}{p_\emptyset^{\ell_\emptyset - 1}} \prod_{k=1}^{\ell_\emptyset - 1} B \left[\frac{\sum_{i=1}^{k} n_k}{p_\emptyset}, \frac{n_{k+1}}{p_\emptyset} \right] = \frac{2^n \Gamma^n \left[\frac{1}{p} \right]}{p^{n-1} \Gamma \left[\frac{n}{p} \right]}.$$

The factor $G_n(\boldsymbol{u}_{\widehat{n}})$ is given by $\left(1 - \sum_{i=1}^{n-1} |u_i|^p\right)^{\frac{1-p}{p}}$ (see the L_p-norm example before), which, after including the factor 2, yields the uniform distribution on the L_p-sphere as defined in Song and Gupta (1997)

$$p(\boldsymbol{u}) = \frac{p^{n-1} \Gamma \left[\frac{n}{p} \right]}{2^{n-1} \Gamma^n \left[\frac{1}{p} \right]} \left(1 - \sum_{i=1}^{n-1} |u_i|^p \right)^{\frac{1-p}{p}}.$$

Example 5 *As a second illustrative example, we consider the uniform density on the L_p-nested unit ball, that is, the set $\{\boldsymbol{x} \in \mathbb{R}^n | \ f(\boldsymbol{x}) \leq 1\}$, and derive its radial distribution ϕ. The density of the uniform distribution on the unit L_p-nested ball does not depend on $\boldsymbol{x}$ and is given by $\rho(\boldsymbol{x}) = 1/\mathcal{V}_f(1)$. Transforming the density into the polar-like coordinates with the determinant from Proposition 3 yields*

$$\frac{1}{\mathcal{V}_f(1)} = \frac{n r^{n-1} \prod_{L \in \mathcal{L}} G_L(\boldsymbol{u}_{\widehat{L}})}{2^{n-1}} \prod_{I \in I} \left(p_I^{\ell_I - 1} \prod_{k=1}^{\ell_I - 1} B \left[\frac{\sum_{i=1}^{k} n_{I,k}}{p_I}, \frac{n_{I,k+1}}{p_I} \right]^{-1} \right).$$

After separating out the uniform distribution on the L_p-nested unit sphere, we obtain the radial distribution

$$\phi(r) = n r^{n-1} \text{ for } 0 < r \leq 1$$

which is a β-distribution with parameters n and 1.

The radial distribution from the preceeding example is of great importance for our sampling scheme derived in Section 6. The idea behind it is the following: First, a sample from a "simple" L_p-nested symmetric distribution is drawn. Since the radial and the uniform component on the L_p-nested unit sphere are statistically independent, we can get a sample from the uniform distribution on the L_p-nested unit sphere by simply normalizing the sample from the simple distribution. Afterwards we can multiply it with a radius drawn from the radial distribution of the L_p-nested symmetric distribution that we actually want to sample from. The role of the simple distribution will be played by the uniform distribution within the L_p-nested unit ball. Sampling from it is basically done by applying the steps in Proposition 4's proof backwards. We lay out the sampling scheme in more detail in Section 6.

4. Marginals

In this section we discuss two types of marginals: First, we demonstrate that, in contrast to L_p-spherically symmetric distributions, marginals of L_p-nested symmetric distributions are not necessarily L_p-nested symmetric again. The second type of marginals we discuss are obtained by collapsing all leaves of a subtree into the value of the subtree's root node. For that case we derive an analytical expression and show that the values of the root node's children follow a special kind of Dirichlet distribution.

Gupta and Song (1997) show that marginals of L_p-spherically symmetric distributions are again L_p-spherically symmetric. This does not hold, however, for L_p-nested symmetric distributions. This can be shown by a simple counterexample. Consider the L_p-nested function

$$f(\boldsymbol{x}) = \left((|x_1|^{p_1} + |x_2|^{p_1})^{\frac{p_\emptyset}{p_1}} + |x_3|^{p_\emptyset} \right)^{\frac{1}{p_\emptyset}}.$$

The uniform distribution inside the L_p-nested ball corresponding to f is given by

$$\rho(\boldsymbol{x}) = \frac{n p_1 p_\emptyset \Gamma\left[\frac{2}{p_1}\right] \Gamma\left[\frac{3}{p_\emptyset}\right]}{2^3 \Gamma^2\left[\frac{1}{p_1}\right] \Gamma\left[\frac{2}{p_0}\right] \Gamma\left[\frac{1}{p_0}\right]}.$$

The marginal $\rho(x_1, x_3)$ is given by

$$\rho(x_1, x_3) = \frac{n p_1 p_\emptyset \Gamma\left[\frac{2}{p_1}\right] \Gamma\left[\frac{3}{p_\emptyset}\right]}{2^3 \Gamma^2\left[\frac{1}{p_1}\right] \Gamma\left[\frac{2}{p_0}\right] \Gamma\left[\frac{1}{p_0}\right]} \left((1 - |x_3|^{p_\emptyset})^{\frac{p_1}{p_\emptyset}} - |x_1|^{p_1} \right)^{\frac{1}{p_1}}.$$

This marginal is not L_p-spherically symmetric. Since any L_p-nested symmetric distribution in two dimensions must be L_p-spherically symmetric, it cannot be L_p-nested symmetric as well. Figure 3 shows a scatter plot of the marginal distribution. Besides the fact that the marginals are not contained in the family of L_p-nested symmetric distributions, it is also hard to derive a general form for them. This is not surprising given that the general form of marginals for L_p-spherically symmetric distributions involves an integral that cannot be solved analytically in general and is therefore not very useful in practice (Gupta and Song, 1997). For that reason we cannot expect marginals of L_p-nested symmetric distributions to have a simple form.

In contrast to single marginals, it is possible to specify the joint distribution of leaves and inner nodes of an L_p-nested tree if all descendants of their inner nodes in question have been integrated

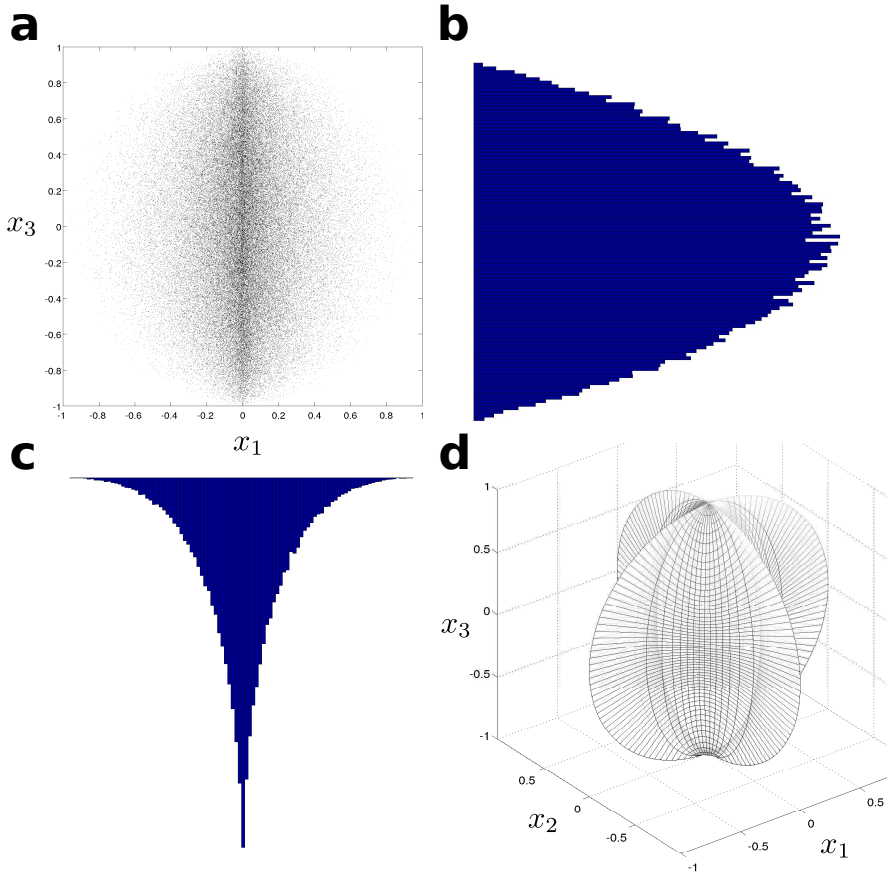


Figure 3: Marginals of L_p-nested symmetric distributions are not necessarily L_p-nested symmetric: Figure (**a**) shows a scatter plot of the (x_1, x_2)-marginal of the counterexample in the text with $p_\emptyset = 2$ and $p_1 = \frac{1}{2}$. Figure (**d**) displays the corresponding L_p-nested sphere. (**b-c**) show the univariate marginals for the scatter plot. Since any two-dimensional L_p-nested symmetric distribution must be L_p-spherically symmetric, the marginals should be identical. This is clearly not the case. Thus, (**a**) is not L_p-nested symmetric.

out. For the simple function above (the same that has been used in Example 1), the joint distribution of x_3 and $v_1 = \|(x_1, x_2)^\top\|_{p_1}$ would be an example of such a marginal. Since marginalization affects

the L_p-nested tree vertically, we call this type of marginals *layer marginals*. In the following, we present their general form.

From the form of a general L_p-nested function and the corresponding symmetric distribution, one might think that the layer marginals are L_p-nested symmetric again. However, this is not the case since the distribution over the L_p-nested unit sphere would deviate from the uniform distribution in most cases if the distribution of its children were L_p-spherically symmetric.

Proposition 7 *Let f be an L_p-nested function. Suppose we integrate out complete subtrees from the tree associated with f, that is, we transform subtrees into radial times uniform variables and integrate out the latter. Let $\mathcal{J}$ be the set of multi-indices of those nodes that have become new leaves, that is, whose subtrees have been removed, and let n_J be the number of leaves (in the original tree) in the subtree under the node J. Let $\boldsymbol{x}_{\tilde{\mathcal{J}}} \in \mathbb{R}^m$ denote those coefficients of $\boldsymbol{x}$ that are still part of that smaller tree and let $\boldsymbol{v}_{\mathcal{J}}$ denote the vector of inner nodes that became new leaves. The joint distribution of $\boldsymbol{x}_{\tilde{\mathcal{J}}}$ and $\boldsymbol{v}_{\mathcal{J}}$ is given by*

$$\rho(\boldsymbol{x}_{\tilde{\mathcal{J}}}, \boldsymbol{v}_{\mathcal{J}}) = \frac{\phi(f(\boldsymbol{x}_{\tilde{\mathcal{J}}}, \boldsymbol{v}_{\mathcal{J}}))}{S_f(f(\boldsymbol{x}_{\tilde{\mathcal{J}}}, \boldsymbol{v}_{\mathcal{J}}))} \prod_{J \in \mathcal{J}} v_J^{n_J - 1}. \tag{10}$$

Proof The proof can be found in the Appendix C. ■

Equation (10) has an interesting special case when considering the joint distribution of the root node's children.

Corollary 8 *The children of the root node $\boldsymbol{v}_{1:\ell_\emptyset} = (v_1, ..., v_{\ell_\emptyset})^\top$ follow the distribution*

$$\rho(\boldsymbol{v}_{1:\ell_\emptyset}) = \frac{p_\emptyset^{\ell_\emptyset - 1} \Gamma\left[\frac{n}{p_\emptyset}\right]}{f(v_1, ..., v_{\ell_\emptyset})^{n-1} 2^m \prod_{k=1}^{\ell_\emptyset} \Gamma\left[\frac{n_k}{p_\emptyset}\right]} \phi(f(v_1, ..., v_{\ell_\emptyset})) \prod_{i=1}^{\ell_\emptyset} v_i^{n_i - 1}$$

where $m \leq \ell_\emptyset$ is the number of leaves directly attached to the root node. In particular, $\boldsymbol{v}_{1:\ell_\emptyset}$ can be written as the product RU, where R is the L_p-nested radius and the single $|U_i|^{p_\emptyset}$ are Dirichlet distributed, that is, $(|U_1|^{p_\emptyset}, ..., |U_{\ell_\emptyset}|^{p_\emptyset}) \sim Dir\left[\frac{n_1}{p_\emptyset}, ..., \frac{n_{\ell_\emptyset}}{p_\emptyset}\right]$.

Proof The joint distribution is simply the application of Proposition (7). Note that $f(v_1, ..., v_{\ell_\emptyset}) = ||\boldsymbol{v}_{1:\ell_\emptyset}||_{p_\emptyset}$. Applying the pointwise transformation $s_i = |u_i|^{p_\emptyset}$ yields

$$(|U_1|^{p_\emptyset}, ..., |U_{\ell_\emptyset - 1}|^{p_\emptyset}) \sim \mathrm{Dir}\left[\frac{n_1}{p_\emptyset}, ..., \frac{n_{\ell_\emptyset}}{p_\emptyset}\right].$$

■

The Corollary shows that the values $f_I(\boldsymbol{x}_I)$ at inner nodes I, in particular the ones directly below the root node, deviate considerably from L_p-spherical symmetry. If they were L_p-spherically symmetric, the $|U_i|^p$ should follow a Dirichlet distribution with parameters $\alpha_i = \frac{1}{p}$ as has been already shown by Song and Gupta (1997). The Corollary is a generalization of their result.

We can use the Corollary to prove an interesting fact about L_p-nested symmetric distributions: The only factorial L_p-nested symmetric distribution must be L_p-spherically symmetric.

Proposition 9 *Let $\boldsymbol{x}$ be L_p-nested symmetric distributed with independent marginals. Then $\boldsymbol{x}$ is L_p-spherically symmetric distributed. In particular, $\boldsymbol{x}$ follows a p-generalized Normal distribution.*

Proof The proof can be found in the Appendix D. ■

One immediate implication of Proposition 9 is that there is no factorial probability model corresponding to mixed norm regularizers which have the form $\sum_{i=1}^{k} \|\boldsymbol{x}_{I_k}\|_p^q$ where the index sets I_k form a partition of the dimensions $1,...,n$ (see, e.g., Zhao et al., 2008; Yuan and Lin, 2006; Kowalski et al., 2008). Many machine learning algorithms are equivalent to minimizing the sum of a regularizer $R(\boldsymbol{w})$ and a loss function $L(\boldsymbol{w},\boldsymbol{x}_1,...,\boldsymbol{x}_m)$ over the coefficient vector $\boldsymbol{w}$. If the $\exp(-R(\boldsymbol{w}))$ and $\exp(-L(\boldsymbol{w},\boldsymbol{x}_1,...,\boldsymbol{x}_m))$ correspond to normalizeable density models, the minimizing solution of the objective function can be seen as the maximum a posteriori (MAP) estimate of the posterior $p(\boldsymbol{w}|\boldsymbol{x}_1,...,\boldsymbol{x}_m) \propto p(\boldsymbol{w}) \cdot p(\boldsymbol{x}_1,...,\boldsymbol{x}_m|\boldsymbol{w}) = \exp(-R(\boldsymbol{w})) \cdot \exp(-L(\boldsymbol{w},\boldsymbol{x}_1,...,\boldsymbol{x}_m))$. In that sense, the regularizer naturally corresponds to the prior and the loss function corresponds to the likelihood. Very often, regularizers are specified as a norm over the coefficient vector $\boldsymbol{w}$ which in turn correspond to certain priors. For example, in Ridge regression (Hoerl, 1962) the coefficients are regularized via $\|\boldsymbol{w}\|_2^2$ which corresponds to a factorial zero mean Gaussian prior on $\boldsymbol{w}$. The L_1-norm $\|\boldsymbol{w}\|_1$ in the LASSO estimator (Tibshirani, 1996), again, is equivalent to a factorial Laplacian prior on $\boldsymbol{w}$. Like in these two examples, regularizers often correspond to a *factorial* prior.

Mixed norm regularizers naturally correspond to L_p-nested symmetric distributions. Proposition 9 shows that there is no factorial prior that corresponds to such a regularizer. In particular, it implies that the prior cannot be factorial between groups and coefficients at the same time. This means that those regularizers implicitly assume statistical dependencies between the coefficient variables. Interestingly, for $q = 1$ and $p = 2$ the intuition behind these regularizers is exactly that whole groups I_k get switched on at once, but the groups are sparse. The Proposition shows that this might not only be due to sparseness but also due to statistical dependencies between the coefficients within one group. The L_p-nested symmetric distribution which implements independence between groups will be further discussed below as a generalization of the p-generalized Normal (see Section 8). Note that the marginals can be independent if the regularizer is of the form $\sum_{i=1}^{k} \|\boldsymbol{x}_{I_k}\|_p^p$. However, in this case $p = q$ and the L_p-nested function collapses to a simple L_p-norm which means that the regularizer is not mixed norm.

5. Maximum Likelihood Estimation of L_p-Nested Symmetric Distributions

In this section, we describe procedures for maximum likelihood fitting of L_p-nested symmetric distributions on data. We provide a toolbox online for fitting L_p-spherically symmetric and L_p-nested symmetric distributions to data. The toolbox can be downloaded at `http://www.kyb.tuebingen.mpg.de/bethge/code/`.

Depending on which parameters are to be estimated, the complexity of fitting an L_p-nested symmetric distribution varies. We start with the simplest case and later continue with more complex ones. Throughout this subsection, we assume that the model has the form $p(\boldsymbol{x}) = \rho(W\boldsymbol{x}) \cdot |\det W| = \frac{\phi(W\boldsymbol{x})}{f(W\boldsymbol{x})^{n-1}S_f(1)} \cdot |\det W|$ where $W \in \mathbb{R}^{n\times n}$ is a complete whitening matrix. This means that given any whitening matrix W_0, the freedom in fitting W is to estimate an orthonormal matrix $Q \in SO(n)$ such that $W = QW_0$. This is analogous to the case of elliptically contoured distributions where the

distributions can be endowed with 2nd-order correlations via W. In the following, we ignore the determinant of W since data points can always be rescaled such that $\det W = 1$.

The simplest case is to fit the parameters of the radial distribution when the tree structure, the values of the p_I, and W are fixed. Due to the special form of L_p-nested symmetric distributions (4), it then suffices to carry out maximum likelihood estimation on the radial component only, which renders maximum likelihood estimation efficient and robust. This is because the only remaining parameters are the parameters $\boldsymbol{\vartheta}$ of the radial distribution and, therefore,

$$\begin{aligned}\operatorname{argmax}_{\boldsymbol{\vartheta}} \log \rho(W\boldsymbol{x}|\boldsymbol{\vartheta}) &= \operatorname{argmax}_{\boldsymbol{\vartheta}} \left(-\log \mathcal{S}_f(f(W\boldsymbol{x})) + \log \phi(f(W\boldsymbol{x})|\boldsymbol{\vartheta})\right) \\ &= \operatorname{argmax}_{\boldsymbol{\vartheta}} \log \phi(f(W\boldsymbol{x})|\boldsymbol{\vartheta}).\end{aligned}$$

In a slightly more complex case, when only the tree structure and W are fixed, the values of the p_I, $I \in \mathcal{I}$ and $\boldsymbol{\vartheta}$ can be jointly estimated via gradient ascent on the log-likelihood. The gradient for a single data point $\boldsymbol{x}$ with respect to the vector $\boldsymbol{p}$ that holds all p_I for all $I \in \mathcal{I}$ is given by

$$\nabla_{\boldsymbol{p}} \log \rho(W\boldsymbol{x}) = \frac{d}{dr} \log \phi(f(W\boldsymbol{x})) \cdot \nabla_{\boldsymbol{p}} f(W\boldsymbol{x}) - \frac{(n-1)}{f(W\boldsymbol{x})} \nabla_{\boldsymbol{p}} f(W\boldsymbol{x}) - \nabla_{\boldsymbol{p}} \log \mathcal{S}_f(1).$$

For i.i.d. data points $\boldsymbol{x}_i$ the joint gradient is given by the sum over the gradients for the single data points. Each of them involves the gradient of f as well as the gradient of the log-surface area of $\mathbb{L}_f$ with respect to $\boldsymbol{p}$, which can be computed via the recursive equations

$$\frac{\partial}{\partial p_J} v_I = \begin{cases} 0 & \text{if } I \text{ is not a prefix of } J \\ v_I^{1-p_I} v_{I,k}^{p_I - 1} \cdot \frac{\partial}{\partial p_J} v_{I,k} & \text{if } I \text{ is a prefix of } J \\ \frac{v_J}{p_J} \left(v_J^{-p_J} \sum_{k=1}^{\ell_J} v_{J,k}^{p_J} \cdot \log v_{J,k} - \log v_J \right) & \text{if } J = I \end{cases}$$

and

$$\begin{aligned}\frac{\partial}{\partial p_J} \log \mathcal{S}_f(1) = &- \frac{\ell_J - 1}{p_J} + \sum_{k=1}^{\ell_J - 1} \Psi\left[\frac{\sum_{i=1}^{k+1} n_{J,k}}{p_J} \right] \frac{\sum_{i=1}^{k+1} n_{J,k}}{p_J^2} \\ &- \sum_{k=1}^{\ell_J - 1} \Psi\left[\frac{\sum_{i=1}^{k} n_{J,k}}{p_J} \right] \frac{\sum_{i=1}^{k} n_{J,k}}{p_J^2} - \sum_{k=1}^{\ell_J - 1} \Psi\left[\frac{n_{J,k+1}}{p_J} \right] \frac{n_{J,k+1}}{p_J^2},\end{aligned}$$

where $\Psi[t] = \frac{d}{dt} \log \Gamma[t]$ denotes the digamma function. When performing the gradient ascent, one needs to set $\mathbf{0}$ as a lower bound for $\boldsymbol{p}$. Note that, in general, this optimization might be a highly non-convex problem.

On the next level of complexity, only the tree structure is fixed, and W can be estimated along with the other parameters by joint optimization of the log-likelihood with respect to $\boldsymbol{p}$, $\boldsymbol{\vartheta}$ and W. Certainly, this optimization problem is also not convex in general. Usually, it is numerically more robust to whiten the data first with some whitening matrix W_0 and perform a gradient ascent on the special orthogonal group $SO(n)$ with respect to Q for optimizing $W = QW_0$. Given the gradient $\nabla_W \log \rho(W\boldsymbol{x})$ of the log-likelihood, the optimization can be carried out by performing line searches along geodesics as proposed by Edelman et al. (1999) (see also Absil et al. (2007)) or by projecting $\nabla_W \log \rho(W\boldsymbol{x})$ on the tangent space $T_W SO(n))$ and performing a line search along $SO(n)$ in that direction as proposed by Manton (2002).

The general form of the gradient to be used in such an optimization scheme can be defined as

$$\begin{aligned}&\nabla_W \log \rho(W\boldsymbol{x})\\ =&\nabla_W \left(-(n-1)\cdot \log f(W\boldsymbol{x}) + \log \phi(f(W\boldsymbol{x}))\right)\\ =&-\frac{(n-1)}{f(W\boldsymbol{x})}\cdot \nabla_{\boldsymbol{y}} f(W\boldsymbol{x})\cdot \boldsymbol{x}^\top + \frac{d\log\phi(r)}{dr}(f(W\boldsymbol{x}))\cdot \nabla_{\boldsymbol{y}} f(W\boldsymbol{x})\cdot \boldsymbol{x}^\top,\end{aligned}$$

where the derivatives of f with respect to $\boldsymbol{y}$ are defined by recursive equations

$$\frac{\partial}{\partial y_i} v_I = \begin{cases} 0 & \text{if } i \notin I \\ \operatorname{sgn} y_i & \text{if } v_{I,k} = |y_i| \\ v_I^{1-p_I}\cdot v_{I,k}^{p_I-1}\cdot \frac{\partial}{\partial y_i} v_{I,k} & \text{for } i \in I,k. \end{cases}$$

Note, that f might not be differentiable at $\boldsymbol{y}=0$. However, we can always define a sub-derivative at zero, which is zero for $p_I \neq 1$ and $[-1,1]$ for $p_I = 1$. Again, the gradient for i.i.d. data points $\boldsymbol{x}_i$ is given by the sum over the single gradients.

Finally, the question arises whether it is possible to estimate the tree structure from data as well. A simple heuristic would be to start with a very large tree, for example, a full binary tree, and to prune out inner nodes for which the parents and the children have sufficiently similar values for their p_I. The intuition behind this is that if they were exactly equal, they would cancel in the L_p-nested function. This heuristic is certainly sub-optimal. Firstly, the optimization will be time consuming since there can be about as many p_I as there are leaves in the L_p-nested tree (a full binary tree on n dimensions will have $n-1$ inner nodes) and due to the repeated optimization after the pruning steps. Secondly, the heuristic does not cover all possible trees on n leaves. For example, if two leaves are separated by the root node in the original full binary tree, there is no way to prune out inner nodes such that the path between those two nodes will not contain the root node anymore.

The computational complexity for the estimation of all other parameters despite the tree structure is difficult to assess in general because they depend, for example, on the particular radial distribution used. While the maximum likelihood estimation of a simple log-Normal distribution only involves the computation of a mean and a variance which are in $O(m)$ for m data points, a mixture of log-Normal distributions already requires an EM algorithm which is computationally more expensive. Additionally, the time it takes to optimize the likelihood depends on the starting point as well as the convergence rate, and we neither have results about the convergence rate nor is it possible to make problem independent statements about a good initialization of the parameters. For this reason we state only the computational complexity of single steps involved in the optimization.

Computation of the gradient $\nabla_{\boldsymbol{p}} \log \rho(W\boldsymbol{x})$ involves the derivative of the radial distribution, the computation of the gradients $\nabla_{\boldsymbol{p}} f(W\boldsymbol{x})$ and $\nabla_{\boldsymbol{p}} \mathcal{S}_f(1)$. Assuming that the derivative of the radial distribution can be computed in $O(1)$ for each single data point, the costly steps are the other two gradients. Computing $\nabla_{\boldsymbol{p}} f(W\boldsymbol{x})$ basically involves visiting each node of the tree once and performing a constant number of operations for the local derivatives. Since every inner node in an L_p-nested tree must have at least two children, the worst case would be a full binary tree which has $2n-1$ nodes and leaves. Therefore, the gradient can be computed in $O(nm)$ for m data points. For similar reasons, $f(W\boldsymbol{x})$, $\nabla_{\boldsymbol{p}} \log \mathcal{S}_f(1)$, and the evaluation of the likelihood can also be computed in $O(nm)$. This means that each step in the optimization of $\boldsymbol{p}$ can be done $O(nm)$ plus the computational costs for the line search in the gradient ascent. When optimizing for $W = QW_0$ as well, the computational

costs per step increase to $O(n^3 + n^2 m)$ since m data points have to be multiplied with W at each iteration (requiring $O(n^2 m)$ steps), and the line search involves projecting Q back onto $SO(n)$ which requires an inverse matrix square root or a similar computation in $O(n^3)$.

For comparison, each step of fast ICA (Hyvärinen and O., 1997) for a complete demixing matrix takes $O(n^2 m)$ when using hierarchical orthogonalization and $O(n^2 m + n^3)$ for symmetric orthogonalization. The same applies to fitting an ISA model (Hyvärinen and Hoyer, 2000; Hyvärinen and Köster, 2006, 2007). A Gaussian Scale Mixture (GSM) model does not need to estimate another orthogonal rotation Q because it belongs to the class of spherically symmetric distributions and is, therefore, invariant under transformations from $SO(n)$ (Wainwright and Simoncelli, 2000). Therefore, fitting a GSM corresponds to estimating the parameters of the scale distribution which is $O(nm)$ in the best case but might be costlier depending on the choice of the scale distribution.

6. Sampling from L_p-Nested Symmetric Distributions

In this section, we derive a sampling scheme for arbitrary L_p-nested symmetric distributions which can for example be used for solving integrals when using L_p-nested symmetric distributions for Bayesian learning. Exact sampling from an arbitrary L_p-nested symmetric distribution is in fact straightforward due to the following observation: Since the radial and the uniform component are independent, normalizing a sample from any L_p-nested symmetric distribution to f-length one yields samples from the uniform distribution on the L_p-nested unit sphere. By multiplying those uniform samples with new samples from another radial distribution, one obtains samples from another L_p-nested symmetric distribution. Therefore, for each L_p-nested function f, a single L_p-nested symmetric distribution which can be easily sampled from is enough. Sampling from all other L_p-nested symmetric distributions with respect to f is then straightforward due to the method we just described. Gupta and Song (1997) sample from the p-generalized Normal distribution since it has independent marginals which makes sampling straightforward. Due to Proposition 9, no such factorial L_p-nested symmetric distribution exists. Therefore, a sampling scheme like that for L_p-spherically symmetric distributions is not applicable. Instead we choose to sample from the uniform distribution inside the L_p-nested unit ball for which we already computed the radial distribution in Example 5. The distribution has the form $\rho(\boldsymbol{x}) = \frac{1}{\mathcal{V}_f(1)}$. In order to sample from that distribution, we will first only consider the uniform distribution in the positive quadrant of the unit L_p-nested ball which has the form $\rho(\boldsymbol{x}) = \frac{2^n}{\mathcal{V}_f(1)}$. Samples from the uniform distributions inside the whole ball can be obtained by multiplying each coordinate of a sample with independent samples from the uniform distribution over $\{-1, 1\}$.

