Complex Cell Pooling and the Statistics of Natural Images

Aapo Hyvärinen, Urs Köster [*]

Basic Research Unit of the Helsinki Institute for Information Technology,
Department of Computer Science, University of Helsinki, Finland

April 25, 2007

Abstract

In previous work, we presented a statistical model of natural images that produced outputs similar to receptive fields of complex cells in primary visual cortex. However, a weakness of that model was that the structure of the pooling was assumed a priori and not learned from the statistical properties of natural images. Here, we present an extended model in which the pooling nonlinearity and the size of the the subspaces are optimized rather than fixed, so we make much fewer assumptions about the pooling. Results on natural images indicate that the best probabilistic representation is formed when the size of the subspaces is relatively large, and that the likelihood is considerably higher than for a simple linear model with no pooling. Further, we show that the optimal nonlinearity for the pooling is squaring. We also highlight the importance of contrast gain control for the performance of the model. Our model is novel in that it is the first to analyze optimal subspace size and how this size is influenced by contrast normalization.

Keywords: Independent Subspace Analysis, Natural Image statistics, L_p-norm spherical distribution, Contrast gain control

1 Introduction

The low-level processing in primary visual cortex (V1) is typically modeled as a linear filtering followed by an energy pooling stage [1]. While there is no clear evidence that such a hierarchical structure is present in the brain [2, 3], this kind of model has been applied successfully to explain a variety of observations [4]. The filtering stage, characterized by receptive fields which are localized, oriented and bandpass can be modelled by Gabor-like filters. In the pooling stage, the squares of the linear filter outputs are summed among filters of similar orientation, frequency, and location.

An important question is the functional utility of the observed receptive field shapes. Here, the currently dominant approach is based on linking the receptive field structure to the statistical structure of ecologically valid stimuli, i.e. natural images. Successful approaches include sparse coding and Independent Component Analysis (ICA) of natural images [5, 6, 7].

The pooling of these responses in the second stage produces features with invariance to spatial phase and exact location. Attempts have been made to model this with Independent Subspace Analysis (ISA) [8]. ISA is an extension to ICA that groups the features into multidimensional subspaces, inside which dependencies are allowed, and minimizes dependencies between the norms of projections onto subspaces. This produces outputs very similar to those of complex cells, where features with similar location and orientation but different phase are pooled. However, some authors [9, 10] have objected that it is not valid to extract complex cell properties from static images by this approach, reasoning that there is no justification for a forced pooling of the ICA features to complex cells.

This paper discusses how ISA can be extended to learn the optimal subspace size and pooling nonlinearity from the data by directly comparing the likelihoods of image models. The extended ISA model is a two layer model which does not simply square and sum fixed groups simple cell outputs, but which estimates the optimal nonlinearity and subspace size. This makes it possible to test the hypothesis that pooling is favorable.

First, we introduce a novel likelihood function which is based on a L_p-spherical probability distribution and covers both ICA and ISA (subspace) type of models. Then we present the results from our simulations that show that a relatively large subspace size is optimal, depending on what kind of nonlinearity accompanies the pooling. We emphasize the importance of contrast gain control, for pooling to be the favorable model.

[*] E-mail: aapo.hyvarinen@helsinki.fi, urs.koster@cs.helsinki.fi. Corresponding author is A.H.

2 Methods

2.1 Independent Subspace Analysis

A fundamental model for natural images is given by ICA [6]. ICA tries to find filters that are as independent as possible. Image patches $I(x,y)$ are modeled as a superposition of features $A_k(x,y)$ as in

$$I(x,y) = \sum_{k=1}^{q} A_k(x,y) s_k \tag{1}$$

where the activities or independent components s_k are assumed independent and non-Gaussian. The index j runs over the q different features. This is a system of linear equations which can be inverted if the number of features is equal to the dimensionality of the data. We can then compute the independent components,

$$s_k = \sum_{x,y} W_k(x,y) I(x,y) = \langle W_k, I \rangle \tag{2}$$

where the filters W_k correspond to the inverse of the features, and $\langle . \rangle$ denotes an inner product. The filters are found by modelling the activities by a supergaussian probability density function (PDF), e.g. a Laplace distribution. Since the q individual components are assumed independent their joint density factorizes,

$$p(s_1, s_2, \dots, s_q) = \prod_{k=1}^{q} p_k(s_k) = \prod_{k=1}^{q} \frac{1}{\sqrt{2}} e^{-\sqrt{2}|s_k|} \tag{3}$$

Then we maximize the log-likelihood given this probability. This can easily be done with gradient methods, but faster alternatives are available [11]. This method has been employed to explain simple cell properties in terms of statistical optimality [5, 6].

ISA is a multidimensional equivalent to ICA where not the outputs of linear filters themselves are assumed independent, but the images are projected onto subspaces and the norms of these projections are assumed independent [8]. In practice the individual filter outputs are squared, divided into groups (subspaces), the members of a subspace are summed and finally the square root of the sum is taken as in

$$u_j = \sqrt{\sum_{i \in S_j} s_i^2} \tag{4}$$

where u now denotes the "output" of one subspace with the index i running over the n constituent filters in the j^{th} subspace whose indices are in the set S_j. In analogy to ICA using the Laplace distribution we can simply define the log-probability as the negative of the norm of the projection summed over the m individual subspaces. Finally, the parameter b is adjusted to make the variance of the s_i equal to one, and Z to normalize the probability distribution:

$$\log p(s_1, ..., s_q) = \sum_{j=1}^{m} \left(-\log Z_j - \frac{\sqrt{\sum_{i \in S_j} s_i^2}}{b} \right) \tag{5}$$

For estimating the model, we maximize the likelihood of the distribution over the data with respect to the features and the parameters of the nonlinearity. This maximizes the independence of the subspace norms. The obvious interpretation of subspaces is that they correspond to complex cells. The pooling reproduces complex cell properties like phase invariance [8]. However, other authors have criticized that using a fixed subspace size and forced pooling is too ad hoc to explain the properties of complex cells. To address this issue, we describe how the results can be obtained without forcing the pooling but estimating the optimal subspace size from the statistical structure of natural images.

