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

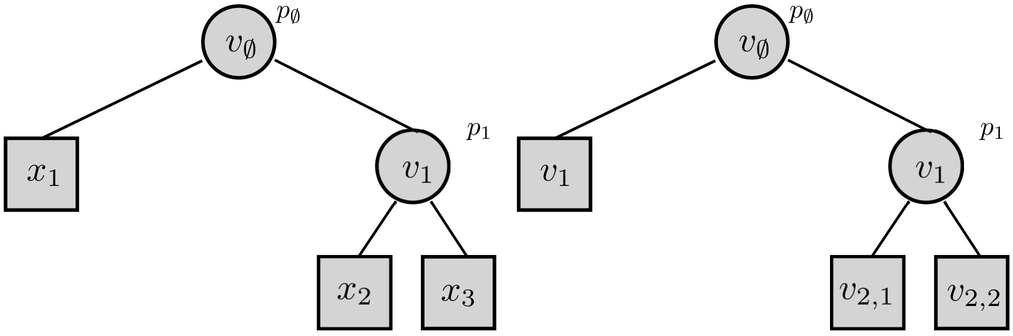
$$f(\boldsymbol{x}) = \left(|x_1|^{p_\emptyset} + (|x_2|^{p_1} + |x_3|^{p_1})^{\frac{p_\emptyset}{p_1}}\right)^{\frac{1}{p_\emptyset}} \tag{1}$$

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

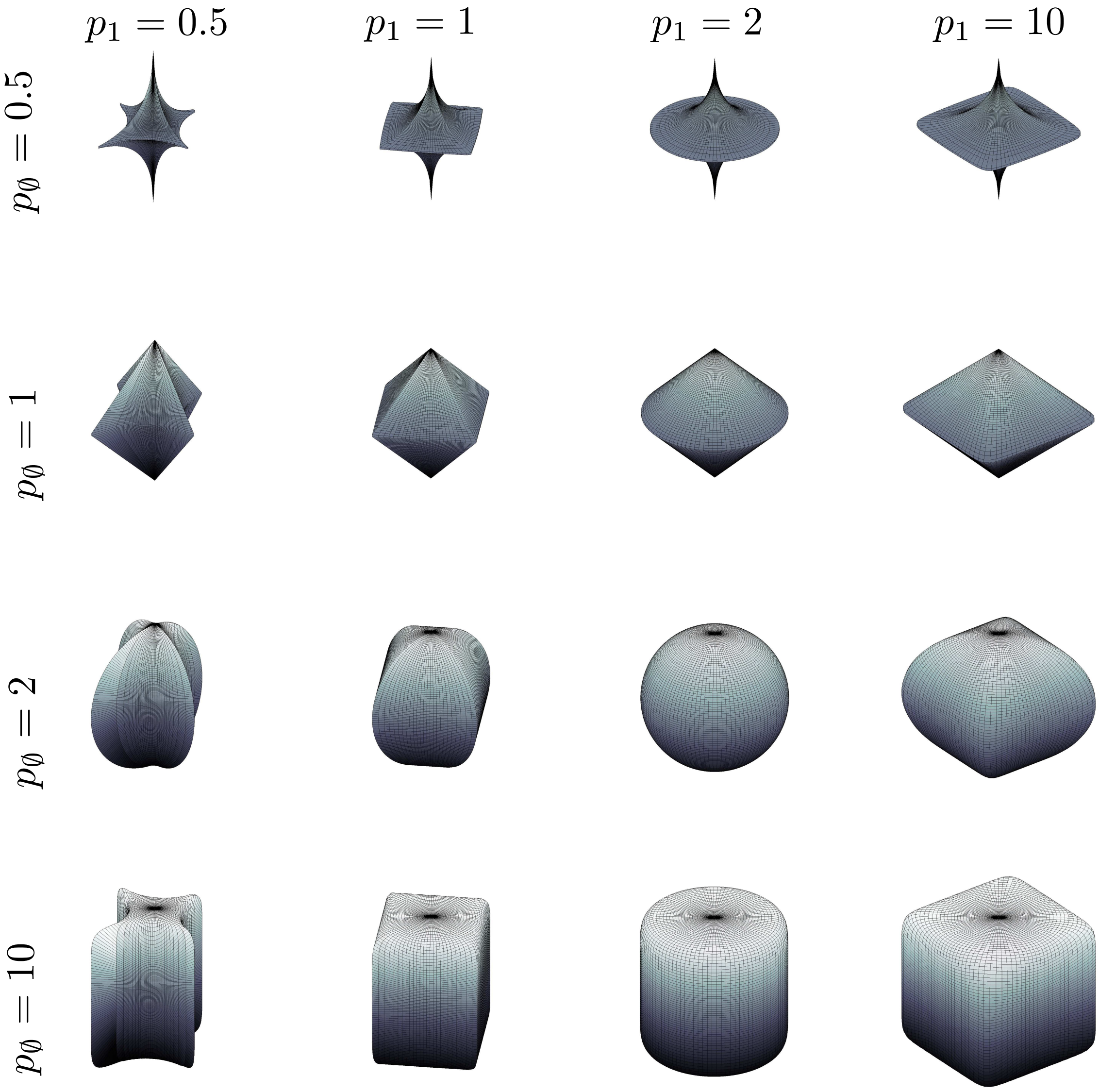


Figure 2: Variety of contours created by the L_p-nested function of Equation (1) for all combinations of $p_\emptyset, p_1 \in \{0.5, 1, 2, 10\}$.