The idea of the sampling scheme for the uniform distribution inside the L_p-nested unit ball is based on the computation of the volume of the L_p-nested unit ball in Proposition 4. The basic mechanism underlying the sampling scheme below is to apply the steps of the proof backwards, which is based on the following idea: The volume of the L_p-unit ball can be computed by computing its volume on the positive quadrant only and multiplying the result with 2^n afterwards. The key is now to not transform the whole integral into radial and uniform coordinates at once, but successively upwards in the tree. We will demonstrate this through a brief example which also should make the sampling scheme below more intuitive. Consider the L_p-nested function

$$f(\boldsymbol{x}) = \left(|x_1|^{p_0} + \left(|x_2|^{p_1} + |x_3|^{p_1} \right)^{\frac{p_0}{p_1}} \right)^{\frac{1}{p_0}}.$$

To solve the integral

$$\int_{\{\boldsymbol{x}: f(\boldsymbol{x})\leq 1 \,\&\, \boldsymbol{x}\in\mathbb{R}^n_+\}} d\boldsymbol{x},$$

we first transform x_2 and x_3 into radial and uniform coordinates only. According to Proposition 3 the determinant of the mapping $(x_2, x_3) \mapsto (v_1, \tilde{u}) = (\|\boldsymbol{x}_{2:3}\|_{p_1}, \boldsymbol{x}_{2:3}/\|\boldsymbol{x}_{2:3}\|_{p_1})$ is given by $v_1(1-\tilde{u}^{p_1})^{\frac{1-p_1}{p_1}}$. Therefore the integral transforms into

$$\int_{\{\boldsymbol{x}: f(\boldsymbol{x})\leq 1 \,\&\, \boldsymbol{x}\in\mathbb{R}^n_+\}} d\boldsymbol{x} = \int_{\{v_1,x_1: f(x_1,v_1)\leq 1 \,\&\, x_1,v_1\in\mathbb{R}_+\}} \int\int v_1(1-\tilde{u}^{p_1})^{\frac{1-p_1}{p_1}} dx_1 dv_1 d\tilde{u}.$$

Now we can separate the integrals over x_1 and v_1, and the integral over $\tilde{u}$, since the boundary of the outer integral does only depend on v_1 and not on $\tilde{u}$:

$$\int_{\{\boldsymbol{x}: f(\boldsymbol{x})\leq 1 \,\&\, \boldsymbol{x}\in\mathbb{R}^n_+\}} d\boldsymbol{x} = \int (1-\tilde{u}^{p_1})^{\frac{1-p_1}{p_1}} d\tilde{u} \cdot \int_{\{v_1,x_1: f(x_1,v_1)\leq 1 \,\&\, x_1,v_1\in\mathbb{R}_+\}} \int v_1 dx_1 dv_1.$$

The value of the first integral is known explicitly since the integrand equals the uniform distribution on the $\|\cdot\|_{p_1}$-unit sphere. Therefore, the value of the integral must be its normalization constant which we can get using Proposition 4:

$$\int (1-\tilde{u}^{p_1})^{\frac{1-p_1}{p_1}} d\tilde{u} = \frac{\Gamma\left[\frac{1}{p_1}\right]^2 \cdot p_1}{\Gamma\left[\frac{2}{p_1}\right]}.$$

An alternative way to arrive at this result is to use the transformation $s = \tilde{u}^{p_1}$ and to notice that the integrand is a Dirichlet distribution with parameters $\alpha_i = \frac{1}{p_1}$. The normalization constant of the Dirichlet distribution and the constants from the determinant of the Jacobian of the transformation yield the same result.

To compute the remaining integral, the same method can be applied again yielding the volume of the L_p-nested unit ball. The important part for the sampling scheme, however, is not the volume itself but the fact that the intermediate results in this integration process equal certain distributions. As shown in Example 5 the radial distribution of the uniform distribution on the unit ball is $\beta[n,1]$, and as just indicated by the example above, the intermediate results can be seen as transformed variables from a Dirichlet distribution. This fact holds true even for more complex L_p-nested unit balls although the parameters of the Dirichlet distribution can be slightly different. Reversing the steps leads us to the following sampling scheme. First, we sample from the β-distribution which gives us the radius $v_\emptyset$ on the root node. Then we sample from the appropriate Dirichlet distribution and exponentiate the samples by $\frac{1}{p_\emptyset}$ which transforms them into the analogs of the variable u from above. Scaling the result with the sample $v_\emptyset$ yields the values of the root node's children, that is, the analogs of x_1 and v_1. Those are the new radii for the levels below them where we simply repeat this procedure with the appropriate Dirichlet distributions and exponents. The single steps are summarized in Algorithm 1.

The computational complexity of the sampling scheme is $O(n)$. Since the sampling procedure is like expanding the tree node by node starting with the root, the number of inner nodes and leaves is the total number of samples that have to be drawn from Dirichlet distributions. Every node in an L_p-nested tree must at least have two children. Therefore, the maximal number of inner nodes and leaves is $2n-1$ for a full binary tree. Since sampling from a Dirichlet distribution is also in $O(n)$, the total computational complexity for one sample is in $O(n)$.

Algorithm 1 Exact sampling algorithm for L_p-nested symmetric distributions

Input: The radial distribution ϕ of an L_p-nested symmetric distribution ρ for the L_p-nested function f.
Output: Sample $\boldsymbol{x}$ from ρ.
Algorithm

1. Sample $v_\emptyset$ from a beta distribution $\beta[n,1]$.
2. For each inner node I of the tree associated with f, sample the auxiliary variable $\boldsymbol{s}_I$ from a Dirichlet distribution $\mathrm{Dir}\left[\frac{n_{I,1}}{p_I},\ldots,\frac{n_{I,\ell_I}}{p_I}\right]$ where $n_{I,k}$ are the number of leaves in the subtree under node I,k. Obtain coordinates on the L_p-nested sphere within the positive orthant by $\boldsymbol{s}_I \mapsto \boldsymbol{s}_I^{\frac{1}{p_I}} = \tilde{\boldsymbol{u}}_I$ (the exponentiation is taken component-wise).
3. Transform these samples to Cartesian coordinates by $v_I \cdot \tilde{\boldsymbol{u}}_I = \boldsymbol{v}_{I,1:\ell_I}$ for each inner node, starting from the root node and descending to lower layers. The components of $\boldsymbol{v}_{I,1:\ell_I}$ constitute the radii for the layer direct below them. If $I = \emptyset$, the radius had been sampled in step 1.
4. Once the two previous steps have been repeated until no inner node is left, we have a sample $\boldsymbol{x}$ from the uniform distribution in the positive quadrant. Normalize $\boldsymbol{x}$ to get a uniform sample from the sphere $\boldsymbol{u} = \frac{\boldsymbol{x}}{f(\boldsymbol{x})}$.
5. Sample a new radius $\tilde{v}_\emptyset$ from the radial distribution of the target radial distribution ϕ and obtain the sample via $\tilde{\boldsymbol{x}} = \tilde{v}_\emptyset \cdot \boldsymbol{u}$.
6. Multiply each entry x_i of $\tilde{\boldsymbol{x}}$ by an independent sample z_i from the uniform distribution over $\{-1,1\}$.

7. Robust Bayesian Inference of the Location

For L_p-spherically symmetric distributions with a location and a scale parameter

$$p(\boldsymbol{x}|\boldsymbol{\mu},\tau) = \tau^n \rho(\|\tau(\boldsymbol{x}-\boldsymbol{\mu})\|_p),$$

Osiewalski and Steel (1993) derived the posterior in closed form using a prior $p(\boldsymbol{\mu},\tau) = p(\mu)\cdot c\cdot\tau^{-1}$, and showed that $p(\boldsymbol{x},\boldsymbol{\mu})$ does not depend on the radial distribution ϕ, that is, the particular type of L_p-spherically symmetric distributions used for a fixed p. The prior on τ corresponds to an improper Jeffrey's prior which is used to represent lack of prior knowledge on the scale. The main implication of their result is that Bayesian inference of the location $\boldsymbol{\mu}$ under that prior on the scale does not depend on the particular type of L_p-spherically symmetric distribution used for inference. This means that under the assumption of an L_p-spherically symmetric distributed variable, for a fixed p, one has to know the exact form of the distribution in order to compute the location parameter.

It is straightforward to generalize their result to L_p-nested symmetric distributions and, hence, making it applicable to a larger class of distributions. Note that when using any L_p-nested symmetric distribution, introducing a scale and a location via the transformation $\boldsymbol{x} \mapsto \tau(\boldsymbol{x}-\boldsymbol{\mu})$ introduces a factor of τ^n in front of the distribution.

Proposition 10 *For fixed values $p_\emptyset, p_1, \ldots$ and two independent priors $p(\boldsymbol{\mu},\tau) = p(\boldsymbol{\mu}) \cdot c\tau^{-1}$ of the location μ and the scale τ where the prior on τ is an improper Jeffrey's prior, the joint distribution $p(\boldsymbol{x},\boldsymbol{\mu})$ is given by*

$$p(\boldsymbol{x},\boldsymbol{\mu}) = f(\boldsymbol{x}-\boldsymbol{\mu})^{-n} \cdot c \cdot \frac{1}{Z} \cdot p(\boldsymbol{\mu}),$$

where Z denotes the normalization constant of the L_p-nested uniform distribution.

Proof Given any L_p-nested symmetric distribution $\rho(f(\boldsymbol{x}))$, the transformation into the polar-like coordinates yields the following relation

$$1 = \int \rho(f(\boldsymbol{x}))d\boldsymbol{x} = \int\int \prod_{L\in\mathcal{L}} G_L(\boldsymbol{u}_{\widehat{L}}) r^{n-1}\rho(r)dr d\boldsymbol{u} = \int \prod_{L\in\mathcal{L}} G_L(\boldsymbol{u}_{\widehat{L}})d\boldsymbol{u} \cdot \int r^{n-1}\rho(r)dr.$$

Since $\prod_{L\in\mathcal{L}} G_L(\boldsymbol{u}_{\widehat{L}})$ is the unnormalized uniform distribution on the L_p-nested unit sphere, the integral must equal the normalization constant which we denote with Z for brevity (see Proposition 6 for an explicit expression). This implies that ρ has to fulfill

$$\frac{1}{Z} = \int r^{n-1}\rho(r)dr.$$

Writing down the joint distribution of $\boldsymbol{x},\boldsymbol{\mu}$ and τ, and using the substitution $s = \tau f(\boldsymbol{x}-\boldsymbol{\mu})$ we obtain

$$\begin{aligned} p(\boldsymbol{x},\boldsymbol{\mu}) &= \int \tau^n \rho(f(\tau(\boldsymbol{x}-\boldsymbol{\mu}))) \cdot c\tau^{-1} \cdot p(\boldsymbol{\mu})d\tau \\ &= \int s^{n-1}\rho(s) \cdot c \cdot p(\boldsymbol{\mu}) f(\boldsymbol{x}-\boldsymbol{\mu})^{-n} ds \\ &= f(\boldsymbol{x}-\boldsymbol{\mu})^{-n} \cdot c \cdot \frac{1}{Z} \cdot p(\boldsymbol{\mu}). \end{aligned}$$

■

Note that this result could easily be extended to ν-spherical distributions. However, in this case the normalization constant Z cannot be computed for most cases and, therefore, the posterior would not be known explicitly.

8. Relations to ICA, ISA and Over-Complete Linear Models

In this section, we explain the relations among L_p-spherically symmetric, L_p-nested symmetric, ICA and ISA models. For a general overview see Figure 4.

The density model underlying ICA models the joint distribution of the signal $\boldsymbol{x}$ as a linear superposition of statistically independent hidden sources $A\boldsymbol{y} = \boldsymbol{x}$ or $\boldsymbol{y} = W\boldsymbol{x}$. If the marginals of the hidden sources belong to the exponential power family, we obtain the p-generalized Normal which is a subset of the L_p-spherically symmetric class. The p-generalized Normal distribution $p(\boldsymbol{y}) \propto \exp(-\tau\|\boldsymbol{y}\|_p^p)$ is a density model that is often used in ICA algorithms for kurtotic natural signals like images and sound by optimizing a demixing matrix W w.r.t. to the model $p(\boldsymbol{y}) \propto \exp(-\tau\|W\boldsymbol{x}\|_p^p)$ (Lee and Lewicki, 2000; Zhang et al., 2004; Lewicki, 2002). It can be

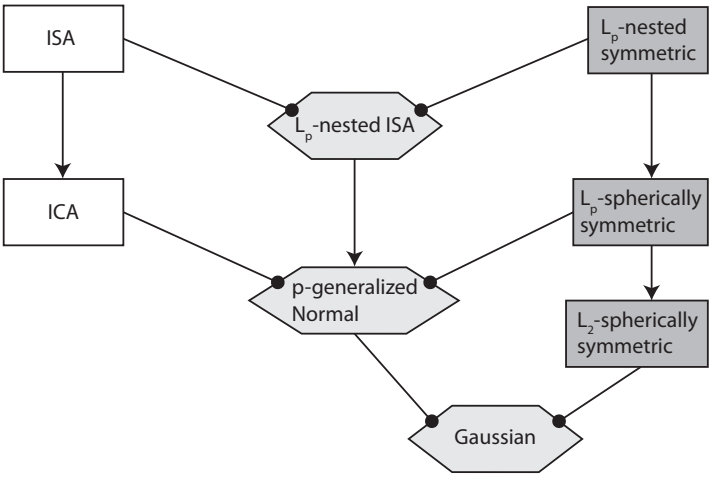


Figure 4: Relations between the different classes of distributions: Arrows indicate that the child class is a specialization (subset) of the parent class. Polygon-shaped classes are intersections of those parent classes which are connected via edges with round arrow-heads. For one-dimensional subspaces ISA is a superclass of ICA. All classes belonging to ISA are colored white or light gray. L_p-nested symmetric distributions are a superclass of L_p-spherically symmetric distributions. All L_p-nested symmetric models are colored dark or light gray. L_p-nested ISA models live in the intersection of L_p-nested symmetric distributions and ISA models. Those L_p-nested ISA models that are L_p-spherically symmetric are also ICA models: This is the class of p-generalized Normal distributions. If p is fixed to two, one obtains the L_2-spherically symmetric distributions. The only class of distributions in the intersection between spherically symmetric distributions and ICA models is the Gaussian.

shown that the p-generalized Normal is the only factorial model in the class of L_p-spherically symmetric models (Sinz et al., 2009a), and, by Proposition 9, also the only factorial L_p-nested symmetric distribution.

An important generalization of ICA is the independent subspace analysis (ISA) proposed by Hyvärinen and Hoyer (2000) and by Hyvärinen and Köster (2007) who used L_p-spherically symmetric distributions to model the single subspaces, that is, each ρ_k below was L_p-spherically symmetric. Like in ICA, ISA models the hidden sources of the signal as a product of multivariate distributions:

$$\rho(\boldsymbol{y}) = \prod_{k=1}^{K} \rho_k(\boldsymbol{y}_{I_k}).$$

Here, $\boldsymbol{y} = W\boldsymbol{x}$ and I_k are index sets selecting the different subspaces from the responses of W to $\boldsymbol{x}$. The collection of index sets I_k forms a partition of $1,\ldots,n$. ICA is a special case of ISA in which

$I_k = \{k\}$ such that all subspaces are one-dimensional. For the ISA models used by Hyvärinen et al. the distribution on the subspaces was chosen to be either spherically or L_p-spherically symmetric.

In its general form, ISA is not a generalization of L_p-spherically symmetric distributions. The most general ISA model for the transformed data $\boldsymbol{y} = W\boldsymbol{x}$ does not assume a certain type of distribution on the single subspace like in Hyvärinen and Köster (2007). While one could say for any non-factorial distribution that a factorial product over subspaces is a generalization, this is certainly a trivial step. Only in this particular sense is the particular ISA model by Hyvärinen and Köster (2007) a generalization of L_p-spherically symmetric distributions.

In contrast to ISA, L_p-nested symmetric distributions generally do not make an independence assumption on the "subspaces". In fact, for most of the models the subspaces will be dependent (see also our diagram in Figure 4). Therefore, not every ISA model is automatically L_p-nested symmetric and vice versa. In fact, in Sinz et al. (2009b) we have demonstrated for natural images that the dependencies *between* subspaces is stronger than the dependencies *within* subspaces on natural image patches. This is in stark contrast to the assumptions underlying ISA.

Note also that the product of L_p-spherically symmetric distributions used by Hyvärinen and Köster (2007) is not an L_p-nested function (Equation (6) in Hyvärinen and Köster, 2007) since the single a_j can be different and, therefore, the overall function is not positively homogeneous in general.

ICA and ISA have been used to infer features from natural signals, in particular from natural images. However, as mentioned by several authors (Zetzsche et al., 1993; Simoncelli, 1997; Wainwright and Simoncelli, 2000) and demonstrated quantitatively by Bethge (2006) and Eichhorn et al. (2009), the assumptions underlying linear ICA are not well matched by the statistics of the pixel intensities of natural images. A reliable parametric way to assess how well the independence assumption is met by a signal at hand is to fit a more general class of distributions that contains factorial as well as non-factorial distributions which both can equally well reproduce the marginals. By comparing the likelihood on held out test data between the best fitting non-factorial and the best-fitting factorial case, one can assess how well the sources can be described by a factorial distribution. For natural images, for example, one can use an arbitrary L_p-spherically symmetric distribution $\rho(\|W\boldsymbol{x}\|_p)$, fit it to the whitened data and compare its likelihood on held out test data to the one of the *p*-generalized Normal distribution (Sinz and Bethge, 2009). Since any choice of radial distribution ϕ determines a particular L_p-spherically symmetric distribution, the idea is to explore the space between factorial and non-factorial models by using a very flexible density ϕ on the radius. Note that having an explicit expression of the normalization constant allows for particularly reliable model comparisons via the likelihood. For many graphical models, for instance, such an explicit and computable expression is often not available.

The same type of dependency-analysis can be carried out for ISA using L_p-nested symmetric distributions (Sinz et al., 2009b). Figure 5 shows the L_p-nested tree corresponding to an ISA with four subspaces. In general, for such trees, each inner node—except the root node—corresponds to a single subspace. When using the radial distribution

$$\phi_\emptyset(v_\emptyset) = \frac{p_\emptyset v_\emptyset^{n-1}}{\Gamma\left[\frac{n}{p_\emptyset}\right] s^{\frac{n}{p_\emptyset}}} \exp\left(-\frac{v_\emptyset^{p_\emptyset}}{s}\right), \tag{11}$$

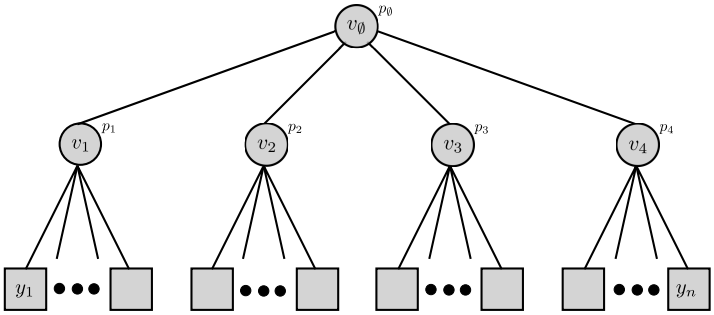


Figure 5: Tree corresponding to an L_p-nested ISA model.

the subspaces $v_1, ..., v_{\ell_\emptyset}$ become independent and one obtains an ISA model of the form

$$\begin{aligned}
\rho(\boldsymbol{y}) &= \frac{1}{Z} \exp\left(-\frac{f(\boldsymbol{y})^{p_\emptyset}}{s}\right) \\
&= \frac{1}{Z} \exp\left(-\frac{\sum_{k=1}^{\ell_\emptyset} \|\boldsymbol{y}_{I_k}\|_{p_k}}{s}\right) \\
&= \frac{p_\emptyset^{\ell_\emptyset}}{s^{\frac{n}{p_\emptyset}} \prod_{i=1}^{\ell_\emptyset} \Gamma\left[\frac{n_i}{p_\emptyset}\right]} \exp\left(-\frac{\sum_{k=1}^{\ell_\emptyset} \|\boldsymbol{y}_{I_k}\|_{p_k}}{s}\right) \prod_{k=1}^{\ell_\emptyset} \frac{p_k^{\ell_k - 1} \Gamma\left[\frac{n_k}{p_k}\right]}{2^{n_k} \Gamma^{n_k}\left[\frac{1}{p_I}\right]},
\end{aligned}$$

which has L_p-spherically symmetric distributions on each subspace. Note that this radial distribution is equivalent to a Gamma distribution whose variables have been raised to the power of $\frac{1}{p_\emptyset}$. In the following we will denote distributions of this type with $\gamma_p(u,s)$, where u and s are the shape and scale parameter of the Gamma distribution, respectively. The particular γ_p distribution that results in independent subspaces has arbitrary scale but shape parameter $u = \frac{n}{p_\emptyset}$. When using any other radial distribution, the different subspaces do not factorize, and the distribution is also not an ISA model. In that sense L_p-nested symmetric distributions are a generalization of ISA. Note, however, that not every ISA model is also L_p-nested symmetric since not every product of arbitrary distributions on the subspaces, even if they are L_p-spherically symmetric, must also be L_p-nested.

It is natural to ask, whether L_p-nested symmetric distributions can serve as a prior distribution $p(\boldsymbol{y}|\boldsymbol{\vartheta})$ over hidden factors in over-complete linear models of the form

$$p(\boldsymbol{x}|W, \sigma, \boldsymbol{\vartheta}) = \int p(\boldsymbol{x}|W\boldsymbol{y}, \sigma) p(\boldsymbol{y}|\boldsymbol{\vartheta}) d\boldsymbol{y},$$

where $p(\boldsymbol{x}|W\boldsymbol{y})$ represents the likelihood of the observed data point $\boldsymbol{x}$ given the hidden factors $\boldsymbol{y}$ and the over-complete matrix W. For example, $p(\boldsymbol{x}|W\boldsymbol{y}, \sigma) = \mathcal{N}(W\boldsymbol{y}, \sigma \cdot I)$ could be a Gaussian like in Olshausen and Field (1996). Unfortunately, such a model would suffer from the same problems as all over-complete linear models: While sampling from the prior is straightforward sampling from the posterior $p(\boldsymbol{y}|\boldsymbol{x}, W, \boldsymbol{\vartheta}, \sigma)$ is difficult because a whole subspace of $\boldsymbol{y}$ leads to the same $\boldsymbol{x}$.

Since parameter estimation either involves solving the high-dimensional integral $p(\boldsymbol{x}|W,\sigma,\boldsymbol{\vartheta}) = \int p(\boldsymbol{x}|W\boldsymbol{y},\sigma)p(\boldsymbol{y}|\boldsymbol{\vartheta})d\boldsymbol{y}$ or sampling from the posterior, learning is computationally demanding in such models. Various methods have been proposed to learn W, ranging from sampling the posterior only at its maximum (Olshausen and Field, 1996), approximating the posterior with a Gaussian via the Laplace approximation (Lewicki and Olshausen, 1999) or using Expectation Propagation (Seeger, 2008). In particular, all of the above studies either do not fit hyper-parameters $\boldsymbol{\vartheta}$ for the prior (Olshausen and Field, 1996; Lewicki and Olshausen, 1999) or rely on the factorial structure of it (Seeger, 2008). Since L_p-nested symmetric distributions do not provide such a factorial prior, Expectation Propagation is not directly applicable. An approximation like in Lewicki and Olshausen (1999) might be possible, but additionally estimating the parameters $\boldsymbol{\vartheta}$ of the L_p-nested symmetric distribution adds another level of complexity in the estimation procedure. Exploring such over-complete linear models with a non-factorial prior may be an interesting direction to investigate, but it will need a significant amount of additional numerical and algorithmical work to find an efficient and robust estimation procedure.

9. Nested Radial Factorization with L_p-Nested Symmetric Distributions

L_p-nested symmetric distribution also give rise to a non-linear ICA algorithm for linearly mixed non-factorial L_p-nested hidden sources $\boldsymbol{y}$. The idea is similar to the radial factorization algorithms proposed by Lyu and Simoncelli (2009) and Sinz and Bethge (2009). For this reason, we call it *nested radial factorization (NRF)*. For a one layer L_p-nested tree, NRF is equivalent to radial factorization as described in Sinz and Bethge (2009). If additionally p is set to $p = 2$, one obtains the radial Gaussianization by Lyu and Simoncelli (2009). Therefore, NRF is a generalization of radial Factorization. It has been demonstrated that radial factorization algorithms outperform linear ICA on natural image patches (Lyu and Simoncelli, 2009; Sinz and Bethge, 2009). Since L_p-nested symmetric distributions are slightly better in likelihood on natural image patches (Sinz et al., 2009b) and since the difference in the average log-likelihood directly corresponds to the reduction in dependencies between the single variables (Sinz and Bethge, 2009), NRF will slightly outperform radial factorization on natural images. For other types of data the performance will depend on how well the hidden sources can be modeled by a linear superposition of—possibly non-independent—L_p-nested symmetrically distributed sources. Here we state the algorithm as a possible application of L_p-nested symmetric distributions for unsupervised learning.

The idea is based on the observation that the choice of the radial distribution ϕ already determines the type of L_p-nested symmetric distribution. This also means that by changing the radial distribution by remapping the data, the distribution could possibly be turned in a factorial one. Radial factorization algorithms fit an L_p-spherically symmetric distribution with a very flexible radial distribution to the data and map this radial distribution ϕ_s (s for source) into the one of a p-generalized Normal distribution by the mapping

$$\boldsymbol{y} \mapsto \frac{(\mathcal{F}_{\perp\!\!\perp}^{-1} \circ \mathcal{F}_s)(\|\boldsymbol{y}\|_p)}{\|\boldsymbol{y}\|_p} \cdot \boldsymbol{y}, \tag{12}$$

where $\mathcal{F}_{\perp\!\!\perp}$ and $\mathcal{F}_s$ are the cumulative distribution functions of the two radial distributions involved. The mapping basically normalizes the demixed source $\boldsymbol{y}$ and rescales it with a new radius that has the correct distribution.

Exactly the same method cannot work for L_p-nested symmetric distributions since Proposition 9 states that there is no factorial distribution into which we could map the data by merely changing the radial distribution. Instead we have to remap the data in an iterative fashion beginning with changing the radial distribution at the root node into the radial distribution of the L_p-nested ISA shown in Equation (11). Once the nodes are independent, we repeat this procedure for each of the child nodes independently, then for their child nodes and so on, until only leaves are left. The rescaling of the radii is a non-linear mapping since the transform in Equation (12) is non-linear. Therefore, NRF is a non-linear ICA algorithm.

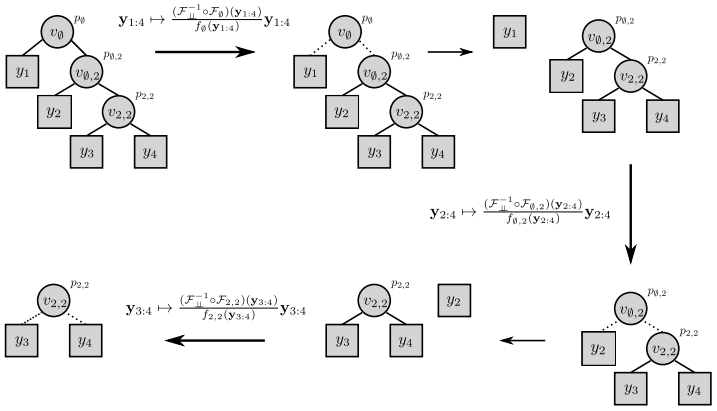


Figure 6: L_p-nested non-linear ICA for the tree of Example 6: For an arbitrary L_p-nested symmetric distribution, using Equation (12), the radial distribution can be remapped such that the children of the root node become independent. This is indicated in the plot via dotted lines. Once the data have been rescaled with that mapping, the children of root node can be separated. The remaining subtrees are again L_p-nested symmetric and have a particular radial distribution that can be remapped into the same one that makes their root nodes' children independent. This procedure is repeated until only leaves are left.

We demonstrate this with a simple example.

