Independent components of natural images under variable compression rate

Akio Utsugi
National Institute of Advanced Industrial Science and Technology (AIST),
1-1-1 Higashi Tsukuba Ibaraki 305-8566, Japan
E-mail: a-utsugi@aist.go.jp

January 28,2002

Abstract

A generalized ICA model allowing overcomplete bases and additive noises in the observables is applied to natural image data. It is well known that such a model produces independent components that resemble simple cells in primary visual cortex or Gabor functions. We adopt a variable-sparsity density on each independent component, given by the mixture of a delta function and a standard Gaussian density. In the experiment, we observe that the aspect ratios of the optimal bases increase with the noise level and the degree of sparsity. The meaning of this phenomenon is discussed.

1 Introduction

The most important contribution of independent component analysis (ICA) to vision research is to find that the independent components of natural images are similar to simple cells in mammalian visual systems or Gabor functions [4][22]. The standard ICA model is extended to generative models allowing more independent components than the data dimension and additive noises in the observables. Although the parameters of such generalized ICA models are difficult to estimate exactly, several approximation estimation methods have been developed using the plug-in of maximum a posteriori (MAP) estimates of components [14], the saddle-point method [11][12], the variational method [2], and Gibbs sampling [15][21]. The generalized ICA models also obtain the independent components similar to simple cells from natural image data [14][11][15]. Moreover, extended ICA models

with vector-valued independent components obtain the independent components similar to complex cells [9][21].

The above generalized ICA models assume data generation by the *linear* transformation of hidden independent components. Furthermore, we need the *nonlinear* extension of the models to obtain higher-level independent components, such as objects in visual environment or concepts [14][3]. Although several kinds of nonlinear ICA models have already been proposed [10][17][18], they are restricted to invertible and noiseless transformations, and require additional constraints on the transformations or information on the source distributions for the uniqueness of the solutions. We have not yet found a proper constraint on the nonlinear transformation for such an ultimate code. We thus need further investigation with the help of the knowledge of brains.

We also need to explore the role of the linear system in the whole sensory process. One possible role is preprocessing for pattern recognition. Such a preprocessor is usually required to compress the input data by eliminating the background noise. It is desirable that the compression rate is variable to control the trade-off between the speed and the precision of the following pattern recognition.

Among linear transformations, the Karhunen-Loeve transformation (KLT) has optimality for data compression under the condition that the data distribution is Gaussian. This transformation is obtained by the principal component analysis (PCA) of the data. The discrete cosine transformation (DCT), used in the most popular image compression algorithm, is asymptotically equivalent to the KLT on the statistics of normal images [16]. Recently, wavelet transformations also become popular in image compression technique. Some kinds of wavelet transformations tend to produce sparse independent codes for natural images [8]. The Gabor transformation emerged in the ICA of natural images is regarded as the optimal linear transformation producing a sparse independent code.

Several researchers claim the superiority of ICA coding over PCA coding in sensory systems. One of the reasons is that ICA uses the higher-order statistics of images, which is ignored in PCA. Since the distribution of natural images is certainly far from Gaussian, PCA coding based only on the covariances loses its optimality. Another reason is that the independent components of natural images are more similar to simple cells than the principal components of those.

Nevertheless, PCA coding has an advantage when the variable compression rate is required. It has a constant ordered basis dictionary regardless of the compression rate. The bases in the dictionary are ordered according to their significance or activity. We can increase the compression rate simply by reducing the number of active bases from the dictionary. This property is maintained in the generative-model version of PCA, that is, probabilistic PCA [19].

In this paper, we observe the shape variation of the optimal bases in ICA coding

according to the compression rate measured by sparsity. In particular, the aspect ratios of the bases increase with the sparsity of the code. Thus, ICA coding cannot maintain a constant ordered basis dictionary, unlike PCA coding. We discuss the meaning of this phenomenon.

2 Linear generative model of images

A linear generative model for an image vector $\boldsymbol{x}_i(p \times 1)$ is given by

$$\boldsymbol{x}_i = \boldsymbol{\Lambda}\boldsymbol{y}_i + \boldsymbol{\epsilon}_i \tag{1}$$