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 $\boldsymbol{x}_{\widehat{I}}$. 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:

$$\begin{aligned} f_I(\boldsymbol{x}_I) &= ||(f_{I,1}(\boldsymbol{x}_{I,1}), \ldots, f_{I,\ell_I}(\boldsymbol{x}_{I,\ell_I}))^\top||_{p_I} \\ &= ||(v_{I,1}, \ldots, v_{I,\ell_I})^\top||_{p_I} = ||\boldsymbol{v}_{I,1:\ell_I}||_{p_I}. \end{aligned}$$

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 $I = i_1, \ldots, i_d$ be the multi-index of a node in the tree. $i_1, \ldots, i_d$ 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 $x_2, \ldots, x_{10}$. Then $\boldsymbol{x}_I = (x_2, \ldots, x_{10})$, $\boldsymbol{x}_{\widehat{I}} = (x_1, x_{11}, x_{12}, \ldots)$, 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 $\boldsymbol{x} \in \mathbb{R}^n$:*

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

The inverse coordinate transformation is given by

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

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):*

$$f(\boldsymbol{x}) = \left((|x_1|^{p_1} + |x_2|^{p_1})^{\frac{p_\emptyset}{p_1}} + |x_3|^{p_\emptyset}\right)^{\frac{1}{p_\emptyset}}.$$

Setting $\boldsymbol{u} = \frac{\boldsymbol{x}}{f(\boldsymbol{x})}$ and solving for u_3 yields

$$f(\boldsymbol{u}) = 1 \Leftrightarrow \;\; u_3 = \left(1 - (|u_1|^{p_1} + |u_2|^{p_1})^{\frac{p_\emptyset}{p_1}}\right)^{\frac{1}{p_\emptyset}}. \tag{3}$$

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

$$G_2(\boldsymbol{u}_{\widehat{2}}) = g_2(\boldsymbol{u}_{\widehat{2}})^{1-p_\emptyset} = \left(1 - (|u_1|^{p_1} + |u_2|^{p_1})^{\frac{p_\emptyset}{p_1}}\right)^{\frac{1-p_\emptyset}{p_\emptyset}},$$

$$F_1(\boldsymbol{u}_1) = f_1(\boldsymbol{u}_1)^{p_\emptyset - p_1} = (|u_1|^{p_1} + |u_2|^{p_1})^{\frac{p_\emptyset - p_1}{p_1}},$$

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

$$\begin{aligned}\frac{1}{r^2}|\mathcal{J}| &= \sum_{k=1}^{2} G_2(\boldsymbol{u}_{\widehat{2}})F_1(\boldsymbol{u}_1)|u_k|^{p_1} + u_3\\ &= G_2(\boldsymbol{u}_{\widehat{2}})\left(F_1(\boldsymbol{u}_1)\sum_{k=1}^{2}|u_k|^{p_1} + 1 - F_1(\boldsymbol{u}_1)F_1(\boldsymbol{u}_1)^{-1}\left(|u_1|^{p_1}+|u_2|^{p_1}\right)^{\frac{p_\emptyset}{p_1}}\right)\\ &= G_2(\boldsymbol{u}_{\widehat{2}})\left(F_1(\boldsymbol{u}_1)\sum_{k=1}^{2}|u_k|^{p_1} + 1 - F_1(\boldsymbol{u}_1)\sum_{k=1}^{2}|u_k|^{p_1}\right)\\ &= G_2(\boldsymbol{u}_{\widehat{2}}).\end{aligned}$$

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 $g_n(\boldsymbol{u}_{\widehat{n}}) = u_n = \left(1 - \sum_{i=1}^{n-1} |u_i|^p\right)^{\frac{1}{p}}$. Thus, the term $G_n(\boldsymbol{u}_{\widehat{n}})$ 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*

$$f(\boldsymbol{x}) = \left(|x_1|^{p_\emptyset} + (|x_2|^{p_1} + |x_3|^{p_1})^{\frac{p_\emptyset}{p_1}}\right)^{\frac{1}{p_\emptyset}}.$$

Normalizing and resolving for the last coordinate yields

$$u_3 = \left((1 - |u_1|^{p_\emptyset})^{\frac{p_1}{p_\emptyset}} - |u_2|^{p_1}\right)^{\frac{1}{p_1}}$$

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

$$\begin{aligned} G_2(\boldsymbol{u}_{\widehat{2}}) &= (1 - |u_1|^{p_\emptyset})^{\frac{p_1 - p_\emptyset}{p_\emptyset}}, \\ G_{2,2}(\boldsymbol{u}_{\widehat{2,2}}) &= \left((1 - |u_1|^{p_\emptyset})^{\frac{p_1}{p_\emptyset}} - |u_2|^{p_1}\right)^{\frac{1-p_1}{p_1}}. \end{aligned}$$

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*

Eqn (7) is not the surface area in general. Please check errata on sinzlab.org. The rest of the results is unaffected.

