Learning Quadratic Forms by Density Estimation and its Applications to Image Coding

Hauke Bartsch, Sepp Hochreiter and Klaus Obermayer

Dept. of Computer Science, Technische Universität Berlin, Germany,
E-mail: hauke@cs.tu-berlin.de

Abstract

We develop a novel method for source separation and apply it to natural images. It is a specialization of independent factor analysis (IFA) but overcomes generic IFA problems and finds many independent sources in few observations. A fast and robust EM learning algorithm produces an over-complete basis. Compared to standard approaches our method generates superior codes in terms of population sparseness and dispersity. The algorithm learns features which possess properties that are observed in simple as well as complex cells found in V1.

1 Introduction

We developed a new algorithm for source separation and applied it to natural images in order to learn efficient codes based on higher order statistics.

Previous approaches for learning sets of basis functions in the domain of natural images established by various techniques that assuming sparse and independent components the appropriate codes resemble spatially localized, oriented and Gabor like basis functions [1–3]. These properties are in accordance with observations of simple cells in V1 [4]. Simple cells which are assumed to have linear response characteristics represent the major coding unit at the early stage of visual information processing of mammals. Some efforts have been made to explain complex cells, i.e. non-linear neurons, found in V1 by independence assumptions [5] or temporal coherence [6–8].

Temporal coherence is based on the assumption that features that are approximately constant when the image is slightly shifted or rotated are the relevant features which reflect the inherent invariances of the data. Using a proper transformation sequence temporal coherence assumptions lead to complex cell

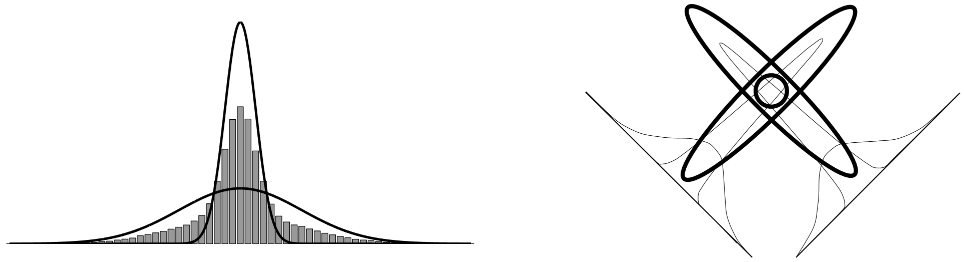


Fig. 1. *Left:* A histogram of data drawn from a mixture of two Gaussian distributions. It shows the properties of a sparse distribution. *Right:* An illustration of a centralized mixture of three Gaussians (bold lines) estimating a linear mixture of two sparse sources in two observations. Shown are also the two (sparse) marginal distributions along the axes

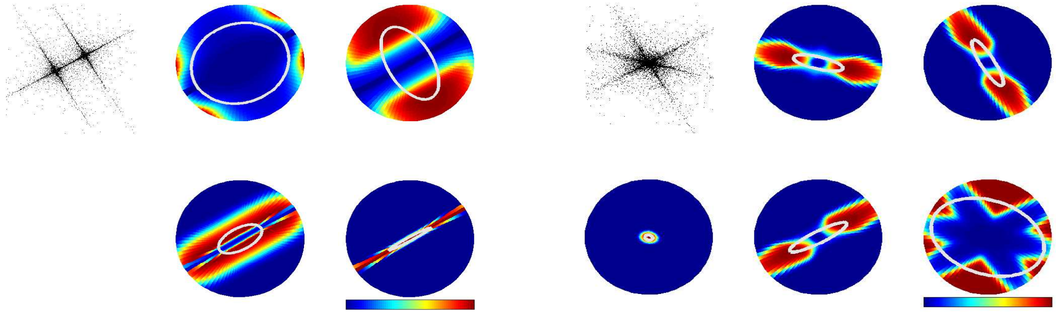
characteristics [8]. The other approach [5] relies on independent component analysis (ICA). The response of complex cells are assumed to be statistically independent and during learning 2D Gabor phase modules emerge and form complex cells. However standard ICA has the drawback that the number of sources and the source densities (priors) have to be known in advance.

