Learning Natural Image Structure with a Horizontal Product Model

Urs Köster[1,2], Jussi T. Lindgren[1,2], Michael Gutmann[1,2]
and Aapo Hyvärinen[1,2,3]

[1] Department of Computer Science
[2] Helsinki Institute for Information Technology
[3] Department of Mathematics and Statistics University of Helsinki, Finland

Abstract. We present a novel extension to Independent Component Analysis (ICA), where the data is generated as the product of two submodels, each of which follow an ICA model, and which combine in a horizontal fashion. This is in contrast to previous nonlinear extensions to ICA which were based on a hierarchy of layers. We apply the product model to natural image patches and report the emergence of localized masks in the additional network layer, while the Gabor features that are obtained in the primary layer change their tuning properties and become less localized. As an interpretation we suggest that the model learns to separate the localization of image features from other properties, since identity and position of a feature are plausibly independent. We also show that the horizontal model can be interpreted as an overcomplete model where the features are no longer independent.

1 Introduction

The study of natural images statistics has recently received a great deal of attention in machine learning as well as in computational neuroscience for its wide applicability from machine vision to the understanding of cortical processing. There is now a large body of evidence suggesting that neural visual systems are adapted to the statistics of the input [1, 2], where the timescale of adaptation can range from evolutionary scale to the scale of seconds. Hence, visual mechanisms reflect the statistical structure of the visual data. For example the features obtained by applying Independent Component Analysis (ICA) to natural images have very similar properties to those of Simple Cells in mammalian primary visual cortex[3–5].

In ICA, the observed data vector $\mathbf{x}$ is assumed to be generated as a linear superposition of features, $\mathbf{x} = \mathbf{A}\mathbf{s}$, where the distribution of the sources is usually assumed to be a known supergaussian probability density function (pdf). Due to the assumption that sources are independent, we can write $p(\mathbf{s}) = \prod_i p_i(s_i)$ or for the log-probability $\log p(\mathbf{s}) = \sum_i \log p_i(s_i)$. If the mixing matrix $\mathbf{A}$ is invertible and has inverse $\mathbf{W}$, consisting of vectors $\mathbf{w}_i$ we can make a transformation of density to obtain the pdf for the data as $\log p(\mathbf{x}) = \sum_i \log p_i(\mathbf{w}_i^T \mathbf{x}) - \log |\det \mathbf{W}|$. This model can easily be optimized with maximum likelihood.

A weakness of ICA is, that as an inherently linear model, it is not able to recover independent sources from data with complex, nonlinear dependencies such as most natural signals. Therefore attempts have been made [6, 7] to extend ICA to model more general densities. Taking these ideas in a different direction, here we try to nonlinearly extend the ICA model to include two classes of sources, which are mixed independently to reflect different aspects of an observed data vector. The two parts are then combined nonlinearly to produce the actual observed data vector. For modelling natural image patches this means that we independently sample from submodels $\mathbf{x}_l$ and $\mathbf{x}_r$, and the actual observed image patch $\mathbf{x}$ is obtained as $\mathbf{x} = \mathbf{x}_l \odot \mathbf{x}_r$, where $\odot$ denotes elementwise multiplication.

This kind of model can be interpreted as taking the basic principle from a linear superposition model such as ICA but generalizing it to a nonlinear superposition of different "sources", where the sources themselves are now generated as ICA-like linear superpositions. As a general example of this idea, one visual subsystem could specialize in 'what' there is in a particular scene, whereas another would code for 'where' in the scene the stimulus is located. These two are plausibly independent in general, but obviously cannot be captured by independent sources in a linear model.

