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

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

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

## Abstract

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

## 1 Introduction

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

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

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

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

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

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

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

## 2 Models and Methods

### 2.1 Decorrelation and Filters

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

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

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

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

### 2.2 $L_p$-spherically Symmetric Distributions

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

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

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

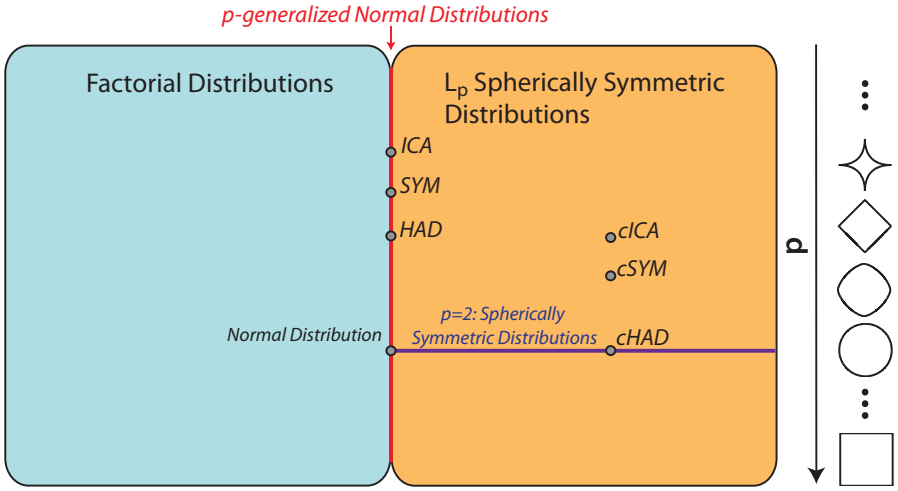


Figure 1: The spherically symmetric distributions are a subset of the $L_p$-spherical symmetric distributions. The right shapes indicate the iso-density lines for the different distributions. The Gaussian is the only $L_2$-spherically symmetric distribution with independent marginals. Like the Gaussian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

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

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

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

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

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

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

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

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

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

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

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

## 3 Experiments and Results

### 3.1 Dataset

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

### 3.2 Evaluation Measure

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

### 3.3 Experiments

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

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

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

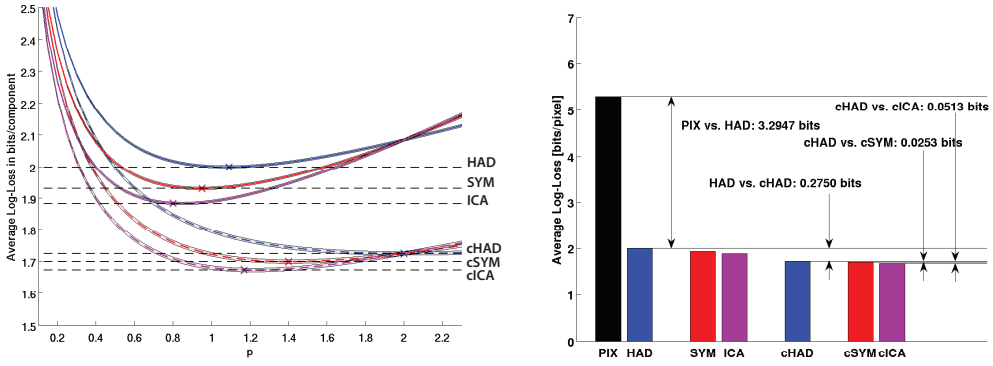


Figure 2: ALL in bits per component as a function of $p$. The linewidth corresponds to the standard deviation over ten pairs of training and test sets. *Left:* ALL for the bases HAD, SYM and ICA under the $p$-generalized Normal (HAD, SYM, ICA) and the factorial $L_p$-spherically symmetric model with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

