Bayesian Estimation of Overcomplete Independent Feature Subspaces for Natural Images

Libo Ma and Liqing Zhang

Department of Computer Science and Engineering,
Shanghai Jiao Tong University
800 Dong Chuan Road, Shanghai 200240, China
malibo@sjtu.edu.cn, zhang-lq@cs.sjtu.edu.cn

Abstract. In this paper, we propose a Bayesian estimation approach to extend independent subspace analysis (ISA) for an overcomplete representation without imposing the orthogonal constraint. Our method is based on a synthesis of ISA [1] and overcomplete independent component analysis [2] developed by Hyvärinen et al. By introducing the variables of dot products (between basis vectors and whitened observed data vectors), we investigate the energy correlations of dot products in each subspace. Based on the prior probability of quasi-orthogonal basis vectors, the MAP (maximum a posteriori) estimation method is used for learning overcomplete independent feature subspaces. A gradient ascent algorithm is derived to maximize the posterior probability of the mixing matrix. Simulation results on natural images demonstrate that the proposed model can yield overcomplete independent feature subspaces and the emergence of phase- and limited shift-invariant features—the principal properties of visual complex cells.

1 Introduction

Recent linear implementations of efficient coding hypothesis [3,4], such as independent component analysis (ICA) [5] and sparse coding [6], have provided functional explanations for the early visual system, especially neurons in the primary visual cortex (V1). Nevertheless, there are many complex nonlinear statistical structures in the natural signals, which are not able to be extracted by a linear model. For instance, Schwartz et al. have observed that, for natural images, there are significant statistical dependencies among the variances of filter outputs [7]. Several algorithms have been proposed to extend the linear ICA model to capture such residual nonlinear dependencies [1,8,7,9]. Hyvärinen et al. developed the independent subspace analysis (ISA) method, which generalizes the assumption of component independence to subspace independence [1]. However, this method is limited to the complete case. The orthogonality requirement of the mixing matrix restricts the generalization to the overcomplete representation. In the overcomplete representation, the dimension of the feature vector is *larger* than the dimension of the input. Overcomplete representations present several potential advantages. High-dimensional representations are more flexible in capturing inherent structures in signals. Overcomplete representations generally provide more efficient representations than the complete case [10]. Furthermore, studies of human visual cortex have shown interesting implications of overcomplete representations in the visual system [11].

M.E. Davies et al. (Eds.): ICA 2007, LNCS 4666, pp. 746–753, 2007.

In this paper, we combine ISA [1] and overcomplete independent component analysis [2] to extend ISA for overcomplete representations. We apply a Bayesian inference to estimating overcomplete independent feature subspaces of natural images. In order to derive the prior probability of the mixing matrix, the quasi-orthogonality of the dot product between two basis vectors is investigated. Moreover, we assume that the probability density of the dot products (between basis vectors and whitened observed data vectors) in one subspace depends only on the norms of the projections of the data onto the subspace. Then, a learning rule based on gradient ascent algorithm is derived to maximize the posterior probability. Simulation results on natural image data are provided to demonstrate the performance of overcomplete representations for independent subspace analysis. Furthermore, our model can lead to the emergence of phase- and limited shift-invariant features—principal properties of visual complex cells as well.

This paper is organized as follows: In section 2, we propose a Bayesian approach to estimate the overcomplete independent feature subspaces. The learning rule is given as well. In section 3, some experimental results on natural images are presented. Finally, some discussions on representation performance of the proposed method are given in section 4.

2 Model

2.1 Bayesian Inference

In this section, we apply Bayesian MAP (maximum a posteriori) approach to estimating overcomplete independent feature subspaces. The basic ICA model can be expressed as:

$$\mathbf{x} = \mathbf{A}\mathbf{s} = \sum_{i=1}^{N} \mathbf{a}_i s_i, \tag{1}$$

where $\mathbf{x} = (x_1, x_2, ..., x_M)^T$ is a vector of observed data, $\mathbf{s} = (s_1, s_2, ..., s_N)^T$ is a vector of components, and $\mathbf{A}$ is the mixing matrix. $\mathbf{a}_i$ is i^{th} the column of $\mathbf{A}$, and it is often called basis function or basis vector. In our model, the observed data vector $\mathbf{x}$ is whitened to vector data $\mathbf{z}$, just as the preprocessing step in most ICA methods. Furthermore, instead of considering the independent components, as in most ICA, we consider the dot product between the i^{th} basis vector and the whitened data vector. For simplicity, it is assumed that the norms of the basis vectors are set to be unity and that the variances of the sources can differ from unity. Then, the dot product is