2 Methods

2.1 The proposed model

We define the generative model for the data as follows:

$$\mathbf{x} = \mathbf{x}_l \odot \mathbf{x}_r = \mathbf{A}\mathbf{s} \odot (\mathbf{B}\mathbf{t} + c) \tag{1}$$

where $\mathbf{x}_l = \mathbf{A}\mathbf{s}$ is the "classical" ICA or sparse coding part and $\mathbf{x}_r = \mathbf{B}\mathbf{t} + c$ codes for aspects of data that cannot be captured by the linear ICA model. The $\odot$ indicates elementwise multiplication, so each pixel is defined by the product of two independent parts. The matrix $\mathbf{A}$ is square and invertible whereas $\mathbf{B}$ is undercomplete, with significantly fewer columns (features) than $\mathbf{A}$. The vectors $\mathbf{s}$ and $\mathbf{t}$ are the independent components of the two subimages. We require both $\mathbf{B}$ and $\mathbf{t}$ to be non-negative to ensure that that $\mathbf{x}_r$ is always positive. c is a small constant that is added for numerical stability, and it was set to $c = 0.1$ for all experiments. The model can be written more succinctly as

$$\mathbf{x} = \mathbb{D}(\mathbf{B}\mathbf{t} + c)\mathbf{A}\mathbf{s} \tag{2}$$

where $\mathbb{D}$ indicates diagonalization of the vector.

2.2 Maximum Likelihood Optimization

As $\mathbf{A}$ is assumed to be invertible, we can solve for the components $\mathbf{s}$ as

$$\mathbf{s} = \mathbf{A}^{-1}\mathbb{D}(\mathbf{B}\mathbf{t} + c)^{-1}\mathbf{x} = \mathbf{W}\mathbb{D}(\mathbf{B}\mathbf{t} + c)^{-1}\mathbf{x} \tag{3}$$

where we define the filter matrix $\mathbf{W} = \mathbf{A}^{-1}$ to be the inverse of the feature matrix $\mathbf{A}$. Now we define a pdf on $\mathbf{s}$ following the ICA model

$$p(\mathbf{s}) = \prod_i \exp\left(g(s_i)\right) = \frac{4\sqrt{3}}{\pi} \prod_i \frac{1}{\cosh^2(\pi/\sqrt{12}s_i)} \tag{4}$$

where the function $g(\mathbf{s})$ defines the normalized log-pdf, which we choose to be the logistic distribution. Now we transform the density to obtain the probability distribution for $\mathbf{x}$ as

$$\log p(\mathbf{x}|\mathbf{W},\mathbf{B},\mathbf{t}) = \sum_i g(\mathbf{w}_i^T \mathbb{D}(\mathbf{B}\mathbf{t}+c)^{-1}\mathbf{x}) + \log|\det \mathbf{W}| - \sum_i \log|\mathbf{b}_i^T\mathbf{t}+c| \tag{5}$$

where the extra terms due to the normalization constant are given by the determinant of the Jacobian of the matrix $\mathbf{W}\mathbb{D}(\mathbf{B}\mathbf{t})^{-1}$. From this we get the log-likelihood of the parameters for a sample of data vectors of size T. We choose a flat prior for $\mathbf{A}$ and $\mathbf{B}$ and a Laplacian prior for $\mathbf{t}$, so the log-likelihood for one data sample becomes:

$$\log p(\mathbf{W},\mathbf{B},\mathbf{t}|\mathbf{x}) = \sum_i g(\mathbf{w}_i^T \mathbb{D}(\mathbf{B}\mathbf{t}+c)^{-1}\mathbf{x}) + \log|\det \mathbf{W}| - \sum_i \log|f(\mathbf{b}_i^T\mathbf{t})| - |\mathbf{t}|_1 \tag{6}$$

This can now be optimized by taking gradients of the sample expectation w.r.t. both the weight matrices and the components $\mathbf{t}$.