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

### 3.4 Results

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

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

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

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

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

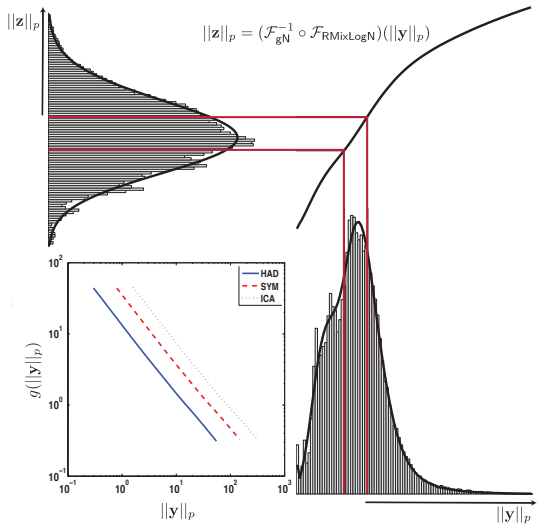


Figure 3: The curve in the upper right corner depicts the transformation $||\boldsymbol{z}||_p = (\mathcal{F}_{\mathsf{gN}}^{-1} \circ \mathcal{F}_{\mathsf{RMixLogN}})(||\boldsymbol{y}||_p)$ of the radial component in the ICA basis for gray scale images. The resulting radial distribution over $||\boldsymbol{z}||_p$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

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

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

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

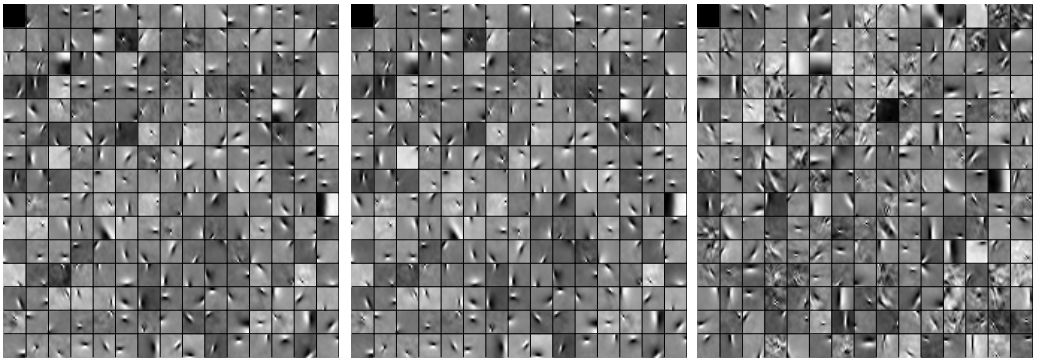

Figure 4: Filters optimized for ICA (*left*) and for the $p$-spherically symmetric model with radial mixture of Log-Normal distributions starting from the ICA solution (*middle*) and from a random basis (*right*). The first filter corresponds to the DC component, the others to the filter shapes under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

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

## 4 Summary

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

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

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

## Acknowledgements

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

## References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[17] S. Lyu and E. P. Simoncelli. Nonlinear extraction of 'independent components' of elliptically symmetric densities using radial Gaussianization. Technical Report TR2008-911, Computer Science Technical Report, Courant Inst. of Mathematical Sciences, New York University, April 2008.

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

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

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

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

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

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

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

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

# Extra Material

## 1. Data Preprocessing

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

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

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

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

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

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

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

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

## 2. Measures Of Redundancy

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

## 3. $L_p$-Spherically Symmetric Distributions

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

**Definition 1.** *$p$-Norm*

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

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

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

**Definition 2.** $p$**-Sphere**

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

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

**Lemma 3.** ***Transformation in Radial and Spherical Coordinates*** [3]

*Let* $\mathbf{y} = (y_1, ...y_n)^\top$ $n \geq 2$ *be a vector in* $\mathbb{R}^n \backslash \{\mathbf{0}\}$. *Consider the transformation*

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

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

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

*are equal and are given by*

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

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

Let

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

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

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

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

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

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

Therefore, the Jacobian, is given by

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

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

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

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

**Theorem 4.** *$p$-Spherical Uniform Distribution* [3]

*Let $Y = (Y_1, ..., Y_n)^\top$ be a random vector. Let the $Y_i$ be i.i.d. distributed with p.d.f.*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Hence,

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

□

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

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

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

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

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

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

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

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

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

3, we can transform the integral into

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

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

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

which is equivalent to

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

Therefore,

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

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

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

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

(see Theorem 4).