$$\mathcal{V}_f(R) = \frac{R^n 2^n}{n} \prod_{I \in I} \left(\frac{1}{p_I^{\ell_I - 1}} \prod_{k=1}^{\ell_I - 1} B\left[\frac{\sum_{i=1}^{k} n_{I,k}}{p_I}, \frac{n_{I,k+1}}{p_I} \right] \right) \tag{5}$$

$$= \frac{R^n 2^n}{n} \prod_{I \in I} \frac{\prod_{k=1}^{\ell_I} \Gamma\left[\frac{n_{I,k}}{p_I}\right]}{p_I^{\ell_I - 1} \Gamma\left[\frac{n_I}{p_I}\right]}, \tag{6}$$

$$\mathcal{S}_f(R) = R^{n-1} 2^n \prod_{I \in I} \left(\frac{1}{p_I^{\ell_I - 1}} \prod_{k=1}^{\ell_I - 1} B\left[\frac{\sum_{i=1}^{k} n_{I,k}}{p_I}, \frac{n_{I,k+1}}{p_I} \right] \right) \tag{7}$$

$$= R^{n-1} 2^n \prod_{I \in I} \frac{\prod_{k=1}^{\ell_I} \Gamma\left[\frac{n_{I,k}}{p_I}\right]}{p_I^{\ell_I - 1} \Gamma\left[\frac{n_I}{p_I}\right]} \tag{8}$$

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*

$$\rho(\mathbf{x}) = \frac{\phi(f(\mathbf{x}))}{f(\mathbf{x})^{n-1} \mathcal{S}_f(1)}$$

$$= \frac{\phi(f(\mathbf{x}))}{2^n f(\mathbf{x})^{n-1}} \prod_{I \in I} \left(p_I^{\ell_I - 1} \prod_{k=1}^{\ell_I - 1} B\left[\frac{\sum_{i=1}^{k} n_{I,k}}{p_I}, \frac{n_{I,k+1}}{p_I} \right]^{-1} \right). \tag{9}$$

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 $\mathbf{u}$. The distribution of $\mathbf{u}$ 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 $\mathbf{u}$ (the factor $f(\mathbf{x})^{n-1} = r$ cancels due to the determinant of the Jacobian). For that reason we call the distribution of $\mathbf{u}$ *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 $\mathbf{u}$.

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 $u_1, \ldots, u_{n-1}$*

$$\rho(u_1, , \ldots, u_{n-1}) = \frac{\prod_{L\in\mathcal{L}} G_L(\boldsymbol{u}_{\widehat{L}})}{2^{n-1}} \prod_{I\in\mathcal{I}} \left(p_I^{\ell_I-1} \prod_{k=1}^{\ell_I-1} B\left[\frac{\sum_{i=1}^{k} n_{I,k}}{p_I}, \frac{n_{I,k+1}}{p_I}\right]^{-1} \right).$$

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 $(u_1, \ldots, u_{n-1}) \in \mathbb{R}^{n-1}$ 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*

$$\frac{1}{\mathcal{V}_f(1)} = \frac{nr^{n-1}\prod_{L\in\mathcal{L}} G_L(\boldsymbol{u}_{\widehat{L}})}{2^{n-1}} \prod_{I\in\mathcal{I}} \left(p_I^{\ell_I-1} \prod_{k=1}^{\ell_I-1} B\left[\frac{\sum_{i=1}^{k} n_{I,k}}{p_I}, \frac{n_{I,k+1}}{p_I}\right]^{-1} \right).$$

After separating out the uniform distribution on the L_p-nested unit sphere, we obtain the radial distribution