Example 6 *Consider the function*

$$f(\boldsymbol{y}) = \left(|y_1|^{p_\emptyset} + \left(|y_2|^{p_{\emptyset,2}} + \left(|y_3|^{p_{2,2}} + |y_4|^{p_{2,2}} \right)^{\frac{p_{\emptyset,2}}{p_{2,2}}} \right)^{\frac{p_\emptyset}{p_{\emptyset,2}}} \right)^{\frac{1}{p_\emptyset}}$$

for $\boldsymbol{y} = W\boldsymbol{x}$ where W has been estimated by fitting an L_p-nested symmetric distribution with a flexible radial distribution to $W\boldsymbol{x}$ as described in Section 5. Assume that the data has already been transformed once with the mapping of Equation (12)*. This means that the current radial distribution*

is given by (11) where we chose $s = 1$ for convenience. This yields a distribution of the form

$$\rho(\boldsymbol{y}) = \frac{p_\emptyset}{\Gamma\left[\frac{n}{p_\emptyset}\right]} \exp\left(-|y_1|^{p_\emptyset} - \left(|y_2|^{p_{\emptyset,2}} + \left(|y_3|^{p_{2,2}} + |y_4|^{p_{2,2}}\right)^{\frac{p_{\emptyset,2}}{p_{2,2}}}\right)^{\frac{p_\emptyset}{p_{\emptyset,2}}}\right) \times \frac{1}{2^n} \prod_{I \in \mathcal{I}} p_I^{\ell_I - 1} \frac{\Gamma\left[\frac{n_I}{p_I}\right]}{\prod_{k=1}^{\ell_I} \Gamma\left[\frac{n_{I,k}}{p_I}\right]}.$$

Now we can separate the distribution of y_1 from the distribution over $y_2, \ldots, y_4$. The distribution of y_1 is a p-generalized Normal

$$p(y_1) = \frac{p_\emptyset}{2\Gamma\left[\frac{1}{p_\emptyset}\right]} \exp\left(-|y_1|^{p_\emptyset}\right).$$

Thus the distribution of $y_2, \ldots, y_4$ is given by

$$\rho(y_2, \ldots, y_4) = \frac{p_\emptyset}{\Gamma\left[\frac{n_{\emptyset,2}}{p_\emptyset}\right]} \exp\left(-\left(|y_2|^{p_{\emptyset,2}} + \left(|y_3|^{p_{2,2}} + |y_4|^{p_{2,2}}\right)^{\frac{p_{\emptyset,2}}{p_{2,2}}}\right)^{\frac{p_\emptyset}{p_{\emptyset,2}}}\right) \times \frac{1}{2^{n-1}} \prod_{I \in \mathcal{I} \setminus \emptyset} p_I^{\ell_I - 1} \frac{\Gamma\left[\frac{n_I}{p_I}\right]}{\prod_{k=1}^{\ell_I} \Gamma\left[\frac{n_{I,k}}{p_I}\right]}.$$

By using Equation (9) we can identify the new radial distribution to be

$$\phi(v_{\emptyset,2}) = \frac{p_\emptyset v_{\emptyset,2}^{n-2}}{\Gamma\left[\frac{n_{\emptyset,2}}{p_\emptyset}\right]} \exp\left(-v_{\emptyset,2}^{p_\emptyset}\right).$$

Replacing this distribution by the one for the p-generalized Normal (for data we would use the mapping in Equation (12)), we obtain

$$\rho(y_2, \ldots, y_4) = \frac{p_{\emptyset,2}}{\Gamma\left[\frac{n_{\emptyset,2}}{p_{\emptyset,2}}\right]} \exp\left(-|y_2|^{p_{\emptyset,2}} - \left(|y_3|^{p_{2,2}} + |y_4|^{p_{2,2}}\right)^{\frac{p_{\emptyset,2}}{p_{2,2}}}\right) \times \frac{1}{2^{n-1}} \prod_{I \in \mathcal{I} \setminus \emptyset} p_I^{\ell_I - 1} \frac{\Gamma\left[\frac{n_I}{p_I}\right]}{\prod_{k=1}^{\ell_I} \Gamma\left[\frac{n_{I,k}}{p_I}\right]}.$$

Now, we can separate out the distribution of y_2 which is again p-generalized Normal. This leaves us with the distribution for y_3 and y_4

$$\rho(y_3, y_4) = \frac{p_{\emptyset,2}}{\Gamma\left[\frac{n_{2,2}}{p_{\emptyset,2}}\right]} \exp\left(-\left(|y_3|^{p_{2,2}} + |y_4|^{p_{2,2}}\right)^{\frac{p_{\emptyset,2}}{p_{2,2}}}\right) \frac{1}{2^{n-2}} \prod_{I \in \mathcal{I} \setminus \{\emptyset, (\emptyset,2)\}} p_I^{\ell_I - 1} \frac{\Gamma\left[\frac{n_I}{p_I}\right]}{\prod_{k=1}^{\ell_I} \Gamma\left[\frac{n_{I,k}}{p_I}\right]}.$$

For this distribution we can repeat the same procedure which will also yield p-generalized Normal distributions for y_3 and y_4.

Algorithm 2 Recursion NRF($\boldsymbol{y}, f, \phi_s$)

Input: Data point $\boldsymbol{y}$, L_p-nested function f, current radial distribution ϕ_s,
Output: Non-linearly transformed data point $\boldsymbol{y}$
Algorithm

1. Set the target radial distribution to be $\phi_{\perp\!\!\!\perp} \leftarrow \gamma_p\left(\frac{n_\emptyset}{p_\emptyset}, \frac{\Gamma\left[\frac{1}{p_\emptyset}\right]^{\frac{p_\emptyset}{2}}}{\Gamma\left[\frac{3}{p_\emptyset}\right]^{\frac{p_\emptyset}{2}}}\right)$
2. Set $\boldsymbol{y} \leftarrow \frac{\mathcal{F}_{\perp\!\!\!\perp}^{-1}(\mathcal{F}_s(f(\boldsymbol{y})))}{f(\boldsymbol{y})} \cdot \boldsymbol{y}$ where $\mathcal{F}$ denotes the cumulative distribution function of the respective ϕ.
3. For all children i of the root node that are not leaves:
   (a) Set $\phi_s \leftarrow \gamma_p\left(\frac{n_{\emptyset,i}}{p_\emptyset}, \frac{\Gamma\left[\frac{1}{p_\emptyset}\right]^{\frac{p_\emptyset}{2}}}{\Gamma\left[\frac{3}{p_\emptyset}\right]^{\frac{p_\emptyset}{2}}}\right)$
   (b) Set $\boldsymbol{y}_{\emptyset,i} \leftarrow$ NRF($\boldsymbol{y}_{\emptyset,i}, f_{\emptyset,i}, \phi_s$). Note that in the recursion $\emptyset, i$ will become the new $\emptyset$.
4. Return $\boldsymbol{y}$

This non-linear procedure naturally carries over to arbitrary L_p-nested trees and distributions, thus yielding a general non-linear ICA algorithm for linearly mixed non-factorial L_p-nested symmetric sources. For generalizing Example 6, note the particular form of the radial distributions involved. As already noted above, the distribution (11) on the root node's values that makes its children statistically independent is that of a Gamma distributed variable with shape parameter $\frac{n_\emptyset}{p_\emptyset}$ and scale parameter s which has been raised to the power of $\frac{1}{p_\emptyset}$. In Section 8 we denoted this class of distributions with $\gamma_p[u,s]$, where u and s are the shape and the scale parameter, respectively. Interestingly, the radial distributions of the root node's children are also γ_p except that the shape parameter is $\frac{n_{\emptyset,i}}{p_\emptyset}$. The goal of the radial remapping of the children's values is hence just changing the shape parameter from $\frac{n_{\emptyset,i}}{p_\emptyset}$ to $\frac{n_{\emptyset,i}}{p_{\emptyset,i}}$. Of course, it is also possible to change the scale parameter of the single distributions during the radial remappings. This will not affect the statistical independence of the resulting variables. In the general algorithm, that we describe now, we choose s such that the transformed data is white.

The algorithm starts with fitting a general L_p-nested model of the form $\rho(W\boldsymbol{x})$ as described in Section 5. Once this is done, the linear demixing matrix W is fixed and the hidden non-factorial sources are recovered via $\boldsymbol{y} = W\boldsymbol{x}$. Afterwards, the sources $\boldsymbol{y}$ are non-linearly made independent by calling the recursion specified in Algorithm 2 with the parameters $W\boldsymbol{x}$, f and ϕ, where ϕ is the radial distribution of the estimated model.

The computational complexity for transforming a single data point is $O(n^2)$ because of the matrix multiplication $W\boldsymbol{x}$. In the non-linear transformation, each single data dimension is not rescaled more that n times which means that the rescaling is certainly also in $O(n^2)$.

An important aspect of NRF is that it yields a probabilistic model for the transformed data. This model is simply a product of n independent exponential power marginals. Since the radial remappings do not change the likelihood, the likelihood of the non-linearly separated data is the

same as the likelihood of the data under L_p-nested symmetric distribution that was fitted to it in the first place. However, in some cases, one might like to fit a different distribution to the outcome of Algorithm 2. In that case the determinant of the transformation is necessary to determine the likelihood of the input data—and not the transformed one—under the model. The following lemma provides the determinant of the Jacobian for the non-linear rescaling.

Lemma 11 (Determinant of the Jacobian) *Let $\boldsymbol{z} = NRF(W\boldsymbol{x}, f, \phi_s)$ as described above. Let $\boldsymbol{t}_I$ denote the values of $W\boldsymbol{x}$ below the inner node I which have been transformed with Algorithm 2 up to node I. Let $g_I(r) = (\mathcal{F}_{\phi_{\perp\!\!\perp}} \circ \mathcal{F}_{\phi_s})(r)$ denote the radial transform at node I in Algorithm 2. Furthermore, let I denote the set of all inner nodes, excluding the leaves. Then, the determinant of the Jacobian $\left(\frac{\partial z_i}{\partial x_j}\right)_{ij}$ is given by*

$$\left|\det \frac{\partial z_i}{\partial x_j}\right| = |\det W| \cdot \prod_{I \in I} \left| \frac{g_I(f_I(\boldsymbol{t}_I))^{n_I - 1}}{f_I(\boldsymbol{t}_I)^{n_I - 1}} \cdot \frac{\phi_s(f_I(\boldsymbol{t}_I))}{\phi_{\perp\!\!\perp}(g_I(f_I(\boldsymbol{t}_I)))} \right|$$

Proof The proof can be found in the Appendix E. ■

10. Conclusion

In this article we presented a formal treatment of the first tractable subclass of ν-spherical distributions which generalizes the important family of L_p-spherically symmetric distributions. We derived an analytical expression for the normalization constant, introduced a coordinate system particularly tailored to L_p-nested functions, and computed the determinant of the Jacobian for the corresponding coordinate transformation. Using these results, we introduced the uniform distribution on the L_p-nested unit sphere and the general form of an L_p-nested symmetric distribution for arbitrary L_p-nested functions and radial distributions. We also derived an expression for the joint distribution of inner nodes of an L_p-nested tree and derived a sampling scheme for an arbitrary L_p-nested symmetric distribution.

L_p-nested symmetric distributions naturally provide the class of probability distributions corresponding to mixed norm priors, allowing full Bayesian inference in the corresponding probabilistic models. We showed that a robustness result for Bayesian inference of the location parameter known for L_p-spherically symmetric distributions carries over to the L_p-nested symmetric class. We discussed the relationship of L_p-nested symmetric distributions to indepedent component (ICA) and independent subspace Analysis (ISA), as well as its applicability as a prior distribution in overcomplete linear models. Finally, we showed how L_p-nested symmetric distributions can be used to construct a non-linear ICA algorithm called nested radial factorization (NRF).

The application of L_p-nested symmetric distribution has been presented in a previous conference paper (Sinz et al., 2009b). Code for training this class of distribution is provided online under `http://www.kyb.tuebingen.mpg.de/bethge/code/`.

Acknowledgments

We would like to thank Eero Simoncelli for bringing up the problem whether the class of L_p-spherical distributions can be generalized to L_p-nested symmetric distributions. Furthermore, we

want to thank Sebastian Gerwinn, Suvrit Sra, Reshad Hosseini, Lucas Theis, Holly Gerhard, and Sina Tootoonian for fruitful discussions and feedback on the manuscript. Finally, we would like to thank the anonymous reviewers for their comments that helped to improve the manuscript.

This work is supported by the German Ministry of Education, Science, Research and Technology through the Bernstein prize to MB (BMBF; FKZ: 01GQ0601), a scholarship to FS by the German National Academic Foundation, and the Max Planck Society.

Appendix A. Determinant of the Jacobian

Proof [Lemma 2] The proof is very similar to the one in Song and Gupta (1997). To derive Equation (2) one needs to expand the Jacobian of the inverse coordinate transformation with respect to the last column using the Laplace's expansion of the determinant. The term Δ_n can be factored out of the determinant and cancels due to the absolute value around it. Therefore, the determinant of the coordinate transformation does not depend on Δ_n.

The partial derivatives of the inverse coordinate transformation are given by:

$$\begin{aligned}
\frac{\partial}{\partial u_k} x_i &= \delta_{ik} r \text{ for } 1 \leq i,k \leq n-1 \\
\frac{\partial}{\partial u_k} x_n &= \Delta_n r \frac{\partial u_n}{\partial u_k} \text{ for } 1 \leq k \leq n-1 \\
\frac{\partial}{\partial r} x_i &= u_i \text{ for } 1 \leq i \leq n-1 \\
\frac{\partial}{\partial r} x_n &= \Delta_n u_n .
\end{aligned}$$

Therefore, the structure of the Jacobian is given by

$$\mathcal{J} = \begin{pmatrix} r & \dots & 0 & u_1 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & r & u_{n-1} \\ \Delta_n r \frac{\partial u_n}{\partial u_1} & \dots & \Delta_n r \frac{\partial u_n}{\partial u_{n-1}} & \Delta_n u_n \end{pmatrix}.$$

Since we are only interested in the absolute value of the determinant and since $\Delta_n \in \{-1,1\}$, we can factor out Δ_n and drop it. Furthermore, we can factor out r from the first $n-1$ columns which yields

$$|\det \mathcal{J}| = r^{n-1} \left| \det \begin{pmatrix} 1 & \dots & 0 & u_1 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & 1 & u_{n-1} \\ \frac{\partial u_n}{\partial u_1} & \dots & \frac{\partial u_n}{\partial u_{n-1}} & u_n \end{pmatrix} \right|.$$

Now we can use the Laplace's expansion of the determinant with respect to the last column. For that purpose, let $\mathcal{J}_i$ denote the matrix which is obtained by deleting the last column and the ith row

from $\mathcal{J}$. This matrix has the following structure

$$\mathcal{J}_i = \begin{pmatrix} 1 & & & 0 & & & \\ & \ddots & & 0 & & & \\ & & 1 & 0 & & & \\ & & & \vdots & 1 & & \\ & & & 0 & & \ddots & \\ & & & 0 & & & 1 \\ \frac{\partial u_n}{\partial u_1} & & & \frac{\partial u_n}{\partial u_i} & & & \frac{\partial u_n}{\partial u_{n-1}} \end{pmatrix}.$$

We can transform $\mathcal{J}_i$ into a lower triangular matrix by moving the column with all zeros and $\frac{\partial u_n}{\partial u_i}$ bottom entry to the rightmost column of $\mathcal{J}_i$. Each swapping of two columns introduces a factor of -1. In the end, we can compute the value of $\det \mathcal{J}_i$ by simply taking the product of the diagonal entries and obtain $\det \mathcal{J}_i = (-1)^{n-1-i}\frac{\partial u_n}{\partial u_i}$. This yields

$$\begin{aligned} |\det \mathcal{J}| &= r^{n-1}\left(\sum_{k=1}^{n}(-1)^{n+k}u_k \det \mathcal{J}_k\right) \\ &= r^{n-1}\left(\sum_{k=1}^{n-1}(-1)^{n+k}u_k \det \mathcal{J}_k + (-1)^{2n}\frac{\partial x_n}{\partial r}\right) \\ &= r^{n-1}\left(\sum_{k=1}^{n-1}(-1)^{n+k}u_k(-1)^{n-1-k}\frac{\partial u_n}{\partial u_k} + u_n\right) \\ &= r^{n-1}\left(-\sum_{k=1}^{n-1}u_k\frac{\partial u_n}{\partial u_k} + u_n\right). \end{aligned}$$

■

Before proving Proposition 3 stating that the determinant only depends on the terms $G_I(\boldsymbol{u}_{\widehat{I}})$ produced by the chain rule when used upwards in the tree, let us quickly outline the essential mechanism when taking the chain rule for $\frac{\partial u_n}{\partial u_q}$: Consider the tree corresponding to f. By definition u_n is the rightmost leaf of the tree. Let L, ℓ_L be the multi-index of u_n. As in the example, the chain rule starts at the leaf u_n and ascends in the tree until it reaches the lowest node whose subtree contains both, u_n and u_q. At this point, it starts descending the tree until it reaches the leaf u_q. Depending on whether the chain rule ascends or descends, two different forms of derivatives occur: while ascending, the chain rule produces $G_I(\boldsymbol{u}_{\widehat{I}})$-terms like the one in the example above. At descending, it produces $F_I(\boldsymbol{u}_I)$-terms. The general definitions of the $G_I(\boldsymbol{u}_{\widehat{I}})$- and $F_I(\boldsymbol{u}_I)$-terms are given by the recursive formulae

$$G_{I,\ell_I}(\boldsymbol{u}_{\widehat{I,\ell_I}}) = g_{I,\ell_I}(\boldsymbol{u}_{\widehat{I,\ell_I}})^{p_{I,\ell_I}-p_I} = \left(g_I(\boldsymbol{u}_{\widehat{I}})^{p_I} - \sum_{j=1}^{\ell_I-1} f_{I,j}(\boldsymbol{u}_{I,j})^{p_I}\right)^{\frac{p_{I,\ell_I}-p_I}{p_I}}$$

and

$$F_{I,i_r}(\boldsymbol{u}_{I,i_r}) = f_{I,i_r}(\boldsymbol{u}_{I,i_r})^{p_I - p_{I,i_r}} = \left(\sum_{k=1}^{\ell_{I,i_r}} f_{I,i_r,k}(\boldsymbol{u}_{I,i_r,k})^{p_{I,i_r}} \right)^{\frac{p_I - p_{I,i_r}}{p_{I,i_r}}}.$$

The next two lemmata are required for the proof of Proposition 3. We use the somewhat sloppy notation $k \in I, i_r$ if the variable u_k is a leaf in the subtree below I, i_r. The same notation is used for $\widehat{I}$.

Lemma 12 *Let $I = i_1, \ldots, i_{r-1}$ and I, i_r be any node of the tree associated with an L_p-nested function f. Then the following recursions hold for the derivatives of $g_{I,i_r}(\boldsymbol{u}_{\widehat{I,i_r}})^{p_{I,i_r}}$ and $f_{I,i_r}^{p_I}(\boldsymbol{u}_{I,i_r})$ w.r.t u_q: If u_q is not in the subtree under the node I, i_r, that is, $k \notin I, i_r$, then*

$$\frac{\partial}{\partial u_q} f_{I,i_r}(\boldsymbol{u}_{I,i_r})^{p_I} = 0$$

and

$$\frac{\partial}{\partial u_q} g_{I,i_r}(\boldsymbol{u}_{\widehat{I,i_r}})^{p_{I,i_r}} = \frac{p_{I,i_r}}{p_I} G_{I,i_r}(\boldsymbol{u}_{\widehat{I,i_r}}) \cdot \begin{cases} \frac{\partial}{\partial u_q} g_I(\boldsymbol{u}_{\widehat{I}})^{p_I} & \text{if } q \in I \\ -\frac{\partial}{\partial u_q} f_{I,j}(\boldsymbol{u}_{I,j})^{p_I} & \text{if } q \in I, j \end{cases}$$

for $q \in I, j$ and $q \notin I, k$ for $k \neq j$. Otherwise

$$\frac{\partial}{\partial u_q} g_{I,i_r}(\boldsymbol{u}_{\widehat{I,i_r}})^{p_{I,i_r}} = 0 \text{ and } \frac{\partial}{\partial u_q} f_{I,i_r}(\boldsymbol{u}_{I,i_r})^{p_I} = \frac{p_I}{p_{I,i_r}} F_{I,i_r}(\boldsymbol{u}_{I,i_r}) \frac{\partial}{\partial u_q} f_{I,i_r,s}(\boldsymbol{u}_{I,i_r,s})^{p_{I,i_r}}$$

for $q \in I, i_r, s$ and $q \notin I, i_r, k$ for $k \neq s$.

Proof Both of the first equations are obvious, since only those nodes have a non-zero derivative for which the subtree actually depends on u_q. The second equations can be seen by direct computation

$$\begin{aligned}
\frac{\partial}{\partial u_q} g_{I,i_r}(\boldsymbol{u}_{\widehat{I,i_r}})^{p_{I,i_r}} &= p_{I,i_r} g_{I,i_r}(\boldsymbol{u}_{\widehat{I,i_r}})^{p_{I,i_r}-1} \frac{\partial}{\partial u_q} G_{I,i_r}(\boldsymbol{u}_{\widehat{I,i_r}}) \\
&= p_{I,i_r} g_{I,i_r}(\boldsymbol{u}_{\widehat{I,i_r}})^{p_{I,i_r}-1} \frac{\partial}{\partial u_q} \left(g_I(\boldsymbol{u}_{\widehat{I}})^{p_I} - \sum_{j=1}^{\ell_I - 1} f_{I,j}(\boldsymbol{u}_{I,j})^{p_I} \right)^{\frac{1}{p_I}} \\
&= \frac{p_{I,i_r}}{p_I} g_{I,i_r}(\boldsymbol{u}_{\widehat{I,i_r}})^{p_{I,i_r}-1} g_{I,i_r}(\boldsymbol{u}_{\widehat{I,i_r}})^{1-p_I} \frac{\partial}{\partial u_q} \left(g_I(\boldsymbol{u}_{\widehat{I}})^{p_I} - \sum_{j=1}^{\ell_I - 1} f_{I,j}(\boldsymbol{u}_{I,j})^{p_I} \right) \\
&= \frac{p_{I,i_r}}{p_I} G_{I,i_r}(\boldsymbol{u}_{\widehat{I,i_r}}) \cdot \begin{cases} \frac{\partial}{\partial u_q} g_I(\boldsymbol{u}_{\widehat{I}})^{p_I} & \text{if } q \in I \\ -\frac{\partial}{\partial u_q} f_{I,j}(\boldsymbol{u}_{I,j})^{p_I} & \text{if } q \in I, j \end{cases}
\end{aligned}$$

Similarly

$$\begin{aligned}\frac{\partial}{\partial u_q} f_{I,i_r}(\boldsymbol{u}_{I,i_r})^{p_I} &= p_I f_{I,i_r}(\boldsymbol{u}_{I,i_r})^{p_I-1}\frac{\partial}{\partial u_q} f_{I,i_r}(\boldsymbol{u}_{I,i_r}) \\ &= p_I f_{I,i_r}(\boldsymbol{u}_{I,i_r})^{p_I-1}\frac{\partial}{\partial u_q}\left(\sum_{k=1}^{\ell_{I,i_r}} f_{I,i_r,k}(\boldsymbol{u}_{I,i_r,k})^{p_{I,i_r}}\right)^{\frac{1}{p_{I,i_r}}} \\ &= \frac{p_I}{p_{I,i_r}} f_{I,i_r}(\boldsymbol{u}_{I,i_r})^{p_I-1} f_{I,i_r}(\boldsymbol{u}_{I,i_r})^{1-p_{I,i_r}}\frac{\partial}{\partial u_q} f_{I,i_r,s}(\boldsymbol{u}_{I,i_r,s})^{p_{I,i_r}} \\ &= \frac{p_I}{p_{I,i_r}} F_{I,i_r}(\boldsymbol{u}_{I,i_r})\frac{\partial}{\partial u_q} f_{I,i_r,s}(\boldsymbol{u}_{I,i_r,s})^{p_{I,i_r}}\end{aligned}$$

for $k \in I, i_r, s$. ■

The next lemma states the form of the whole derivative $\frac{\partial u_n}{\partial u_q}$ in terms of the $G_I(\boldsymbol{u}_{\widehat{I}})$- and $F_I(\boldsymbol{u}_I)$-terms.

Lemma 13 *Let $|u_q| = v_{\ell_1,\ldots,\ell_m,i_1,\ldots,i_t}$, $|u_n| = v_{\ell_1,\ldots,\ell_d}$ with $m < d$. The derivative of u_n w.r.t. u_q is given by*

$$\begin{aligned}\frac{\partial}{\partial u_q} u_n = &-G_{\ell_1,\ldots,\ell_d}(\boldsymbol{u}_{\widehat{\ell_1,\ldots,\ell_d}})\cdot\ldots\cdot G_{\ell_1,\ldots,\ell_{m+1}}(\boldsymbol{u}_{\ell_1,\widehat{\ldots,\ell_{m+1}}}) \\ &\times F_{\ell_1,\ldots,\ell_m,i_1}(\boldsymbol{u}_{\ell_1,\ldots,\ell_m,i_1})\cdot F_{\ell_1,\ldots,\ell_m,i_1,\ldots,i_{t-1}}(\boldsymbol{u}_{\ell_1,\ldots,\ell_m,i_1,\ldots,i_{t-1}})\cdot\Delta_q|u_q|^{p_{\ell_1,\ldots,\ell_m,i_1,\ldots,i_{t-1}}-1}\end{aligned}$$

with $\Delta_q = \operatorname{sgn} u_q$ and $|u_q|^p = (\Delta_q u_q)^p$. In particular

$$\begin{aligned}u_q\frac{\partial}{\partial u_q} u_n = &-G_{\ell_1,\ldots,\ell_d}(\boldsymbol{u}_{\widehat{\ell_1,\ldots,\ell_d}})\cdot\ldots\cdot G_{\ell_1,\ldots,\ell_{m+1}}(\boldsymbol{u}_{\ell_1,\widehat{\ldots,\ell_{m+1}}}) \\ &\times F_{\ell_1,\ldots,\ell_m,i_1}(\boldsymbol{u}_1)\cdot F_{\ell_1,\ldots,\ell_m,i_1,\ldots,i_{t-1}}(\boldsymbol{u}_{\ell_1,\ldots,\ell_m,i_1})\cdot|u_q|^{p_{\ell_1,\ldots,\ell_m,i_1,\ldots,i_{t-1}}}.\end{aligned}$$

Proof Successive application of Lemma (12). ■

Proof [Proposition 3] Before we begin with the proof, note that $F_I(\boldsymbol{u}_I)$ and $G_I(\boldsymbol{u}_{\widehat{I}})$ fulfill following equalities

$$\begin{aligned}G_{I,i_m}(\boldsymbol{u}_{\widehat{I,i_m}})^{-1} g_{I,i_m}(\boldsymbol{u}_{\widehat{I,i_m}})^{p_{I,i_m}} &= g_{I,i_m}(\boldsymbol{u}_{\widehat{I,i_m}})^{p_I} \\ &= g_I(\boldsymbol{u}_{\widehat{I}})^{p_I} - \sum_{k=1}^{\ell_I-1} F_{I,k}(\boldsymbol{u}_{I,k})\, f_{I,k}(\boldsymbol{u}_{I,k})^{p_{I,k}} \qquad (13)\end{aligned}$$

and

$$f_{I,i_m}(\boldsymbol{u}_{I,i_m})^{p_{I,i_m}} = \sum_{k=1}^{\ell_{I,im}} F_{I,i_m,k}(\boldsymbol{u}_{I,i_m,k})\, f_{I,i_m,k}(\boldsymbol{u}_{I,i_m,k})^{p_{I,im,k}}. \qquad (14)$$

Now let $L = \ell_1,\ldots,\ell_{d-1}$ be the multi-index of the parent of u_n. We compute $\frac{1}{r^{n-1}}|\det \mathcal{J}|$ and obtain the result by solving for $|\det \mathcal{J}|$. As shown in Lemma (2) $\frac{1}{r^{n-1}}|\det \mathcal{J}|$ has the form

$$\frac{1}{r^{n-1}}|\det \mathcal{J}| = -\sum_{k=1}^{n-1}\frac{\partial u_n}{\partial u_k}\cdot u_k + u_n.$$

By definition $u_n = g_{L,\ell_d}(\boldsymbol{u}_{\widehat{L,\ell_d}}) = g_{L,\ell_d}(\boldsymbol{u}_{\widehat{L,\ell_d}})^{p_{L,\ell_d}}$. Now, assume that $u_m,...,u_{n-1}$ are children of L, that is, $u_k = v_{L,I,i_t}$ for some $I, i_t = i_1,...,i_t$ and $m \le k < n$. Remember, that by Lemma (13) the terms $u_q \frac{\partial}{\partial u_q} u_n$ for $m \le q < n$ have the form

$$u_q \frac{\partial}{\partial u_q} u_n = -G_{L,\ell_d}(\boldsymbol{u}_{\widehat{L,\ell_d}}) \cdot F_{L,i_1}(\boldsymbol{u}_{L,i_1}) \cdot ... \cdot F_{L,I}(\boldsymbol{u}_{L,I}) \cdot |u_q|^{p_{\ell_1,...,\ell_{d-1},i_1,...,i_{t-1}}}.$$