2.2 Extensions to Independent Subspace Analysis

We propose a generalized form of ISA where the data is modeled by a log-PDF of the form

$$\log p(s_1, ..., s_q) = \sum_{j=1}^{m} \left(-\log Z_j - \frac{(\sum_{i \in S_j} |s_i|^d)^{a_j}}{b_j^{a_j}} \right) \tag{6}$$

Given the observed images $I_1 ... I_t$ we obtain the log-likelihood of the model

$$\log L(W_1 ... W_k, d, a, b | I_1 ... I_T) = \sum_{t=1}^{T} \sum_{j=1}^{m} \left(-\log Z_j - \frac{(\sum_{i \in S_j} |\langle W_i, I_t \rangle|^d)^{a_j}}{.} .. b_j^{a_j} \right) \tag{7}$$

for orthonormal W_k. We sum over T observations, j runs over the different subspaces which all have dimensionality n, and the index i runs over the filters inside one subspace. This corresponds to a distribution whose isocontours have the form of n-dimensional spheres under the L_p norm, i.e. we are using an L_p-spherical distribution [12]. The subspace size n can be varied including both extremes of one filter per subspace and all filters in one subspace, so both the ICA and general ISA case are included. Further, the use of the L_p norm lifts the contraint of the squaring nonlinearity (we had a spherical subgaussian probability density with fixed $d = 2$ and $a = 1/2$ in our previous implementations of ISA) but it is still possible to manipulate the expression algebraically. In particular the normalization constant Z_j can be calculated in closed form, as shown in appendix A, to give

$$Z_j = \frac{2^n {b_j}^{n/d} n \Gamma(\frac{n}{a_j d}) \Gamma(1/d)^n}{a_j d^{n+1} \Gamma(\frac{n}{d}+1)} \tag{8}$$

The filters $W_k(x,y)$ are then optimized for maximum likelihood over the data with a gradient algorithm. The gradient step for the likelihood with respect to one filter $W_k(x,y)$ in the j^{th} subspace is straightforward to compute, yielding

$$\Delta W_k(x,y) = \gamma \frac{\partial \log L}{\partial W_k(x,y)} = \gamma b_j^{-a_j} d \left(\sum_{i \in S_j} |\langle W_i, I \rangle|^d \right)^{a_j - 1} I(x,y) |\langle W_k, I \rangle|^{d-1} \tag{9}$$

where γ is the gradient step size. This sets the frame for computing a set of independent feature subspaces from image data. For initial experiments, a brute force search was used to determine the optimal parameters d, a_j and b_j. In later experiments, we fixed $d = 2$, so the variance parameter b_j could be computed analytically.

$$b = \sqrt{\frac{n\Gamma(\frac{n}{2a})}{\Gamma(\frac{n+2}{2a})}} \tag{10}$$

Therefore, only a remained to be determined by a brute force parameter search, considerably speeding up the estimation procedure.

2.3 Identifiability of the model

To show that the estimation of the model by maximization of likelihood is valid, we created artificial data with a known subspace size and then analyzed it using L_p-ISA. The data was created as follows, using 8-dimensional subspaces as an example: First, we take 10000 samples from a 128-dimensional Gaussian distribution. Since we have $128/8 = 16$ subspaces, we also take 10000 samples from a 16-dimensional uniform distribution. Now we introduce dependencies into the subspaces. This is done by multiplying the group of 8 Gaussians generated for the subspace by the random variable from the uniform distribution. The product of the Gaussian and uniform gives a supergaussian distribution, and due to the common sample from the uniform distribution, the variables in the subspace have dependencies. This gave us the components s_i. We then created a random 128×128 mixing matrix and multiplied the vector of the s_i by that matrix, thus obtaining simulated data.

2.4 Preprocessing and Contrast Gain Control

The higher order statistical structure in natural images is very complex and cannot be fully captured by a simple two-layer model. Therefore considerable preprocessing is usually applied to simplify the statistical structure of the images and make the assumption of independence hold better. This is of particular importance with the new model presented here, since it is not forced to model the images in terms of independent sources, but is also given the option to model arbitrary dependencies when all features fall in one single subspace. We have adopted the following procedure for preprocessing, following [13]: First, the frequency spectrum of the images, which usually falls off with the second power, is normalized (i.e. the data is whitened), but cut off at high frequencies to remove sampling artifacts. In the second step, the contrast of the images is normalized. This is done by computing the activity in the neighborhood of the pixel under consideration, which is given by the weighted sum of squares of pixel values, and dividing by this activity. Since the choice of the neighborhood is an important factor for the success of the model, we have used Gaussian neighborhoods of varying size, specified by σ^2 which we varied from 12 to 32 pixels, so we could analyze the effect of the neighborhood size. See figure 1 for a visualization of these processing stages. Whitening and Contrast Gain Control (CGC) can be considered as a simple model of visual processing in the retina and lateral geniculate nucleus. The importance is evident from previous work like [14] and backed up physiologically[15].

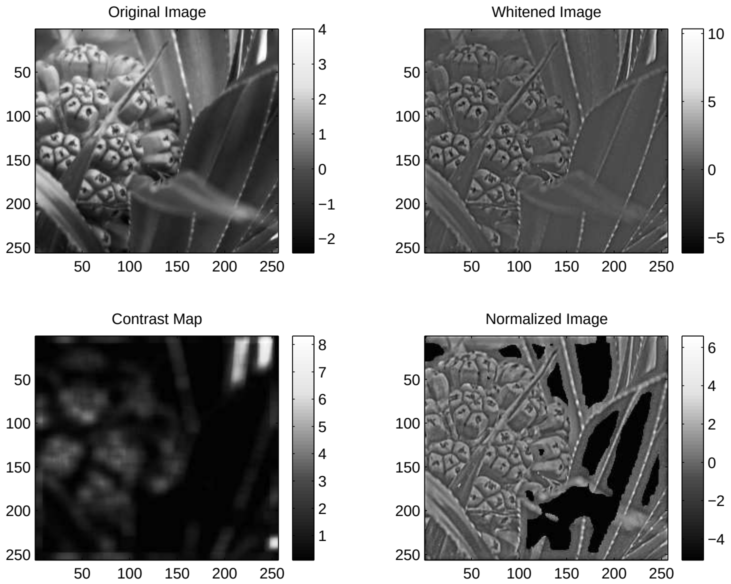


Figure 1: Example of preprocessing on a natural image: We first flatten the frequency spectrum, which puts more emphasis on the high frequency components of the image. Then we compute a contrast map, and normalize the image by dividing each pixel by its associated contrast. Note that in order to avoid over-amplification of noise, we cut out image regions with too little contrast. After drawing random patches from the images, these are whitened again to remove residual first-order correlations that are introduced by the normalization. The image shown here is for illustrative purposes only and does not belong to the original data set.

We compared this to a different method of contrast gain control, where image patches are sampled first, then whitened and divided by their norms. This method is computationally easier to perform but has the drawback that the contrast neighborhood size is tied to the image patch size.