$$\phi(r) = nr^{n-1} \text{ for } 0 < r \leq 1$$

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

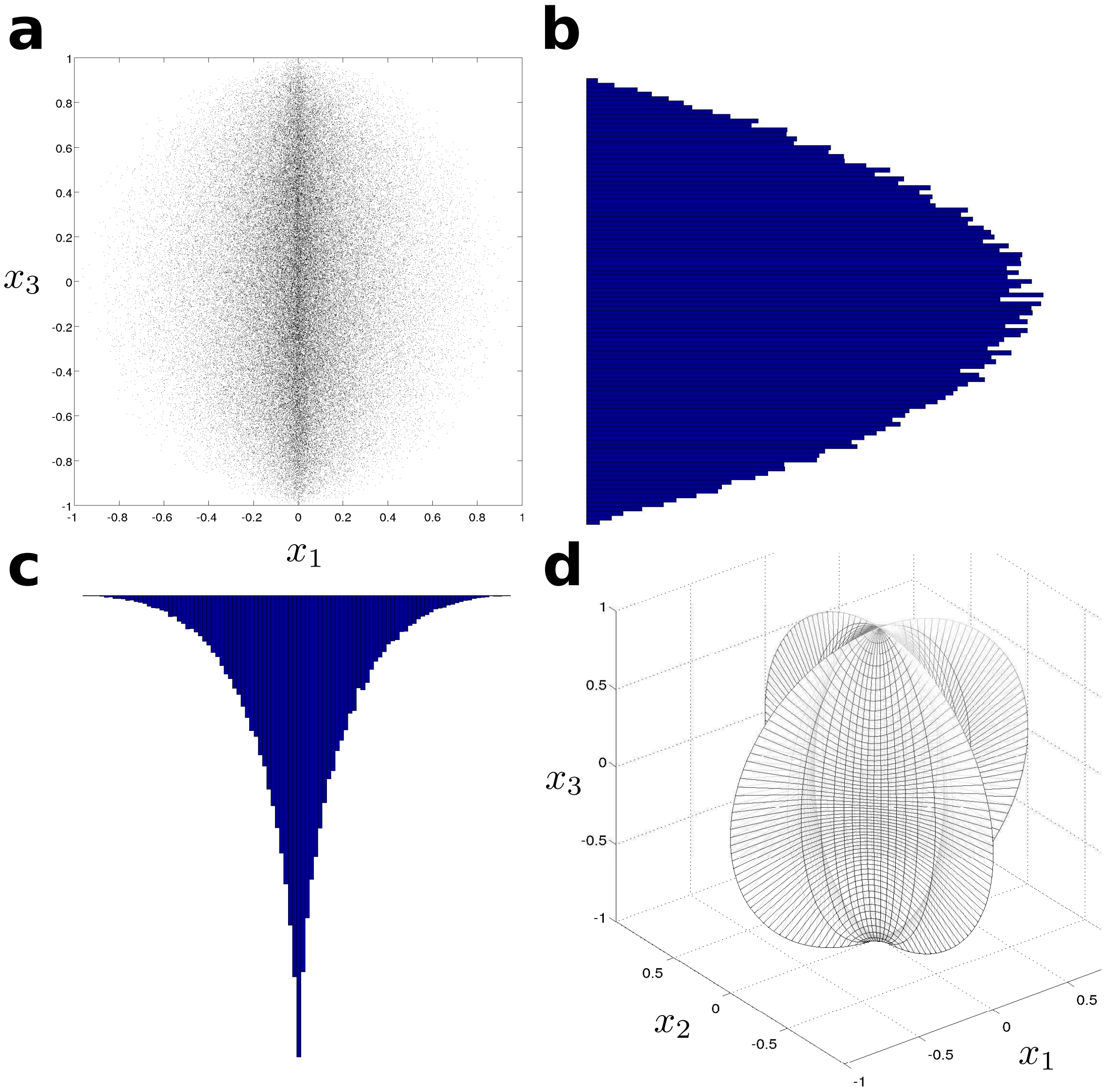


Figure 3: Marginals of L_p-nested symmetric distributions are not necessarily L_p-nested symmetric: Figure (**a**) shows a scatter plot of the (x_1, x_2)-marginal of the counterexample in the text with $p_\emptyset = 2$ and $p_1 = \frac{1}{2}$. Figure (**d**) displays the corresponding L_p-nested sphere. (**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 $\boldsymbol{x}_{\widehat{\mathcal{J}}} \in \mathbb{R}^m$ 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 $\boldsymbol{x}_{\widehat{\mathcal{J}}}$ and $\boldsymbol{v}_{\mathcal{J}}$ is given by*

$$\rho(\boldsymbol{x}_{\widehat{\mathcal{J}}}, \boldsymbol{v}_{\mathcal{J}}) = \frac{\phi(f(\boldsymbol{x}_{\widehat{\mathcal{J}}}, \boldsymbol{v}_{\mathcal{J}}))}{\mathcal{S}_f(f(\boldsymbol{x}_{\widehat{\mathcal{J}}}, \boldsymbol{v}_{\mathcal{J}}))} \prod_{J \in \mathcal{J}} v_J^{n_J - 1}. \tag{10}$$

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 $p(\boldsymbol{x}) = \rho(W\boldsymbol{x}) \cdot |\det W| = \frac{\phi(W\boldsymbol{x})}{f(W\boldsymbol{x})^{n-1}\mathcal{S}_f(1)} \cdot |\det W|$ 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

$$\begin{aligned}
&\nabla_W \log \rho(W\boldsymbol{x})\\
=&\nabla_W \left(-(n-1)\cdot \log f(W\boldsymbol{x}) + \log \phi(f(W\boldsymbol{x}))\right)\\
=&-\frac{(n-1)}{f(W\boldsymbol{x})}\cdot \nabla_{\boldsymbol{y}} f(W\boldsymbol{x})\cdot \boldsymbol{x}^\top + \frac{d\log\phi(r)}{dr}\left(f(W\boldsymbol{x})\right)\cdot \nabla_{\boldsymbol{y}} f(W\boldsymbol{x})\cdot \boldsymbol{x}^\top,
\end{aligned}$$

where the derivatives of f with respect to $\boldsymbol{y}$ are defined by recursive equations

$$\frac{\partial}{\partial y_i} v_I = \begin{cases} 0 & \text{if } i \not\in I\\ \operatorname{sgn} y_i & \text{if } v_{I,k} = |y_i|\\ v_I^{1-p_I}\cdot v_{I,k}^{p_I-1}\cdot \frac{\partial}{\partial y_i} v_{I,k} & \text{for } i \in I, k. \end{cases}$$

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 $O(n^3+n^2m)$ since m data points have to be multiplied with W at each iteration (requiring $O(n^2m)$ 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 $O(n^2m)$ when using hierarchical orthogonalization and $O(n^2m+n^3)$ 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

$$f(\boldsymbol{x}) = \left(|x_1|^{p_\emptyset} + (|x_2|^{p_1} + |x_3|^{p_1})^{\frac{p_\emptyset}{p_1}}\right)^{\frac{1}{p_\emptyset}}.$$