Using Equation (13), we can expand the determinant as follows

$$\begin{aligned}
&-\sum_{k=1}^{n-1} \frac{\partial u_n}{\partial u_k} \cdot u_k + g_{L,\ell_d}(\boldsymbol{u}_{\widehat{L,\ell_d}})^{p_{L,\ell_d}} \\
=&-\sum_{k=1}^{m-1} \frac{\partial u_n}{\partial u_k} \cdot u_k - \sum_{k=m}^{n-1} \frac{\partial u_n}{\partial u_k} \cdot u_k + g_{L,\ell_d}(\boldsymbol{u}_{\widehat{L,\ell_d}})^{p_{L,\ell_d}} \\
=&-\sum_{k=1}^{m-1} \frac{\partial u_n}{\partial u_k} \cdot u_k \\
&+ G_{L,\ell_d}(\boldsymbol{u}_{\widehat{L,\ell_d}}) \left(-\sum_{k=m}^{n-1} G_{L,\ell_d}(\boldsymbol{u}_{\widehat{L,\ell_d}})^{-1} \frac{\partial u_n}{\partial u_k} \cdot u_k + G_{L,\ell_d}(\boldsymbol{u}_{\widehat{L,\ell_d}})^{-1} g_{L,\ell_d}(\boldsymbol{u}_{\widehat{L,\ell_d}})^{p_{L,\ell_d}} \right) \\
=&-\sum_{k=1}^{m-1} \frac{\partial u_n}{\partial u_k} \cdot u_k \\
&+ G_{L,\ell_d}(\boldsymbol{u}_{\widehat{L,\ell_d}}) \left(-\sum_{k=m}^{n-1} G_{L,\ell_d}(\boldsymbol{u}_{\widehat{L,\ell_d}})^{-1} \frac{\partial u_n}{\partial u_k} \cdot u_k + g_L(\boldsymbol{u}_{\widehat{L}})^{p_L} - \sum_{k=1}^{\ell_d - 1} F_{L,k}(\boldsymbol{u}_{L,k}) f_{L,k}(\boldsymbol{u}_{L,k})^{p_{L,k}} \right).
\end{aligned}$$

Note that all terms $G_{L,\ell_d}(\boldsymbol{u}_{\widehat{L,\ell_d}})^{-1} \frac{\partial u_n}{\partial u_k} \cdot u_k$ for $m \le k < n$ now have the form

$$G_{L,\ell_d}(\boldsymbol{u}_{\widehat{L,\ell_d}})^{-1} u_k \frac{\partial}{\partial u_k} u_n = -F_{L,i_1}(\boldsymbol{u}_{L,i_1}) \cdot ... \cdot F_{L,I}(\boldsymbol{u}_{L,I}) \cdot |u_q|^{p_{\ell_1,...,\ell_{d-1},i_1,...,i_{t-1}}}$$

since we constructed them to be neighbors of u_n. However, with Equation (14), we can further expand the sum $\sum_{k=1}^{\ell_d - 1} F_{L,k}(\boldsymbol{u}_{L,k}) f_{L,k}(\boldsymbol{u}_{L,k})^{p_{L,k}}$ down to the leaves $u_m,...,u_{n-1}$. When doing so we end up with the same factors $F_{L,i_1}(\boldsymbol{u}_{L,i_1}) \cdot ... \cdot F_{L,I}(\boldsymbol{u}_{L,I}) \cdot |u_q|^{p_{\ell_1,...,\ell_{d-1},i_1,...,i_{t-1}}}$ as in the derivatives $G_{L,\ell_d}(\boldsymbol{u}_{\widehat{L,\ell_d}})^{-1} u_q \frac{\partial}{\partial u_q} u_n$. This means exactly that

$$-\sum_{k=m}^{n-1} G_{L,\ell_d}(\boldsymbol{u}_{\widehat{L,\ell_d}})^{-1} \frac{\partial u_n}{\partial u_k} \cdot u_k = \sum_{k=1}^{\ell_d - 1} F_{L,k}(\boldsymbol{u}_{L,k}) f_{L,k}(\boldsymbol{u}_{L,k})^{p_{L,k}}$$

and, therefore,

$$
\begin{aligned}
&-\sum_{k=1}^{m-1}\frac{\partial u_n}{\partial u_k}\cdot u_k \\
&+G_{L,\ell_d}(\boldsymbol{u}_{\widehat{L,\ell_d}})\left(-\sum_{k=m}^{n-1}G_{L,\ell_d}(\boldsymbol{u}_{\widehat{L,\ell_d}})^{-1}\frac{\partial u_n}{\partial u_k}\cdot u_k+g_L(\boldsymbol{u}_{\widehat{L}})^{p_L}-\sum_{k=1}^{\ell_d-1}F_{L,k}(\boldsymbol{u}_{L,k})\,f_{L,k}(\boldsymbol{u}_{L,k})^{p_{L,k}}\right)\\
=&-\sum_{k=1}^{m-1}\frac{\partial u_n}{\partial u_k}\cdot u_k \\
&+G_{L,\ell_d}(\boldsymbol{u}_{\widehat{L,\ell_d}})\left(\sum_{k=1}^{\ell_d-1}F_{L,k}(\boldsymbol{u}_{L,k})\,f_{L,k}(\boldsymbol{u}_{L,k})^{p_{L,k}}+g_L(\boldsymbol{u}_{\widehat{L}})^{p_L}-\sum_{k=1}^{\ell_d-1}F_{L,k}(\boldsymbol{u}_{L,k})\,f_{L,k}(\boldsymbol{u}_{L,k})^{p_{L,k}}\right)\\
=&-\sum_{k=1}^{m-1}\frac{\partial u_n}{\partial u_k}\cdot u_k+G_{L,\ell_d}(\boldsymbol{u}_{\widehat{L,\ell_d}})g_L(\boldsymbol{u}_{\widehat{L}})^{p_L}.
\end{aligned}
$$

By factoring out $G_{L,\ell_d}(\boldsymbol{u}_{\widehat{L,\ell_d}})$ from the equation, the terms $\frac{\partial u_n}{\partial u_k}\cdot u_k$ loose the G_{L,ℓ_d} in front and we get basically the same equation as before, only that the new leaf (the new "u_n") is $g_L(\boldsymbol{u}_{\widehat{L}})^{p_L}$ and we got rid of all the children of L. By repeating that procedure up to the root node, we successively factor out all $G_{L'}(\boldsymbol{u}_{\widehat{L'}})$ for $L'\in\mathcal{L}$ until all terms of the sum vanish and we are only left with $v_\emptyset=1$. Therefore, the determinant is

$$\frac{1}{r^{n-1}}|\det \mathcal{J}|=\prod_{L\in\mathcal{L}}G_L(\boldsymbol{u}_{\widehat{L}})$$

which completes the proof. ■

Appendix B. Volume and Surface of the L_p-Nested Unit Sphere

Proof [Proposition 4] We obtain the volume by computing the integral $\int_{f(\boldsymbol{x})\le R}d\boldsymbol{x}$. Differentiation with respect to R yields the surface area. For symmetry reasons we can compute the volume only on the positive quadrant $\mathbb{R}^n_+$ and multiply the result with 2^n later to obtain the full volume and surface area. The strategy for computing the volume is as follows. We start with inner nodes I that are parents of leaves only. The value v_I of such a node is simply the L_{p_I} norm of its children. Therefore, we can convert the integral over the children of I with the transformation of Gupta and Song (1997). This maps the leaves $\boldsymbol{v}_{I,1:\ell_I}$ into v_I and "angular" variables $\tilde{\boldsymbol{u}}$. Since integral borders of the original integral depend only on the value of v_I and not on $\tilde{\boldsymbol{u}}$, we can separate the variables $\tilde{\boldsymbol{u}}$ from the radial variables v_I and integrate the variables $\tilde{\boldsymbol{u}}$ separately. The integration over $\tilde{\boldsymbol{u}}$ yields a certain factor, while the variable v_I effectively becomes a new leaf.

Now suppose I is the parent of leaves only. Without loss of generality let the ℓ_I leaves correspond to the last ℓ_I coefficients of $\boldsymbol{x}$. Let $\boldsymbol{x} \in \mathbb{R}^n_+$. Carrying out the first transformation and integration yields

$$\begin{aligned}\int_{f(\boldsymbol{x})\le R} d\boldsymbol{x} &= \int_{f(\boldsymbol{x}_{1:n-\ell_I}, v_I)\le R} \int_{\tilde{\boldsymbol{u}}\in \mathcal{V}_+^{\ell_I-1}} v_I^{\ell_I-1} \left(1-\sum_{i=1}^{\ell_I-1} \tilde{u}_i^{p_I}\right)^{\frac{1-p_I}{p_I}} dv_I d\tilde{\boldsymbol{u}} d\boldsymbol{x}_{1:n-\ell_I} \\ &= \int_{f(\boldsymbol{x}_{1:n-\ell_I}, v_I)\le R} v_I^{n_I-1} dv_I d\boldsymbol{x}_{1:n-\ell_I} \times \int_{\tilde{\boldsymbol{u}}\in \mathcal{V}_+^{\ell_I-1}} \left(1-\sum_{i=1}^{\ell_I-1} \tilde{u}_i^{p_I}\right)^{\frac{n_{I,\ell_I}-p_I}{p_I}} d\tilde{\boldsymbol{u}}.\end{aligned}$$

where $\mathcal{V}_+$ denotes the intersection of the positive quadrant and the L_{p_I}-norm unit ball. For solving the second integral we make the pointwise transformation $s_i = \tilde{u}_i^{p_I}$ and obtain

$$\begin{aligned}\int_{\tilde{\boldsymbol{u}}\in \mathcal{V}_+^{\ell_I-1}} \left(1-\sum_{i=1}^{\ell_I-1} \tilde{u}_i^{p_I}\right)^{\frac{n_{I,\ell_I}-p_I}{p_I}} d\tilde{\boldsymbol{u}} &= \frac{1}{p_I^{\ell_I-1}} \int_{\sum s_i\le 1} \left(1-\sum_{i=1}^{\ell_I-1} s_i\right)^{\frac{n_{I,\ell_I}}{p_I}-1} \prod_{i=1}^{\ell_I-1} s_i^{\frac{1}{p_I}-1} d\boldsymbol{s}_{\ell_I-1} \\ &= \frac{1}{p_I^{\ell_I-1}} \prod_{k=1}^{\ell_I-1} B\left[\frac{\sum_{i=1}^{k} n_{I,k}}{p_I}, \frac{n_{I,k+1}}{p_I}\right] \\ &= \frac{1}{p_I^{\ell_I-1}} \prod_{k=1}^{\ell_I-1} B\left[\frac{k}{p_I}, \frac{1}{p_I}\right]\end{aligned}$$

by using the fact that the transformed integral has the form of an unnormalized Dirichlet distribution and, therefore, the value of the integral must equal its normalization constant.

Now, we solve the integral

$$\int_{f(\boldsymbol{x}_{1:n-\ell_I}, v_I)\le R} v_I^{n_I-1} dv_I d\boldsymbol{x}_{1:n-\ell_I}. \tag{15}$$

We carry this out in exactly the same manner as we solved the previous integral. We need only to make sure that we only contract nodes that have only leaves as children (remember that radii of contracted nodes become leaves) and we need to find a formula describing how the factors $v_I^{n_I-1}$ propagate through the tree.

For the latter, we first state the formula and then prove it via induction. For notational convenience let $\mathcal{J}$ denote the set of multi-indices corresponding to the contracted leaves, $\boldsymbol{x}_{\hat{\mathcal{J}}}$ the remaining coefficients of $\boldsymbol{x}$ and $\boldsymbol{v}_{\mathcal{J}}$ the vector of leaves resulting from contraction. The integral which is left to solve after integrating over all $\tilde{\boldsymbol{u}}$ is given by (remember that n_J denotes real leaves, that is, the ones corresponding to coefficients of $\boldsymbol{x}$):

$$\int_{f(\boldsymbol{x}_{\hat{\mathcal{J}}}, \boldsymbol{v}_{\mathcal{J}})\le R} \prod_{J\in\mathcal{J}} v_J^{n_J-1} d\boldsymbol{v}_{\mathcal{J}} d\boldsymbol{x}_{\hat{\mathcal{J}}}.$$

We already proved the first induction step by computing Equation (15). For computing the general induction step suppose I is an inner node whose children are leaves or contracted leaves. Let $\mathcal{J}'$ be the set of contracted leaves under I and $\mathcal{K} = \mathcal{J}\backslash\mathcal{J}'$. Transforming the children of I into radial

coordinates by Gupta and Song (1997) yields

$$\begin{aligned}
\int_{f(\boldsymbol{x}_{\widehat{\mathcal{J}}},\boldsymbol{v}_{\mathcal{J}})\leq R}\prod_{J\in\mathcal{J}} v_J^{n_J-1}d\boldsymbol{v}_{\mathcal{J}}d\boldsymbol{x}_{\widehat{\mathcal{J}}} &= \int_{f(\boldsymbol{x}_{\widehat{\mathcal{J}}},\boldsymbol{v}_{\mathcal{J}})\leq R}\left(\prod_{K\in\mathcal{K}} v_K^{n_K-1}\right)\cdot\left(\prod_{J'\in\mathcal{J}'} v_{J'}^{n_{J'}-1}\right)d\boldsymbol{v}_{\mathcal{J}}d\boldsymbol{x}_{\widehat{\mathcal{J}}}\\
&= \int_{f(\boldsymbol{x}_{\widehat{\mathcal{K}}},\boldsymbol{v}_{\mathcal{K}},v_I)\leq R}\int_{\tilde{\boldsymbol{u}}_{\ell_I-1}\in\mathcal{V}_+^{\ell_I-1}}\left(\left(1-\sum_{i=1}^{\ell_I-1}\tilde{u}_i^{p_I}\right)^{\frac{1-p_I}{p_I}} v_I^{\ell_I-1}\right)\cdot\left(\prod_{K\in\mathcal{K}} v_K^{n_K-1}\right)\\
&\times\left(\left(v_I\left(1-\sum_{i=1}^{\ell_I-1}\tilde{u}_i^{p_I}\right)^{\frac{1}{p_I}}\right)^{n_{\ell_I}-1}\prod_{k=1}^{\ell_I-1}(v_I\tilde{u}_k)^{n_k-1}\right)d\boldsymbol{x}_{\widehat{\mathcal{K}}}d\boldsymbol{v}_{\mathcal{K}}dv_Id\tilde{\boldsymbol{u}}_{\ell_I-1}\\
&= \int_{f(\boldsymbol{x}_{\widehat{\mathcal{K}}},\boldsymbol{v}_{\mathcal{K}},v_I)\leq R}\int_{\tilde{\boldsymbol{u}}_{\ell_I-1}\in\mathcal{V}_+^{\ell_I-1}}\left(\prod_{K\in\mathcal{K}} v_K^{n_K-1}\right)\\
&\times\left(v_I^{\ell_I-1+\sum_{i=1}^{\ell_I}(n_i-1)}\left(1-\sum_{i=1}^{\ell_I-1}\tilde{u}_i^{p_I}\right)^{\frac{n_{\ell_I}-p_I}{p_I}}\prod_{k=1}^{\ell_I-1}\tilde{u}_k^{n_k-1}\right)d\boldsymbol{x}_{\widehat{\mathcal{K}}}d\boldsymbol{v}_{\mathcal{K}}dv_Id\tilde{\boldsymbol{u}}_{\ell_I-1}\\
&= \int_{f(\boldsymbol{x}_{\widehat{\mathcal{K}}},\boldsymbol{v}_{\mathcal{K}},v_I)\leq R}\left(\prod_{K\in\mathcal{K}} v_K^{n_K-1}\right)v_I^{n_I-1}d\boldsymbol{x}_{\widehat{\mathcal{K}}}d\boldsymbol{v}_{\mathcal{K}}dv_I\\
&\times\int_{\tilde{\boldsymbol{u}}_{\ell_I-1}\in\mathcal{V}_+^{\ell_I-1}}\left(1-\sum_{i=1}^{\ell_I-1}\tilde{u}_i^{p_I}\right)^{\frac{n_{\ell_I}-p_I}{p_I}}\prod_{k=1}^{\ell_I-1}\tilde{u}_k^{n_k-1}d\tilde{\boldsymbol{u}}_{\ell_I-1}.
\end{aligned}$$

Again, by transforming it into a Dirichlet distribution, the latter integral has the solution

$$\int_{\tilde{\boldsymbol{u}}_{\ell_I-1}\in\mathcal{V}_+^{\ell_I-1}}\left(1-\sum_{i=1}^{\ell_I-1}\tilde{u}_i^{p_I}\right)^{\frac{n_{\ell_I}-p_I}{p_I}}\prod_{k=1}^{\ell_I-1}\tilde{u}_k^{n_k-1}d\tilde{\boldsymbol{u}}_{\ell_I-1}=\prod_{k=1}^{\ell_I-1}B\left[\frac{\sum_{i=1}^{k}n_{I,k}}{p_I},\frac{n_{I,k+1}}{p_I}\right]$$

while the remaining former integral has the form

$$\int_{f(\boldsymbol{x}_{\widehat{\mathcal{K}}},\boldsymbol{v}_{\mathcal{K}},v_I)\leq R}\left(\prod_{K\in\mathcal{K}} v_K^{n_K-1}\right)v_I^{n_I-1}d\boldsymbol{x}_{\widehat{\mathcal{K}}}d\boldsymbol{v}_{\mathcal{K}}dv_I=\int_{f(\boldsymbol{x}_{\widehat{\mathcal{J}}},\boldsymbol{v}_{\mathcal{J}})\leq R}\prod_{J\in\mathcal{J}} v_J^{n_J-1}d\boldsymbol{v}_{\mathcal{J}}d\boldsymbol{x}_{\widehat{\mathcal{J}}}$$

as claimed.

By carrying out the integration up to the root node, the remaining integral becomes

$$\int_{v_\emptyset\leq R} v_\emptyset^{n-1}dv_\emptyset=\int_0^R v_\emptyset^{n-1}dv_\emptyset=\frac{R^n}{n}.$$

Collecting the factors from integration over the $\tilde{\boldsymbol{u}}$ proves the Equations (5) and (7). Using $B[a,b]=\frac{\Gamma[a]\Gamma[b]}{\Gamma[a+b]}$ yields Equations (6) and (8). ■

Appendix C. Layer Marginals

Proof [Proposition 7]

$$\begin{aligned}\rho(\boldsymbol{x}) &= \frac{\phi(f(\boldsymbol{x}))}{\mathcal{S}_f(f(\boldsymbol{x}))} \\ &= \frac{\phi(f(\boldsymbol{x}_{1:n-\ell_I}, v_I, \tilde{\boldsymbol{u}}_{\ell_I-1}, \Delta_n))}{\mathcal{S}_f(f(\boldsymbol{x}))} \cdot v_I^{\ell_I-1} \left(1 - \sum_{i=1}^{\ell_I-1} |\tilde{u}_i|^{p_I}\right)^{\frac{1-p_I}{p_I}}\end{aligned}$$

where $\Delta_n = \operatorname{sign}(x_n)$. Note that f is invariant to the actual value of Δ_n. However, when integrating it out, it yields a factor of 2. Integrating out $\tilde{\boldsymbol{u}}_{\ell_I-1}$ and Δ_n now yields

$$\begin{aligned}\rho(\boldsymbol{x}_{1:n-\ell_I}, v_I) &= \frac{\phi(f(\boldsymbol{x}_{1:n-\ell_I}, v_I))}{\mathcal{S}_f(f(\boldsymbol{x}))} \cdot v_I^{\ell_I-1} \frac{2^{\ell_I}\Gamma^{\ell_I}\left[\frac{1}{p_I}\right]}{p_I^{\ell_I-1}\Gamma\left[\frac{\ell_I}{p_I}\right]} \\ &= \frac{\phi(f(\boldsymbol{x}_{1:n-\ell_I}, v_I))}{\mathcal{S}_f(f(\boldsymbol{x}_{1:n-\ell_I}, v_I))} \cdot v_I^{\ell_I-1}\end{aligned}$$

Now, we can go on and integrate out more subtrees. For that purpose, let $\boldsymbol{x}_{\hat{\mathcal{J}}}$ denote the remaining coefficients of $\boldsymbol{x}$, $\boldsymbol{v}_{\mathcal{J}}$ the vector of leaves resulting from the kind of contraction just shown for v_I, and $\mathcal{J}$ the set of multi-indices corresponding to the "new leaves", that is, node v_I after contraction. We obtain the following equation

$$\rho(\boldsymbol{x}_{\hat{\mathcal{J}}}, \boldsymbol{v}_{\mathcal{J}}) = \frac{\phi(f(\boldsymbol{x}_{\hat{\mathcal{J}}}, \boldsymbol{v}_{\mathcal{J}}))}{S_f(f(\boldsymbol{x}_{\hat{\mathcal{J}}}, \boldsymbol{v}_{\mathcal{J}}))} \prod_{J \in \mathcal{J}} v_J^{n_J-1}.$$

where n_J denotes the number of leaves in the subtree under the node J. The calculations for the proof are basically the same as the one for proposition (4). ■

Appendix D. Factorial L_p-Nested Distributions

Proof [Proposition 9] Since the single x_i are independent, $f_1(\boldsymbol{x}_1), \ldots, f_{\ell_\emptyset}(\boldsymbol{x}_{\ell_\emptyset})$ and, therefore, $v_1, \ldots, v_{\ell_\emptyset}$ must be independent as well ($\boldsymbol{x}_i$ are the elements of $\boldsymbol{x}$ in the subtree below the ith child of the root node). Using Corollary 8 we can write the density of $v_1, \ldots, v_{\ell_\emptyset}$ as (the function name g is unrelated to the usage of the function g above)

$$\rho(\boldsymbol{v}_{1:\ell_\emptyset}) = \prod_{i=1}^{\ell_\emptyset} h_i(v_i) = g(\|\boldsymbol{v}_{1:\ell_\emptyset}\|_{p_\emptyset}) \prod_{i=1}^{\ell_\emptyset} v_i^{n_i-1}$$

with

$$g(\|\boldsymbol{v}_{1:\ell_\emptyset}\|_{p_\emptyset}) = \frac{p_\emptyset^{\ell_\emptyset-1}\Gamma\left[\frac{n}{p_\emptyset}\right]}{\|\boldsymbol{v}_{1:\ell_\emptyset}\|_{p_\emptyset}^{n-1} 2^m \prod_{k=1}^{\ell_\emptyset} \Gamma\left[\frac{n_k}{p_\emptyset}\right]} \phi(\|\boldsymbol{v}_{1:\ell_\emptyset}\|_{p_\emptyset})$$

Since the integral over g is finite, it follows from Sinz et al. (2009a) that g has the form $g(\|\boldsymbol{v}_{1:\ell_\emptyset}\|_{p_\emptyset}) = \exp(a_\emptyset \|\boldsymbol{v}_{1:\ell_\emptyset}\|_{p_\emptyset}^{p_\emptyset} + b_\emptyset)$ for appropriate constants $a_\emptyset$ and $b_\emptyset$. Therefore, the marginals have the form

$$h_i(v_i) = \exp(a_\emptyset v_i^{p_\emptyset} + c_\emptyset) v_i^{n_i - 1}. \tag{16}$$

On the other hand, the particular form of g implies that the radial density has the form $\phi(f(\boldsymbol{x})) \propto f(\boldsymbol{x})^{(n-1)} \exp(a_\emptyset f(\boldsymbol{x})^{p_\emptyset} + b_\emptyset)^{p_\emptyset}$. In particular, this implies that the root node's children $f_i(\boldsymbol{x}_i)$ ($i = 1, \ldots, \ell_\emptyset$) are independent and L_p-nested symmetric again. With the same argument as above, it follows that their children $\boldsymbol{v}_{i,1:\ell_i}$ follow the distribution $\rho(v_{i,1}, \ldots, v_{i,\ell_i}) = \exp(a_i \|\boldsymbol{v}_{i,1:\ell_i}\|_{p_i}^{p_i} + b_i) \prod_{j=1}^{\ell_i} v_{i,j}^{n_{i,j}-1}$. Transforming that distribution to L_p-spherically symmetric polar coordinates $v_i = \|\boldsymbol{v}_{i,1:\ell_i}\|_{p_i}$ and $\tilde{\boldsymbol{u}} = \boldsymbol{v}_{i,1:\ell_i-1} / \|\boldsymbol{v}_{i,1:\ell_i}\|_{p_i}$ as in Gupta and Song (1997), we obtain the form

$$\begin{aligned}
\rho(v_i, \tilde{\boldsymbol{u}}) &= \exp(a_i v_i^{p_i} + b_i) v_i^{\ell_i - 1} \left(1 - \sum_{j=1}^{\ell_i - 1} |\tilde{u}_j|^{p_i}\right)^{\frac{1-p_i}{p_i}} \left(v_i \left(1 - \sum_{j=1}^{\ell_i - 1} |\tilde{u}_j|^{p_i}\right)^{\frac{1}{p_i}}\right)^{n_{i,\ell_i} - 1} \prod_{j=1}^{\ell_i - 1} (\tilde{u}_j v_i)^{n_{i,j}-1} \\
&= \exp(a_i v_i^{p_i} + b_i) v_i^{n_i - 1} \left(1 - \sum_{j=1}^{\ell_i - 1} |\tilde{u}_j|^{p_i}\right)^{\frac{n_{i,\ell_i} - p_i}{p_i}} \prod_{j=1}^{\ell_i - 1} \tilde{u}_j^{n_{i,j}-1},
\end{aligned}$$

where the second equation follows the same calculations as in the proof of Proposition 4. After integrating out $\tilde{\boldsymbol{u}}$, assuming that the x_i are statistically independent, we obtain the density of v_i which is equal to (16) if and only if $p_i = p_\emptyset$. However, if $p_\emptyset$ and p_i are equal, the hierarchy of the L_p-nested function shrinks by one layer since p_i and $p_\emptyset$ cancel themselves. Repeated application of the above argument collapses the complete L_p-nested tree until one effectively obtains an L_p-spherical function. Since the only factorial L_p-spherically symmetric distribution is the p-generalized Normal (Sinz et al., 2009a) the claim follows. ■

Appendix E. Determinant of the Jacobian for NRF

Proof [Lemma 11] The proof is a generalization of the proof of Lyu and Simoncelli (2009). Due to the chain rule the Jacobian of the entire transformation is the multiplication of the Jacobians for each single step, that is, the rescaling of a subset of the dimensions for one single inner node. The Jacobian for the other dimensions is simply the identity matrix. Therefore, the determinant of the Jacobian for each single step is the determinant for the radial transformation on the respective dimensions. We show how to compute the determinant for a single step.

Assume that we reached a particular node I in Algorithm 2. The leaves, which have been rescaled by the preceding steps, are called $\boldsymbol{t}_I$. Let $\boldsymbol{\xi}_I = \frac{g_I(f_I(\boldsymbol{t}_I))}{f_I(\boldsymbol{t}_I))} \cdot \boldsymbol{t}_I$ with $g_I(r) = (\mathcal{F}_\perp^{-1} \circ \mathcal{F}_s)(r)$. The general form of a single Jacobian is

$$\frac{\partial \boldsymbol{\xi}_I}{\partial \boldsymbol{t}_I} = \boldsymbol{t}_I \cdot \frac{\partial}{\partial \boldsymbol{t}_I} \left(\frac{g_I(f_I(\boldsymbol{t}_I))}{f_I(\boldsymbol{t}_I)}\right) + \frac{g_I(f_I(\boldsymbol{t}_I))}{f_I(\boldsymbol{t}_I)} I_{n_I},$$

where

$$\frac{\partial}{\partial \boldsymbol{t}_I} \left(\frac{g_I(f_I(\boldsymbol{t}_I))}{f_I(\boldsymbol{t}_I)}\right) = \left(\frac{g_I'(f_I(\boldsymbol{t}_I))}{f_I(\boldsymbol{t}_I)} - \frac{g_I(f_I(\boldsymbol{t}_I))}{f_I(\boldsymbol{t}_I)^2}\right) \frac{\partial}{\partial \boldsymbol{t}_I} f_I(\boldsymbol{t}_I).$$

Let y_i be a leave in the subtree under I and let $I, J_1, ..., J_k$ be the path of inner nodes from I to y_i, then

$$\frac{\partial}{\partial y_i} f_I(\boldsymbol{t}_I) = v_I^{1-p_I} v_{J_1}^{p_I - p_{J_1}} \cdot ... \cdot v_k^{p_{J_{k-1}} - p_{J_k}} |y_i|^{p_{J_k}-1} \cdot \operatorname{sgn} y_i.$$

If we denote $r = f_I(\boldsymbol{t}_I)$ and $\zeta_i = v_{J_1}^{p_I - p_{J_1}} \cdot ... \cdot v_k^{p_{J_{k-1}} - p_{J_k}} |y_i|^{p_{J_k}-1} \cdot \operatorname{sgn} y_i$ for the respective J_k, we obtain

$$\det\left(\boldsymbol{t}_I \cdot \frac{\partial}{\partial \boldsymbol{t}_I}\left(\frac{g_I(f_I(\boldsymbol{t}_I))}{f_I(\boldsymbol{t}_I)}\right) + \frac{g_I(f_I(\boldsymbol{t}_I))}{f_I(\boldsymbol{t}_I)} I_{n_I}\right) = \det\left(\left(g_I'(r) - \frac{g_I(r)}{r}\right) r^{-p_I} \boldsymbol{t}_I \cdot \boldsymbol{\zeta}^\top + \frac{g_I(r)}{r} I_{n_I}\right).$$

Now we can use Sylvester's determinant formula $\det(I_n + b\boldsymbol{t}_I\boldsymbol{\zeta}^\top) = \det(1 + b\boldsymbol{t}_I^\top\boldsymbol{\zeta}) = 1 + b\boldsymbol{t}_I^\top\boldsymbol{\zeta}$ or equivalently

$$\begin{aligned}
\det(aI_n + b\boldsymbol{t}_I\boldsymbol{\zeta}^\top) &= \det\left(a \cdot \left(I_n + \frac{b}{a}\boldsymbol{t}_I\boldsymbol{\zeta}^\top\right)\right) \\
&= a^n \det\left(I_n + \frac{b}{a}\boldsymbol{t}_I\boldsymbol{\zeta}^\top\right) \\
&= a^{n-1}(a + b\boldsymbol{t}_I^\top\boldsymbol{\zeta}),
\end{aligned}$$

as well as $\boldsymbol{t}_I^\top\boldsymbol{\zeta} = f_I(\boldsymbol{t}_I)^{p_I} = r^{p_I}$ to see that

$$\begin{aligned}
\det\left(\left(g_I'(r) - \frac{g_I(r)}{r}\right) r^{-p_I}\boldsymbol{t}_I \cdot \boldsymbol{\zeta}^\top + \frac{g_I(r)}{r} I_n\right) &= \frac{g_I(r)^{n-1}}{r^{n-1}} \det\left(\left(g_I'(r) - \frac{g_I(r)}{r}\right) r^{-p_I}\boldsymbol{t}_I^\top \cdot \boldsymbol{\zeta} + \frac{g_I(r)}{r}\right) \\
&= \frac{g_I(r)^{n-1}}{r^{n-1}} \det\left(g_I'(r) - \frac{g_I(r)}{r} + \frac{g_I(r)}{r}\right) \\
&= \frac{g_I(r)^{n-1}}{r^{n-1}} \frac{d}{dr} g_I(r).
\end{aligned}$$

$\frac{d}{dr} g_I(r)$ is readily computed via $\frac{d}{dr} g_I(r) = \frac{d}{dr}(\mathcal{F}_{\perp\perp}^{-1} \circ \mathcal{F}_s)(r) = \frac{\phi_s(r)}{\phi_{\perp\perp}(g_I(r))}$.