Finally, we randomly sampled patches of various size (12×12 ,16×16 and 24×24 pixels) from the images, and whitened these patches. Whitening correspond to decorrelation, i.e. it removes all linear dependencies from the data. Simultaneously we used PCA to reduce the dimensionality of the data, which corresponds to low-pass filtering. For 12×12 pixels we retained 120, for 16×16 pixels 240 and for 24×24 pixels 480 dimensions. Removing the highest frequency components makes sure sampling artifacts do not affect the results. The exact number of dimensions was chosen so it can be factorized into a large number of different possible subspace sizes. All image patches were randomly selected from natural images taken from P.O. Hoyer's *imageica* package [16].

2.5 Simplifications of the model

2.5.1 Initializing the algorithm

Our original estimation procedure used a gradient algorithm to simultaneously update the features and the parameters of the nonlinearity. However, the algorithm converged to local minima indicated by multiple Gabor features in some of the receptive fields. This problem could not be alleviated using a stochastic gradient method, i.e. simulated annealing, so the full gradient estimation had to be abandoned.

Therefore we decided to use an iterative method which proceeded in two stages. In the first stage, classical ISA (i.e. clamping $d = 2$ and $a = 1/2$) is used to compute a set of feature vectors W as a starting point. In the second stage, we estimated the nonlinearity parameters d, a and b with a brute force method, while keeping the features fixed. To show that this is sufficient for learning both the optimal features as well as the optimal nonlinearity parameters, we evaluated the effects of alternating training of the two layers. This did not change the results quantitatively or qualitatively in a significant way, so we could procede with the simpler problem of two separate optimizations. To further test the validity of this approach we performed experiments with 120-dimensional data from 16×16 image patches, and a subspace size of 4. Here the log-likelihood of the original method had a mean of -1.36 with a low variance, compared to of -1.34 (see results for more details) with the improved method.

The need for this brute force procedure arose because gradient descent is not a suitable algorithm for determining the correct values of the parameters a, b and d. The error surface for the optimization has a very narrow maximum where the error changes by several orders of magnitude within a small volume of parameter space, and the log-likelihood is far from being concave. Gradient algorithms are not suitable for this sort of optimization problem, so we had to resort to an iterative search of parameter space, which does not suffer from these problems, but is considerably slower.

2.5.2 ICA inside subspaces

Our estimations starting with classical ISA introduces the problem that the estimation of d may not be reliable because the ISA estimation with fixed $d = 2$ does not take into account the direction of the feature vectors inside the subspaces, but only determines them up to an orthogonal transform. Thus, estimating d after ordinary ISA estimation of the subspace may bias the value towards 2. To alleviate this problem, we decided to integrate another stage of ICA which was done inside the subspaces, to rotate the subspaces into the most supergaussian components, so we could remove the imprint of the L_2-ISA and get a better estimate of the norm parameter d. We did this using the FastICA algorithm with a tanh-nonlinearity.

To verify that this method would produce the same results as a full estimation, we also estimated features with L_p-ISA and values of d ranging from 1 to 3. We did this with image patches of 12×12 pixels. After this we estimated the optimal parameter of d by brute force again, to test if the features estimated differently would lead to a different optimum in the later parameter estimation. The estimation of the L_p-ISA model with a fixed value of d did not produce any stability problems.

3 Results

3.1 Identifiability of the model

When testing the L_p-ISA Algorithm with artificial data, we could confirm that the highest likelihood was reached for the subspace size that was embedded in the data. This shows that the algorithm is correctly identifying any subspace structure hidden in the data. As expected, when the algorithm was run with plain Gaussian data the likelihood was completely flat.

The results from these test runs are shown in figure 2.

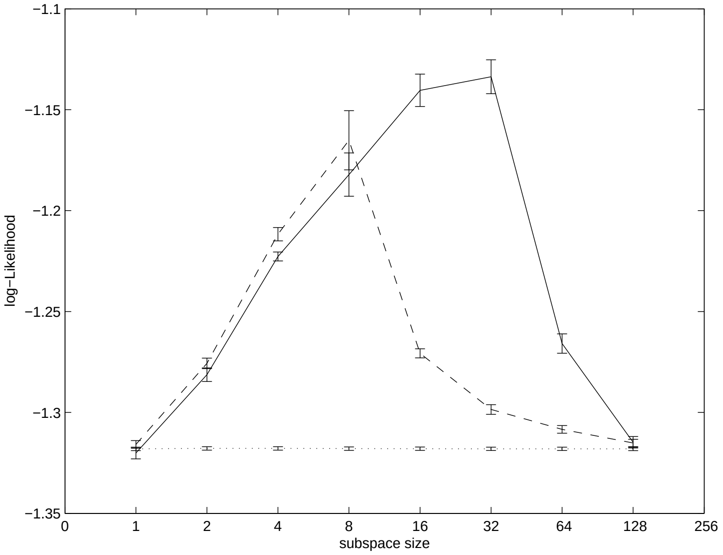


Figure 2: Test of identifiability of the model: The subspace dependencies we introduced into the Gaussian random data are correctly identified by the L_p-ISA model. The dotted line is Gaussian data, dashed line has embedded subspaces of size 8 and the solid line of 32. Error bars indicate standard error on the mean.

3.2 Pooling nonlinearity

The first result we present concerns the nature of the nonlinearity that is associated with the pooling. Our model predicts that it is a squaring operation ($d = 2$), as can be seen from the maximum likelihood estimation in Fig 3. To rule out that this value of d is merely an artifact due to our estimation of the features by ISA using a squaring

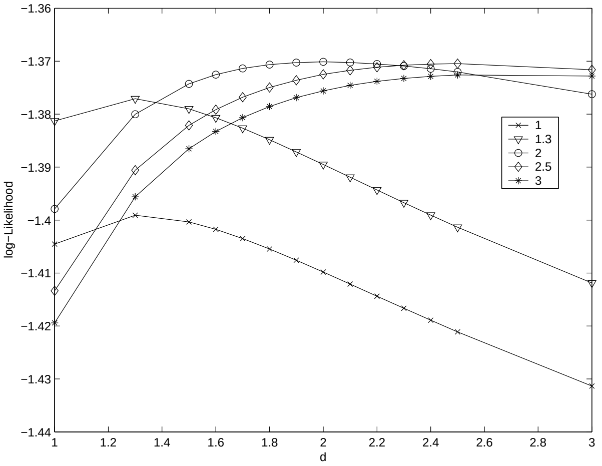


Figure 3: The superimposed plots show how the likelihood varies with norm parameter d for L_p-ISA estimations with five different norms, as indicated in the legend, corresponding to the five connected curves. The overall maximum occurs when $d = 2$ for both the ISA estimation and the brute force likelihood maximization algorithm. This shows that subspaces are for all practical purposes spherical. The experiment was performed on 8×8 image patches. Each line is the mean of five random trials.