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

$$\int_{\{\boldsymbol{x}:f(\boldsymbol{x})\leq 1\,\&\,\boldsymbol{x}\in\mathbb{R}^n_+\}} d\boldsymbol{x} = \int_{\{v_1,x_1:f(x_1,v_1)\leq 1\,\&\,x_1,v_1\in\mathbb{R}_+\}}\int\int v_1(1-\tilde{u}^{p_1})^{\frac{1-p_1}{p_1}}dx_1dv_1d\tilde{u}.$$

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 $\tilde{u}$:

$$\int_{\{\boldsymbol{x}:f(\boldsymbol{x})\leq 1\,\&\,\boldsymbol{x}\in\mathbb{R}^n_+\}} d\boldsymbol{x} = \int(1-\tilde{u}^{p_1})^{\frac{1-p_1}{p_1}}d\tilde{u}\cdot\int_{\{v_1,x_1:f(x_1,v_1)\leq 1\,\&\,x_1,v_1\in\mathbb{R}_+\}}\int v_1dx_1dv_1.$$

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 $\beta\left[n,1\right]$.
2. For each inner node I of the tree associated with f, sample the auxiliary variable $\boldsymbol{s}_I$ from a Dirichlet distribution $\mathrm{Dir}\left[\frac{n_{I,1}}{p_I},\dots,\frac{n_{I,\ell_I}}{p_I}\right]$ 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

$$1 = \int \rho(f(\boldsymbol{x}))d\boldsymbol{x} = \int\int \prod_{L\in\mathcal{L}} G_L(\boldsymbol{u}_{\widehat{L}})r^{n-1}\rho(r)drd\boldsymbol{u} = \int \prod_{L\in\mathcal{L}} G_L(\boldsymbol{u}_{\widehat{L}})d\boldsymbol{u} \cdot \int r^{n-1}\rho(r)dr.$$

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

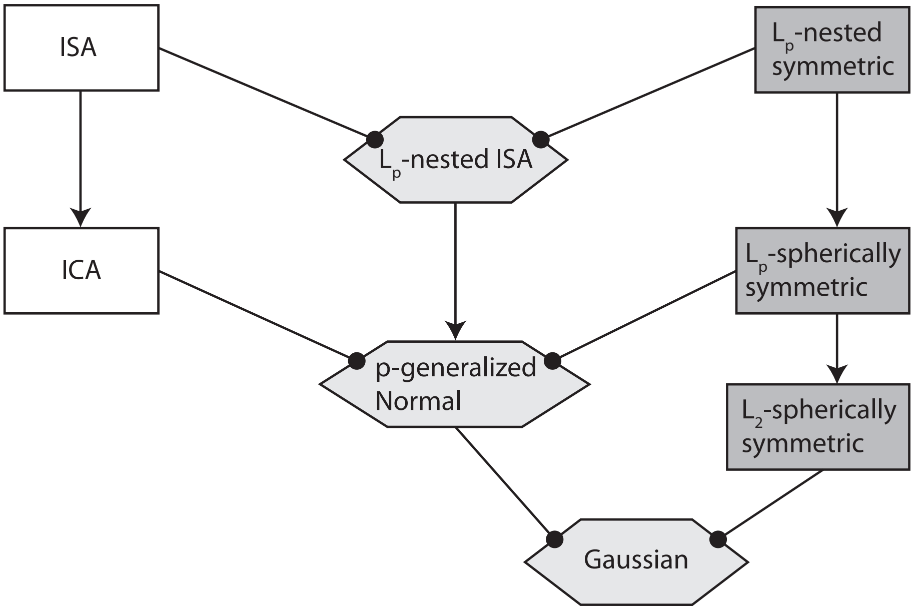


Figure 4: Relations between the different classes of distributions: Arrows indicate that the child class is a specialization (subset) of the parent class. Polygon-shaped classes are intersections of those parent classes which are connected via edges with round arrow-heads. For one-dimensional subspaces 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

$$\phi_\emptyset(v_\emptyset) = \frac{p_\emptyset v_\emptyset^{n-1}}{\Gamma\left[\frac{n}{p_\emptyset}\right] s^{\frac{n}{p_\emptyset}}} \exp\left(-\frac{v_\emptyset^{p_\emptyset}}{s}\right), \tag{11}$$

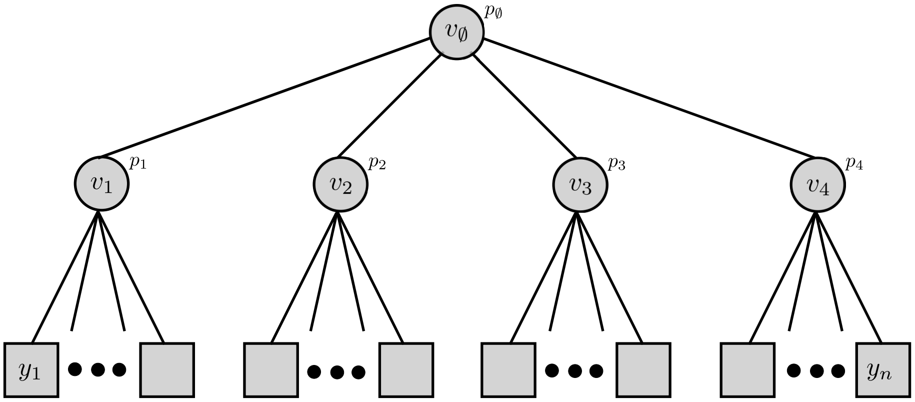


Figure 5: Tree corresponding to an L_p-nested ISA model.