$$y_i = \mathbf{a}_i^T \mathbf{z} = \mathbf{a}_i^T \mathbf{A} s = s_i + \sum_{j \neq i} \mathbf{a}_i^T \mathbf{a}_j s_j, \tag{2}$$

where s_i is the i^{th} independent component. Given the overcomplete representations of our model (there is a large number of components in a high-dimensional space), the second term approximately follows Gaussian distribution. Moreover there is no component whose variance is considerably larger than others. Therefore the marginal distributions of dot products should be maximally sparse (super-Gaussian). And maximizing the non-Gaussianities of these dot products is sufficient to provide an approximation

of basis vectors. Thus, we we can replace the component s_i by the dot product y_i to estimate independent feature subspaces. Considering the dot product vector $\mathbf{y} = (y_1, ..., y_N)^T = \mathbf{A}^T\mathbf{z}$, the probability for $\mathbf{z}$ given $\mathbf{A}$ can be approximated by

$$p(\mathbf{z}(t)|\mathbf{A}) = p(\mathbf{y}) \approx C \prod_{i=1}^{N} p_{y_i}(y_i) = C \prod_{i=1}^{N} p_{y_i}(\mathbf{a}_i^T z(t)), \tag{3}$$

where C is a constant. Obviously, the accuracy of the prior probability p_{y_i} is important, especially for overcomplete representations [10]. Several choices of prior on the basis coefficients $P(\mathbf{s})$ have been applied in classical linear models respectively. Bell and Sejnowski utilize the prior $P(s_i) \propto \text{sech}(s_i)$, which is corresponding to the hyperbolic tangent nonlinearity [5]. Olshausen and Field use a generalized Cauchy prior [6]. Whereas van Hateren and van der Schaaf simply explore non-Gaussianity [12]. Nevertheless, all these choices of prior is derived under a single-layer network of linear model. Surely, it is desirable to capture nonlinear dependencies by a second or third stage in a hierarchical fashion.

In our model, we apply the prior probability p_{y_i} proposed in the ISA algorithm, in which the basis function coefficients in each subspace have the energy correlations [1]. A diagram of feature subspaces is given in Figure 1.

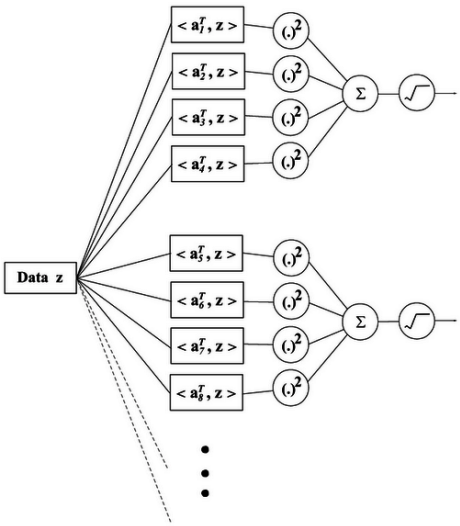


Fig. 1. Illustration of feature subspaces. The dot products between basis vectors and whitened observed data vectors are taken. Then, they are squared respectively and summed inside the same feature subspace. Square roots are taken for normalization.

The dot product (neuronal response) y_i is assumed to be divided into n-tuples, so that y_i inside a given n-tuple may be dependent on each other, but different n-tuples are mutually independent. The subspaces model introduces a certain dependency structure

for different components. Let $\Omega_j, j = 1, ..., J$ denote the set of independent feature subspaces, where J is the number of subspaces. The probability distributions for n-tuples of y_i are spherically symmetric. In other words, the probability density $p_{y_j}(.)$ of n-tuple can be expressed as a function of the sum of the squares of y_i, $i \in \Omega_j$ only. And, for simplicity, we assume $p_{y_j}(.)$ are identical for all subspaces. Therefore, the probability density inside the j^{th} n-tuple of y_i can be calculated:

$$p_{y_j}(y_j) = \exp\Big(G\Big(\sum_{i \in \Omega_j} y_i^2\Big)\Big), \tag{4}$$

where the function $G(y)$ should be convex for non-negative y. For example, one could use the form of $G(.)$ as: $G(y) = -\alpha_1\sqrt{y} + \beta_1$, where α_1 is the scaling constant and β_1 is the normalization constant. These constants are unimportant for the learning process.