where $\boldsymbol{y}_i(m \times 1)$ is a hidden source vector (a component vector), $\boldsymbol{\Lambda}(p \times m)$ is a linear transformation matrix and $\boldsymbol{\epsilon}_i(p \times 1)$ is an additive noise vector following a Gaussian distribution $N(\boldsymbol{0}, \boldsymbol{\Psi})$. All elements of $\boldsymbol{y}_i$ and $\boldsymbol{\epsilon}_i$ are assumed to be independent. Thus, the data density $f(\boldsymbol{x}_i|\boldsymbol{y}_i, \boldsymbol{\Lambda}, \boldsymbol{\Psi})$ is given by a Gaussian distribution $N(\boldsymbol{\Lambda}\boldsymbol{y}_i, \boldsymbol{\Psi})$ with a diagonal covariance matrix $\boldsymbol{\Psi}$. In this paper, we focus on spherical Gaussian noises: $\boldsymbol{\Psi} = \psi \boldsymbol{I}_p$. When $m \leq p$, the columns of $\boldsymbol{\Lambda} = [\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_m]$ are the bases of an intrinsic subspace in the data space. When $m > p$, they are called *overcomplete* bases.

2.1 Probabilistic PCA model

If the prior on the source vector $\boldsymbol{y}_i$ is a Gaussian distribution $N(\boldsymbol{0}, \boldsymbol{I}_m)$, the linear generative model is called a *probabilistic PCA model* [19]. The parameters $\boldsymbol{\Lambda}$ and ψ are estimated properly on a data set $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$, if $m < p$. The likelihood of the parameters is given by

$$f(\boldsymbol{X}|\boldsymbol{\Lambda}, \psi) = \int f(\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{\Lambda}, \psi) f(\boldsymbol{Y}) d\boldsymbol{Y} \tag{2}$$

where $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n]$ is regarded as *missing data* and integrated out. This is calculated easily, because the densities in the integrand are Gaussian. The maximum likelihood (ML) estimates of the parameters can be obtained from the eigenvalue decomposition of the data covariance. The estimated bases have the same shapes as the first m principal eigenvectors. The size of noise ψ is estimated by the mean of eigenvalues corresponding to the discarded eigenvectors. However, the number of the components m is difficult to determine only from data, because actual images do not follow strictly the generative model. In the application to data compression, m is set according to a desired compression rate. The ML estimate of ψ decreases monotonically with increasing m. This is normal rate-distortion relationship.

In fact, the ML estimates of the bases have rotational indefiniteness, that is, if $\hat{\boldsymbol{\Lambda}}$ is an ML estimate, $\hat{\boldsymbol{\Lambda}}\boldsymbol{R}$ is also an ML estimate for any orthogonal matrix $\boldsymbol{R}$. Nevertheless, the eigenvectors are special, because they construct an ordered basis dictionary, which is fixed regardless of m.

2.2 Generalized ICA model

If the source prior is the product of univariate non-Gaussian densities, the linear generative model (1) is called a *generalized ICA model*. In this case, the likelihood (2) is generally difficult to calculate exactly, unlike the case of the Gaussian prior. Thus, various approximation methods for the ML estimation are employed. In the generative models of natural images, the univariate non-Gaussian densities are assumed to be sparse, that is, unimodal, supergaussian and peaked at zero. Several sparse densities, such as a Cauchy distribution and a Laplace distribution, produce roughly similar results [14][11].

The shapes of the sparse densities can be learned from data. For example, a mixture-of-Gaussian density is employed as an adaptive univariate density [2]. The generalized ICA model with such a prior was applied to natural image data and obtained the independent components similar to simple cells [15]. In this experiment, each learned density ultimately became the mixture of a pulse and a broad Gaussian around the origin. Thus, we can use the mixture of a delta function and a standard Gaussian

$$f(y_{ki}|\tau_k) = (1-\tau_k)\delta(y_{ki}) + \frac{\tau_k}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}y_{ki}^2\Big) \tag{3}$$

as a sparse density for natural images from the beginning [21]. This univariate density has only one parameter $\tau_k \in [0,1]$, which is related to the degree of sparsity and estimated from data.

The generalized ICA model with the above prior is simply represented as

$$\boldsymbol{x}_i = \sum_{k=1}^{m} \boldsymbol{\lambda}_k z_{ki} \tilde{y}_{ki} + \boldsymbol{\epsilon}_i. \tag{4}$$

The binary variable z_{ki} expresses the activation of the kth component and follows a Bernoulli distribution

$$f(z_{ki}|\tau_k) = \tau_k^{z_{ki}}(1-\tau_k)^{1-z_{ki}} \tag{5}$$

where τ_k is the activation probability of the kth component. The real variable $\tilde{y}_{ki}$ follows a standard Gaussian distribution $N(0,1)$. All $z_{ki}, \tilde{y}_{ki}$ and $\boldsymbol{\epsilon}_i$ are independent. The independent components are given by $y_{ki} = z_{ki}\tilde{y}_{ki}$. This model is a

special case of *the ensemble of independent factor analyzers* (EIFA) [21] with one-dimensional factors. Moreover, it relates to a Bayesian wavelet shrinkage model [7], where $\mathbf{\Lambda}$ is fixed to predefined wavelet bases.