Multiplying the single determinants along with $\det W$ for the final step of the chain rule completes the proof. ■

References

P-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton Univ Pr, Dec 2007. ISBN 0691132984.

M. Bethge. Factorial coding of natural images: How effective are linear model in removing higher-order dependencies? *J. Opt. Soc. Am. A*, 23(6):1253–1268, June 2006.

A. Edelman, T. A. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, 20(2):303–353, 1999. ISSN 0895-4798.

J. Eichhorn, F. Sinz, and M. Bethge. Natural image coding in v1: How much use is orientation selectivity? *PLoS Comput Biol*, 5(4), Apr 2009.

K. T. Fang, S. Kotz, and K. W. Ng. *Symmetric Multivariate and Related Distributions*. Chapman and Hall New York, 1990.

C. Fernandez, J. Osiewalski, and M.F.J. Steel. Modeling and inference with ν-spherical distributions. *Journal of the American Statistical Association*, 90(432):1331–1340, Dec 1995. URL http://www.jstor.org/stable/2291523.

A.K. Gupta and D. Song. L_p-norm spherical distribution. *Journal of Statistical Planning and Inference*, 60:241–260, 1997.

A.E. Hoerl. Application of ridge analysis to regression problems. *Chemical Engineering Progress*, 58(3):54–59, 1962.

A. Hyvärinen and P. Hoyer. Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Comput.*, 12(7):1705–1720, 2000.

A. Hyvärinen and U. Köster. Fastisa: A fast fixed-point algorithm for independent subspace analysis. In *Proc. of ESANN*, pages 371–376, 2006.

A. Hyvärinen and U. Köster. Complex cell pooling and the statistics of natural images. *Network: Computation in Neural Systems*, 18(2):81–100, 2007.

A. Hyvärinen and Erkki O. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492, Oct 1997. doi: 10.1162/neco.1997.9.7.1483.

D. Kelker. Distribution theory of spherical distributions and a location-scale parameter generalization. *Sankhya: The Indian Journal of Statistics, Series A*, 32(4):419–430, Dec 1970. doi: 10.2307/25049690. URL http://www.jstor.org/stable/25049690.

M. Kowalski, E. Vincent, and R. Gribonval. Under-determined source separation via mixed-norm regularized minimization. In *Proceedings of the European Signal Processing Conference*, 2008.

TW. Lee and M. Lewicki. The generalized gaussian mixture model using ica. In P. Pajunen and J. Karhunen, editors, *ICA' 00*, pages 239–244, Helsinki, Finland, june 2000.

M. S. Lewicki. Efficient coding of natural sounds. *Nat Neurosci*, 5(4):356–363, Apr 2002. doi: 10.1038/nn831.

M.S. Lewicki and B.A. Olshausen. Probabilistic framework for the adaptation and comparison of image codes. *J. Opt. Soc. Am. A*, 16:1587–1601, 1999.

S. Lyu and E. P. Simoncelli. Nonlinear extraction of independent components of natural images using radial gaussianization. *Neural Computation*, 21(6):1485–1519, Jun 2009. doi: 10.1162/neco.2009.04-08-773.

J. H. Manton. Optimization algorithms exploiting unitary constraints. *IEEE Transactions on Signal Processing*, 50:635 – 650, 2002.

B.A. Olshausen and D.J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:560–561, 1996.

J. Osiewalski and M. F. J. Steel. Robust bayesian inference in l_q-spherical models. *Biometrika*, 80 (2):456–460, Jun 1993. URL http://www.jstor.org/stable/2337215.

M. W. Seeger. Bayesian inference and optimal design for the sparse linear model. *Journal of Machine Learning Research*, 9:759–813, 04 2008. URL http://www.jmlr.org/papers/volume9/seeger08a/seeger08a.pdf.

E.P. Simoncelli. Statistical models for images: compression, restoration and synthesis. In *Conference Record of the Thirty-First Asilomar Conference on Signals, Systems & Computers, 1997.*, volume 1, pages 673–678 vol.1, 1997. doi: 10.1109/ACSSC.1997.680530.

F. Sinz and M. Bethge. The conjoint effect of divisive normalization and orientation selectivity on redundancy reduction. In D. Schuurmans Y. Bengio L. Bottou Koller, D., editor, *Twenty-Second Annual Conference on Neural Information Processing Systems*, pages 1521–1528, Red Hook, NY, USA, 06 2009. Curran. URL http://nips.cc/Conferences/2008/.

F. Sinz, S. Gerwinn, and M. Bethge. Characterization of the p-generalized normal distribution. *Journal of Multivariate Analysis*, 100(5):817–820, May 2009a. doi: 10.1016/j.jmva.2008.07.006.

F. Sinz, E. P. Simoncelli, and M. Bethge. Hierarchical modeling of local image features through L_p-nested symmetric distributions. In *Twenty-Third Annual Conference on Neural Information Processing Systems*, pages 1–9, 12 2009b. URL http://nips.cc/Conferences/2009/.

D. Song and A.K. Gupta. L_p-norm uniform distribution. *Proceedings of the American Mathematical Society*, 125:595–601, 1997.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 00359246. URL http://www.jstor.org/stable/2346178.

M.J. Wainwright and E.P. Simoncelli. Scale mixtures of Gaussians and the statistics of natural images. In S.A. Solla, T.K. Leen, and K.-R. Müller, editors, *Adv. Neural Information Processing Systems (NIPS*99)*, volume 12, pages 855–861, Cambridge, MA, May 2000. MIT Press.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*, 68(1):49–67, 2006.

C. Zetzsche, B. Wegmann, and E. Barth. Nonlinear aspects of primary vision: entropy reduction beyond decorrelation. In *Int'l Symposium, Soc. for Information Display*, volume XXIV, pages 933–936. 1993.

L. Zhang, A. Cichocki, and S. Amari. Self-adaptive blind source separation based on activation functions adaptation. *Neural Networks, IEEE Transactions on*, 15:233–244, 2004.

P. Zhao, G. Rocha, and B. Yu. Grouped and hierarchical model selection through composite absolute penalties. *Annals of Statistics*, 2008.

4.9 Lower bounds on the redundancy of natural images: Original Article

Vision Research 50 (2010) 2213–2222

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

Lower bounds on the redundancy of natural images

Reshad Hosseini, Fabian Sinz, Matthias Bethge *

Max Planck Institute for Biological Cybernetics, Spemannstraße 41, 72076 Tübingen, Germany

ARTICLE INFO

Article history:
Received 23 October 2009
Received in revised form 28 July 2010

Keywords:
Redundancy
Natural image statistics
Multi-information rate
Information theory

ABSTRACT

The light intensities of natural images exhibit a high degree of redundancy. Knowing the exact amount of their statistical dependencies is important for biological vision as well as compression and coding applications but estimating the total amount of redundancy, the multi-information, is intrinsically hard. The common approach is to estimate the multi-information for patches of increasing sizes and divide by the number of pixels. Here, we show that the limiting value of this sequence—the multi-information rate—can be better estimated by using another limiting process based on measuring the *mutual information* between a pixel and a causal neighborhood of increasing size around it. Although in principle this method has been known for decades, its superiority for estimating the multi-information rate of natural images has not been fully exploited yet. Either method provides a lower bound on the multi-information rate, but the *mutual information* based sequence converges much faster to the multi-information rate than the conventional method does. Using this fact, we provide improved estimates of the multi-information rate of natural images and a better understanding of its underlying spatial structure.

1. Introduction

Natural images contain an abundance of structure and regularities which can be quantified as statistical dependencies or redundancy between image pixels. Coding and compression algorithms for photographic images exploit these dependencies for achieving a good performance. Besides technical applications, the statistical regularities in natural images also play an important role for our understanding of sensory coding in the mammalian brain. In a wide range of studies it has been shown that many response properties of neurons in the early visual system such as color opponency, bandpass filtering, contrast gain control and orientation selectivity can be interpreted as mechanisms for removing these redundancies in natural images (Atick & Redlich, 1992; Barlow, 1959; Buchsbaum & Gottschalk, 1983; Karklin & Lewicki, 2008; Linsker, 1990; Olshausen & Field, 1996; Schwartz & Simoncelli, 2001; Simoncelli & Olshausen, 2001; Sinz & Bethge, 2009; Srinivasan, Laughlin, & Dubs, 1982). Quantitative comparisons have shown that these response properties are not all equally effective in removing statistical dependencies. Mechanisms removing second-order correlations in natural images such as color opponency and bandpass filtering yield a large reduction of redundancy. Less pronounced but still substantial is the effect of contrast gain control (Lyu & Simoncelli, 2009; Sinz & Bethge, 2009). For orientation selectivity, however, the potential for redundancy reduction turns out to be much smaller (Bethge, 2006). Since the emergence of orientation selectivity is the most prominent difference in the response properties of V1 neurons compared to the retina it can serve as an important witness on whether neural response properties in cortex can still be interpreted convincingly in terms of redundancy reduction (Eichhorn, Sinz, & Bethge, 2009).

An important unknown that is critical to judging this case is the true total amount of redundancy in natural images. A principled way of quantifying redundancy is to measure the *multi-information* of a distribution (Perez, 1977). The multi-information of a multivariate random variable is the difference between the sum of its marginal entropies and its joint entropy

$$I[X_1 : \ldots : X_n] = \sum_{i=1}^{n} H[X_i] - H[X_1, \ldots, X_n].$$

It equals zero if and only if the individual components are statistically independent and is positive otherwise. It measures the information gain caused by statistical dependencies between the single variables. Unlike differential entropy, the multi-information is invariant against arbitrary component-wise transformations both for linear mappings, such as scaling, and nonlinear mappings, such as taking the logarithm.

The conventional approach for estimating the redundancy per pixel—the *multi-information rate*—is to estimate the multi-information for patches of increasing sizes and divide by the number of pixels (Bethge, 2006; Chandler & Field, 2007; Eichhorn et al., 2009; Lee, Wachtler, & Sejnowski, 2002; Lewicki & Olshausen, 1999; Lewicki & Sejnowski, 2000; Lyu & Simoncelli, 2009; Sinz & Bethge, 2009; Wachtler, Lee, & Sejnowski, 2001). In this way we

* Corresponding author.
E-mail address: mbethge@tuebingen.mpg.de (M. Bethge).

0042-6989/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.visres.2010.07.025

obtain a monotonically increasing sequence converging to the multi-information rate

$$I_\infty = \lim_{n\to\infty} \frac{1}{n} I[X_1 : \ldots : X_n].$$

There is an important trade-off between two different kinds of errors that affect the outcome of this limiting process: On the one hand, the earlier we stop the sequence of increasing patch sizes, the more we ignore long-range dependencies between image pixels and, hence, underestimate the redundancy of natural images. On the other hand, the larger the patch sizes get, the more difficult it becomes to estimate the multi-information reliably due to the increase in dimensionality. Multi-information estimation strongly resembles the problem of estimating the joint density and similarly suffers from the curse of dimensionality: The number of states that need to be estimated grows exponentially with the number of dimensions. This means that more and more regularization is needed to avoid overfitting in high dimensions. As a consequence, with increasing dimensionality it becomes increasingly unlikely to capture all the structure of the density.

The trade-off between ignoring long range correlations for small n and the increasing difficulty to estimate $I[X_1 : \ldots : X_n]$ for large n suggests that the estimation of the multi-information rate can be improved substantially if one manages to construct sequences other than $\{\frac{1}{n} I[X_1 : \ldots : X_n]\}_{n=1}^{\infty}$ which converge faster to the same limiting value I_∞.

In this paper, we show that it is possible to construct such a sequence. The basic idea can be illustrated in the case of one-dimensional stationary stochastic processes. From information theory it is known that the conditional entropy converges to the entropy rate of such processes[1] (Cover & Thomas, 2006; Shannon, 1948)

$$\lim_{n\to\infty} \frac{1}{n} H[X_1, \ldots, X_n] = \lim_{n\to\infty} H[X_n | X_{n-1}, \ldots, X_1].$$

Multiplying this equation by (-1) and adding the marginal entropy of the stationary process $H[X_1] = \frac{1}{n}\sum_{k=1}^{n} H[X_k]$ at both sides, yields an analogous relationship for the multi-information rate

$$\begin{aligned} I_\infty &= \lim_{n\to\infty} \frac{1}{n} I[X_1 : \ldots : X_n] = \lim_{n\to\infty} I[X_n : X_{n-1}, \ldots, X_1] \\ &= \lim_{n\to\infty} H[X_n] - H[X_n | X_{n-1}, \ldots, X_1]. \end{aligned} \tag{1}$$

Note that the sequence on the left hand side of Eq. (1) reflects the *multi*-information[2] between all the variables $X_1, \ldots, X_n$ while the sequence on the right hand side reflects the *mutual* information between X_n and $(X_1, \ldots, X_{n-1})$. The mutual information is the special case of the multi-information which measures the statistical dependencies between two random variables only, while it is possible that the dimensionality of the two random variables is different. For example, in our case X_n is a univariate random variable and $(X_1, \ldots, X_{n-1})$ is $(n-1)$-dimensional. The chain rule for the multi-information (Cover & Thomas, 2006)

$$I[X_1 : \ldots : X_n] = \sum_{k=2}^{n} I[X_k : X_{k-1}, \ldots, X_1],$$

shows that the multi-information can be decomposed into a sum of mutual information terms. This suggests that the mutual information based sequence $\left\{I_n^{inc}\right\}_{n=1}^{\infty}$ with $I_n^{inc} := I[X_n : X_{n-1}, \ldots, X_1]$ quantifies the asymptotic *increment* in the multi-information while the conventionally used multi-information based sequence $\{I_n^{cum}\}_{n=1}^{\infty}$ with $I_n^{cum} := \frac{1}{n} I[X_n : \ldots : X_1]$ constitutes a *cumulative* approach which averages over these increments.

[1] For continuous random variables it is necessary to additionally assume that the limit exists.

[2] More precisely the multi-information divided by n.

Inspired by an early study in the fifties (Schreiber, 1956), an incremental approach for estimating I_∞ has already been used before in Petrov and Zhaoping (2003) but did not reveal its full potential. Our work elucidates a couple of points that have not been addressed in those papers: First, we revise the mathematical justification for using the incremental approach in case of two-dimensional random fields rather than one-dimensional processes as it is necessary for modeling images. Second, we show that the mutual information based method yields significantly better estimates of I_∞ than the conventional method does while Petrov and Zhaoping (2003) did not provide any comparisons with previous methods. Third, we show how particularly reliable multi-information estimators can be constructed for the incremental approach such that one obtains conservative lower bounds to the multi-information rate. This allows us, fourth, to systematically investigate how the two approaches perform on natural images for different number of dimensions n also far beyond the case of $n = 7$ pixels that was studied in Petrov and Zhaoping (2003). Our best lower bound on the multi-information rate for the van Hateren data set exceeds their estimate by more than 20% and slightly outperforms the bound obtained with the L_p-spherical model (Sinz & Bethge, 2009). It is obtained when using a causal neighborhood of only 25 pixels.

The remaining part of the paper is structured as follows: In Section 2, we introduce the *multi-information* based and the *mutual information* based method for estimating the multi-information rate. In particular, we present a proof for the convergence of the two methods to the same limiting value I_∞ for two-dimensional stationary stochastic processes. In Section 3, we perform experiments on artificial images in order to demonstrate the validity of the method, and apply it to natural images afterwards. Our results show that the incremental method based on conditional distributions performs significantly better and indicates that the multi-information rate of natural images contains a substantial contribution from higher-order moments. We further corroborate this finding by a second set of experiments where we first pre-whiten the images before we fit the local image statistics. In this way, we not only confirm our previous estimates for the multi-information rate but we can also show that the predominant statistical dependencies captured by current models of natural images are of very limited spatial extent. In particular, the increase in the multi-information rate observed for the cumulative method for increasing patch size does not reflect a meaningful contribution of long range correlations but rather an artifact caused by the pixels at the boundary. Finally, in Section 5, we discuss the significance of our results and compare them to existing work.

2. Methods

In order to describe the statistical regularities of natural images, they are often modeled as two-dimensional stationary random fields. For the present study, stationarity is crucial as it is provides the critical link between the *cumulative* and the *incremental* method for computing the multi-information rate. Stationarity means that the random field is invariant under translations with respect to the x- and y-coordinates of the image intensities. In the following, we will first depict the mathematical underpinnings for using the incremental approach in case of two-dimensional stationary random fields. After that we will show that the incremental method is generally superior to the cumulative method, and then we will describe how to construct reliable multi-information and mutual information estimators for the cumulative and the incremental method, respectively. In particular, we will construct conservative estimators such that also the empirical quantities become reliable lower bounds to the multi-information rate.

2.1. Mathematical underpinnings

Throughout the paper, we use uppercase letters to denote random variables, bold font to indicate vectors sometimes equipped with an subindex denoting the dimensionality. In particular, we write $I[\mathbf{X}_{1:n}]$ to refer to the multi-information $I[X_1 : \ldots : X_n]$ and $I[X_1 : \mathbf{X}_{2:n}]$ to refer to the mutual information between X_1 and $(X_2, \ldots, X_n)$.

For the incremental method, we estimate the multi-information rate I_∞ via the mutual information between X_n and $\mathbf{X}_{1:(n-1)}$ for increasing n

$$I_n^{inc} = I[X_n : \mathbf{X}_{1:(n-1)}] = H[X_n] - H[X_n|\mathbf{X}_{1:(n-1)}]. \quad (2)$$

As mentioned in the introduction, I_n^{inc} and I_n^{cum} converge to the true multi-information rate I_∞ for one-dimensional stationary stochastic processes. One subtle complication, hidden in the expression $H[X_n|\mathbf{X}_{1:(n-1)}]$, is that the proof for the one-dimensional case (see Cover & Thomas, 2006) uses stationarity to replace all conditional entropy terms $H[X_k|\mathbf{X}_{1:k-1}]$ in the chain rule decomposition of the joint entropy

$$H[\mathbf{X}_{1:n}] = H[X_1] + \sum_{k=2}^{n} H[X_k|\mathbf{X}_{1:k-1}] = H[X_n] + \sum_{k=2}^{n} H[X_n|\mathbf{X}_{n-k+1:n-1}],$$

with shifted versions $H[X_n|\mathbf{X}_{n-k+1:n-1}]$ where the index of each component is shifted by $(n-k)$. For two-dimensional Markov chains, however, the two-dimensional shape of the causal neighborhood (see Fig. 1) implies that there are always conditional entropy terms $H[X_k|\mathbf{X}_{1:k-1}]$ that cannot be matched by index shifting. Nevertheless, it is possible to show that $\left\{I_n^{inc}\right\}_{n=1}^{\infty}$ converges to the same limiting value I_∞ as $\{I_n^{cum}\}_{n=1}^{\infty}$ for all stationary random fields of arbitrary dimensions (Föllmer, 1973). In order to make this theorem more assessable we provide a simple proof for the special case of two dimensions in Appendix A.

2.2. Superiority of the incremental approach over the cumulative approach

Both types of limiting processes, the *cumulative*, multi-information based sequence $\{I_n^{cum}\}_{n=1}^{\infty}$ and the *incremental*, mutual information based sequence $\left\{I_n^{inc}\right\}_{n=1}^{\infty}$, grow monotonically with n and converge to the true multi-information rate from below. In other words, each sequence defines a lower bound on the multi-information rate that becomes increasingly tighter for large n and in the limit converges to the same value for the multi-information rate. Using the chain rule for the multi-information together with the fact that conditioning reduces entropy we further obtain the following relations

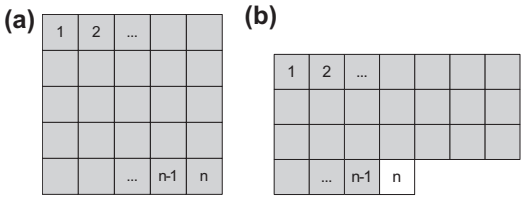


Fig. 1. Illustration of the shape of the image regions for the two different entropy estimation methods: (a) The square shaped patch used for estimating I_n^{cum}. (b) The causal neighborhood used for estimating I_n^{inc}. In this approach we compute the conditional distribution of the white pixel given the gray ones.

$$I_n^{cum} = \frac{1}{n} I[\mathbf{X}_{1:n}] < \frac{1}{n-1} I[\mathbf{X}_{1:n}] \equiv I_n^{cum*} = \frac{1}{n-1} \sum_{k=2}^{n} I[X_k : \mathbf{X}_{1:k-1}]$$
$$\leqslant \frac{1}{n-1} \sum_{k=2}^{n} I[X_n : \mathbf{X}_{1:n-1}] = I_n^{inc} \leqslant I_\infty. \quad (3)$$

First, this demonstrates that I_n^{cum*}, for which the total multi-information is divided by $(n-1)$, is a uniformly better approximation to I_∞ than the conventionally used I_n^{cum}, for which the multi-information is divided by n. While the difference between the two sequences decays very fast, $(I_n^{cum*} - I_n^{cum}) \sim 1/n^2$, the difference between the cumulative and the incremental sequence

$$I_n^{inc} - I_n^{cum*} = \frac{1}{n-1} \sum_{k=2}^{n} \left(I[X_n : \mathbf{X}_{1:n-1}] - I[X_k : \mathbf{X}_{1:k-1}]\right),$$

can be quite substantial also for moderately large n. It is zero if and only if $I[X_n{:}\mathbf{X}_{1:n-1}] = I[X_{n-1}{:}\mathbf{X}_{1:n-2}] = \cdots = I[X_2{:}X_1]$ which is equivalent to saying that the process is a stationary Markov process of order one. For all other processes, both cumulative sequences, I_n^{cum} and I_n^{cum*}, always underestimate the true multi-information rate for any finite n. In contrast, for the incremental, mutual information based sequence $\left\{I_n^{inc}\right\}_{n=1}^{\infty}$ it holds $I_n^{inc} = I_\infty$ for any Markov chain model if only the neighborhood $\mathbf{X}_{1:n-1}$ is sufficiently large (i.e. X_n conditioned on $\mathbf{X}_{1:n-1}$ is statistically independent of all other variables). In summary, for any given number of dimensions n, the incremental, mutual information based sequence in general yields better estimates of I_∞ than the cumulative, multi-information based one.

2.3. Cumulative (multi-information based) method

The cumulative method is commonly used for estimating the multi-information rate of natural images. For the sequence of the multi-information of image patches of increasing size we have

$$I_n^{cum} = \frac{1}{n} I[\mathbf{X}_{1:n}] = \frac{1}{n} \sum_{i=1}^{n} H[X_i] - H[\mathbf{X}_{1:n}]$$
$$= H[X_1] + \frac{1}{n} \langle \log p(\mathbf{x}_{1:n}) \rangle_{\mathbf{X}_{1:n}}$$
$$\geqslant -\langle \log p(x_1) \rangle_{X_1} + \frac{1}{n} \langle \log \hat{p}(\mathbf{x}_{1:n}) \rangle_{\mathbf{X}_{1:n}} \equiv \widehat{I}_n^{cum}, \quad (4)$$

where $\hat{p}$ denotes a particular model distribution.

In order to obtain an empirical estimate of I_n^{cum} we use the lower bound given by Eq. (4). The first term is the entropy $H[X_i]$ of the univariate marginal distribution over the pixel intensities which is the same for all $i = 1, \ldots, n$ due to stationarity. Since the problem of estimating this term is identical for both cases, the cumulative as well as the incremental approach, we will discuss it separately at the end of the method section.

The second term in the definition of our estimator $\widehat{I}_n^{cum}$ reflects the *average log-loss* (Bernardo, 1979)

$$-\langle \log \hat{p}(\mathbf{x}) \rangle_{\mathbf{X}_{1:n}} = H[\mathbf{X}_{1:n}] + D_{KL}[p\|\hat{p}] \geqslant H[\mathbf{X}_{1:n}],$$

where D_{KL} denotes the Kullback–Leibler divergence, a positive quantity that measures the mismatch between the true and the model distribution. Therefore, the average log-loss has the desirable property that any systematic mismatch between the model distribution $\hat{p}$ and the true distribution p will lead to overestimation of the joint entropy. In this way, we obtain a conservative estimate of the true multi-information rate I_∞.

For estimating the average log-loss, we follow (Eichhorn et al., 2009; Lewicki & Olshausen, 1999; Lewicki & Sejnowski, 2000) and use Monte-Carlo sampling

$$-\langle \log \hat{p}(\mathbf{x}) \rangle_{\mathbf{x}_{1:n}} \approx -\frac{1}{m} \sum_{i=1}^{m} \log \hat{p}(\mathbf{x}_i),$$

over a large ensemble of m samples $\mathbf{x}_i$ which differs from the training set used for fitting the parameters of $\hat{p}$.

2.4. Incremental (mutual information based) method

For the incremental approach we employ the same strategy as for the cumulative method: We use the average log-loss of a parametric density for estimating the conditional entropy in Eq. (2) in order to obtain a conservative estimator for I_n^{inc}. In principle, it would be nice to rewrite the conditional entropy in terms of the joint entropy again

$$\begin{aligned} H[X_n|\mathbf{X}_{1:(n-1)}] &= H[\mathbf{X}_{1:n}] - H[\mathbf{X}_{1:(n-1)}] \\ &\approx \frac{1}{n} \langle \log \hat{p}(\mathbf{x}_{1:(n-1)}) \rangle_{\mathbf{x}_{1:(n-1)}} - \frac{1}{n} \langle \log \hat{p}(\mathbf{x}_{1:n}) \rangle_{\mathbf{x}_{1:n}}, \end{aligned}$$

as it would allow one to use exactly the same parametric density model like in the cumulative method to estimate the joint entropies. The caveat, however, is that the upward bias in the error induced by using the average log-loss when estimating entropies can now occur in both directions.

Therefore, we resort to a different strategy, using the average log-loss directly for estimating the conditional entropy which again yields a lower bound

$$\widehat{I}_n^{inc} \equiv H[X_n] + \langle \log \hat{p}(x_n|\mathbf{x}_{1:(n-1)}) \rangle_{\mathbf{x}_{1:n}} \tag{5}$$

$$\leqslant H[X_n] - H[X_n|\mathbf{X}_{1:(n-1)}] = I_n^{inc}. \tag{6}$$

Therefore, we have to fit a conditional density model $\hat{p}(x_n|\mathbf{x}_{1:(n-1)})$ rather than a joint density model $\hat{p}(\mathbf{x}_{1:n})$ like in the cumulative approach.

2.5. Parametric density model

For the sake of better comparison, we will use the same Gaussian scale mixture (GSM) model to serve as the parametric model for the average log-loss estimators in both approaches. The GSM model is a rich subfamily of elliptical contoured distributions (Wainwright & Simoncelli, 2000) which have recently been demonstrated to provide a good fit to local patches of natural images (Eichhorn et al., 2009; Lyu & Simoncelli, 2009).

We use a variant of the GSM model which is defined as a mixture of a finite number of zero mean Gaussians with differently scaled versions of the same covariance matrix Σ:

$$p(\mathbf{x}) = \mathrm{GSM}(\mathbf{x}|\boldsymbol{s}, \Sigma, \boldsymbol{\lambda}) = \sum_{k=1}^{K} \lambda_k \cdot \mathcal{N}(\mathbf{x}|s_k \cdot \Sigma), \quad \boldsymbol{\lambda}, \boldsymbol{s} \in \mathbb{R}^K,$$

where the class probabilities λ_k sum up to one.