We propose a new ICA method where the number of sources must not be know a priori. Our method is able to extract the sources if the number of sources is larger than the number of observations without restrictions on the source distributions. Such ICA techniques are relevant also for other applications because for many real world problems the number of sources is unknown. Additionally the underlying physical conditions lead often to few observations which contain many sources. Using our method in the domain of image coding we propose an alternative to explain complex cell properties.

2 The Method

We use the idea of independent factor analysis (IFA) to estimate the unknown source distributions together with the mixing matrix. The ability to learn the densities allows the model to adjust its parameters accordingly to the special requirements of the data which results in better separation performance. The model of IFA also allows to handle the case where the number of mixtures differs from the number of sources and the case where the data are noisy.

A disadvantage of IFA is that the number of sources has to be set in advance and the true number of sources cannot be obtained because by increasing the number of estimated sources the generative model can better explain the data independent on the true number of sources. For the high dimensional case with many sources the original IFA algorithm becomes computationally intractable because density estimation in high dimensional spaces is a challenging problem.



(a) (b)

Fig. 2. Mixtures of two (a) respectively three (b) independent sources learned by a mixture of centralized Gaussians. A scatter plot of the data (ICA source model) is shown together with plots of the probability of a data point to be generated by the Gaussian (red codes for high probability). Each Gaussian is depicted additionally by its eigen-ellipse to make plain the direction of the principal axis of the Gaussian

We overcome these problems of IFA by pining the Gaussians at zero. By this we focus on the shape of the source densities. Super-Gaussians are easily represented by peaky Gaussians at zero and high-variance Gaussians which match the active coefficients of the sources (see Fig. 1). This constrain reflects the sparseness conditions on the sources which are obtained in [9] for natural images. For sub-Gaussian source distributions the correct sources are detected although the densities cannot be learned.

The basis functions are obtained as the directions of the largest elongation, i.e. the principal axes of the high variance Gaussians (see Fig. 1 and 3). In higher dimensional spaces we propose that for each multivariate Gaussian additional principal axes are learned which constitute an orthogonal basis for each (statistically independent) source.

3 Results

3.1 Test case

We first apply the model for 2-dimensional toy problems. Two respectively three sources are linearly mixed which results in the data shown as scatter plots in Fig. 2. We trained mixtures of four respectively five Gaussians by EM. To illustrate the partitionings of the data by the generative model we evaluated the probability of a specific data point to be generated by the Gaussian and displayed them for each Gaussian separately in color. In both cases (a) and (b) the model correctly predicts the source densities by high variance distributions.

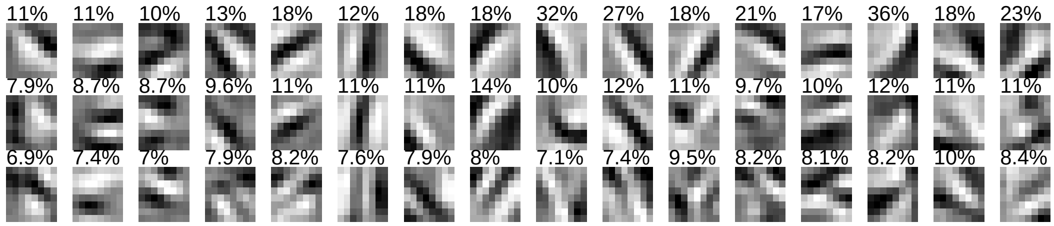


Fig. 3. Three principal axes (column wise) which explain the largest variances are shown for each mixture. On number on-top of each axis represents the percentage of variance that is explained by this component

3.2 Natural Images

We applied the algorithm to pre-whited natural images obtained from the home page of Bruno Olshausen [1]. From these images we extracted 8×8 patches and learned (by EM [10]) a two times over-complete set of 128 centralized Gaussians each parameterized by their full covariance matrices. For 16 arbitrarily selected Gaussians (i.e. coding units) in Fig. 3 we plotted the first three principal axes as 8×8 patches. The response of a coding unit to an image $\mathbf{x}$ is computed as a quadratic form by first applying on $\mathbf{x}$ the spatial filters obtained as principal axes (by dot-products) and afterwards summing the squares of the obtained values weighted by the respective variance of the principal axis (depicted above the patches in Fig. 3 in percentages). Units that display in successive principal axes similar oriented but phase shifted edge detectors will respond to stimuli containing these orientations similar to complex cells.