The ML estimates of $\mathbf{\Lambda}$ and $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_m)$ are obtained using a Monte Carlo expectation-maximization (MCEM) algorithm with a Gibbs sampler on the missing data z_{ki} and $\tilde{y}_{ki}$ [21]. The noise level ψ cannot be determined by the ML estimation in the case of complete or overcomplete bases. Moreover, ψ is regarded as temperature in simulated annealing for the MCEM algorithm. Thus, we should start the algorithm from a large value of ψ and reduce it gradually to an allowable noise level in order to avoid poor estimation.

The inactivation probabilities of the components $\bar{\tau}_k = 1 - \tau_k$ represent the sparsity of the univariate densities straightforwardly. Under a fixed m, we can show that the entropy of the source prior decreases with the sparsity of each univariate density. Thus, the attainable compression rate of internal code on this prior increases with the sparsity. Moreover, we can increase the degree of sparsity by increasing the allowable noise level ψ.

In the following experiment, we adopt the above model to observe the independent components of natural images under different degrees of sparsity.

3 Independent components of natural images

In our simulation experiment, the same images and preprocessing as Hyvärinen and Hoyer [9] are used. The images consist of 13 pictures of natural scenes including wild life. First, 40000 patches of 12×12 pixels are randomly extracted from the images ($n = 40000$). These patches constitute an original data set. Each data vector is then subtracted its DC component, and is projected onto a subspace spanned by 90 principal component vectors of the data ($p = 90$). The dimension of the subspace is determined such that its ratio to the original data dimension is the same as in [9], though its exact value is not important. On this subspace, the coordinates of the data vectors are scaled such that they have unit variance. This preprocessing corresponds to low-pass filtering and whitening, which are regarded as retinal process [1].

The generalized ICA models (4) with $m = 90$ and 135 are applied to the preprocessed data set. The ML estimates of $\mathbf{\Lambda}$ and $\boldsymbol{\tau}$ are obtained using the MCEM algorithm. In the first session (cooling), the noise variance ψ is initially set to one, and then reduced with a schedule given by $\psi = \exp(-0.001(t-1)), t = 1, \ldots, 1000$, where t is the number of iterations in the MCEM algorithm. The entries of $\boldsymbol{\tau}$ are initially set to 0.1. The initial value of $\mathbf{\Lambda}$ is set by a standard Gaussian random generator. In the second session (warming), the parameters and the missing data

are initially set to their values at $t = 500$ in the first session. Then, ψ is increased with a schedule given by $\psi = \exp(-0.001(500 - t + 1)), t = 1, \ldots, 1000$.

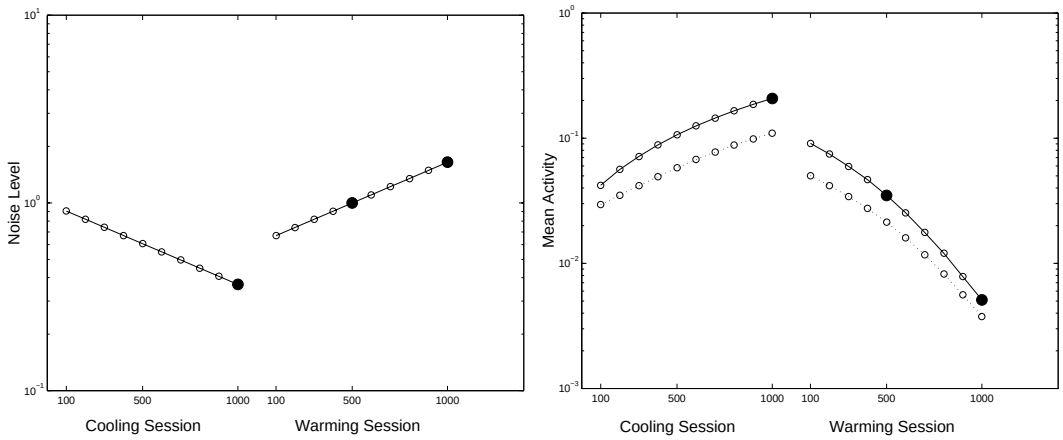

Figure 1: The variations of noise level and mean activity, in the case of $m = 90$ (solid lines) and $m = 135$ (dashed lines). The black circles indicate points corresponding to Figure 2–4.