3 Identifiability with Artificial Data

To create random data following the model, we sample from a logistic distribution for $\mathbf{s}$ and from an exponential distribution for $\mathbf{t}$. The mixing matrices are also generated randomly, with the restriction that the matrices have to be well-conditioned for the algorithm to converge. We arbitrarily constrained the condition number of $\mathbf{A}$ and $\mathbf{B}$ to be no larger than ten. Furthermore, $\mathbf{B}$ is constrained to be non-negative, following the model definition. The independent mixtures $\mathbf{x}_l = \mathbf{A}\mathbf{s}$ and $\mathbf{x}_r = \mathbf{B}\mathbf{t} + c$ are multiplied elementwise to obtain data following the model distribution. We generated 20,000 samples with a dimensionality of 60, and with 4 and 8 features in $\mathbf{B}$. Then, we attempted to estimate the model parameters $\mathbf{A}$ and $\mathbf{B}$ from the data. Like in ICA, the order and the sign of the components cannot be determined, so given the true mixing matrix $\tilde{\mathbf{A}}$ we expect the product $\tilde{\mathbf{A}}\mathbf{A}^{-1}$ to be a permuted diagonal matrix with random sign. Similarly, for the second part of the model we expect $\tilde{\mathbf{B}}\mathbf{B}^{\dagger}$ to be a permuted identity matrix. Here the pseudo-inverse is used, since $\mathbf{B}$ is not a square matrix.

The results for our experiments on artificial data are given in Fig. 1. Up to some noise, both $\mathbf{A}$ and $\mathbf{B}$ are correctly identified. The product $\tilde{\mathbf{A}}\mathbf{A}^{-1}$ shows that the vectors in $\mathbf{A}$ and $\tilde{\mathbf{A}}$ are identical up to randomly flipped signs, but the order of the vectors is randomized. Since the vectors in $\mathbf{B}$ are constrained to be non-negative, there is no sign indeterminacy but only the order of the vectors is shuffled. This shows that the parameters of the proposed model can be identified for a range of different sizes of $\mathbf{B}$.

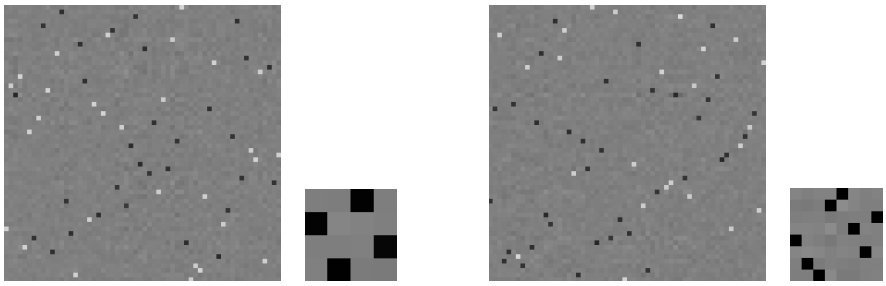

Fig. 1. Both parameter matrices $\mathbf{A}$ and $\mathbf{B}$ can be identified up to order and sign indeterminacies. We show the product of the true and the inverse of the estimated matrices, which results in permuted diagonal matrices. In the plots we code 0 as gray, 1 as black and -1 as pure white. The two plots on the left are for data generated with 4 vectors in $\mathbf{B}$, on the right there are 8 vectors. The larger plots show $\tilde{\mathbf{A}}\mathbf{A}^{-1}$, the smaller ones $\tilde{\mathbf{B}}\mathbf{B}^{\dagger}$, the product of the true parameter matrix and the pseudoinverse of the estimated matrix.

4 Experiments on Natural Images

4.1 Preprocessing

Experiments were performed on natural image patches sampled from natural images available in P. O. Hoyer's ImageICA package[4]. We used 20,000 patches of size 16×16 pixels for all experiments and performed zero phase whitening on the data [8]. The dimensionality was not reduced, and the DC-component was retained. We discarded 20% of the patches with the lowest variance, which correspond to blank image regions and do not significantly affect the gradient.