For parameter fitting we use an expectation maximization (EM) algorithm. To this end, we define the hidden variable Z indicating which scale is picked for a specific data point $\mathbf{x}$:

$$p_{\mathbf{X}|Z}(\mathbf{x}|k) = \mathcal{N}(\mathbf{x}|s_k \cdot \Sigma) \quad \text{and} \quad p_Z(k) = \lambda_k.$$

For the E-step, we need to compute the probability t_i^k that $Z = k$ given the ith data point

$$t_i^k = p_{Z|\mathbf{X}}(k|\mathbf{x}_i) = \frac{\lambda_k \mathcal{N}(\mathbf{x}_i|s_k \cdot \Sigma)}{\sum_{k=1}^{K} \lambda_k \mathcal{N}(\mathbf{x}_i|s_k \cdot \Sigma)}.$$

In the M-step, for given λ_k and t_i^k, $1 \leqslant k \leqslant K$, $1 \leqslant i \leqslant m$ we obtain

$$\lambda_k = \frac{\sum_{i=1}^{m} t_i^k}{\sum_{i=1}^{m} \sum_{k=1}^{K} t_i^k}.$$

For computing the scales and the covariance in the M-step, we need to maximize

$$\mathscr{L}(\boldsymbol{s}, \Sigma) = \sum_{i=1}^{m} \sum_{k=1}^{K} t_i^k \log \mathcal{N}(\mathbf{x}_i|s_k, \cdot \Sigma).$$

Since the maximum cannot be calculated analytically, we use a block coordinate descent approach. In the first step, we fix $\boldsymbol{s}$ and calculate Σ, in the second step, we fix Σ and calculate $\boldsymbol{s}$, using the equations

$$\Sigma = \frac{1}{m} \sum_{k=1}^{K} \sum_{i=1}^{m} \frac{t_i^k}{s_k} \mathbf{x}_i \mathbf{x}_i^\top \quad \text{and} \quad s_k = \frac{\sum_{i=1}^{m} t_i^k \mathbf{x}_i^\top \Sigma^{-1} \mathbf{x}_i}{K \sum_{i=1}^{m} t_i^k}.$$

In our simulations, we find that one or two iteration are enough for the covariance matrix and scale parameters to converge.

In order to use the same distribution for the second method, we calculate the conditional distribution from the GSM model for fixed parameters. This can be done analytically in the GSM model: Let the covariance matrix of Σ of $\mathrm{GSM}(\mathbf{x}_{1:n}|\boldsymbol{s}, \Sigma, \boldsymbol{\lambda})$ be

$$\Sigma = \begin{bmatrix} \Sigma_{1:(n-1),1:(n-1)} & \Sigma_{1:(n-1),n} \\ \Sigma_{1:(n-1),n} & \Sigma_{n,n} \end{bmatrix}.$$

Marginalizing out the random variable X_n again yields a GSM with parameters

$$\mathrm{GSM}\big(\mathbf{x}_{1:(n-1)}|\boldsymbol{s}_{1:(n-1)}, \Sigma_{1:(n-1),1:(n-1)}, \boldsymbol{\lambda}_{1:(n-1)}\big).$$

Then the conditional distribution is just the ratio between the original joint and the marginalized distribution:

$$p_{X_n|\mathbf{X}_{1:(n-1)}}(x_n|\mathbf{x}_{1:(n-1)}) = \frac{\mathrm{GSM}(\mathbf{x}_{1:n}|\boldsymbol{s}_{1:n}, \Sigma_{1:n,1:n}, \boldsymbol{\lambda}_{1:n})}{\mathrm{GSM}(\mathbf{x}_{1:(n-1)}|\boldsymbol{s}_{1:(n-1)}, \Sigma_{1:(n-1),1:(n-1)}, \boldsymbol{\lambda}_{1:(n-1)})}.$$

2.6. Estimation of the univariate pixel entropy

In order to minimize the risk of overestimating the univariate marginal entropy in either of the two approaches, we aim at using a very precise nonparametric approach. To this end we use a histogram based jackknifed maximum likelihood estimator (see e.g. Paninski, 2003). Given m samples with a marginal standard deviation of σ we chose the bin width Δ according to the heuristic proposed by Scott (1979): $\Delta = 3.49\sigma m^{-\frac{1}{3}}$. Since the discrete entropy asymptotically equals the differential entropy plus $-\log \Delta$, we obtain an estimate of the marginal entropy by adding the log of the bin width Δ. Using that method we reliably obtain a value of 1.57 bits per pixel for the univariate pixel entropy. Note that this number like all differential entropies depends on the scale of the pixel intensities. The multi-information rate, however, is independent of the scale as it is computed from differences between differential entropies.

3. Experiments

3.1. Experiment on artificial data

In order to illustrate the two estimation methods, we first compare the cumulative and the incremental approach on an artificial stationary Gaussian random field using the autocorrelation of natural images. To this end, we generated 10.000 images of 60×60 pixels by applying a linear transformation A to Gaussian white noise ξ such that the covariance matrix of the resulting Gaussian distribution $\Sigma = AA^\top$ resembles the covariance matrix of the van Hateren data set. We estimated the covariance matrix from samples of 60×60 patches using the fact that due to stationarity the covariance between two pixels at location (x, y) and location (x', y'), respectively, must only depend on their relative distance

$(x - x', y - y')$, which results in a symmetric block-Toeplitz covariance matrix.

From those images we sampled ten pairs of training and test sets of 1.000.000 image patches each, for a range of different patch sizes. Fig. 1 shows the shape of the image patch shape used in our two approaches. For the cumulative approach, we use patch sizes $2 \times 2, \ldots, 12 \times 12$. For the incremental approach, we use causal neighborhoods of sizes 5, 13, 25, 41, 61, 85, 113, 145. The parameter estimation for the models was done in exactly the same way as for the natural images below.

As a stationary Gaussian random field is completely defined by the autocorrelation function we can compute the multi-information and the mutual information analytically from $AA^\top$. Fig. 2a shows the result for the full range from 1 to 3600 pixels.

Fig. 2b shows the empirical results obtained for $\widehat{I}_n^{cum}$, $\widehat{I}_n^{cum*}$ and $\widehat{I}_n^{inc}$ as a function of the dimension N when using the average log-loss of a Gaussian model distribution. For comparison, the dashed black lines indicate the true multi-information rate I_∞ obtained analytically from the relevant submatrices Σ_n^{cum} and Σ_n^{inc} of the covariance matrix C needed to compute the multi-information bounds

$$I_n^{cum} = \frac{1}{2n}\left(\sum_{k=1}^{n} \log_2 (\Sigma_n^{cum})_{k,k} - \log_2 |\det (\Sigma_n^{cum})|\right),$$

$$I_n^{inc} = \frac{1}{2}\left(\log_2 \sigma_n^2 - \log_2 \sigma^2_{n|1:(n-1)}\right),$$

respectively, where

$$\sigma_n^2 := \left(\Sigma_n^{inc}\right)_{n,n},$$

$$\sigma^2_{n|1:(n-1)} := \sigma_n^2 - \left(\Sigma_n^{inc}\right)_{n,1:(n-1)} \left(\Sigma_n^{inc}\right)^{-1}_{1:(n-1),1:(n-1)} \left(\Sigma_n^{inc}\right)_{1:(n-1),n}.$$

The example visualizes the superiority of the incremental method over the cumulative method. The agreement between the analytical and empirical curves illustrates that the difference between the two methods is not caused by insufficient amount of data or by wrong model assumptions but solely by an unavoidable downward bias of the cumulative method. As apparent from Eq. (3), this downward bias originates from the fact that pixels close to the boundaries suffer from an incomplete neighborhood. Therefore, they do not contribute the full amount of redundancy to the multi-information rate and it requires very large image patches until the pixels in the interior can sufficiently outnumber the pixels at the boundaries. Even at a patch size of 60×60 the cumulative method still underestimates the asymptotic information rate of this stationary Gaussian random field by 0.02 bits per pixel. In other words, the convergence of the cumulative methods is extremely slow even though we are using the correct density model for the evaluation of the average log-loss.

3.2. Natural image dataset and parameter estimation

We perform two blocks of experiments with natural images. In the first block, we use images whose pixel values encode log-intensities. In the second block, we use pre-whitened images generated by a predictive coding scheme that subtracts from each pixel the optimal linear prediction from a causal neighborhood around it. In order to compute the multi-information for the original pixels, we have to account for the whitening transformation. As this whitening step can be described by a linear transform which has vanishing log-Jacobian in the limit, we can lower bound the multi-information rate by the difference of the marginal entropy (1.57 bits) on the pixel domain and the ALLs on the whitened domain:

$$\begin{aligned}
\widehat{I}_n^{inc} &= H[X_n] + \langle \log \hat{p}(y_n | \boldsymbol{y}_{1:(n-1)}) \rangle \\
&\leqslant H[X_n] - H[Y_n | \mathbf{Y}_{1:(n-1)}] \\
&\leqslant H[X_n] - \lim_{k\to\infty} H[Y_k | \mathbf{Y}_{1:(k-1)}] \\
&= H[X_n] - \lim_{k\to\infty} \frac{1}{k} H[\mathbf{Y}_{1:k}] \\
&\log |\underline{J}| = 0 H[X_n] - \lim_{k\to\infty} \frac{1}{k} H[\mathbf{X}_{1:k}] = I_\infty
\end{aligned}$$

where $\mathbf{Y}_{1:n}$ denotes the whitened pixels. This lower bound is equal to the multi-information estimate after whitening the images, plus the difference of original marginal pixel entropy and marginal pixel entropy of pre-whitened data, i.e.

$$\widehat{I}_n^{inc} = \underbrace{H[Y_n] - \langle \log \hat{p}(y_n | \boldsymbol{y}_{1:(n-1)}) \rangle}_{\text{MI estimate in second layer}} + \underbrace{H[X_n] - H[Y_n]}_{\text{marginal entropy difference}} . \tag{7}$$

The difference between the marginal entropies for the van Hateren dataset is equal to 2.9 bits.

For the experiments on natural images we used exactly the same amount of data as in the artificial example described above. That is for each patch size we sampled ten pairs of training and test sets of 1.000.000 log-intensity image patches from the van Hateren database (van Hateren & van der Schaaf, 1998). Again, for the cumulative approach, we use patch sizes $2 \times 2, \ldots, 12 \times 12$. For the incremental approach, we use causal neighborhoods of sizes

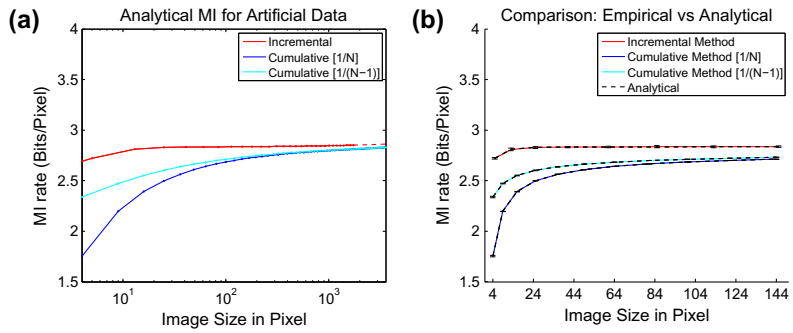


Fig. 2. Verification of the estimation methods on artificial data: Multi-information rate in bits per pixel as estimated by our two methods as a function of the number of pixels. The blue and cyan curves show the result for the cumulative method and the red curve shows the result of the incremental method which significantly outperforms the cumulative ones. The left figure (a) shows the analytic results for the full range of up to $n = 3600$ dimensions using a logarithmic x-axis. The right figure (b) shows an excellent agreement between the analytical and the empirically estimated lower bounds for both methods.

5, 13, 25, 41, 61, 85, 113, 145. For each patch size we run different versions of the GSM model with $K = 1, 4, 7, 10$ scale mixture components. All results shown for $\widehat{I}^{cum}$ and $\widehat{I}^{inc}$ are evaluations on the test set. Importantly, all evaluations on the training set yield identical results so that potential effects due to overfitting can be safely excluded. The error bars in all figures indicate *three* standard deviations over the ensemble of ten different test sets, apart from Fig. 6b where we used *two* standard deviations because of the smaller range of the *y*-axis.

4. Results

Fig. 3 shows the multi-information rate computed with the two different methods for the SGM with $K = 10$ scale mixture components. One can see from the figure that the incremental method significantly outperforms the cumulative one and provides a tighter lower bound.

Fig. 4 shows the estimated multi-information rates for the different methods and different numbers of scale mixture components. The performance seems to saturate for about seven mixture components.

For the incremental method, the lower bound takes a maximum at a neighborhood size of 25 pixels, whereas the cumulative method still exhibits a tiny increase of the lower bound at 144 pixels. This raises two questions:

(1) How can it be that the amount of dependencies captured with the incremental method is decreasing with increasing patch size?
(2) Could it be that the cumulative method is able to better capture long range interactions between pixels and hence at some point can yield a tighter lower bound when using very large image patches?

The first question is motivated by the fact that I_n^{cum} and I_n^{inc} can only increase with increasing patch or neighborhood size. As one can see from the Eqs. (4) and (5), however, the lower bounds $\widehat{I}_n^{cum}$ and $\widehat{I}_n^{inc}$ can still decrease with increasing n if the inequalities $\widehat{I}_n^{cum} \leqslant I_n^{cum}$ and $\widehat{I}_n^{inc} \leqslant I_n^{inc}$ become less and less tight. The differences between the true and the estimated quantities $I_n^{cum} - \widehat{I}_n^{cum}$ and $I_n^{inc} - \widehat{I}_n^{inc}$ equal the Kullback–Leibler distance between the true distribution and the model distribution. If the mismatch of the model distribution becomes larger for increasing patch size, this can result in a lower bound which decreases with increasing patch size.

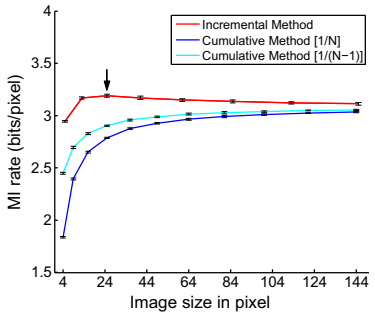


Fig. 3. Comparison of the cumulative and the incremental approach on natural images with $K = 1, 4, 7, 10$ scale mixture components. The blue and cyan curves show the result for the cumulative method and the red curve shows the result of the incremental method. Analogous to the results for artificial data, the incremental method significantly outperforms the cumulative ones. The arrow shows the maximum amount of multi-information estimated by the incremental method.

This is what we see in case of the incremental method. In case of small patch sizes, the GSM model can exploit higher-order correlations to model contrast dependencies between nearby pixels. In case of large image patches, however, the GSM model has to compromise between strong higher-order correlations between nearby pixels and weak higher-order correlations between distant pixels. Therefore, the model fit of the GSM becomes worse for larger patch sizes which causes the decrease in $\widehat{I}_n^{inc}$. In other words, the limited flexibility of the GSM model to capture the structure of higher-order correlations becomes increasingly severe with increasing dimensionality. For second-order correlations, however, this is different, because with a Gaussian distribution one can always fit any possible pattern of second-order correlations. Since in contrast to a general GSM, a single Gaussian distribution is always entirely ignorant against higher-order correlation, we do not see the effects of imperfect fitting of higher-order correlations in case of $K = 1$. For a Gaussian model, the lower bound can therefore only increase. This is nicely reflected in Fig. 4b: for $K = 1$ the lower bound always increases, whereas for $K \geqslant 4$ the lower bound decreases for large patch sizes.

Given that we explained the decrease of the lower bound for the incremental method with the limited flexibility of the GSM model, why do we not see a decrease for the cumulative method? We can explain this with the downward bias caused by the reduced contribution to the multi-information from pixels close to the patch boundaries. It is important to note that the persistent increase in case of the cumulative method does not originate from a better image model. Like in the artificial example, we fitted the same model distribution to optimally fit the *joint* distribution over the image pixels for the cumulative as well as for the incremental method. The crucial difference lies only in the way how we compute the lower bound to the asymptotic information rate from it. In one case we divide the total multi-information by the number of pixels and in the other case we compute the mutual information between one pixel and the rest by computing the conditional from the joint model. Therefore, the persistent increase up to $N = 144$ for the cumulative method does not reflect a better fit to the data but merely shows that the downward bias of the cumulative method for small image patches, for which the ratio of boundary to interior pixels is still large enough, is so substantial that it easily outbalances the decrease caused by degradation in the model fit.

Our second set of experiments on the pre-whitened images (see Section 3) further corroborates this explanation. The redundancy reduction caused by the pre-whitening is assessed as explained above and is the same for both methods. Therefore, after pre-whitening, all differences between the two methods can only originate from differences in assessing the contribution of higher-order correlations. Without the large contribution of second-order correlations, the downward bias for the cumulative method for small image patches becomes much smaller and hence, the effect of degradation in the model fit on the lower bound becomes more visible for the cumulative method as well. As can be seen in Fig. 5, the cumulative method now has a maximum as well at a patch size of 7×7 pixels. For the incremental method, the optimal neighborhood size is further reduced to $n = 13$. The type of higher-order correlations that can be captured by the GSM model are limited to variance (contrast) correlations between the different pixels. The fact that the lower bound takes its maximum for a very small neighborhood size shows that this type of correlations can be explained (away) by short range couplings.

Note that the curves shown include the contribution of second-order correlations that were removed during the pre-whitening step. The second-order contribution equals the lower bound obtained with the Gaussian distribution ($K = 1$) and is about 2.7 bits per pixel. Remarkably, the maximum lower bound determined with the pre-whitened images yields the same estimate for the

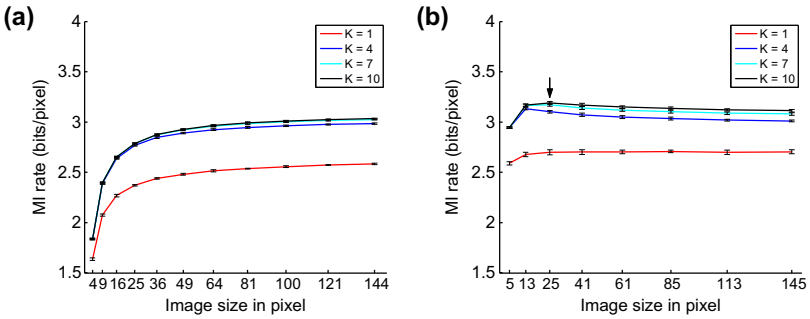


Fig. 4. Comparison of the multi-information rate estimates for different numbers of components ($K = 1, 4, 7, 10$). (a) Multi-Information rate estimated by the cumulative approach. (b) The same result using the incremental approach. In both cases the number of scale mixture components have similar effects and the performance seems to saturate for seven components. The arrow indicates the maximum amount of multi-information estimated by the incremental method.

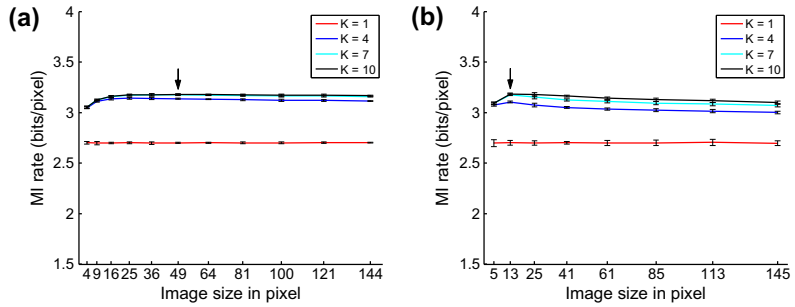


Fig. 5. Comparison of the multi-information rate estimates for different numbers of components ($K = 1, 4, 7, 10$) based on pre-whitened image data set. (a) Multi-Information rate estimated by the cumulative approach. (b) The same result using the incremental approach. Since the pre-whitening removes the downward bias of the cumulative method for the second-order contribution to the multi-information, it now has substantially improved and its lower bound—similarly to the incremental method—now takes a maximum for a relatively small patch size as well. The arrows indicate the maxima for both methods.

multi-information rate as the maximum lower bound obtained on the original images. This nicely underlines the reliability of our estimates.

As a final result we show how the incremental method can be further improved by improving the parameter fitting. As explained in Section 2, we always optimized the likelihood for the joint distribution and not for the conditional one. However, maximizing the likelihood for the joint model does not necessarily also maximize the likelihood for the conditional distribution which would be equivalent to minimizing the average log-loss of the conditional distribution. Based on Jebara's work on conditional expectation maximization (Jebara, 2002) we developed a new algorithm (see Appendix B) that we used to optimize the conditional likelihood for the GSM model. The result of this optimization is shown in Fig. 6a. In this way we obtained our best lower bound of 3.26 bits per pixel which is almost 0.6 bits larger than the multi-information rate obtained for a single Gaussian.

Fig. 6b shows the residual multi-information rate (see Eq. (7)) achieved by optimizing the conditional likelihood after pre-whitening (solid red). For comparison we also show the residual multi-information rate when optimizing for the joint likelihood (dashed) and the cumulative method (solid).

The large difference between the GSM using only a single mixture component and the GSMs with several ones is particularly interesting. Since the GSM in case of $K = 1$ is a plain Gaussian which is completely determined by its mean and its covariance matrix, the entropy rate of this GSM shows the contribution of the second-order moments to the total entropy of the signal. The fact that this difference is large shows the highly non-Gaussian behavior of natural images and, therefore, a substantial amount of higher-order correlations (Eichhorn et al., 2009; Chandler & Field, 2007; Ruderman & Bialek, 1994).

5. Summary and discussion

Measuring the total redundancy of natural images is a challenging task. In this paper we showed that the conventionally used cumulative method suffers from an unfavorable downward bias for small image patches. This problem can be avoided by using the incremental method. We compared the two methods for both artificial data and natural images, and demonstrated that the incremental method always yields a better lower bound on the multi-information rate.

As our method yields a conservative lower bound on the multi-information rate, we can safely conclude from our results that $I_\infty \geqslant 3.26$ bits per pixel for the van Hateren data set. This number is substantially larger than the 2.7 bits per pixel previously estimated by Petrov and Zhaoping (2003) who used very small neighborhoods only ($n = 7$). While they concluded that the total amount of higher-order correlations in natural images is small, the difference in the performance of the Gaussian model ($K = 1$) and the full GSM model ($K = 10$) suggests that the amount of higher-order correlations is at least 0.6 bits per pixel which we think is quite substantial.

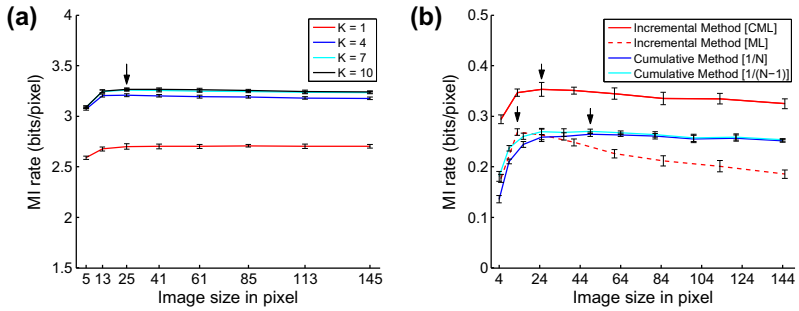


Fig. 6. Further improvement of the lower bound by optimizing the GSM model for the conditional likelihood. The arrow indicates the maximal amount of multi-information that was estimated. (a) shows the total multi-information rate while (b) shows the residual multi-information rate after the pre-whitening step (first term of Eq. (7)). The optimization of the conditional likelihood leads to a better fit of the conditional distribution and, hence, less degradation in the incremental method (solid vs. dashed red curve). It further corroborates the superiority of the incremental method above the cumulative method also for the pre-whitened data (red vs. other solid curves).

Using a less conservative nearest neighbor estimation method, Chandler and Field (2007) arrived at an information rate similar to ours. Taking the difference between the data points for Gaussian white noise and natural scenes in Fig. 14 in Chandler and Field (2007) would yield a multi-information rate estimate of about 3.1–3.3 bits per pixel. From their extrapolation in the same figure one obtains a multi-information rate of 3.7 bits per pixel in the limit.

In previous studies, we used the cumulative method together with an L_p-spherically symmetric model and an ICA model to estimate the redundancy reduction achieved by different neural response properties (Sinz & Bethge, 2009). The multi-information reported for ICA and the L_p-spherically symmetric model are 3.41 and 3.62 bits per pixel. Given that the multi-information estimates in Sinz and Bethge (2009) were obtained on a different dataset (Bristol Hyperspectral), the results are reasonably similar. We repeated the experiments of Sinz and Bethge (2009) and computed the values for the van Hateren dataset for 144 dimensions. We obtained 2.92 bits per pixel for ICA, 3.05 bits per pixels for the joint GSM, and 3.17 bits per pixel for the L_p-spherically symmetric model. This is better than the result of the cumulative method for the GSM but about 0.1 bits per pixel worse than the result of the incremental method. Thus, again the incremental method provides a better bound by using only 25 dimensions. The differences between the results for the Bristol Hyperspectral dataset and the van Hateren dataset are within the typical variations one observes for different image libraries. They mainly originate from variations in the second-order redundancies. In particular, the difference between the L_p-spherically symmetric model and ICA is very similar for both data sets: 0.21 bits per pixel for Bristol Hyperspectral and 0.25 for van Hateren.

In this study, we used the Gaussian scale mixture model for both the cumulative and the incremental approach for the sake of comparison. In the future we can make further advantage by using more sophisticated conditional density models that are optimally tailored to the incremental approach. It is interesting to note that the conditional distribution has a close link to the inverse of the auto-covariance matrix of random processes (the so called *precision matrix*). Typically, the precision matrix is much sparser and hence captures the conditional dependency structures much more efficiently than the covariance matrix (Rue & Held, 2005). In fact, for a Gaussian Markov random field, an entry of the precision matrix is non-zero if and only if the two points are conditionally dependent. When looking at the precision matrix for natural images, the number of components that have a value significantly larger than zero is typically very small and restricted to a very small neighborhood around that pixel.

In summary, we expect that the incremental method combined with an appropriate conditional density model will lead to major improvements in statistical modeling of natural images.

Acknowledgments

This work is supported by the German Ministry of Education, Science, Research and Technology through the Bernstein award to MB (BMBF; FKZ: 01GQ0601) and a scholarship of the German National Academic Foundation to FS. We thank Sebastian Gerwinn for comments on the manuscript.

Appendix A

Definition 1. (**Causal points**). Let the *causal points* of a particular point in a random field be all points that are above that particular point or at its left in the same row.

Definition 2. (**Causal neighborhood of radius** l). Let the *causal neighbors of radius* l of a particular point be all causal points which their horizontal and vertical distance from that particular point being smaller or equal to l (see Fig. 1b for an example of a causal neighborhood of radius 3).

Theorem 1. (**Convergence of entropy rate for 2D stationary process**). *The sequence of conditional entropies with causal neighborhoods converges to the entropy rate of a stationary random process.*

Proof. Consider a sequence of sections $\mathbf{X}$ with increasing size which is taken from a 2D stationary process (see Fig. 7). Each section is parametrized by a parameter l which determines the extent of the section. The width of the section is chosen to be $w = l^3$ and its height is equal to $h = l^2 + l - 2$. □

The pixels are enumerated from top-left to button-right as it is shown in the Fig. 7. Let G and $\overline{G}$ be the sets that contain the indices which are shaded in gray and white colors, respectively. Furthermore, let n denote the total amount of pixels in the section, and let n_G and $n_{\overline{G}}$ be the number of pixels in the gray and white regions, respectively.

If we let the size of the sections go to infinity by letting l go to infinity, they will cover the whole plane and the number of white pixels will become negligible, i.e.