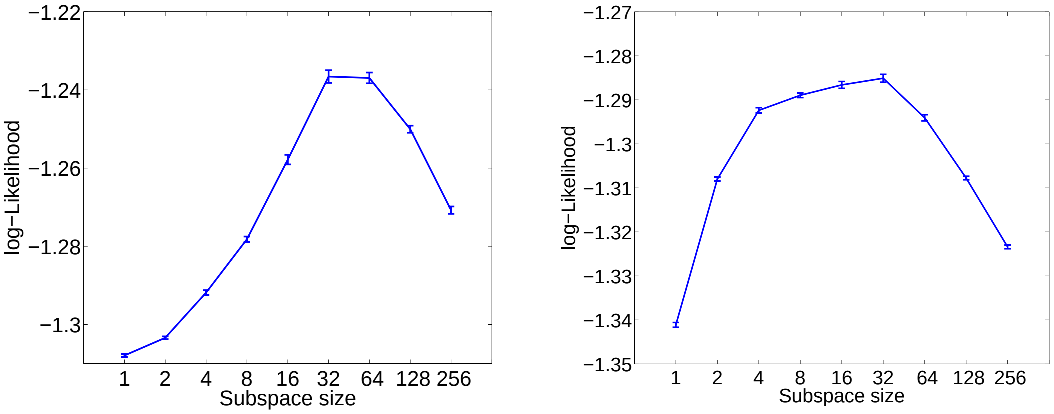
operation, we computed the features using five different values of d in the initial ISA estimation, as explained in section 2.5. The overall maximum was obtained when d was equal to two in the algorithm estimating the features, in which case the parameter optimization method also found the value $d = 2$. Therefore we decided to fix $d = 2$ in the following experiments, which considerably simplified the brute-force parameter search, as the value for b can also be calculated in closed form in this special case of spherical subspaces. The proof for this can be found in Appendix B.

3.3 Finite optimal subspace size

The most notable aspect of our work is that we can directly compare the likelihood of ICA and ISA models. We found that for a range of subspace dimensionalities ISA produces a higher likelihood and is therefore a better model of the data. The optimal subspace size is strongly dependent on the size of the image windows and wether CGC is performed, but it is not influenced by the details of the CGC procedure. We tested this on images patches ranging in size from 8×8 pixels to 24×24 pixels. This finding provides evidence that some of the dependencies of natural images, that cannot be removed by a linear transform, can be captured by subspaces or complex cells. This gives complex cell receptive fields a statistical justification in terms of efficient coding. In figures 4 and 5 we show how the likelihood changes with subspace size, and that a maximum is reached for a subspace size that depends on the size of the image window. As the individual estimation of a for each subspace leads to a larger number of free parameters for smaller subspace size, which would naturally favor small subspaces, we investigated if the results reproduced if we clamped the number of parameters by fixing a. We found that for a limited range of values this was indeed the case. fig 4 b) was computed with a arbitrarily fixed to 0.2, a value which was typically encountered for the optimal subspace size. While the overall likelihood plummets, the maximum is still at 32, which proves that the maximum was not an artifact due to the variable number of parameters. An intuitive explanation for the approximately inversely proportional relation between a and the subspace size we observed stems from the the relatively larger probability volume far away from the origin for higher dimensions. This is compensated by small values of a moving probability volume towards the origin. For this reason, a was estimated for each subspace unless otherwise noted.

3.4 Influence of Contrast Gain Control

Our experiments with different kinds of CGC on images patches of various size indicate that preprocessing to reduce dependencies has a significant influence on the performance of the ISA model. If no CGC is performed, the dependencies especially in small image patches are so strong that the highest likelihood is given to the case where all linear filters are pooled into one big subspace as shown in figure 5 for 8×8 pixel windows. Only by performing



(a) estimating *a*

(b) fixed *a*=0.2

Figure 4: Dependency of the likelihood of natural image data on subspace dimensionality. A more pronounced optimum appears when *a* is optimized, but it is persistent for fixed *a* as well. Error bars indicate standard error on the mean from 10 trials with different random seeds. Image patches were of size 24 × 24 for this experiment, and CGC was performed on individual image patches.

contrast gain control, subspaces of finite size are formed for the small image patches. With bigger image windows, e.g. 16 × 16 even without CGC a significant maximum is found for a finite subspace size. In both cases, however, a more significant maximum is obtained if divisive CGC is performed.

To analyze how the dependencies between filters are affected by CGC, we plotted the correlations of squares (energy dependencies) of the filter outputs as shown in figure 6. Since linear correlations are removed by the whitening, energy correlations are the dominant term. It can be seen that the histogram moves towards the origin when CGC is performed, i.e. the overall amount of correlations is reduced considerably. We found that the exact method of CGC does not have a significant effect on the results. However, a tradeoff considering the strength of the normalization, which is controlled by the contrast neighborhood size, has to be made: Too strong CGC causes negative energy correlations, instead of just moving them towards zero.

4 Discussion

4.1 Related work

Our results show that the pooling in ISA can be justified in terms of statistical modelling of natural images, and that Complex Cell responses can be modelled in terms of the statistical properties of static natural images. Here we compare our results to similar experiments, and then highlight the key features that are unique to our model.

Firstly we would like to draw attention to the recent work by Karklin and Lewicki [17], which also describes a two-layer model of natural images. The main difference is that in their model the second layer is a general linear transform, whereas we constrain the second layer to pool inputs for computing a norm. The cost for the more general second layer is that the model is not normalizable and therefore requires the use of approximations for the optimization. Another important difference is that fact that the first layer of the Karklin and Lewicki model is identical to ICA, and the authors claim that the presence of the second layer does not affect the optimal features in the first layer. On a more technical level, the key idea of the work is to model common variances within groups of variables. The authors take the variance to be the higher order feature underlying the data. It seems fair to link this to the squaring operation in our model, because ”energy” or ”activation” cannot clearly be distinguished from variance.

The results are quite different from ours in that the authors find global structure in the images in the form of spatially extended and diverse higher order units. They are able to distinguish areas of low and high contrast, and they can classify image patches into visually different groups based on variance patterns. This is strikingly different from our results showing higher order features describing local invariances. The reason for this difference is likely to be only partly due to the different technical constraints of the two models, but is probably related to our preprocessing with CGC. Conceivably it removes the global dependencies modelled by Karklin and Lewicki and leaves local structure which replicate the properties of Complex Cells. Therefore our model tries to learn a sparse pooling of only a few local features, which are themselves optimized to allow for this sparse pooling.