We performed experiments with $\mathbf{B}$ having a varying number of features between 2, 4, 8 and 16. The estimation was started with $\mathbf{A}$ initialized randomly, and $\mathbf{B}$ to a matrix of all ones divided by the number of elements. The hidden variables $\mathbf{t}$ were also initialized randomly, but each vector $\mathbf{t}$ was then normalized to unit L_1-norm. This had the effect that, with $c = 0.1$, each pixel of $\mathbf{x}_r$ was close to one initially and not influencing the $\mathbf{x}_l$ part of the model. The estimation was then started by learning only the matrix $\mathbf{A}$ for $\mathbf{x}_l$ with a stepsize of 0.1, until visual inspection showed that it had converged to an ICA basis set characterized by Gabor-like receptive fields. After this initialization, $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{t}$ were estimated simultaneously. The stepsize for $\mathbf{t}$ was chosen to be 1, while the stepsizes for $\mathbf{A}$ and $\mathbf{B}$ were both 0.1. The non-negativity of the components $\mathbf{x}_r$ was ensured by forcing both $\mathbf{B}$ and $\mathbf{t}$ to be non-negative after every update step.

[4] available at http://www.cis.hut.fi/projects/ica/imageica/

4.2 Separation into Gabors and localized masks

Since the novel model presented here is a generalization of ICA, and in fact feature matrix $\mathbf{A}$ is initialized with an ICA basis, it should not be surprising that the "independent components" recovered by the model are Gabor-like filters that are localized, oriented and band-pass, as shown in Fig. 2(a) for different numbers of columns in $\mathbf{B}$. However there are important differences that emerge once the modulation due to the $\mathbf{Bt}$ component is taken into account. While for a small number of columns in $\mathbf{B}$ the features look like the Gabor filters familiar from the classical ICA model, they become less localized as the number increases. In all cases the filters in $\mathbf{B}$ learn to perform a localized modulation, that dampens some of the image to create a patch with blank areas. The vectors in $\mathbf{B}$ evenly tile all of the pixel space, but selectively boost or mask regions of individual patches. This is shown in Fig. 2(b).

4.3 Dependence of tuning properties on the number of filters

To investigate the change in appearance of the features in $\mathbf{A}$, we parametrized them with a least-squares fit to Gabor functions, i.e. Sinusoids with a Gaussian envelope. We then analyzed the tuning statistics of the Gabors in terms of frequency and size. As we show in Fig. 3, there is a significant change in aspect ratio and modulation (number of zero crossings of the sinusoid) of the Gabors as the number of filters in $\mathbf{B}$ is increased.

5 Discussion

5.1 Separation of structure and position

The most striking aspect of the results is that with an increasing number of vectors in $\mathbf{B}$, the appearance of the features starts to differ significantly from the Gabor-like features that are obtained by most other ICA or Sparse Coding models. All features become less localized, and especially the highest frequency features, which tend to be very localized in the classical ICA model, loose all localization and cover the whole image patch. In an ICA model, this would clearly be less than optimal because most natural images patches have only localized structure. In the nonlinear model the situations is quite different though: Depending on the structure of $\mathbf{x}_r$, the localization properties can be recovered by "masking off" the part of the reconstruction $\mathbf{x}_l$ that does not contribute to the total image patch $\mathbf{x} = \mathbf{x}_l \odot \mathbf{x}_r$ that is being coded. This is conceivable since most image patches have blank regions and only localized structure such as edges or textured objects. Rather than having a set of features that can code for arbitrary image patches, it is advantageous to independently specify the region of the image patch that contains structure, and the kind of structure. Our new model can be viewed as accomplishing this by coding image structure in $\mathbf{x}_l$ and location in $\mathbf{x}_r$. By having the ICA reconstruction $\mathbf{x}_l = \mathbf{As}$ matched just to the structure, and discarding most localization information from the basis $\mathbf{A}$, a representation