To evaluate the quality of the obtained code for natural images we analyzed its population sparseness (measured as fourth order cumulant) and how well the variance is spread amongst the population of coding units (dispersity). Population sparseness is measured as a property of the distribution of responses of the coding units in the population. This is in contrast to the measure of lifetime sparseness which is a property of the response distribution of a single coding through time [11]. Both measures are not equivalent and in the context of efficient coding it is preferable to obtain codes with high population sparseness and high dispersity. In such codes for any given stimulus only a small subset of all neurons is active at one time. Averaged over many stimulus presentations no specific set of neurons is preferred.

Compared with the results obtained in [11] for pre-whitened images and various linear codes (e.g. Gabor, PCA, [1], etc.) the centralized Gaussian mixtures perform superior in terms of population sparseness (8.83) with a high dispersity (small variance of ± 0.21). The best code obtained in [11] for pre-whitened images was the one from a Gabor filter bank which results in a code with a population sparseness of 5.37.

4 Discussion

The presented algorithm solves the problem of estimating a set of basis functions for high dimensional data with more sources than observations. It can be interpreted as a mapping of the data into a high dimensional feature space. This feature space is constructed from the 2-point correlations of the data which are the subject of analysis of the multivariate Gaussians. The obtained code represents therefore a partitioning of the correlation structure of the data. In [12] the same feature space was used but a factorizing distribution was learned by standard ICA. Because the feature space is high dimensional this also solved the problem of obtaining more sources than observations. In contrast to [12] in the algorithm using centralized Gaussian mixtures the number of sources can be larger than the dimension of the feature space, i.e. the number of two-point correlations in the quadratic form.

References

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

[2] Anthony J. Bell and Terrence J. Sejnowski, "Edges are the 'independent components' of natural scenes", *in Advances in Neural Information Processing Systems 9*, 1996.

[3] J. H. van Hateren and A. van der Schaaff, "Independent components of natural images are edge filters", *Proc. Royal Soc. Lond. B*, vol. 265, pp. 359–366, 1997.

[4] J. P. Jones and L. A. Palmer, "An evaluation of the two-dimensional gabor filter model of simple receptive fields in the cat striate cortex", *J. Neurophysiology*, vol. 58, pp. 1,233–1,258, 1987.

[5] A. Hyvärinen and P. Hoyer, "Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces", *Neural Comput.*, vol. 12, pp. 1705–1720, 2000.

[6] P. Földiàk, "Learning invariance from transformation sequences", *Neural Comput.*, vol. 3, pp. 194–200, 1991.

[7] L. Wiskott and T. Sejnowski, "Slow feature analysis: unsupervised learning of invariances", *Neural Computation*, vol. 14, pp. 715–770, 2002.

[8] P. Berkes and L. Wiskott, "Applying slow feature analysis to image sequences yields a rich repertoire of complex cell properties", in José R. Dorronsoro, editor, *Proc. Int'l Conf. on Artificial Neural Networks, ICANN'02*, Lecture Notes in Computer Science, pp. 81–86. Springer-Verlag, 2002.

[9] B. A. Olshausen and K. J. Millman, "Learning sparse codes with a mixture-of-gaussians prior", in S. A. Solla, T. K. Leen, and K. R. Müller, editors, *Advances in Neural Information Processing Systems*, vol. 12, pp. 841–847. MIT Press, 2000.

[10] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the em algorithm", *J. Royal Star. Soc., Series B*, vol. 39, pp. 1–38, 1977.

[11] B. Willmore and D. J. Tolhurst, "Characterizing the sparseness of neuronal code", *Network: Computations in Neural Systems*, vol. 12, pp. 255–270, 2001.

[12] H. Bartsch and K. Obermayer, "A structure preserving image transformation as the goal of visual sensory coding", *Neurocomputing*, vol. 44-46, pp. 729–734, 2002.