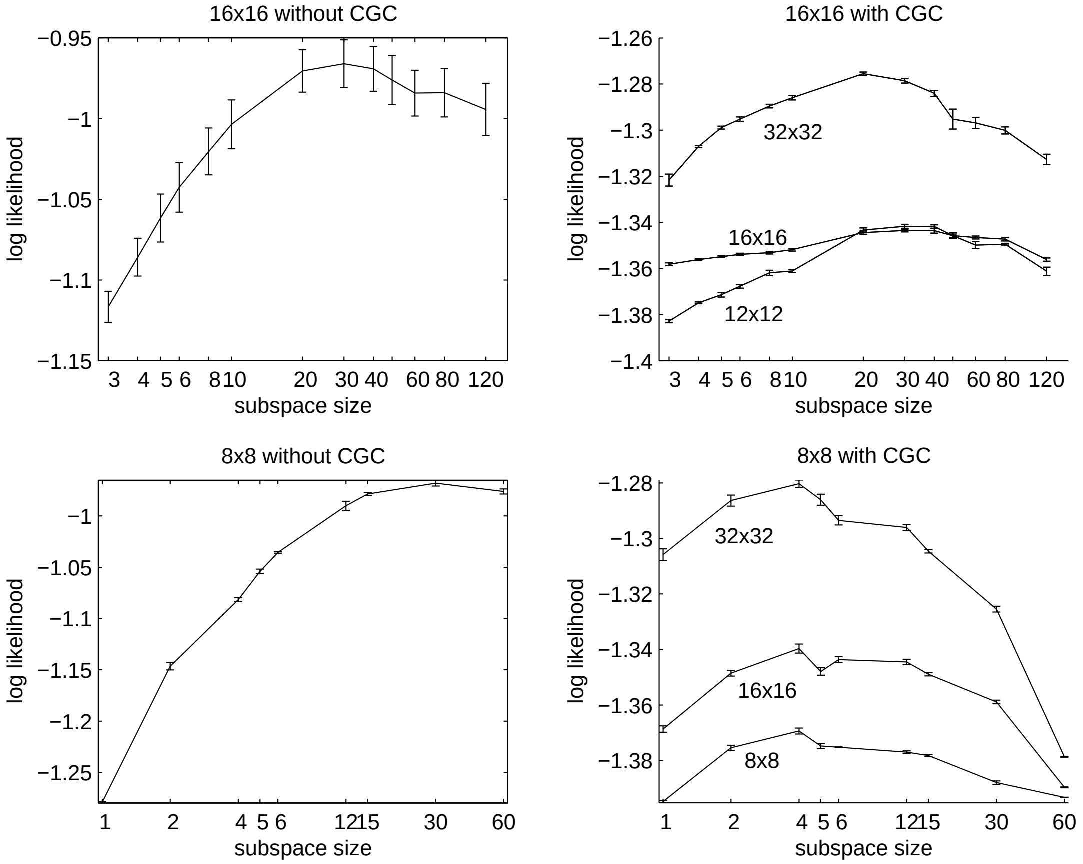


Figure 5: Dependency of optimal subspace size on the size of the image window and the CGC neighborhood. The top row shows the likelihood functions computed from 240-dimensional data (16 × 16 image patches), bottom row is for 60-dimensional data (8 × 8 image patches). The graphs on the left hand side show the resulting likelihood without CGC, the graphs on the right with CGC. CGC was performed with Gaussian neighborhoods of size as indicated in the legend, where the uppermost plot corresponds to the uppermost legend item etc. Error bars indicate standard error on the mean from 6 trials with different random seeds. Please note that only the position of the maximum of the individual curves can be compared, but not the overall likelihood value. In order to compare these, it would be necessary to compute the likelihood of the *original data* in a model where the data is transformed by CGC and linear transform into independent subspaces. Here we ignore the influence of the CGC procedure entirely, so there is a significant offset between the lines.

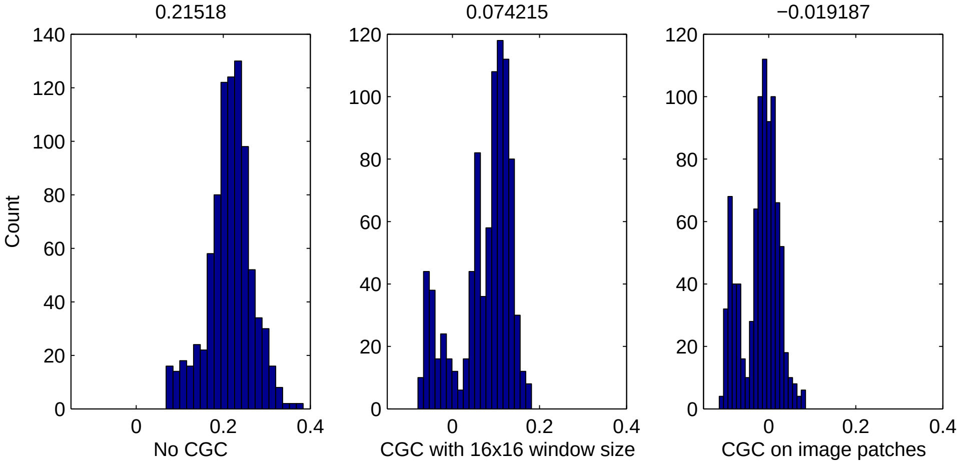


Figure 6: Effect of CGC on energy correlations: Before performing CGC, the average energy correlations of filter outputs are 0.2, as plotted above the corresponding graph. With CGC performed over a neighborhood of 16×16 pixels, the correlations are reduced and shifted towards zero. The average now is 0.07 and there are some slight negative correlations. In the last plot, CGC on individual image patches (12×12) is shown to be almost identical to the former method. Because of the smaller neighborhood, more negative correlations appear. These plots corresponds to 12×12 image patches.

Also closely related to our work is the ”Product of Student-t Models” of Osindero et al. [18]. The model described in this publication is an alternative to ICA-based approaches, and uses Contrastive Divergence [19] to optimize a hierarchical ”product of experts” type distribution. By clamping the second layer to the identity matrix, this model produces Simple Cell-like responses, and subsequent learning of the second layer gives rise to connections between cells in a similar fashion to topographic ICA. By grouping the units with the strongest connections, the authors are able to produce Complex Cell-like receptive fields with phase-invariance, while spatial frequency and orientation tuning remain unchanged. In contrast to our model, no limitation on the number of simple cells feeding one complex cell is enforced. This might be considered as an advantage since it is likely to reflect the properties of biological neural networks, which is in contrast to the advantage of a principled estimation that our model offers.

Finally, [20] shows how modelling the energy dependencies between neighboring Gabor-wavelets can be used for state-of-the-art denoising of images.