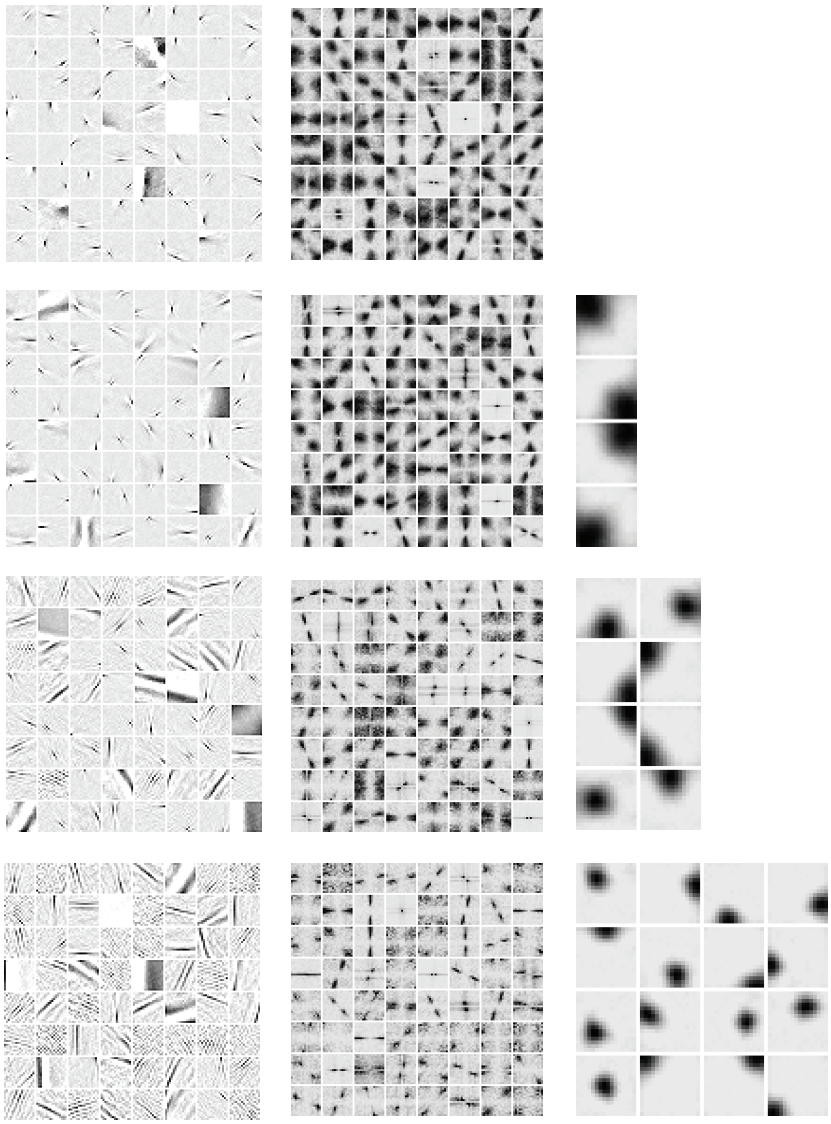

Fig. 2. Comparison of the features obtained with ICA (top) and the new product model (bottom three rows). We show a subset of 64 randomly chosen feature vectors of **A** in the first column, and their Fourier power in the second column. For the product model the with 4, 8 and 16 secondary features, the vectors of **B** are shown in the third column. While **A** converges to the familiar ICA features, **B** produces localized masks. As the number of features in **B** increases, the Gabors in **A** spread out to cover more of the image patch, this is particularly evident for 16 columns in **B**. Intuitively, this can be explained as a masking, where combining one feature from **A** with different masks from **B** can produce new Gabors in various positions. The Fourier transforms show how the features become more localized in Fourier space as the number of vectors of **B** increases, but also helps to identify the unlocalized highest frequency features as aliasing artifacts: All the Fourier power should be confined to a cicle around the origin, the four corners are artifacts due to the rectangular sampling grid.

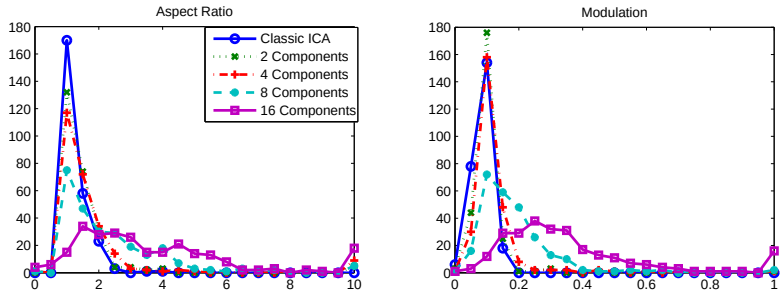

Fig. 3. Change in tuning of the basis functions in $\mathbf{A}$ with an increasing number of local fields in $\mathbf{B}$. The aspect ratio increases for more fields, i.e. the Gabors become more elongated, filling most of the patch rather than just a small portion. The number of sidelobes of the Gabors also increases, making the basis functions less localized and more similar to a Fourier basis.