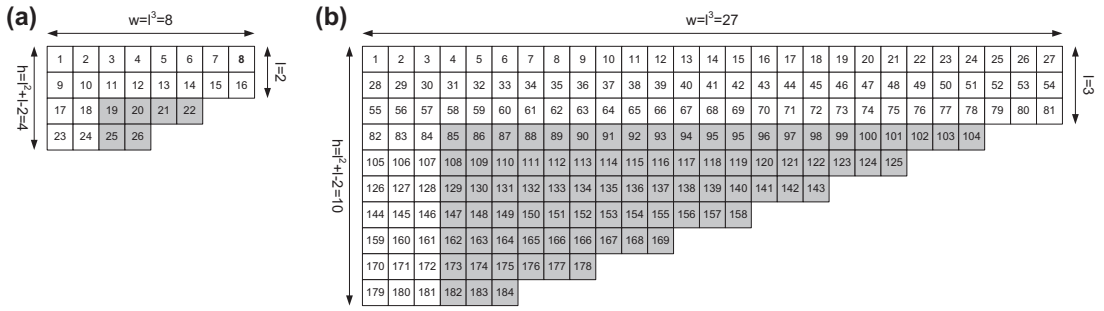


Fig. 7. Enumeration of the pixels in a 2D stationary process.

$$\lim_{l\to\infty}\frac{n_G(l)}{n(l)}=\lim_{l\to\infty}\frac{\frac{1}{2}(w(l)-l)\cdot(h(l)-l)}{w(l)\cdot l+h(l)\cdot l-l^2+\frac{1}{2}(w(l)-l)\cdot(h(l)-l)}=1, \tag{8}$$

$$\lim_{l\to\infty}\frac{n_{\overline{G}}(l)}{n(l)}=\lim_{l\to\infty}\frac{w(l)\cdot l+h(l)\cdot l-l^2}{w(l)\cdot l+h(l)\cdot l-l^2+\frac{1}{2}(w(l)-l)\cdot(h(l)-l)}=0. \tag{9}$$

Since the sections **X** will cover the whole plane in the limit, the sequence of entropies of the single sections converges to the entropy of the stationary process:

$$h=\lim_{l\to\infty}\frac{1}{n(l)}H[\mathbf{X}_{1:n(l)}]=\lim_{l\to\infty}\frac{1}{n(l)}\sum_{k=1}^{n(l)}H[X_k|\mathbf{X}_{1:k-1}].$$

If we split the sum into two sums for the pixels in the gray, and white region, respectively, we obtain

$$h=\lim_{l\to\infty}\frac{1}{n(l)}\left(\sum_{k\in G}H[X_k|\mathbf{X}_{1:k-1}]\right)+\lim_{l\to\infty}\frac{1}{n(l)}\left(\sum_{k\in\overline{G}}H[X_k|\mathbf{X}_{1:k-1}]\right). \tag{10}$$

Define $H_\alpha[X]$ to be the conditional entropy of X given a causal neighborhood of radius α (see Fig. 1b). Since conditioning decreases the entropy we obtain the following inequalities for stationary processes:

$$H_w\leqslant H[X_k|\mathbf{X}_{1:k-1}]\leqslant H_l,\quad\forall\ k\in G,$$
$$H_w\leqslant H[X_k|\mathbf{X}_{1:k-1}]\leqslant H_0,\quad\forall\ k\in\overline{G}.$$

Using these inequalities in Eq. (10) we obtain:

$$\lim_{l\to\infty}H_{w(l)}\leqslant h\leqslant\lim_{l\to\infty}\frac{n_G(l)}{n(l)}H_l+\lim_{l\to\infty}\frac{n_{\overline{G}}(l)}{n(l)}H_0. \tag{11}$$

Using Eqs. (8) and (9) in Eq. (11) we get:

$$\lim_{l\to\infty}H_{w(l)}\leqslant h\leqslant\lim_{l\to\infty}H_l.$$

The sequences $H_{w(l)}$ and H_l will converge to the same limit, since $\{H_{w(l)}\}_{l=1,2,\dots}$ is a proper subsequence of $\{H_l\}_{l=1,2,\dots}$. Hence, using the sandwich theorem the sequence of conditional entropies $\{H_l\}_{l=1,2,\dots}$ converges to the true entropy rate from above.

Appendix B

Minimizing the conditional average log-loss for a given model is equal to maximizing the conditional likelihood. Given the observed data $\{\mathbf{x}_i\}_{i=1}^m$, the conditional log-likelihood is given by:

$$\mathscr{L}(\boldsymbol{s},\Sigma,\boldsymbol{\lambda})=\underbrace{\sum_{i=1}^m\log\mathrm{GSM}(x_{1:n,i}|\boldsymbol{s},\Sigma,\boldsymbol{\lambda})}_{\mathscr{L}_1(\boldsymbol{s},\Sigma,\boldsymbol{\lambda})}-\underbrace{\sum_{i=1}^m\log\mathrm{GSM}(x_{1:(n-1),i}|\boldsymbol{s},\Sigma_{1:(n-1),1:(n-1)},\boldsymbol{\lambda})}_{\mathscr{L}_2(\boldsymbol{s},\Sigma_{1:(n-1),1:(n-1)},\boldsymbol{\lambda})}.$$

The conditional log-likelihood is the difference between the joint log-likelihood $\mathscr{L}_1$ and the marginal log-likelihood $\mathscr{L}_2$. Commonly, the EM algorithm is used to estimate mixture distributions. It constitutes a variational approach which maximizes a lower bound on the *joint* log-likelihood based on the Jensen inequality. In each iteration the maximum of the bound is computed. Since here $\mathscr{L}_2$ enters the conditional log-likelihood with a negative sign, the normal Jensen inequality is not useful to bound this function. Jebara derived a reversed form of the Jensen inequality for the exponential family (Jebara, 2002).

We used Jebara's method for deriving a conditional EM algorithm for the scale mixture of Gaussians. In the E-step the following coefficients are computed:

$$t_i^k=\frac{\lambda_k\mathscr{N}(\mathbf{x}_i|s_k\cdot\Sigma)}{\sum_{k=1}^K\lambda_k\mathscr{N}(\mathbf{x}_i|s_k\cdot\Sigma)},$$
$$r_i^k=\frac{\lambda_k\mathscr{N}(x_{1:(n-1),i}|s_k\cdot\Sigma_{1:(n-1),1:(n-1)})}{\sum_{k=1}^K\lambda_k\mathscr{N}(x_{1:(n-1),i}|s_k\cdot\Sigma_{1:(n-1),1:(n-1)})},$$
$$w_i^{k'}=\max\left[0,\frac{r_i^k}{x_{1:(n-1),i}^T s_k^{-1}\Sigma_{1:(n-1),1:(n-1)}^{-1}x_{1:(n-1),i}}-1\right],$$
$$w_i^k=2G\left(\frac{r_i^k}{2}\right)\left(\left(x_{1:(n-1),i}^T s_k^{-1}\Sigma_{1:(n-1),1:(n-1)}^{-1}x_{1:(n-1),i}-1\right)^2+n-2\right)+w_i^{k'},$$
$$G(\xi)=\begin{cases}\xi+\frac{1}{4\log(6)}+\frac{25}{36\log(6)^2}-\frac{1}{6} & \text{if } \xi\geqslant\frac{1}{6};\\ \frac{(\xi-1)^2}{\log(\xi)^2}-\frac{1}{4\log(\xi)} & \text{if } \xi\leqslant\frac{1}{6}.\end{cases}$$

Using these coefficients we get the following update rule for the scale and marginal covariance parameters.

$$\Sigma_{1:(n-1),1:(n-1)}=\frac{1}{m+\sum_{k=1}^K\sum_{i=1}^m w_i^k}\left(\sum_{k=1}^K\sum_{i=1}^m t_i^k s_k^{-1}x_{1:(n-1),i}x_{1:(n-1),i}^T-\sum_{k=1}^K\sum_{i=1}^m r_i^k s_k^{-1}x_{1:(n-1),i}x_{1:(n-1),i}^T\right)+\Sigma_{1:(n-1),1:(n-1)},$$

$$s_k=\frac{1}{K\sum_{i=1}^m t_i^k+(K-1)\sum_{i=1}^m w_i^k}\left(\sum_{i=1}^m t_i^k\mathbf{x}_i^T\Sigma^{-1}\mathbf{x}_i-\sum_{i=1}^m r_i^k x_{1:(n-1),i}^T\Sigma_{1:(n-1),1:(n-1)}^{-1}x_{1:(n-1),i}\right)+\frac{(K-1)\sum_{i=1}^m(w_i^k+r_i^k)}{K\sum_{i=1}^m t_i^k+(K-1)\sum_{i=1}^m w_i^k}s_k$$

The conditional prediction matrix $\Gamma = \Sigma^{-1}_{1:(n-1),1:(n-1)}\Sigma_{1:(n-1),n}$ and the conditional variance $\gamma = \Sigma_{n,n} - \Sigma_{n,1:(n-1)}\Sigma^{-1}_{1:(n-1),1:(n-1)}\Sigma_{1:(n-1),n}$ only depend on the joint log-likelihood $\mathscr{L}_1$ and their estimations in the M-step are given by:

$$\mathbf{M} = \frac{1}{m}\sum_{k=1}^{K}\sum_{i=1}^{m} t_i^k s_k^{-1}\mathbf{x}\mathbf{x}^T,$$
$$\Gamma = \mathbf{M}^{-1}_{1:(n-1),1:(n-1)}\mathbf{M}_{1:(n-1),n},$$
$$\gamma = \mathbf{M}_{n,n} - \mathbf{M}_{n,1:(n-1)}\mathbf{M}^{-1}_{1:(n-1),1:(n-1)}\mathbf{M}_{1:(n-1),n}.$$

Similar to the normal EM algorithm for optimizing the joint likelihood of theGSM model, one needs to iterate between estimating $\mathbf{s}$ and Σ for maximizing the bound.

The derivation before was for the case of fixed weighting coefficients $\boldsymbol{\lambda}$. For updating the weighting coefficients one can derive another EM update rule. Define a $(K-1)\times(K-1)$ matrix $\mathbf{N}$ with the following entries

$$N_{ij} = \begin{cases} \lambda_i - \lambda_i^2 & \text{if } i = j; \\ -\lambda_i\lambda_j & \text{if } i \neq j. \end{cases}$$

Consider a $K-1$-dimensional vector $\boldsymbol{z}_k$,$0 < k < K$ for which all entries are zero except the kth one, which equals one. Furthermore, let $\boldsymbol{z}_K$ a zero vector with $K-1$ elements. Using those vectors, we get the following update rules for the E-step

$$v_i^k = 4G(r_i^k/2)(\boldsymbol{z}_k - \lambda_{1:(K-1)})^T\mathbf{N}^{-1}(\boldsymbol{z}_k - \lambda_{1:(K-1)}),$$

and the M-step

$$\lambda_k = \frac{\sum_{i=1}^m t_i^k - \sum_{i=1}^m r_i^k}{m + \sum_{k=1}^K\sum_{i=1}^m v_i^k} + \lambda_k.$$

We observed that in practice the conditional EM algorithm converges very slowly. We found out that this is because the reverse Jensen inequality for the covariance is a very loose bound which becomes even looser for higher dimensions since the coefficient w increases rapidly with increasing dimensionality. As a consequence of this, we observed empirically that the EM algorithm increases the log-likelihood slower than gradient ascend with line search.

We accelerated the EM algorithm by using the Quasi-Newton method (algorithm QN2 in Jamshidian & Jennrich, 1997). The idea behind this method is to approximate the Newton update $\boldsymbol{H}^{-1}\mathbf{g}(\boldsymbol{\theta})$, where $\boldsymbol{H}$ is the Hessian and $\mathbf{g}$ is the gradient at $\boldsymbol{\theta}$ with the update $\hat{\mathbf{g}}(\boldsymbol{\theta}) - \boldsymbol{S}\mathbf{g}(\boldsymbol{\theta})$ where $\hat{\mathbf{g}}$ is EM gradient. In other words, the difference of two EM steps and $\boldsymbol{S}$ is a matrix that needs to be updated as well. The authors modify BFGS Quasi-Newton method to get the update for $\boldsymbol{S}$:

$$\Delta\boldsymbol{S} = \left(1 + \frac{\Delta\mathbf{g}^T\Delta\boldsymbol{\theta}^\star}{\Delta\mathbf{g}^T\Delta\boldsymbol{\theta}}\right)\frac{\Delta\boldsymbol{\theta}\Delta\boldsymbol{\theta}^T}{\Delta\mathbf{g}^T\Delta\boldsymbol{\theta}} - \frac{\Delta\boldsymbol{\theta}^\star\Delta\boldsymbol{\theta}^T + (\Delta\boldsymbol{\theta}^\star\Delta\boldsymbol{\theta}^T)^T}{\Delta\mathbf{g}^T\Delta\boldsymbol{\theta}}.$$

where $\Delta\boldsymbol{\theta}^\star = -\Delta\hat{\mathbf{g}} + \boldsymbol{S}\Delta\mathbf{g}$ while $\Delta\boldsymbol{\theta}$ and $\Delta\mathbf{g}$ show the amount of change in variables $\boldsymbol{\theta}$ and $\mathbf{g}$ after each iteration, respectively. In the implementation one initializes with $\boldsymbol{S} = 0$ and then updates $\boldsymbol{S}$ according to the update rule. If the line search is not successful $\boldsymbol{S}$ is reset to zero.

In practice we observed that this Quasi-Newton acceleration significantly increases the convergence speed but it still remains slow. In the future, this may be substantially improved by exploiting the Quasi-Newton and Newton method directly on the log-likelihood.

References

Atick, J., & Redlich, A. (1992). What does the retina know about natural scenes. *Neural Computation, 4*, 196–210.

Barlow, H. (1959). Sensory mechanisms, the reduction of redundancy, and intelligence. In *The mechanisation of thought processes* (pp. 535–539). London: Her Majesty's Stationery Office.

Bernardo, J. M. (1979). Expected information as expected utility. *The Annals of Statistics, 7*(May), 686–690.

Bethge, M. (2006). Factorial coding of natural images: How effective are linear models in removing higher-order dependencies? *Journal of the Optical Society of America A, 23*(6), 1253–1268.

Buchsbaum, G., & Gottschalk, A. (1983). Trichromacy, opponent colours coding and optimum colour information transmission in the retina. *Proceedings of the Royal Society of London. Series B, Biological Sciences, 220*(November), 89–113.

Chandler, D. M., & Field, D. J. (2007). Estimates of the information content and dimensionality of natural scenes from proximity distributions. *Journal of the Optical Society of America A, 24*(4), 922–941.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). Wiley & Sons. Sep.

Eichhorn, J., Sinz, F., & Bethge, M. (2009). Natural image coding in v1: How much use is orientation selectivity? *PLoS Computational Biology, 5*(4), e1000336.

Föllmer, H. (1973). On entropy and information gain in random fields. *Probability Theory and Related Fields, 26*(3), 207–217.

Jamshidian, M., & Jennrich, R. I. (1997). Acceleration of the EM algorithm by using Quasi-Newton methods. *Journal of the Royal Statistical Society. Series B (Methodological), 59*(3), 569–587.

Jebara, T. (2002). Discriminative, generative, and imitative learning. Thesis. Massachusetts Institute of Technology.

Karklin, Y., & Lewicki, M. (2008). Emergence of complex cell properties by learning to generalize in natural scenes. *Nature, 457*, 83–86.

Lee, T.-W., Wachtler, T., & Sejnowski, T. J. (2002). Color opponency is an efficient representation of spectral properties in natural scenes. *Vision Research, 42*(17), 2095–2103.

Lewicki, M. S., & Olshausen, B. A. (1999). Probabilistic framework for the adaptation and comparison of image codes. *Journal of the Optical Society of America A, 16*(July), 1587–1601.

Lewicki, M., & Sejnowski, T. (2000). Learning overcomplete representations. *Neural Computation, 12*, 337–365.

Linsker, R. (1990). Perceptual neural organization: Some approaches based on network models and information theory. *Annual Review of Neuroscience, 13*(1), 257–281.

Lyu, S., & Simoncelli, E. P. (2009). Reducing statistical dependencies in natural signals using radial Gaussianization. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.). *Advances in neural information processing systems* (Vol. 21, pp. 1009–1016). Cambridge, MA: MIT Press.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381*(6583), 607–609.

Paninski, L. (2003). Estimation of entropy and mutual information. *Neural Computation, 15*(6), 1191–1253.

Perez, A. (1977). ε-Admissible simplification of the dependence structure of a set of random variables. *Kybernetika, 13*, 439–444.

Petrov, Y., & Zhaoping, L. (2003). Local correlations, information redundancy, and sufficient pixel depth in natural images. *Journal of the Optical Society of America A, 20*(1), 56–66.

Ruderman, D. L., & Bialek, W. (1994). Statistics of natural images: Scaling in the woods. *Physical Review Letters, 73*(6), 814. copyright (C) 2009 The American Physical Society. Please report any problems to prola@aps.org..

Rue, H., & Held, L. (2005). *Gaussian Markov random fields: Theory and applications*. Chapman & Hall/CRC.

Schreiber, W. (1956). The measurement of third order probability distributions of television signals. *IRE Transactions on Information Theory, 2*(3), 94–105.

Schwartz, O., & Simoncelli, E. P. (2001). Natural signal statistics and sensory gain control. *Nature Neuroscience, 4*(8), 819–825.

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika, 66*(3), 605–610.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*, 379–423. and 623–656.

Simoncelli, E., & Olshausen, B. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience, 24*, 1193–1216.

Sinz, F., & Bethge, M. (2009). The conjoint effect of divisive normalization and orientation selectivity on redundancy reduction. In *Neural information processing systems, 2008* (p. 8).

Srinivasan, M., Laughlin, S., & Dubs, A. (1982). Predictive coding: A fresh view of inhibition in the retina. *Proceedings of the Royal Society of London. Series B, Biological Sciences, 216*(1205), 427–459.

van Hateren, J. H., & van der Schaaf, A. (1998). Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society of London. Series B, Biological Sciences, 265*(1394), 1724–1726.

Wachtler, T., Lee, T. W., & Sejnowski, T. J. (2001). Chromatic structure of natural scenes. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision, 18*, 65–77. pMID: 11152005.

Wainwright, M. J., & Simoncelli, E. P. (2000). Scale mixtures of gaussians and the statistics of natural images. In: *Advances in neural information processing systems* (Vol. 12, pp. 855–861).

4.10 Temporal adaptation enhances efficient contrast gain control on natural images: Submitted Draft

Temporal adaptation enhances efficient contrast gain control on natural images

Fabian Sinz [*] [†] **and Matthias Bethge** [*] [†] [‡]

[*]Max Planck Institute for Biological Cybernetics, Tübingen, Germany, [†]Werner Reichardt Centre for Integrative Neuroscience, University of Tübingen, Germany, and [‡]Bernstein Center for Computational Neuroscience, Tübingen, Germany

Submitted to Proceedings of the National Academy of Sciences of the United States of America

Divisive normalization in primary visual cortex has been linked to adaptation to natural image statistics via Barlow's redundancy reduction hypothesis. Using recent advances in natural image modeling, we quantify the residual redundancy after divisive normalization in a population of linear-nonlinear neurons. We find that static divisive normalization is rather inefficient in reducing local contrast correlations and demonstrate that a simple temporal contrast adaptation mechanism can substantially increase the efficiency. Our findings highlight the importance of adaptation to the local contrast statistics via shifts in the contrast response curve of neurons.

sensory coding | primary visual cortex | redundancy reduction | natural image statistics

Abbreviations: V1, primary visual cortex;

It has been a long-standing hypothesis that the computational goal of the early stages of visual processing is to reduce the redundancies which are abundantly present in natural sensory signals [1, 2]. Redundancy reduction is a general information theoretic principle that subsumes many possible goals of sensory systems like maximizing the amount of information between stimulus and neural response [3], obtaining a probabilistic model of sensory signals [4], or learning a representation of hidden causes [3, 5]. For a population of neurons, redundancy reduction predicts that neuronal responses should be statistically independent from each other [2].

Previous work has linked redundancy reduction of natural signals to divisive normalization contrast gain control in primary visual cortex by demonstrating that correlations in the variances of neuronal responses are removed [7, see also Figure 1a]. Divisive normalization is a nonlinear mechanism that non-linearly rescales the response of a population of neurons by dividing the activity y_i of a single neuron by the activity of an inhibitory pool of other neurons [6].

In this study we compare the redundancy reduction achieved by a static divisive normalization mechanism in a model population of V1 neurons to a recently developed optimal divisive transformation, called *radial factorization* or *radial Gaussianization* [8, 9], to assess whether divisive normalization is powerful enough to capture the rich dependencies of natural images. The model population receives an input image patch $\mathbf{x}$ which is filtered by linear receptive fields $\mathbf{w}_i$. The resulting responses $y_i = \mathbf{w}_i^\top \mathbf{x}$ are transformed with divisive normalization. The essential mechanism in divisive normalization is a rescaling $\|\mathbf{z}\| = \kappa\|\mathbf{y}\|/\sqrt{\sigma + \|\mathbf{y}\|^2}$ of the norm of the population response $\mathbf{y}$. Under reasonable assumptions about the statistics of natural image patches, radial factorization is the optimal mechanism in terms of redundancy reduction acting on the norm $\|\mathbf{y}\|$ (see Methods).

Experiments and Results

We compared the amount of redundancies removed by divisive normalization in the response of a population of model neurons to natural image patches to the amount removed by radial factorization and find that divisive normalization leaves a substantial amount of residual redundancies (Figure 1b). While both divisive normalization and radial factorization remove correlations in the variances of the neural responses, the residual amount of dependencies for divisive normalization is still approximately 34% of the total redundancies present in the unnormalized population response (Figure 1a-b). This demonstrates that the underlying assumption of divisive normalization about the statistics of natural image patches misses important regularities.

To understand this in more detail, we derived what distribution the linear filter responses $\|\mathbf{y}\|$ would have if divisive normalization were the optimal redundancy reducing mechanism (referred to as *Naka-Rushton distribution* in the following), and compared it to the empirical distribution represented by a large collection of uniformly sampled patches from natural images (Figure 1c). The only free parameter of the Naka-Rushton distribution is the semi-saturation constant σ^2 of the divisive normalization function which determines the horizontal position of the contrast response curve in neurons. We fitted σ^2 via maximum likelihood (see Methods) and found that even for the best fitting σ^2 there is a substantial mismatch. This explains the insufficient redundancy reduction because the Naka-Rushton distribution expects most of the responses $\|\mathbf{y}\|$ to fall into a much narrower range than responses to natural images do in reality (Figure 1c).

We explored two options how the visual system could potentially increase the flexibility and, therefore, the redundancy reduction performance of divisive normalization: enhancing static divisive normalization with more parameters or allowing for a temporal adaptation of σ^2.

We find that an extended divisive normalization transform $\|\mathbf{z}\| = \kappa\|\mathbf{y}\|^{\frac{\gamma}{2}+\delta}/\sqrt{\sigma^2 + \|\mathbf{y}\|^\gamma}$ achieves substantially more redundancy reduction and that its corresponding distribution on $\|\mathbf{y}\|$ fits significantly better (Figure 1b-c). However, we also find that the corresponding shape of the population contrast response exhibits a physiologically unreasonable shape (Figure 1c inset).

Exploring the second option, we found that the distribution on $\|\mathbf{y}\|$ predicted by a temporally adapting σ^2 closely matches the empirical distribution of responses to patches sampled with simulated eye movements, and yields a substantial reduction in redundancy (Figure 2a-b). Our tempo-

Reserved for Publication Footnotes

rally adapting model relies on correlations between the contrast at different time steps to choose the current σ^2 based on the recent stimulation history. Previous studies on redundancy reduction with divisive normalization [7, 9, 8] ignored the structure caused by fixations between saccades in natural viewing conditions. Contrast response curves of neurons in primary visual cortex are known to adapt to the ambient contrast level [10] by adapting σ^2. A temporally adapting σ under redundancy reduction predicts that the joint population response $\|\mathbf{y}\|$ should be well modeled by a mixture of Naka-Rushton distributions each of which corresponding to a different value of σ^2. For a fixed history of responses $H_k = (r_{t-1}, ..., r_{t-k})$ preceding r_t the normalized response $\kappa r_t / \sqrt{\sigma(H_k)^2 + r_t^2}$ would follow a truncated χ-distribution, which is equivalent to a Naka-Rushton distribution on r_t conditioned on H_k

$$r_t | H_k \sim \nu(r_t | \sigma(H_k)).$$

Averaged over all histories the distribution of r_t is a mixture of Naka-Rushton distributions

$$r_t \sim \varrho(r_t) = \int \nu(r_t|\sigma(H_k))\rho(H_k)dH_k = \int \nu(r_t|\sigma)\rho(\sigma)d\sigma. \quad [1]$$

We used a simple model of saccades and micro-saccades to simulate eye movements on natural images and fitted such a mixture to the responses in our model. In order to quantify the amount of redundancy reduction, we then estimated σ^2 for the present patch from the immediately preceding one using this mixture of distributions. We found that a simple strategy for choosing σ given the immediate history significantly decreased the amount of residual redundancies to 1.1%.

We also verified that σ cannot be chosen randomly but the correct utilization of temporal correlations is crucial for this improvement. If that was the case σ could be chosen independently of the preceding history at each time step, and be used to transform the current response with $r_t \mapsto \kappa r_t / \sqrt{\sigma_t^2 + r_t^2}$ such that the result still yield a truncated χ-distribution. This is the same as saying that a truncated χ-distribution could be described as a mixture of the distributions that result from transforming r_t with Naka-Rushton functions with different values of σ. We transformed the input distribution with Naka-Rushton functions that differed in the value of σ (Figure 2c, colored lines). Different colors in Figure 2c refer to different values of σ. If σ could be drawn independently, a positively weighted average of the colored distributions should be able to yield a truncated χ-distribution (Figure 2c, dashed line). One can immediately see that this is not possible. Every component will either add a tail to the left of the χ-distribution or a peak to the right of it. Since distributions can only be added with non-negative weight in a mixture there is no way that one distribution can make up for a tail or peak introduced by another. Therefore, σ cannot be chosen independently of the preceding stimulation.

Discussion

Our results suggest a very specific link between the adaptation of neurons to the ambient contrast level and redundancy reduction for natural images. Our analysis does not commit to a certain physiological implementation or biophysical constraints, but it demonstrates that the statistics of natural images require more degrees of freedom for redundancy reduction in a population response than a static divisive normalization model can offer, and that the temporal adaptation of σ might be necessary for a flexible adaptation to the statistics of natural images.

Compared to extended divisive normalization, the main reason for the worse performance of divisive normalization with static σ^2 is that the interval containing most of the probability mass is too narrow and too close to zero compared to the empirical distribution. To visualize that, we sought after a general signature that could depict whether an adaptation mechanism is powerful enough for substantial redundancy reduction. To that end, we plotted the median of the different empirical distributions and the ones implied by the models against the width of the interval between the 10% and the 90% percentile (Figure 3). We also included a dataset from real human eye movements by Kienzle et al. to ensure the generality of this signature [11]. Real fixations could introduce a change in the statistics because real observers tend to look at patches with higher contrasts [14]. The empirical data and all models that yield strong redundancy reduction exhibit a ratio greater than 1.5. This signature can be used for future physiological experiments to test the suggested link between redundancy reduction and contrast gain control.

Methods

The code and the data are available online under
`http://www.bethgelab.org/code/sinz2012`.

Data.

van Hateren data For the static experiments, we used randomly sampled 17×17 patches from the van Hateren database [12]. For all experiments we used the logarithm of the raw light intensities. We sampled 10 pairs of training and test sets of $500,000$ patches for which we employed the preprocessing of Eichhorn et al. by centering all patches on the pixel mean and rescaling them such that whitening became volume conserving [13].

For the simulated eye movements, we also used 4 pairs of training and test sets. For the sampling procedure, we repeated the following steps until $500,000$ samples were drawn: We first drew an image randomly from the van Hateren database. For each image, we simulated ten saccades to random locations in that image. For each saccade location which was uniformly drawn over the entire image, we determined the number m of patches to be sampled from around that location by $m = \lceil \nu \cdot \tau \rceil$ where $\nu = 50Hz$ was the assumed sampling frequency and τ was a sample from an exponential distribution with average fixation time $0.2s$. The actual locations of the patches were determined by Brownian motion with standard deviation $\sigma = 30$ starting at the saccade location

Kienzle data While the van Hateren database is a standard dataset for static natural image statistics. To make sure that our results also hold for real fixations, we sampled data from the images used by Kienzle et al. [11]. We computed the 10% and 90% percentiles as well as the width of the interval between them for both datasets for Figure 3.

We constructed two datasets: One where the patches were uniformly drawn from the images, and one where we used Brownian motion with standard deviation of $\sigma \approx 35$ around human fixation spots to simulate human fixational data. We applied the same preprocessing as for the van Hateren data: centering, rescaling such that whitening is volume conserving, and whitening

Models. Both the divisive normalization model and the optimal radial factorization consist of two steps: a linear filtering

step and a divisive normalization step (Table 1). In the following, we describe the different steps in more detail.

Note that, for modeling neural responses, both models' responses would be mapped into firing rates afterwards by an elementwise rectification step and possibly a nonlinearity. Since the positive and the negative part of each filter response can be encoded by two neurons with opposite rectifiers and since elementwise nonlinearities do not change the redundancy (i.e. the multi-information), we did not explicitly model the rectification step in our analyses.

Filters The receptive fields of our model neurons, i.e. the linear filters of our models, are given by the rows of a matrix $W = Q\Lambda^{-\frac{1}{2}}U^\top A$. A is an 288×289 matrix with mutually orthogonal rows with mean zero. This matrix projects out the DC component of the data [13]. U contains the first $n = 72$ principal components of $A\mathbf{x}$ in its columns, and Λ is a diagonal matrix with the corresponding eigenvalues. Therefore, $\Lambda^{-\frac{1}{2}}U^\top$ is a whitening matrix. We used only $n = 72$ filters corresponding to the first 72 principle components in order to exclude high spatial frequencies.