4.2 Likelihood and Sparseness

Our experiments showed that the likelihood increases as one goes from ICA to the subspace model, and then decreases again as very large subspaces are reached. For our experiments with 24×24 pixels in particular, subspaces with a size of 16 to 32 have a significantly larger likelihood than the simple ICA model. This shows that the generalized ISA model provides a better statistical description for natural images than ICA, and hence provides an adequate model of observed complex cell properties. It directly justifies pooling simple cell outputs from the statistics of static images, and shows that the dynamics of image sequences are not required as claimed by other authors[9, 10].

It is important to note that as the size of the subspaces grows, their number decreases. This means that with large subspaces, only a few highly complex features were estimated. In principle this could have been avoided by working with a fixed number of subspaces, but this would require an overcomplete set of linear filters, which is difficult to estimate with ICA models. In any case it is natural to expect that experiments of this kind would lead to pooling over an even large number of filters. In plot 7 it can be seen that the features at a subspace size of 16 are much less clean and localized than results that are known from ISA with 2 or 4 features per subspace. This is presumably because there are only 16 complex features which have to represent the whole image. It is interesting to note that the maximum likelihood still occurs at this rather large subspace size, again indicating that strong residual dependencies make image data difficult to approximate with ICA models.

It should be noted that the goal of the estimation is not to maximize sparseness but this is merely a vehicle for maximizing likelihood. The sparseness cannot be compared meaningfully between different subspace sizes because the number of cells is different. As the number of cells decreases with increasing subspace size, it is

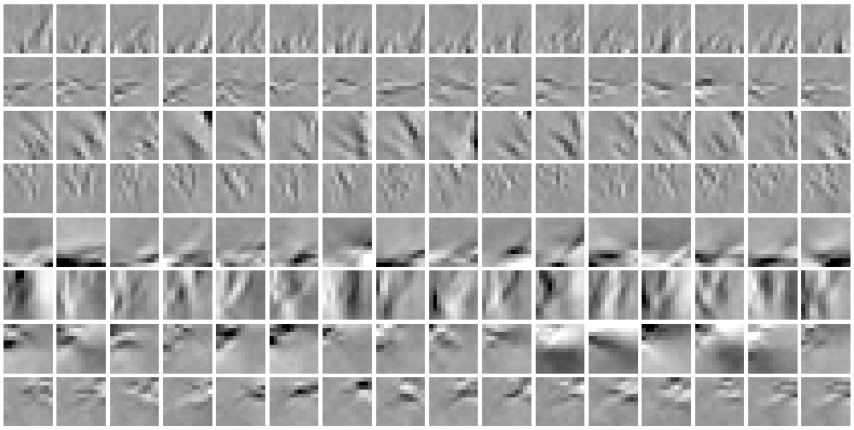


Figure 7: A random selection of features at maximum likelihood with a subspace size of 16. One row corresponds to one subspace. Note that there is a relatively small number of features, so they are less selective than ICA features.

natural that a larger percentage of these cells need to be activated, but the total number of active cells may still be less than with the linear features. Thus the sparseness would have to be weighted in some way, which would make comparisons difficult. Therefore likelihood is the only principled method of comparing ICA and ISA models.

4.3 Pooling nonlinearity

Our results indicate that the parameter d is always close to 2, which means that the subspaces are in fact spherical in the Euclidean sense. This is in accordance with the results from physiological experiments, which agree that nonlinearities are best modelled by squaring and not e.g. by taking the absolute value[15]. Other recent theoretical work such as [21] also supports the case for a squaring nonlinearity. The authors in [21] describe a somewhat simpler model were a pooling of two subunits is forced. Since they allow for an individual estimation of the optimal nonlinearity for each complex cell, they are able to show that in addition to the predominant squaring, some units compute the L_∞-norm, corresponding to selecting the maximum input. The main difference to our model is that the authors use the principle of temporal coherence on movie sequences to obtain the results that we achieve from efficient coding on static images. Since they mainly obtain an exponent of two this result further supports the notion that this is not merely an artifact of our particular method, but indeed an underlying feature of the data.

The parameter a however, which s related to the sparseness of the distribution, varied over a considerable range, so it was important to determine its optimal value individually for each subspace size. There was a strong trend of decreasing a with increasing subspace size.The fact that we find a close to 0.5 for small subspaces confirms that the standard ISA model which fixes a to 0.5 - and has been used with subspace size of two or four - is in good agreement with our new model. Only for larger subspaces dimensionalities, as those we found to have maximum likelihood, we find a to be considerably smaller.

4.4 Contrast Gain Control

Our experiments showed that the way CGC is performed on the data has a strong impact on the performance of the model. It is widely accepted that one important function of the visual pathway up to V1 is normalization and gain control[14]. This is not only required in real world perception, where luminance changes over six orders of magnitude and more occur frequently, but it is also crucial in our model to make the underlying assumption of ICA more realistic, i.e. that there exist independent sources which are mixed linearly. However it is not clear whether there is a ’correct’ way to perform CGC, and what this would be. It should be noted that unlike whitening, CGC is a nonlinear process in which information is lost, so the gain from a simpler statistical structure has to be weighted against a loss in image contents.

5 Conclusion

We have demonstrated that our proposed method of L$_p$-ISA is sufficient to learn complex cell responses from static natural images.

After it was established that simple cell receptive fields can be obtained by imposing statistical constraints on image data [5, 6], it was natural to investigate if complex cell properties could also be obtained from the statistical properties of static images. Basic ISA indicated that forced pooling of simple cells gives complex cells [8], but the question remained if the pooling could also be estimated from the statistics of static images. With the present results we have put forward strong evidence that this is the case, by directly comparing the likelihood of ICA (simple cell) and ISA (complex cell) models, and finding that subspaces do in fact provide a description of natural images with a higher likelihood than single independent features do. This is not only of theoretical interest, as we also found evidence about the nature of the nonlinearity that is used for the pooling: We have shown that from a statistical point of view squaring and not e.g. absolute value rectification fits image data better. For the future we propose to exploit the idea of learned pooling further by adding another stage of pooling to the model, in order to predict what the next stages of processing, for example in V2, might be [22].

6 Acknowledgments

Urs Köster is kindly supported by a studentship from the *”Alfried Krupp von Bohlen und Halbach-Stiftung”*. Thanks to Patrik Hoyer for comments on the manuscript.

A The multivariate L_p-norm spherical distribution

A.1 Likelihood function

We consider the following probability density function in an n-dimensional (sub)space:

$$p(\mathbf{s}) = \frac{1}{c}\exp(-\frac{(\sum_{i=1}^{n}|s_i|^d)^a}{b^a}) \tag{11}$$

We want to determine the constant c (previously referred to as Z_j) as a function of a, b, and d so that this is a proper probability density (integrates to one). Closely related results were already derived in Lemma 2.3. of [12], and Eq.(3)-(4) of [23]. Here we provide a simpler proof based on geometrical considerations.