with higher likelihood can be achieved. The additional part of the model $\mathbf{x}_r$ then simply masks off *where* in the image patch the particular structure occurs, leaving the rest of the image patch blank. In this way, is possible to encode a particular image patch with fewer basis functions than with classical ICA, since the features can become more specialized for orientation and frequency, while the localization in preserved in the second part of the model.

Along these lines, it is also possible to view the novel model as an implicitly overcomplete version of ICA. By multiplying each of the n features in $\mathbf{A}$ with each of the m features in $\mathbf{B}$, a new set of features of size mn is obtained. For a large number of secondary features, e.g. $m = 16$ the vectors in $\mathbf{A}$ are close to sinusoids and the vectors of $\mathbf{B}$ are nearly Gaussian. Each of the sinusoids is masked with Gaussians at different positions, which corresponds to constructing a new set of Gabor features. It is important to note that the weights of the new features will no longer be independent, since the "mask" $\mathbf{x}_r$ chosen for one of the features in $\mathbf{A}$ will also be applied to each other features.

5.2 Relation to Contrast Gain Control

One of the initial motivations for the way the model was specified, in particular the nonnegativity of $\mathbf{x}_r$, was that the secondary features would perform divisive Contrast Gain Control (CGC) on the image patches. This can be easily seen by rewriting 1 as

$$\frac{\mathbf{x}}{\mathbf{Bt} + c} = \mathbf{As} \tag{7}$$

where with slight abuse of notation the fraction is taken to be elementwise. Models of divisive normalization in various ways are abundant in the literature [9] and are motivated from the observation that natural images are not stationary, and the statistics vary considerably from one image region to another [10].

However, in preliminary experiments (results not shown) we could not confirm a significant reduction in energy dependencies in our model compared to the classical ICA model.

6 Conclusion

We have presented an extension of ICA with a second layer, where, in contrast to previous work, the layers are horizontal rather than hierarchical. After showing the identifiability on simulated data, we have applied the novel model to natural images. We report the emergence of localized "masks" in the additional layer, while the Gabor-like features in the primary layer become less localized than in classical ICA. As a possible interpretation we suggest that the model learns to separately code for the structure and position of features in image patches. This gives the features an implicit position invariance, with one feature in **A** being able to code for many different positions conditional on **B**. This is a powerful principle which is outside the scope of linear models but may be of great importance in neural visual systems.

Acknowedgements We wish to thank David C.J. Senne for comments on the manuscript. Urs Köster is kindly supported by a Scholarship from the *Alfried Krupp von Bohlen und Halbach-foundatation.*

References

1. B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37(23):3311–3325, 1997.
2. 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*, 265:359–366, 1998.
3. C. Jutten and J. Herault. Blind separation of sources part i: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1–10, 1991.
4. P. Comon. Independent component analysis – a new concept? *Signal Processing*, 36:287–314, 1994.
5. A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis.* Wiley, 2001.
6. A. Hyvärinen and P. Hoyer. Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation, 12(7):1705-1720.*, 2000.
7. A. Hyvärinen, P. Hoyer, and M. Inki. Topographic independent component analysis. *Neural Computation.*, 2001.
8. J.J. Atick. Could information theory provide an ecological theory of sensory processing? *Network: Computation in neural systems*, 3:213–251, 1992.
9. O. Schwartz and E. P. Simoncelli. Natural signal statistics and sensory gain control. *Nature Neuroscience*, 4(8):819–825, 2001.
10. Y. Karklin and M. S. Lewicki. A hierarchical bayesian model for learning nonlinear statistical regularities in non-stationary natural signals. *Neural Computation*, 17(2):397–423, 2005.