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 $\underline{s}$ource) 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.

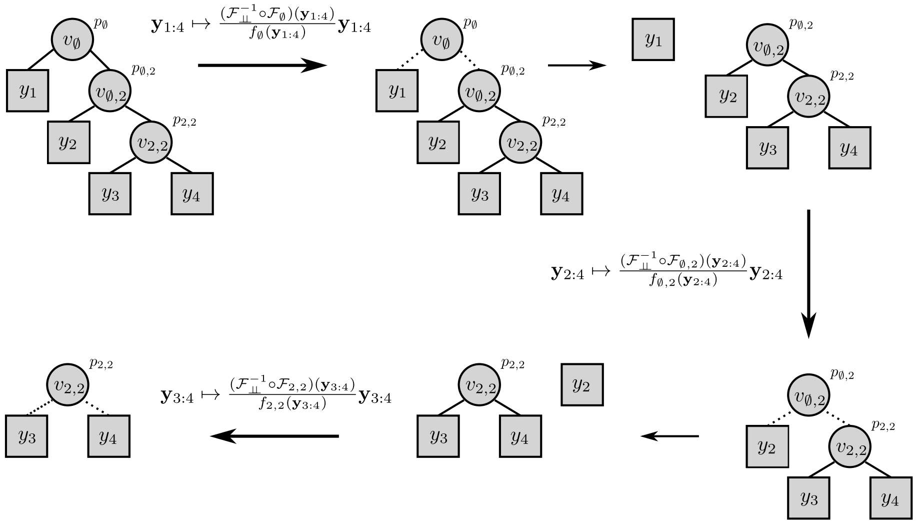

$\mathbf{y}_{1:4} \mapsto \frac{(\mathcal{F}_{\perp\!\!\!\perp}^{-1}\circ\mathcal{F}_{\emptyset})(\mathbf{y}_{1:4})}{f_{\emptyset}(\mathbf{y}_{1:4})}\mathbf{y}_{1:4}$
$\mathbf{y}_{2:4} \mapsto \frac{(\mathcal{F}_{\perp\!\!\!\perp}^{-1}\circ\mathcal{F}_{\emptyset,2})(\mathbf{y}_{2:4})}{f_{\emptyset,2}(\mathbf{y}_{2:4})}\mathbf{y}_{2:4}$
$\mathbf{y}_{3:4} \mapsto \frac{(\mathcal{F}_{\perp\!\!\!\perp}^{-1}\circ\mathcal{F}_{2,2})(\mathbf{y}_{3:4})}{f_{2,2}(\mathbf{y}_{3:4})}\mathbf{y}_{3:4}$
$v_\emptyset$
$p_\emptyset$
$v_{\emptyset,2}$
$p_{\emptyset,2}$
$v_{2,2}$
$p_{2,2}$
y_1
y_2
y_3
y_4

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*

$$\rho(\boldsymbol{y}) = \frac{p_\emptyset}{\Gamma\left[\frac{n}{p_\emptyset}\right]} \exp\left(-|y_1|^{p_\emptyset} - \left(|y_2|^{p_{\emptyset,2}} + \left(|y_3|^{p_{2,2}} + |y_4|^{p_{2,2}}\right)^{\frac{p_{\emptyset,2}}{p_{2,2}}}\right)^{\frac{p_\emptyset}{p_{\emptyset,2}}}\right)$$
$$\times \frac{1}{2^n} \prod_{I \in I} p_I^{\ell_I - 1} \frac{\Gamma\left[\frac{n_I}{p_I}\right]}{\prod_{k=1}^{\ell_I} \Gamma\left[\frac{n_{I,k}}{p_I}\right]}.$$

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

$$\rho(y_2, \ldots, y_4) = \frac{p_\emptyset}{\Gamma\left[\frac{n_{\emptyset,2}}{p_\emptyset}\right]} \exp\left(-\left(|y_2|^{p_{\emptyset,2}} + \left(|y_3|^{p_{2,2}} + |y_4|^{p_{2,2}}\right)^{\frac{p_{\emptyset,2}}{p_{2,2}}}\right)^{\frac{p_\emptyset}{p_{\emptyset,2}}}\right)$$
$$\times \frac{1}{2^{n-1}} \prod_{I \in I \setminus \emptyset} p_I^{\ell_I - 1} \frac{\Gamma\left[\frac{n_I}{p_I}\right]}{\prod_{k=1}^{\ell_I} \Gamma\left[\frac{n_{I,k}}{p_I}\right]}.$$