We define the d-sphere of radius r_0, denoted by $S_n^d(r_0)$ as a generalization of an ordinary sphere as

$$S_n^d(r) = \{\mathbf{s} \in \mathbb{R}^n \mid \sum_{i=1}^{n}|s_i|^d \leq r^d\} \tag{12}$$

i.e. the set where the d-norm is smaller than a given quantity. We will need to know the volume of the d-sphere. The following lemma is proven below.

Lemma 1 *The volume of S_n^d is given by*

$$V(S_n^d(r)) = r^n 2^n \Gamma(1/d)^n d^{-n} \Gamma(n/d+1)^{-1} \tag{13}$$

In order for p to be a proper probability density, we need to have

$$c = \int \exp(-\frac{(\sum_{i=1}^{n}|s_i|^d)^a}{b^a})d\mathbf{s} \tag{14}$$

As the probability density only depends on the d-norm, we make a transformation that is somewhat similar to the n-dimensional spherical coordinates. We define

$$r = (\sum_{i=1}^{n}|s_i|^d)^{1/d}. \tag{15}$$

We can evaluate the integral by using the differential of V as follows:

$$\int \exp(-\frac{(\sum_{i=1}^{n}|s_i|^d)^a}{b^a})d\mathbf{s} = \int \exp(-\frac{r^{ad}}{b^a})dV \tag{16}$$

where dV is the differential element of V, i.e. the infinitesimal change in volume induced by r. We have by definition of differentials and by the Lemma

$$dV = \frac{dV}{dr}dr = r^{n-1}n2^n\Gamma(1/d)^n d^{-n}\Gamma(n/d+1)^{-1}dr \tag{17}$$

so we need to have

$$c = \int \exp(-\frac{r^{ad}}{b^a})\frac{dV}{dr}dr \tag{18}$$

$$= \int \exp(-\frac{r^{ad}}{b^a})r^{n-1}n2^n\Gamma(1/d)^n d^{-n}\Gamma(n/d+1)^{-1}dr \tag{19}$$

Now we make the following transformation to q:

$$q = \frac{r^{ad}}{b^a} \Leftrightarrow r = b^{1/d}q^{1/(ad)} \tag{20}$$

for which the differential element can be computed as

$$\frac{dr}{dq} = \frac{b^{1/d}}{ad}q^{1/(ad)-1} \tag{21}$$

which gives

$$c = n2^n\Gamma(1/d)^n d^{-n}\Gamma(n/d+1)^{-1}\int \exp(-q)q^{(n-1)/(ad)}b^{(n-1)/d}\frac{b^{1/d}}{ad}q^{1/(ad)-1}dq \tag{22}$$

$$= n2^n\Gamma(1/d)^n d^{-n}\frac{b^{n/d}}{ad}\Gamma(n/d+1)^{-1}\int \exp(-q)q^{n/(ad)-1}dq \tag{23}$$

$$= \frac{2^n b^{n/d} n\Gamma(n/(ad))\Gamma(1/d)^n}{ad^{n+1}\Gamma(n/d+1)} \tag{24}$$

A.1.1 Proof of Lemma 1

First note that we only need to prove the lemma for $r = 1$. Due to the homogeneity of the definition of the set, choosing an $r \neq 1$ simply expands the set by a factor of r in every dimension, and thus multiplies the volume by r^n.

The proof for $r = 1$ proceeds by induction with respect to n. For simplicity of notation, we denote the set by S_n although it does still depend on d. For any d, S_1 is simply the line segment $[-1,1]$ for any d. Thus,

$$V(S_1) = 2 \tag{25}$$

For an arbitrary n, note that the cross-section of S_n for a fixed s_n, defined by

$$S_n(s_n) = \{(s_1,\ldots,s_n) \in \mathbb{R}^{n-1} | \sum_{i=1}^{n-1}|s_i|^d \leq 1 - s_n^d\} \tag{26}$$

is simply an $n-1$-dimensional d-sphere of radius $r = (1-s_n^d)^{1/d}$. Thus, the volume of the crosssection equals $(1-s_n^d)^{(n-1)/d}V(S_{n-1})$. We can compute the volume of S_n simply by letting s_n take all the values in $[-1,1]$ and summing the volumes of these crosssections together. This gives

$$V(S_n) = \int_{-1}^{1}(1-|s_n|^d)^{(n-1)/d}V(S_{n-1})ds_n = 2V(S_{n-1})\int_0^1(1-s_n^d)^{(n-1)/d}ds_n \tag{27}$$

To evaluate this integral, we make the change of variables to y:

$$y = s_n^d \iff s_n = y^{1/d} \iff \frac{ds_n}{dy} = \frac{1}{d}y^{1/d-1} \tag{28}$$

which gives

$$V(S_n) = V(S_{n-1})\frac{2}{d}\int_0^1 (1-y)^{(n-1)/d} y^{1/d-1} dy \tag{29}$$

The integral in this equation turns out to be equal to the definition of the classic beta function, which can be expressed using the gamma function as

$$\int_0^1 (1-y)^{\alpha} y^{\beta} dy = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{30}$$

Thus, the recursive formula for $V(S_n)$ can be expressed as

$$V(S_n) = V(S_{n-1})\frac{2}{d}\frac{\Gamma(\frac{n-1}{d}+1)\Gamma(\frac{1}{d})}{\Gamma(\frac{n}{d}+1)} \tag{31}$$

The formula given in the lemma fulfills this recursive equation, as well as the initial value in (25). Thus the lemma is proven.

B The multivariate spherical distribution

We consider the following probability density function in an n-dimensional (sub)space:

$$p(\mathbf{s}) = \frac{1}{c}\exp(-\frac{(\sum_{i=1}^n s_i^2)^a}{b^{2a}}) \tag{32}$$

We want to determine the constants c and b as a function of a so that this is a proper probability density (integrates to one) and the s_i have unit variance. By symmetry, the s_i have zero mean.

Let us take the n-dimensional polar coordinates by taking

$$r = \sqrt{\sum_{i=1}^n s_i^2} \tag{33}$$

and $\mathbf{u}$ which is an isometric parameterization of the unit sphere S_n, i.e. the set where $r = 1$, see e.g. [24, 25]. It is not necessary here to explicitly construct such a parameterization. The determinant of the Jacobian of the transformation is given by r^{n-1}.