**Lemma 7.** ***Probability Density Functions*** [2]
*Let $Y = (Y_1, ..., Y_n)^\top$ be an $n$-dimensional random variable with $P\{Y = \mathbf{0}\} = 0$ and a density of the form $Y \sim \tilde{\varrho}(||\mathbf{y}||_p^p)$. Then the following three statements hold:*

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

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

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

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

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

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

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

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

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

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

From that, we can immediately derive

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

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

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

Thus,

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

and

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

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

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

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

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

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

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

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

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

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

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

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

Therefore,

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

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

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

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

### 3.2. **Distributions.**

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

This yields

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

## 5. ALL Scores For Color and Gray Value Images

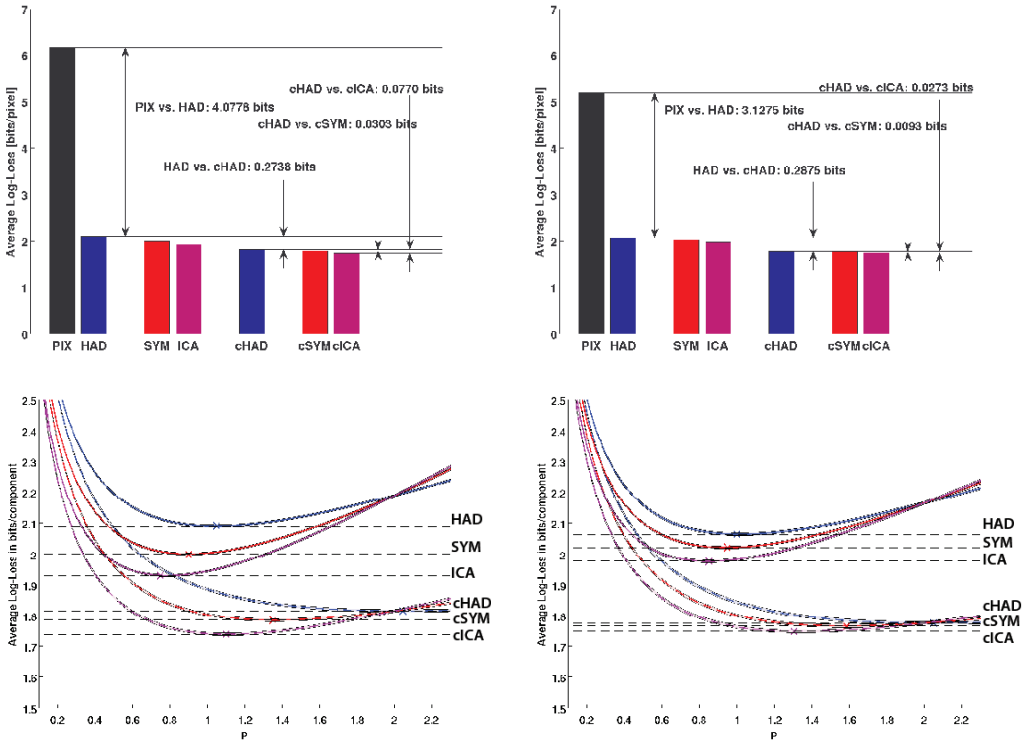


FIGURE 5.1. ALL in Bits per component as a function of $p$ for achromatic (*right*) and chromatic (left) images. For computational efficiency both plots have been computed on patches of size $7 \times 7$. The slightly brighter envelope depicts the standard deviation over ten pairs of training and test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

## References

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

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

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