Overcomplete representations mean that there is a large number of basis vectors. In other words, the basis vectors are randomly distributed in a high-dimensional space. In order to approximate the prior probability of basis vectors, we employ a result presented by Hecht-Nielsen [13]: the number of almost orthogonal directions is much larger than that of orthorgonal directions. This property is called quasi-orthogonality [2]. Therefore, in a high-dimensional space even vectors having random directions might be sufficiently close to orthogonality. Thus, the prior probability of the mixing matrix $\mathbf{A}$ can be obtained in terms of the quasi-orthogonality as follows:

$$p(\mathbf{A}) = c_m \prod_{i<j} \big(1 - (\mathbf{a}_i^T \mathbf{a}_j)^2\big)^{\frac{m-3}{2}}, \tag{5}$$

where c_m is a constant. The detailed derivation of Equation (5) can be obtained in [2].

Bayes' Theorem allows one to describe the probability of the model in terms of the likelihood of the data and the prior probability of the model. Thus, given observation $\mathbf{z}$, the posterior probability $p(\mathbf{A}|\mathbf{z})$ can be derived as follows:

$$p(\mathbf{A}|\mathbf{z}) = \frac{p(\mathbf{z}|\mathbf{A})p(\mathbf{A})}{p(\mathbf{z})}, \tag{6}$$

where $p(\mathbf{z})$ is constant with respect to $\mathbf{A}$.

It is easier to estimate the mixing matrix that maximize the logarithm of posterior probability $p(\mathbf{A}|\mathbf{z})$. Thus, taking the logarithm of Equation (6) and combining Equation (5) with Equation (3) and (4), we obtain the approximation of log-probability of the posterior:

$$\log p(\mathbf{A}|\mathbf{z}(t), t = 1, .., T) \propto \sum_{t=1}^{T}\sum_{j=1}^{J} G\big(\sum_{i \in \Omega_j} y_i^2\big) + \alpha T \sum_{i<j} \log(1 - (\mathbf{a}_i^T \mathbf{a}_j)^2) + C \tag{7}$$

where α is a constant related to c_m.

2.2 Learning Rule

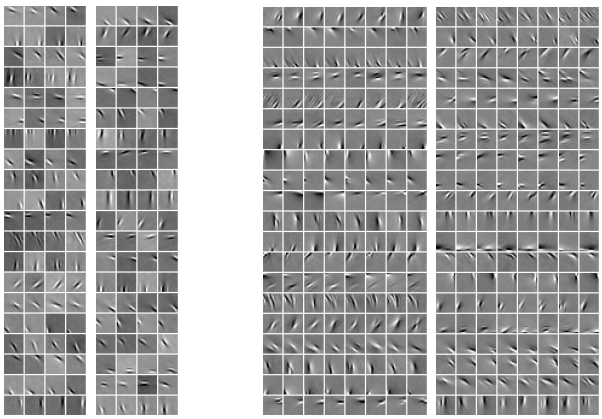
Gradient ascent maximization of posterior probability over basis vector $\mathbf{a}_k$ yields the following learning rule:

$$\Delta\mathbf{a}_k \propto \eta\Big(\sum_{t=1}^{T}\mathbf{z}(t)\big(\mathbf{a}_k^T\mathbf{z}(t)\big)g\Big(\sum_{i\in\Omega_{j(k)}}(\mathbf{a}_i^T\mathbf{z}(t))^2\Big) + \alpha T\sum_{i<j}\frac{-2\mathbf{a}_i^T\mathbf{a}_j}{1-(\mathbf{a}_i^T\mathbf{a}_j)^2}\mathbf{b}_k\Big), \quad (8)$$

where η is the learning rate, and $\Omega_{j(k)}$ is the subspace to which $\mathbf{a}_k$ belongs. $\mathbf{b}_k$ is the k^{th} column vector of matrix $\mathbf{B} = [0, ..., \mathbf{a}_j, ..., \mathbf{a}_i, ...0]$, $\mathbf{a}_j$ is the i^{th} column vector, and $\mathbf{a}_i$ is the j^{th} column vector. The function g is the derivative of G. After each iteration in equation (8), the norm of the basis vector $\mathbf{a}_k$ needs to be set to unity. This is different from ordinary ISA, where the mixing matrix is orthonormalized.