Now, we can compute the normalizing constant c as follows.

$$\int \exp(-\frac{(\sum_{i=1}^n s_i^2)^a}{b^{2a}}) d\mathbf{s} = \int_{S_n}\int_0^{\infty} \exp(-\frac{r^{2a}}{b^{2a}}) r^{n-1} dr\, d\mathbf{u} \tag{34}$$

$$= \int_0^{\infty} \exp(-\frac{r^{2a}}{b^{2a}}) r^{n-1} dr \int_{S_n} d\mathbf{u} \tag{35}$$

The latter integral equals the surface of the unit sphere S_n, which in n dimension is equal to

$$\frac{2\pi^{n/2}}{\Gamma(n/2)} \tag{36}$$

Next, let us make the transformation of variables

$$t = \frac{r^{2a}}{b^{2a}} \Leftrightarrow r = bt^{1/(2a)} \tag{37}$$

for which the volume element can be computed as

$$\frac{dr}{dt} = \frac{b}{2a} t^{1/(2a)-1} \tag{38}$$

So, we have

$$\int \exp(-\frac{(\sum_{i=1}^n s_i^2)^a}{b^{2a}}) d\mathbf{s} = \frac{2\pi^{n/2}}{\Gamma(n/2)} \int_0^{\infty} \exp(-t)(bt^{1/(2a)})^{n-1} \frac{b}{2a} t^{1/(2a)-1} dt \tag{39}$$

$$= \frac{2\pi^{n/2}}{\Gamma(n/2)} \frac{b^n}{2a} \int_0^{\infty} \exp(-t) t^{n/(2a)-1} dr \tag{40}$$

$$= \frac{\pi^{n/2} b^n \Gamma(\frac{n}{2a})}{a\Gamma(n/2)} \tag{41}$$

Thus, to make p a proper probability density function, we must have

$$c = \frac{\pi^{n/2} b^n \Gamma(\frac{n}{2a})}{a\Gamma(n/2)} \tag{42}$$

Next, we compute the value that b should take to make the probability density standardized, i.e. $E\{s_i^2\} = 1$ for all i. We have by symmetry

$$E\{s_i^2\} = \frac{1}{n} E\{\sum_{i=1}^{n} s_i^2\} \tag{43}$$

So, we can use the same transformation of variables to compute

$$\begin{aligned}
\frac{1}{c}\int \frac{1}{n} E\{\sum_{i=1}^{n} s_i^2\} \exp(-\frac{(\sum_{i=1}^{n} s_i^2)^a}{b^{2a}}) d\mathbf{s} &= \frac{1}{c}\int_0^\infty \frac{1}{n} r^2 \exp(-\frac{r^{2a}}{b^{2a}}) r^{n-1} dr \int_{S_n} d\mathbf{u} && (44)\\
&= \frac{1}{c}\int_{S_n} d\mathbf{u} \int_0^\infty \frac{1}{n} \exp(-t)(bt^{1/(2a)})^{n+1} \frac{b}{2a} t^{1/(2a)-1} dt && (45)\\
&= \frac{2a}{nb^n\Gamma(\frac{n}{2a})} \int_0^\infty \frac{1}{n} \exp(-t) \frac{b^{n+2}}{2a} t^{(n+2)/(2a)-1} dt && (46)\\
&= b^2 \frac{\Gamma(\frac{n+2}{2a})}{n\Gamma(\frac{n}{2a})} && (47)
\end{aligned}$$

In order for this to be equal to one for a positive b, we must have

$$b = \sqrt{\frac{n\Gamma(\frac{n}{2a})}{\Gamma(\frac{n+2}{2a})}} \tag{48}$$

References

[1] E. H. Adelson and J. R. Bergen. Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A*, 2(2):284–299, 1985.

[2] F. Mechler and D. L. Ringach. On the classification of simple and complex cells. *Vision Research*, 42(8):1017–33, 2002.

[3] F .S. Chance, S. B. Nelson, and L .F. Abbott. Complex cells as cortically amplified simple cells. *Nature Neuroscience*, 2(3):277–282, 1999.

[4] J.-M. Alonso and L. M. Martinez. Functional connectivity between simple cells and complex cells in cat striate cortex. *Nature Neuroscience*, 1(5):395–403, 1998.

[5] 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, 1996.

[6] A. J. Bell and T. J. Sejnowski. The ‘independent components’ of natural scenes are edge filters. *Vision Research*, 37:3327–3338, 1997.

[7] J. H. van Hateren and A. van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc. Royal Society ser. B*, 265:359–366, 1998.

[8] A. Hyvärinen and P. O. Hoyer. Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, 12(7):1705–1720, 2000.

[9] W Hashimoto. Quadratic forms in natural images. *Network: Computation in Neural Systems*, 14(4):765–88, 2003.

[10] K. Körding, C. Kayser, W. Einhäuser, and P. König. How are complex cell properties adapted to the statistics of natural stimuli? *Journal of Neurophysiology*, 91(1):206–12, 2004.

[11] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. on Neural Networks*, 10(3):626–634, 1999.

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

[13] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37:3311–3325, 1997.

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

[15] A. Anzai, I. Ohzawa, and R.D. Freeman. Neural mechanisms for processing binocular information i. simple cells. *J. Neurophysiol.*, 82:891–908, 1999.

[16] The package can be downloaded at http://www.cs.helsinki.fi/patrik.hoyer/.

[17] Y. Karklin and M. S. Lewicki. Learning higher-order structures in natural images. *Network: Computation in Neural Systems*, 14:483–499, 2003.

[18] S. Osindero, M. Welling, and G. E. Hinton. Topographic product models applied to natural scene statistics. *Neural Computation*, 18:381–414, 2006.

[19] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14:1771–1800, 2002.

[20] P. V. Gehler and M. Welling. Product of edgeperts. In *Advances in Neural Information Processing System 18*, 08 2005.

[21] Christoph Kayser, Konrad P. Körding, and Peter König. Learning the nonlinearity of neurons from natural visual stimuli. *Neural Comput.*, 15(8):1751–1759, 2003.

[22] A. Hyvärinen, M. Gutmann, and P. Hoyer. Statistical model of natural stimuli predicts edge-like pooling of spatial frequency channels in V2. *BMC Neuroscience*, 6(12), 2005.

[23] Jacek Osiewalski and Mark F. J. Steel. Robust bayesian inference in l_q-spherical models (in miscellanea). *Biometrika*, 80(2):456–460, 1993.

[24] K. T. Fang, S. Kotz, and K. W. Ng. *Symmetric Multivariate and Related Distributions*. Chapman and Hall, London and New York., 1990.

[25] O. G. Guleryuz, E. Lutwak, D. Yang, and G. Zhang. Information theoretic inequalities for contoured probability distributions. *IEEE Transactions on Information Theory*, 48(8):2377–2383, 2002.