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

$$\rho(y_2, \ldots, y_4) = \frac{p_{\emptyset,2}}{\Gamma\left[\frac{n_{\emptyset,2}}{p_{\emptyset,2}}\right]} \exp\left(-|y_2|^{p_{\emptyset,2}} - \left(|y_3|^{p_{2,2}} + |y_4|^{p_{2,2}}\right)^{\frac{p_{\emptyset,2}}{p_{2,2}}}\right)$$
$$\times \frac{1}{2^{n-1}} \prod_{I \in I \setminus \emptyset} p_I^{\ell_I - 1} \frac{\Gamma\left[\frac{n_I}{p_I}\right]}{\prod_{k=1}^{\ell_I} \Gamma\left[\frac{n_{I,k}}{p_I}\right]}.$$

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

$$\rho(y_3, y_4) = \frac{p_{\emptyset,2}}{\Gamma\left[\frac{n_{2,2}}{p_{\emptyset,2}}\right]} \exp\left(-\left(|y_3|^{p_{2,2}} + |y_4|^{p_{2,2}}\right)^{\frac{p_{\emptyset,2}}{p_{2,2}}}\right) \frac{1}{2^{n-2}} \prod_{I \in I \setminus \{\emptyset, (\emptyset,2)\}} p_I^{\ell_I - 1} \frac{\Gamma\left[\frac{n_I}{p_I}\right]}{\prod_{k=1}^{\ell_I} \Gamma\left[\frac{n_{I,k}}{p_I}\right]}.$$

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 $\phi_s$,
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 $\phi$.
    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 $\mathcal{I}$ 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*

$$\left|\det \frac{\partial z_i}{\partial x_j}\right| = |\det W| \cdot \prod_{I \in \mathcal{I}} \left| \frac{g_I(f_I(\boldsymbol{t}_I))^{n_I - 1}}{f_I(\boldsymbol{t}_I)^{n_I - 1}} \cdot \frac{\phi_s(f_I(\boldsymbol{t}_I))}{\phi_{\perp\perp}(g_I(f_I(\boldsymbol{t}_I)))} \right|$$

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

$$\mathcal{J}_i = \left(\begin{array}{ccccccc} 1 & & & 0 & & & \\ & \ddots & & 0 & & & \\ & & 1 & 0 & & & \\ & & & \vdots & 1 & & \\ & & & 0 & & \ddots & \\ & & & 0 & & & 1 \\ \frac{\partial u_n}{\partial u_1} & & & \frac{\partial u_n}{\partial u_i} & & & \frac{\partial u_n}{\partial u_{n-1}} \end{array}\right).$$

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 $|u_q| = v_{\ell_1,\dots,\ell_m,i_1,\dots,i_t}$, $|u_n| = v_{\ell_1,\dots,\ell_d}$ with $m < d$. The derivative of u_n w.r.t. u_q is given by*

$$\begin{aligned}\frac{\partial}{\partial u_q} u_n = &-G_{\ell_1,\dots,\ell_d}(\boldsymbol{u}_{\widehat{\ell_1,\dots,\ell_d}})\cdot\ldots\cdot G_{\ell_1,\dots,\ell_{m+1}}(\boldsymbol{u}_{\widehat{\ell_1,\dots,\ell_{m+1}}})\\ &\times F_{\ell_1,\dots,\ell_m,i_1}(\boldsymbol{u}_{\ell_1,\dots,\ell_m,i_1})\cdot F_{\ell_1,\dots,\ell_m,i_1,\dots,i_{t-1}}(\boldsymbol{u}_{\ell_1,\dots,\ell_m,i_1,\dots,i_{t-1}})\cdot\Delta_q|u_q|^{p_{\ell_1,\dots,\ell_m,i_1,\dots,i_{t-1}}-1}\end{aligned}$$

with $\Delta_q = \operatorname{sgn} u_q$ and $|u_q|^p = (\Delta_q u_q)^p$. In particular

$$\begin{aligned}u_q\frac{\partial}{\partial u_q} u_n = &-G_{\ell_1,\dots,\ell_d}(\boldsymbol{u}_{\widehat{\ell_1,\dots,\ell_d}})\cdot\ldots\cdot G_{\ell_1,\dots,\ell_{m+1}}(\boldsymbol{u}_{\widehat{\ell_1,\dots,\ell_{m+1}}})\\ &\times F_{\ell_1,\dots,\ell_m,i_1}(\boldsymbol{u}_1)\cdot F_{\ell_1,\dots,\ell_m,i_1,\dots,i_{t-1}}(\boldsymbol{u}_{\ell_1,\dots,\ell_m,i_1})\cdot|u_q|^{p_{\ell_1,\dots,\ell_m,i_1,\dots,i_{t-1}}}.\end{aligned}$$

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

$$f_{I,i_m}(\boldsymbol{u}_{I,i_m})^{p_{I,i_m}} = \sum_{k=1}^{\ell_{I,i_m}} F_{I,i_m,k}(\boldsymbol{u}_{I,i_m,k})\, f_{I,i_m,k}(\boldsymbol{u}_{I,i_m,k})^{p_{I,i_m,k}}. \qquad (14)$$

Now let $L = \ell_1,\dots,\ell_{d-1}$ 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 $m \leq k < n$. Remember, that by Lemma (13) the terms $u_q \frac{\partial}{\partial u_q} u_n$ for $m \leq q < n$ 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 $m \leq k < n$ 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. ■

The derivative of the volume is not the surface area in general.
See errata on sinzlab.org