3 Simulations

We tested the algorithm for overcomplete independent subspace analysis on natural image data. The training set of images consists of 50,000 patches of size 16×16 that were randomly extracted from thirteen 256×512 pixel gray images. We use the natural images in [1], which is available on http://www.cis.hut.fi/projects/ica/data/images/. The mean gray-scale value of data (i.e., the DC component) was removed. The dimension of data was reduced by principle component analysis, projecting onto the leading 160 eigenvectors of the data covariance matrix. Then, the data vectors were whitened as in most ICA methods. The log posterior probability was maximized by an ordinary gradient method to estimate $\mathbf{A}$, using the averaged version of the learning rule in equation (8). Note that there was no constraint of orthogonality of basis vectors during each



(a) (b)

Fig. 2. Learned bases from natural images. (a) complete case (40 subspaces and 4 basis vectors in each subspace) (b) $2\times$ overcomplete case (40 subspaces and 8 basis vectors in each subspace).

iteration. Only the norms of basis vectors were set to unity. The random initial value was set for mixing matrix.

The effects of varying the level of overcompleteness and the dimension of subspaces were investigated in depth. The basis was set to be complete and $2\times$ overcomplete. The dimension of components is 160 and 320, respectively. Figure 2 shows the estimated basis vectors, which is the complete case of four-dimensional subspaces and $2\times$ overcomplete case of eight-dimensional subspaces.

To analyze the tiling properties of the estimated basis vectors, we fitted each basis vector with a Gabor function by minimizing the squared error between the estimated basis vectors and the model Gabor. Figure 3 shows the distribution of parameters obtained by fitting Gabor functions to complete and $2\times$ overcomplete basis vectors. We can see that, with the increasing of the level of overcompleteness, the scattering points in the plot of location, spatial frequency and orientation become denser and more uniform. And the distribution of phase is much closer to uniform.

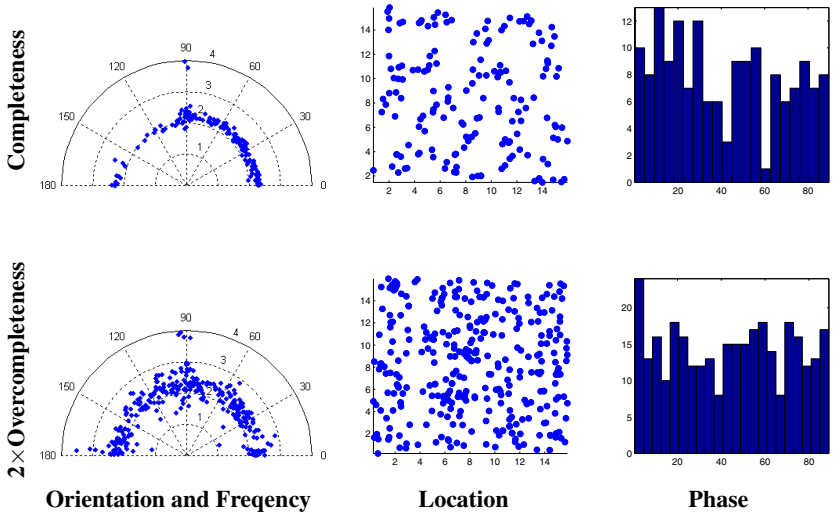


Fig. 3. The distributions of parameters derived by fitting Gabor functions with completeness and $2\times$ overcompleteness. (a) Center location of Gabor fitted within a patch. (b) Joint distribution of orientation and spatial frequency (plotted in the upper-half plane) (c) Histogram of phase of Gabor fitted (mapped to range $0\,^{\circ}$~ $90\,^{\circ}$).

Furthermore, we compare the responses of all the feature subspace and the corresponding linear filters for different stimulus cases. First, an optimal stimulus for the feature subspace was computed in the set of Gabor filters. The tested stimuli for the subspace was calculated in the set of Gabor functions. In each time, only one stimuli parameter was changed to see how the response changes. The tested parameters were location (shift), orientation, and phase. Figure 4 shows the median responses of the whole population of 40 subspace and 320 linear filters corresponding to $2\times$ overcomplete case. The top row shows the absolute responses of the linear filters, and the

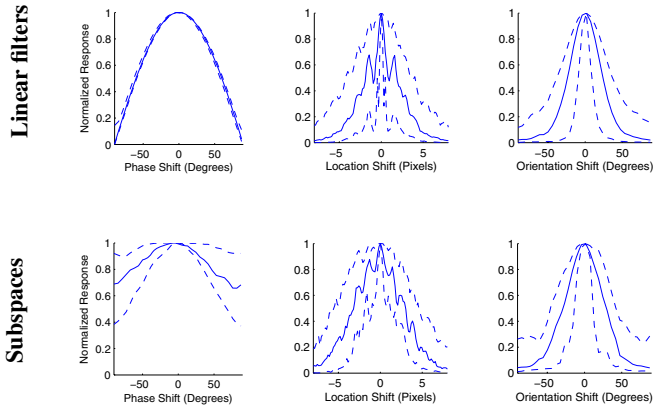


Fig. 4. Statistical curves for whole population and linear filers while shifting different Gabor parameters: orientation, frequency, and phase with $2\times$ overcompleteness. The solid line gives the median response in the population of all filters or subspaces. The dashed lines give the 90% and 10% percentiles of the responses.