Q is an orthogonal matrix, which was trained with independent subspace analysis with two-dimensional subspaces [15]:

$$\rho(\mathbf{y}) = \prod_{k=1}^{n/2} \rho_k(y_{2k}, y_{2k+1}|\vartheta_k) \text{ with } \mathbf{y} = W\mathbf{x} \qquad [2]$$

where ϑ denotes the list of free parameters for each ϱ_k. This yields filter pairs that resemble quadrature pairs like in the energy model of complex cells [17, 18]. Each single ρ_k was chosen to be a two-dimensional L_p-spherically symmetric distribution [16]

$$\rho_k(\mathbf{y}_{2k:2k+1}|\vartheta_k) = \frac{\varrho_k(\|\mathbf{y}_{2k:2k+1}\|_p|\vartheta_k)}{\|\mathbf{y}_{2k:2k+1}\|_p^{K-1}\mathcal{S}_p^2}$$

$$\|\mathbf{y}\|_p = \left(\sum_{i=1}^{2} |y_i|^p\right)^{\frac{1}{p}},\ p > 0$$

with a radial γ-distribution $\varrho(r|u,s) = \gamma(u,s)$ with shape u and scale s. $\mathcal{S}_p^2$ denotes the surface area of the L_p-norm unit sphere in two dimensions [16]. During training, we first fixed $p = u = 1$; after initial convergence, we retrained the model with free p and u.

The likelihood of the data under equation (2) was optimized by alternating between optimizing Q for fixed ϑ_k, and optimizing the ϑ_k for fixed Q. The gradient ascent on the log-likelihood of Q over the orthogonal group used the backprojection method by Manton [19, 20, 21].

Normalization

Optimal contrast gain control: radial factorization Radial factorization is the optimal redundancy reduction mechanism for L_p-spherically symmetric distributed data [22, 16]. L_p-spherical symmetry assumes that all data points of a given L_p-norm $r = \|\mathbf{y}\|_p = \left(\sum_{i=1}^n |y_i|^p\right)^{\frac{1}{p}}$ are uniformly distributed on the L_p-sphere with that radius. A radial distribution $\varrho(r)$ determines how likely it is that a data point is drawn from an L_p-sphere with that specific radius. Since the distribution on the sphere is always uniform, the radial distribution ϱ determines the specific type of distribution. For example, for $p = 2$ and $\varrho(r) = \chi(r)$ one obtains a isotropic Gaussian, since the Gaussian distribution is spherically symmetric ($p = 2$) and has a radial χ-distribution ($\varrho(r) = \chi(r)$). One can show that, for a fixed value of p, there is only one type of radial distribution such that the joint distribution is factorial [23]. For $p = 2$ this radial distribution is the χ-distribution which corresponds to a joint Gaussian distribution. For $0 < p \neq 2$, the radial distribution is a generalization of the χ-distribution and the joint distribution is the so called p-generalized Normal [24]. Radial factorization is a mapping on the L_p-norm $\|\mathbf{y}\|_p$ of the data points that transforms a given source L_p-spherically symmetric distribution into a p-generalized Normal. Since the p-generalized Normal is factorial, radial factorization is a nonlinear redundancy reduction mechanism.

The reason why radial factorization is a very strong redundancy reduction mechanism on natural images is that the filter responses of whitening filters to natural image patches are well modeled by L_p-spherically symmetric distributions [22]. It models the distribution of $r = \|\mathbf{y}\|_p$ with a flexible distribution and non-linearly rescales the radius r such that the radial distribution becomes a generalized χ-distribution and, hence, the joint distribution becomes factorial. If the flexible distribution is denoted by ϱ the new χ_p-distributed radius can be computed via $\left(\mathcal{F}_{\chi_p}^{-1} \circ \mathcal{F}_\varrho\right)(\|\mathbf{y}\|)$. This mapping, also known under the name *histogram equalization*, transforms ϱ-distributed radii in χ_p-distributed one. χ_p denotes the generalized χ-distribution and $\mathcal{F}$ denote cumulative distribution functions of the respective distributions. On the joint responses $\mathbf{y}$, radial factorization first divides out the radius and rescales it with the new radius:

$$\mathbf{y} \mapsto \frac{\left(\mathcal{F}_{\chi_p}^{-1} \circ \mathcal{F}_\varrho\right)(\|\mathbf{y}\|_p)}{\|\mathbf{y}\|_p}\mathbf{y}$$

In our case ϱ was chosen to be a mixture of five γ-distributions. $\mathcal{F}_{\chi_p}^{-1}$ is the inverse cumulative distribution function of a χ_p-distribution which is the radial distribution of a p-generalized Normal distribution [24].

When determining the optimal redundancy reduction performance on the population response, we set $p = 2$ in order to use the same norm as the divisive normalization model. Only when estimating the redundancy of the linear filter responses, we use $p = 1.3$ [22].

Divisive normalization model and Naka-Rushton distribution We use the following divisive normalization transform

$$\|\mathbf{y}\|_2 \mapsto \frac{\kappa\|\mathbf{y}\|_2}{\sqrt{\sigma^2 + \|\mathbf{y}\|_2^2}}$$

which is the standard model for neural contrast gain control [6].

Divisive normalization acts on the Euclidean norm of the filter responses $\mathbf{y}$. While in radial factorization the target and source distribution were fixed, and the goal was to find a mapping that transforms one into the other, we now fix the mapping to divisive normalization, the target distribution on the normalized response $\mathbf{z}$ to be Gaussian ($\|\mathbf{z}\|_2$ to be χ-distributed) and search for the corresponding source distribution. Since divisive normalization saturates at κ, we will actually have to use a truncated χ-distribution on $\|\mathbf{z}\|_2$. κ becomes the truncation threshold. Note that radial truncation actually introduces some dependencies, but we keep them small by choosing the truncation threshold κ to be the 99% percentile of the radial χ-distribution which is approximately $\kappa \approx 10.14$. Note also that choosing a Gaussian target distribution does not contradict the finding that cortical firing rates are found to be exponentially distributed [25], since each

single response z_i can always be transformed again to be exponentially distributed without changing the redundancy of $\mathbf{z}$.

The distribution on $r = \|\mathbf{y}\|_2$ such that

$$\|\mathbf{z}\|_2 = \frac{\kappa\|\mathbf{y}\|_2}{\sqrt{\sigma^2 + \|\mathbf{y}\|_2^2}}$$

is truncated χ-distributed can be derived by a simple change of variables. In the resulting distribution

$$\varrho\left(r\right) = \frac{2\kappa^n\sigma^2 r^{n-1}}{\mathfrak{G}\left(\frac{n}{2}, \frac{\kappa^2}{2s}\right)\Gamma\left(\frac{n}{2}\right)(2s)^{\frac{n}{2}}\left(\sigma^2 + r^2\right)^{\frac{n+2}{2}}} \exp\left(-\frac{\kappa^2 r^2}{2s\left(\sigma^2 + r^2\right)}\right),$$

the truncation threshold κ, the semi-saturation constant σ, and the scale of the χ-distribution become parameters of the model. The parameter s of the Naka-Rushton distribution controls the variance of the corresponding Gaussian and was always chosen such that the Gaussian was white with variance one. The only free parameter of the Naka-Rushton distribution is σ which couples shape and scale. $\mathfrak{G}$ is the regularized-incomplete-gamma function which accounts for the truncation at κ. We call the distribution *Naka-Rushton distribution* and denote it with $\nu\left(\kappa, \sigma, s\right)$.

To derive the distribution on $\|\mathbf{y}\|$ for which the extended divisive normalization transformation $\frac{\kappa\|\mathbf{y}\|^{\frac{\gamma}{2}+\delta}}{\sqrt{\sigma^2+\|\mathbf{y}\|_2^\gamma}}$ yields a (truncated) χ-distribution, the steps are exactly the same as for the standard divisive normalization transform above. Note that extended divisive normalization saturates only for $\delta = 0$. Therefore, the distribution on $\|\mathbf{z}\|_2$ has to be a χ-distribution if $\delta > 0$ and a truncated χ-distribution if $\delta = 0$. This yields

$$\varrho\left(r\right) = \frac{p\kappa^n r^{\frac{n\gamma+2n\delta-2}{2}}\left(2\delta\left(r^\gamma + \sigma^2\right) + \gamma\sigma^2\right)}{\Gamma\left(\frac{n}{p}\right)s^{\frac{n}{p}}2^{\frac{n+p}{p}}\left(r^\gamma + \sigma^2\right)^{\frac{n+2}{2}}} \times \exp\left(-\frac{\kappa^p r^{\frac{p\gamma}{2}+p\delta}}{2s\left(\sigma^2 + r^\gamma\right)^{\frac{p}{2}}}\right)$$

for $\delta > 0$ and

$$\varrho\left(r\right) = \frac{p\kappa^n r^{\frac{n\gamma-2}{2}}\gamma\sigma^2}{\mathfrak{G}\left(\frac{n}{p}, \frac{\kappa^p}{2s}\right)\Gamma\left(\frac{n}{p}\right)s^{\frac{n}{p}}2^{\frac{n+p}{p}}\left(r^\gamma + \sigma^2\right)^{\frac{n+2}{2}}} \times \exp\left(-\frac{\kappa^p r^{\frac{p\gamma}{2}}}{2s\left(\sigma^2 + r^\gamma\right)^{\frac{p}{2}}}\right)$$

for $\delta = 0$. The parameters of the distribution are now $\sigma, \delta, \kappa, \gamma$ and s.

The parameters for all divisive normalization transforms were estimated via maximum likelihood of the Naka-Rushton distribution on the Euclidean norms $\{r_i\}_{i=1}^m = \{\|\mathbf{y}_i\|_2\}_{i=1}^m$ of the filter responses to natural image patches. We did not optimize for s in the extended Naka-Rushton distribution but fixed it such that the corresponding Gaussian was white.

Dynamically adapting σ For the model with dynamically adapting σ, we first model the Euclidean norms $r_i = \|\mathbf{y}_i\|_2$ of the filter responses to the patches from the simulated eye movement data with a mixture of 500 Naka-Rushton distributions

$$\varrho(r) = \sum_{i=1}^{500} \nu(r|\sigma_i)\pi_i,$$

using EM [26]. π_i denotes the probability that $\sigma = \sigma_i$. The values of σ_i where chosen in 500 equidistant steps from 0.01 to 12.

How much redundancy reduction can be achieved with a dynamically adapting σ, depends on the dynamics according to which it is selected based on the recent history. While there might be many strategies, we chose a parsimonious one. To that end, we evaluated the posterior

$$\varrho(\sigma = \sigma_i|r) = \frac{\pi_i\nu(r|\sigma_i)}{\sum_{j=1}^{500}\nu(r|\sigma_j)\pi_j}.$$

of the mixture distribution at 100 equidistant locations between 10^{-12} and 35, computed the posterior mean and standard deviation at those locations, rescaled the standard deviation by $1/\sqrt{2}$, and fitted a piecewise linear function on the intervals $[0, 1), [1, 2), \ldots, [30, \infty)$ to each set of values. In the first interval, the linear function was constraint to start at zero. From these two functions $\mu(r)$ and $\sigma(r)$, we computed two functions for the scale θ and the shape u of a γ-distribution

$$u(r) = \frac{\mu(r)^2}{\sigma(r)^2} \text{ and } \theta(r) = \frac{\sigma(r)^2}{\mu(r)}$$

via moment matching.

In order to obtain a value σ for the Naka-Rushton function for transforming a value r_t based on the value of its predecessor r_{t-1}, we sampled σ from a γ-distribution with shape and scale determined by $u(r_{t-1})$ and $\theta(r_{t-1})$.

Percentiles For the dynamically adapting σ^2 in Figure 3, we sampled from

$$p(r) = \int\int \nu(r|\sigma, \kappa, s)\gamma(\sigma|u(r_i), \theta(r_i))p(r_i)d\sigma dr_i$$

and computed the percentiles based on the sampled dataset. For the sampling procedure, we drew σ from the γ-distribution $\gamma(\sigma|u(r_i), \theta(r_i))$ with shape and scale computed from r_i and then sampled r from the Naka-Rushton distribution $\nu(r|\sigma, \kappa, s)$ with that σ. We repeated that for all r_i from a test set of simulated eye movement radii. This procedure was carried out for all pairs of training and test sets, and the distributions fitted to them.

For the static case, we sampled data from single Naka-Rushton distributions for different values of σ and computed the percentiles from the samples.

Multi-information estimation We use the *multi-information* to quantify the statistical dependencies between the filter responses $\mathbf{y}$ [27]. The multi-information is the n-dimensional generalization of the *mutual-information*. It is defined as the Kullback-Leibler divergence between the joint distribution and the product of its marginals or, equivalently, the difference between the sum of the marginal entropies and the joint entropy

$$I[\mathbf{Y}] = D_{KL}\left(\rho(\mathbf{y})\|\prod_{i=1}^n \rho_i(y_i)\right) = \sum_{i=1}^n H[Y_i] - H[\mathbf{Y}]. \quad [3]$$

The multi-information is zero if and only if the different dimensions of the random vector $\mathbf{Y}$ are independent. Since the joint entropy $H[\mathbf{Y}]$ is hard to estimate we employ a semi-parametric estimate of the multi-information that is conservative in the sense that it is downward biased.

For the marginal entropies $H[Y_i]$, we use a jackknifed estimator for the discrete entropy on the binned values [28]. We

chose the bin size with the heuristic proposed by Scott [29]. We obtain an estimate for the differential entropy by correcting with the logarithm of the bin width (see e.g. [13]).

In order to estimate the joint entropy, we use the average log-loss to get an upper bound

$$A[\hat{\rho}(\mathbf{y})] := -\left\langle \log \hat{\rho}(\mathbf{y}) \right\rangle_{\mathbf{Y}\sim\rho(\mathbf{y})} = H[\mathbf{Y}] + D_{KL}\left(\rho(\mathbf{y}) \| \hat{\rho}(\mathbf{y})\right).$$

Since the average log-loss overestimates the true entropy, replacing the joint entropy by A in equation (3) underestimates the multi-information. Therefore, we sometimes get estimates smaller than zero. Since the multi-information is always positive, we set the value to zero in that case. For computing errorbars on the multi-information estimations, we use the negative values but a mean zero in such cases, which effectively increases the standard deviation of the error.

Since we want commit ourselves as little as possible to a particular model, we estimate $A[\hat{\rho}(\mathbf{y})]$ by making the assumption that $\mathbf{y}$ is L_p-spherically symmetric distributed but estimating everything else with non-parametric estimators. If $\mathbf{y}$ is L_p-spherically symmetric distributed, the radial component is independent from the directional component [16] and we can write

$$\hat{H}\left[\mathbf{Y}\right] = \hat{H}\left[R\right] + (n-1)\left\langle \log r \right\rangle_R + \log \mathcal{S}_p. \qquad [4]$$

The entropy $H[R]$ of the radial component is again estimated via a histogram estimator. The term $(n-1)\left\langle \log r \right\rangle_R$ is approximated by the empirical mean.

Putting all the equations together yields our estimator for the multi-information under the assumption of L_p-spherically symmetric distributed $\mathbf{Y}$

$$\hat{I}[\mathbf{Y}] = \sum_{i=1}^{n} \hat{H}[Y_i] - \hat{H}\left[R\right] - \frac{(n-1)}{m}\sum_{j=1}^{m} \log r_j - \log \mathcal{S}_p,$$

where $\hat{H}[\cdot]$ are the univariate entropies estimated via binning.

Since the optimal value of p for filter responses $\mathbf{y}$ to natural image patches is approximately $p \approx 1.3$ we use that value to estimate the multi-information of $\mathbf{y}$.

When estimating the multi-information of the responses $\mathbf{z}$ of either divisive normalization or radial factorization, we use the fact that

$$I[\mathbf{Z}] = \sum_{i=1}^{n} H[\mathbf{Z}_i] - H[\mathbf{Z}] = \sum_{i=1}^{n} H[\mathbf{Z}_i] - H[\mathbf{Y}] - \left\langle \log\det\left|\frac{d\mathbf{z}}{d\mathbf{y}}\right| \right\rangle_{\mathbf{Y}}$$

where $\frac{d\mathbf{z}}{d\mathbf{y}}$ is the Jacobian of the normalization transformation. The mean is estimated by averaging over data points. The determinants of radial factorization, divisive normalization, and extended divisive normalization are given by

$$\begin{aligned}
\det\left|\frac{d\mathbf{z}}{d\mathbf{y}}\right| &= \frac{\|\mathbf{z}\|_p^{n-1}\varrho(\|\mathbf{y}\|_p)}{\|\mathbf{y}\|_p^{n-1}\chi_p(\|\mathbf{z}\|_p)} \\
\det\left|\frac{d\mathbf{z}}{d\mathbf{y}}\right| &= \kappa^n\left(\sigma^2 + \|\mathbf{y}\|_2^2\right)^{-\frac{n+2}{2}}\sigma^2 \\
\det\left|\frac{d\mathbf{z}}{d\mathbf{y}}\right| &= \frac{\|\mathbf{z}\|_p^{n-1}}{\|\mathbf{y}\|_p^{n-1}}\frac{\kappa r^{\frac{\gamma}{2}+\delta-1}\left(2\delta\left(r^\gamma+\sigma^2\right)+\gamma\sigma^2\right)}{2\left(r^\gamma+\sigma^2\right)^{\frac{3}{2}}}.
\end{aligned}$$

All multi-information values were computed on test data.

ACKNOWLEDGMENTS. We thank P. Berens, L. Busse, S. Katzner and L. Theis for fruitful discussions and comments on the manuscript.

1. Barlow HB (1961) Possible Principles Underlying the Transformations of Sensory Messages. Rosenblith WA, ed. Sensory Communication: 217-234.
2. Simoncelli EP, Olshausen BA (2003) Natural Image Statistics and Neural Representation. Annual Review of Neuroscience 24:1193-1216.
3. Bell AJ, Sejnowski T.J. (1997) The "independent component" of natural scenes are edge filters. Vision Research 37(23):3327-3338
4. Barlow HB (1989) Unsupervised Learning Neural Computation 1(3):295-311
5. Lewicki MS, Olshausen BA (1999) Probabilistic framework for the adaptation and comparison of image codes Journal of the Optical Society of America A 16:1587-1601
6. Heeger DJ (1992) Normalization of cell responses in cat striate cortex Vis Neurosci 9(2):181-197.
7. Schwartz O, Simoncelli EP (2001) Natural signal statistics and sensory gain control Nat. Neurosci. 4(8):819-825.
8. Lyu S, Simoncelli EP (2009)Nonlinear extraction of independent components of natural images using radial gaussianization. Neural Computation 21(6):1485-1519.
9. Sinz F, Bethge M (2009) The Conjoint Effect of Divisive Normalization and Orientation Selectivity on Redundancy Reduction. Advances in neural information processing systems 21 : 22nd Annual Conference on Neural Information Processing Systems 2008 :1521-1528.
10. Bonds AB (1991) Temporal dynamics of contrast gain in single cells of the cat striate cortex. Vis. Neurosci 6(3):239-255
11. Kienzle W, Franz MO, Schölkopf B, Wichmann FA (2009) Center-surround patterns emerge as optimal predictors for human saccade targets. Journal of Vision 9(5):7.1-15.
12. Hateren JH van, Der Schaaf A van (1998) Independent component filters of natural images compared with simple cells in primary visual cortex. Proceedings of the Royal Society B Biological Sciences 265(1394):359-366
13. Eichhorn J, Sinz F, Bethge M (2009) Natural Image Coding in V1: How Much Use Is Orientation Selectivity? PLoS Comput Biol. 5(4).
14. Reinagel P, Zador AM (1999) Natural scene statistics at the centre of gaze. Network10(4):341-350.
15. Hyvärinen A, Köster U (2007) Complex cell pooling and the statistics of natural images. Network: Computation in Neural Systems 18(2):81-100.
16. Gupta AK, Song D (1997) L_p-norm spherical distribution. Journal of Statistical Planning and Inference 60(2):241-260.
17. Pollen D, Ronner S (1981) Phase relationships between adjacent simple cells in the visual cortex Science 212(4501):1409-1411
18. Adelson EH, Bergen JR (1985) Spatiotemporal energy models for the perception of motion. Journal of the Optical Society of America A 2(2):284-299.
19. Manton JH (2002) Optimization algorithms exploiting unitary constraints. Signal Processing, IEEE Transactions on.50(3):635-650.
20. Sinz F, Simoncelli EP, Bethge M (2009) Hierarchical Modeling of Local Image Features through L_p-Nested Symmetric Distributions. Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009 1696-1704.
21. Sinz F, Bethge M (2010) L_p-Nested Symmetric Distributions Journal of Machine Learning Research 11:3409-345.
22. Sinz F, Bethge M (2009) The Conjoint Effect of Divisive Normalization and Orientation Selectivity on Redundancy Reduction. Advances in neural information processing systems 21: 22nd Annual Conference on Neural Information Processing Systems 2008 1521-1528.
23. Sinz F, Gerwinn S, Bethge M (2009) Characterization of the p-generalized normal distribution.Journal of Multivariate Analysis 100:817820.
24. Goodman IR, Kotz S (1973) Multivariate θ-generalized normal distributions. Journal of Multivariate Analysis 3(2):204-219.
25. Baddeley R, Abbott LF, Booth MC, et al. (1997) Responses of neurons in primary and inferior temporal visual cortices to natural scenes. Proceedings of the Royal Society B Biological Sciences 264(1389):1775-1783.
26. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society Series B Methodological 39(1):1-38.
27. Perez A (1977) ε-admissible simplification of the dependence structure of a set of random variables. Kybernetika 13:439-444.
28. Paninski L (2003) Estimation of Entropy and Mutual Information. Neural Computation 15(6):1191-1253.
29. Scott DW (1979) On optimal and data-based histograms. Biometrika 66(3):605-610.

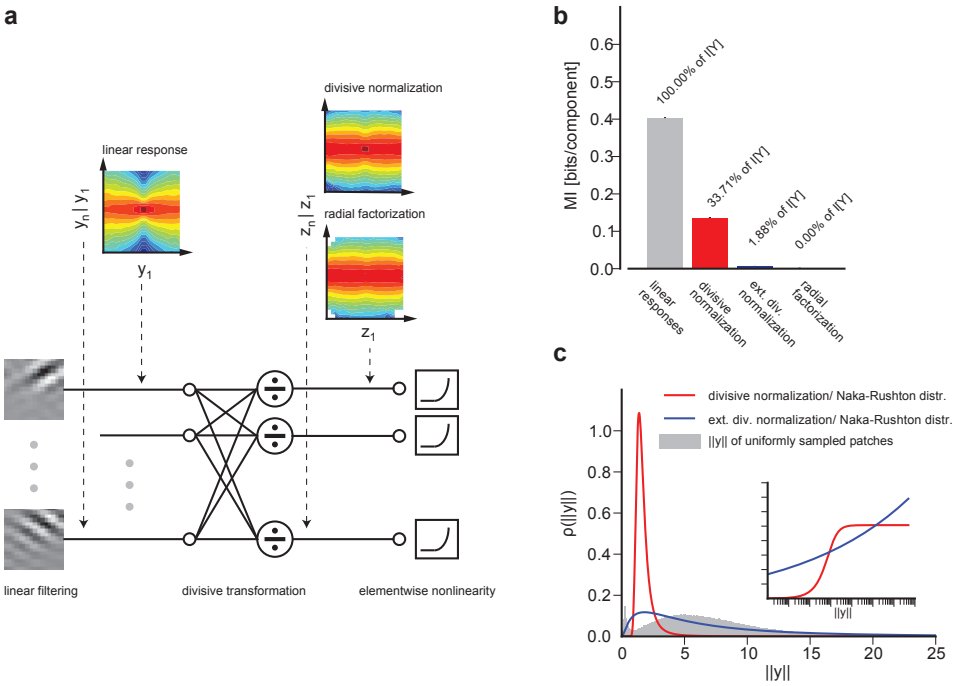


Fig. 1. **a:** Divisive normalization model used in this study: Natural image patches are linearly filtered. These responses are nonlinearly transformed by divisive normalization or radial factorization (see text). After linear filtering the width of the conditional distribution $p(y_j|y_i)$ of two filter responses depends on the value of y_i (conditional histograms as contour plots). This demonstrates the presence of variance correlations. These dependencies are decreased by divisive normalization and radial factorization. **b:** Redundancy measured by multi-information after divisive normalization, extended divisive normalization, and radial factorization: divisive normalization leaves a substantial amount of residual redundancy (error bars show standard deviation over different datasets). **c:** Distributions on the norm of the filter responses $\|\mathbf{y}\|$ for which divisive normalization (red) and extended divisive normalization (blue) are the optimal redundancy reducing mechanisms. While extended divisive normalization achieves good redundancy reduction, it exhibits a physiologically implausible shape of the population contrast response curve $\|\mathbf{y}\|^{\frac{\gamma}{2}+\delta}/\sqrt{\sigma^2+\|\mathbf{y}\|^\gamma}$ (inset, blue curve). The population contrast response curve of divisive normalization is shown for comparison (inset, red curve).

Table 1. Model components of the divisive normalization and radial factorization model: Natural image patches are filtered by a set of linear oriented band-pass filters. The filter responses are normalized and their norm is rescaled in the normalization step.

	divisive normalization model	radial factorization
filtering	$\mathbf{y} = W\mathbf{x}$	$\mathbf{y} = W\mathbf{x}$
normalization	$\mathbf{z} = \frac{\kappa\|\mathbf{y}\|_2^{\frac{\gamma}{2}+\delta}}{\sqrt{\sigma^2+\|\mathbf{y}\|_2^\gamma}} \frac{\mathbf{y}}{\|\mathbf{y}\|_2}$ (static case $\delta = 0$ and $\gamma = 2$)	$\mathbf{z} = \frac{\left(\mathcal{F}_{\chi_p}^{-1} \circ \mathcal{F}_\varrho\right)\left(\|\mathbf{y}\|_p\right)}{\|\mathbf{y}\|_p}\mathbf{y}$

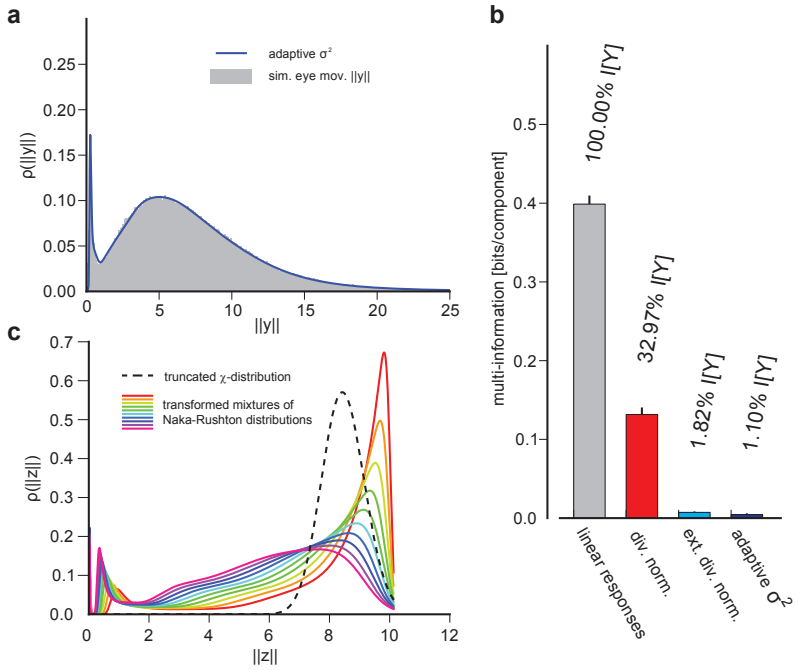


Fig. 2. **a:** Histogram of $\|\mathbf{y}\|$ for natural image patches sampled with simulated eye movements: The distribution predicted by the dynamically adapting model closely matches the empirical distribution. **b:** Redundancy measured by multi-information between the linear filter responses $\mathbf{y}$ without divisive normalization, for divisive normalization with static σ^2, extended divisive normalization, and dynamically adapting σ^2 for simulated eye movement data. The dynamically adapting σ^2 achieves almost the same performance as the optimal radial factorization transform. **c:** Each colored line is a mixture of Naka-Rushton distributions like in (a) transformed with a Naka-Rushton function. Different colors correspond to different values of σ. The dashed curve corresponds to a truncated χ-distribution. A mixture of the colored distributions cannot resemble the truncated χ-distribution since there will either be peaks on the left or the right of the dashed distribution that cannot be canceled by other mixture components.

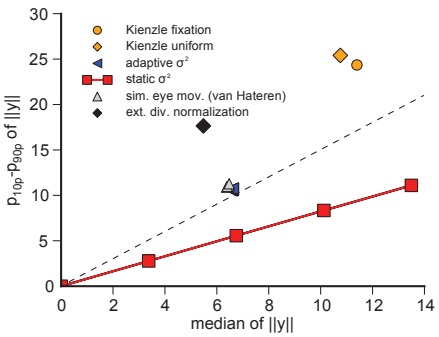


Fig. 3. Median vs. width of 10% to 90% percentile interval of the models from Figure 2b. The red line corresponds to a static σ^2 for different values of σ^2, blue corresponds to the temporally adapting σ^2, the orange markers correspond to uniformly sampled (diamond) and fixational image patches with Brownian motion micro-saccades (circle) from Kienzle et al.[11], the gray markers to simulated eye movement datasets from van Hateren image data [12], and the black marker to the optimal extended divisive normalization model. All transforms that yield a strong redundancy reduction have models that exhibit a ratio greater than 1.5 (dashed lines).