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

$$\begin{aligned}\int_{f(\boldsymbol{x})\leq R} d\boldsymbol{x} &= \int_{f(\boldsymbol{x}_{1:n-\ell_I},v_I)\leq R}\int_{\tilde{\boldsymbol{u}}\in\mathcal{V}_+^{\ell_I-1}} v_I^{\ell_I-1}\left(1-\sum_{i=1}^{\ell_I-1}\tilde{u}_i^{p_I}\right)^{\frac{1-p_I}{p_I}} dv_I d\tilde{\boldsymbol{u}} d\boldsymbol{x}_{1:n-\ell_I}\\ &= \int_{f(\boldsymbol{x}_{1:n-\ell_I},v_I)\leq R} v_I^{n_I-1} dv_I d\boldsymbol{x}_{1:n-\ell_I} \times \int_{\tilde{\boldsymbol{u}}\in\mathcal{V}_+^{\ell_I-1}}\left(1-\sum_{i=1}^{\ell_I-1}\tilde{u}_i^{p_I}\right)^{\frac{n_{I,\ell_I}-p_I}{p_I}} d\tilde{\boldsymbol{u}}.\end{aligned}$$

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

$$\begin{aligned}\int_{\tilde{\boldsymbol{u}}\in\mathcal{V}_+^{\ell_I-1}}\left(1-\sum_{i=1}^{\ell_I-1}\tilde{u}_i^{p_I}\right)^{\frac{n_{I,\ell_I}-p_I}{p_I}} d\tilde{\boldsymbol{u}} &= \frac{1}{p_I^{\ell_I-1}}\int_{\sum s_i\leq 1}\left(1-\sum_{i=1}^{\ell_I-1}s_i\right)^{\frac{n_{I,\ell_I}}{p_I}-1}\prod_{i=1}^{\ell_I-1}s_i^{\frac{1}{p_I}-1}d\boldsymbol{s}_{\ell_I-1}\\ &= \frac{1}{p_I^{\ell_I-1}}\prod_{k=1}^{\ell_I-1}B\left[\frac{\sum_{i=1}^k n_{I,k}}{p_I},\frac{n_{I,k+1}}{p_I}\right]\\ &= \frac{1}{p_I^{\ell_I-1}}\prod_{k=1}^{\ell_I-1}B\left[\frac{k}{p_I},\frac{1}{p_I}\right]\end{aligned}$$

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 solve the integral

$$\int_{f(\boldsymbol{x}_{1:n-\ell_I},v_I)\leq R} v_I^{n_I-1} dv_I d\boldsymbol{x}_{1:n-\ell_I}. \tag{15}$$

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, $\boldsymbol{x}_{\widehat{\mathcal{J}}}$ 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 $\boldsymbol{x}$):

$$\int_{f(\boldsymbol{x}_{\widehat{\mathcal{J}}},\boldsymbol{v}_\mathcal{J})\leq R}\prod_{J\in\mathcal{J}} v_J^{n_J-1} d\boldsymbol{v}_\mathcal{J} d\boldsymbol{x}_{\widehat{\mathcal{J}}}.$$

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}\left(v_I\tilde{u}_k\right)^{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 $\boldsymbol{x}_{\widehat{\mathcal{J}}}$ 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

$$\rho(\boldsymbol{x}_{\widehat{\mathcal{J}}}, \boldsymbol{v}_{\mathcal{J}}) = \frac{\phi(f(\boldsymbol{x}_{\widehat{\mathcal{J}}}, \boldsymbol{v}_{\mathcal{J}}))}{\mathcal{S}_f(f(\boldsymbol{x}_{\widehat{\mathcal{J}}}, \boldsymbol{v}_{\mathcal{J}}))}\prod_{J\in\mathcal{J}} v_J^{n_J-1}.$$

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

$$g(\|\boldsymbol{v}_{1:\ell_\emptyset}\|_{p_\emptyset}) = \frac{p_\emptyset^{\ell_\emptyset-1}\Gamma\left[\frac{n}{p_\emptyset}\right]}{\|\boldsymbol{v}_{1:\ell_\emptyset}\|_{p_\emptyset}^{n-1} 2^m \prod_{k=1}^{\ell_\emptyset}\Gamma\left[\frac{n_k}{p_\emptyset}\right]}\phi\left(\|\boldsymbol{v}_{1:\ell_\emptyset}\|_{p_\emptyset}\right)$$

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 $g_I(r) = (\mathcal{F}_{\perp\perp}^{-1} \circ \mathcal{F}_s)(r)$. 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 $I, J_1, \ldots, J_k$ be the path of inner nodes from I to y_i, then

$$\frac{\partial}{\partial y_i} f_I(\boldsymbol{t}_I) = v_I^{1-p_I} v_{J_1}^{p_I - p_{J_1}} \cdot \ldots \cdot v_k^{p_{J_{k-1}} - p_{J_k}} |y_i|^{p_{J_k}-1} \cdot \mathrm{sgn} y_i .$$

If we denote $r = f_I(\boldsymbol{t}_I)$ and $\zeta_i = v_{J_1}^{p_I - p_{J_1}} \cdot \ldots \cdot v_k^{p_{J_{k-1}} - p_{J_k}} |y_i|^{p_{J_k}-1} \cdot \mathrm{sgn} y_i$ 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.