bottom row shows the results of the feature subspaces. We can see that the responses of subspaces are considerably invariant to phase, and somewhat invariant to position. The sharpness of tuning to orientation and spatial frequency remains roughly unchanged. Thus it can be observed that invariance with respect to phases is a strong property of the feature subspaces. It is closely related to the response properties of complex cells in V1, which are based on location, frequency, and orientation and independent of phase. In contrast, the responses of the linear filters show no invariance with respect to any of these parameters.

4 Discussions and Conclusions

We have demonstrated in this paper how the Bayesian approach can be employed for learning overcomplete representations by utilizing the quasi-orthogonal property of basis vectors in a high-dimensional space, whereas ordinary ISA can only provide complete representations of basis functions. In addition, we examine the dot products (between basis vectors and whitened observed data vectors) instead of the basis function coefficients. Furthermore, our model need not impose the constraint of orthogonality on basis vectors. Only the norms of basis vectors were set to unity during the learning process. In contrast, basis vectors have to be orthogonal in ordinary ISA. Compared with the methods for estimating overcomplete bases by using maximum likelihood estimation, our method is as computationally effective as basic ICA estimation.

Another issue addressed in this paper is the relevance of the learned codes to neurobiological plausibilities. Both complete and overcomplete basis functions adapted to natural images suggest functional similarities to neurons of V1 receptive fields. Simulation results on natural image data demonstrate that our model can lead to the emergence of phase- and shift-invariant features—principal properties of visual complex cells as

well. This method shows promising prospects in extended applications of our method to higher levels of cortical representations.

An important concern in our model is the accuracy of the coefficient prior probability. Our overcomplete ISA algorithm can capture the underlying statistical structure of images, i.e., the energy correlations of coefficients in each subspace. However, a Laplacian prior probability as in overcomplete ICA algorithms can not capture well higher-order statistics, such as dependencies among the variances of filter outputs. This method finds compact descriptions of overcomplete representation and has the potential in a wide varieties of applications, such as image processing and pattern recognition.

Acknowledgments

The work was supported by the National Basic Research Program of China (Grant No. 2005CB724301) and the National High-Tech Research Program of China (Grant No.252006AA01Z125).

References

1. Hyvärinen, A., Hoyer, P.: Emergence of phase-and shift-invariant features by decomposition of natural images into independent feature subspaces. Neural Computation 12(7), 1705–1720 (2000)
2. Hyvärinen, A., Inki, M.: Estimating Overcomplete Independent Component Bases for Image Windows. Journal of Mathematical Imaging and Vision 17(2), 139–152 (2002)
3. Attneave, F.: Some informational aspects of visual perception. Psychol. Rev. 61(3), 183–193 (1954)
4. Barlow, H.B.: Possible principles underlying the transformation of sensory messages. Sensory Communication, 217–234 (1961)
5. Bell, A.J., Sejnowski, T.J.: The independent components of natural scenes are edge filters. Vision Research 37(23), 3327–3338 (1997)
6. Olshausen, B., Field, D.: Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381(6583), 607–609 (1996)
7. Schwartz, O., Simoncelli, E.: Natural signal statistics and sensory gain control. Nature Neuroscience 4, 819–825 (2001)
8. Hyvärinen, A., Hoyer, P.O., Inki, M.: Topographic Independent Component Analysis. Neural Computation 13(7), 1527–1558 (2001)
9. Karklin, Y., Lewicki, M.S.: A Hierarchical Bayesian Model for Learning Nonlinear Statistical Regularities in Nonstationary Natural Signals (2005)
10. Lewicki, M., Olshausen, B.: Probabilistic framework for the adaptation and comparison of image codes. Journal of the Optical Society of America A 16(7), 1587–1601 (1999)
11. Popovic, Z., Sjostrand, J.: Resolution, separation of retinal ganglion cells, and cortical magni fication in humans. Vision Research 41(10-11), 1313–1319 (2001)
12. van Hateren, J.H.: Independent component filters of natural images compared with simple cells in primary visual cortex. Proceedings: Biological Sciences 265(1394), 359–366 (1998)
13. Hecht-Nielsen, R.: Context vectors: general purpose approximate meaning representations self-organized from raw data. Computational Intelligence: Imitating Life, 43–56 (1994)