Figure 1 illustrates the variations of ψ and mean τ_k. These graphs show the monotonic relationship between the noise level and the degree of sparsity. Such a rate-distortion relationship is general in optimally tuned coding schemes.

Figure 2–4 display the learned bases viewed on the original data space at $t = 1000$ in the cooling session and $t = 500, 1000$ in the warming session. We can observe that the shapes of the bases elongate with the degree of sparsity. To examine this phenomenon quantitatively, we fit a parametric Gabor function to the bases. The same method of Gabor fitting as Lewicki and Olshausen [11] is adopted, though the center of the Gabor function is constrained softly within the image range. From the fitted Gabor functions, the aspect ratios of the learned bases are calculated. The variation of the trimmed mean (10%) of the aspect ratios is illustrated in Figure 5. This graph shows the monotonic relationship between the degree of sparsity and the aspect ratios of the bases. The bases under the sparsest condition (Figure 4) are less stable than the other conditions. This is because the effective size of data for learning is very small under such a condition.

4 Discussion

We observed that the shapes of the optimal bases elongate as the compression rate grows with the allowable noise level. This phenomenon can be interpreted

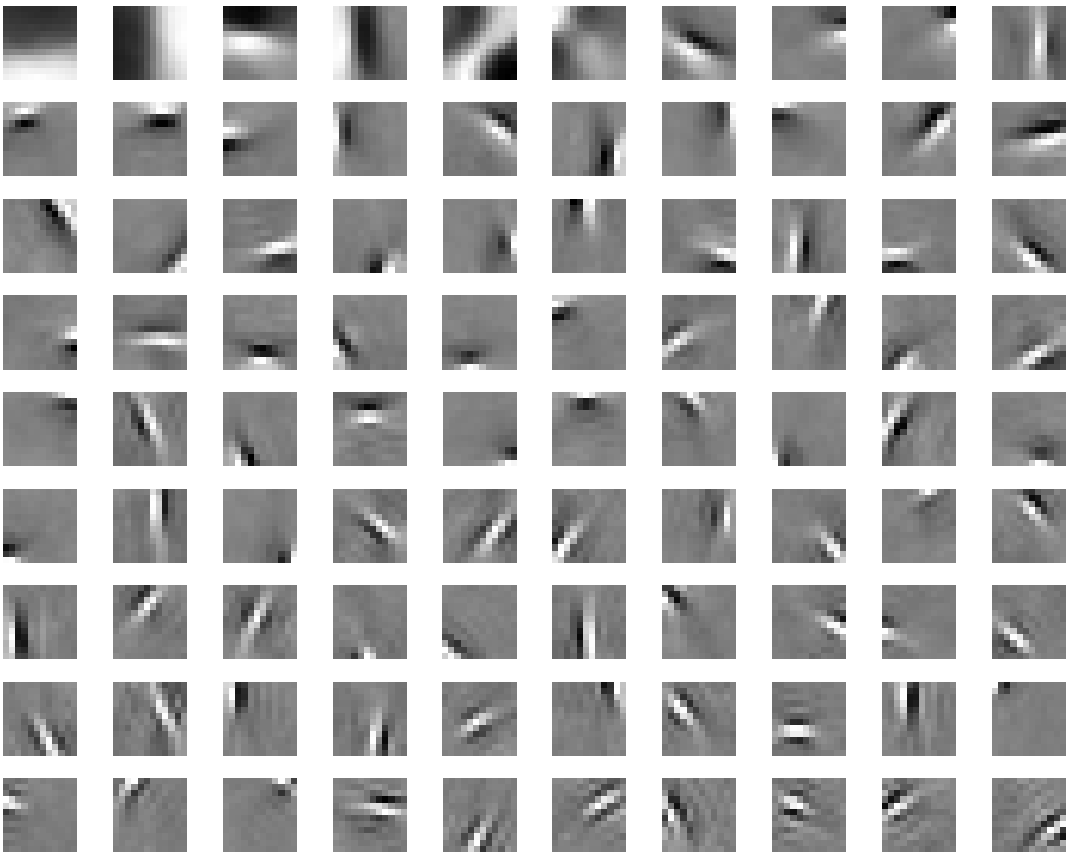

Figure 2: Learned bases of a generalized ICA model ($m = 90$) at $t = 1000$ in the cooling session (the densest condition). They are arranged in decreasing order of the energies of the components. The energy of a component is defined by the product of its activation probability and the squared Euclidean norm of the basis.

as follows. In an early visual stage, contours and textures are considered two main features of natural images. The textures are distributed all over images but have small intensity, while the contours are localized and appear rarely but have large intensity. Thus, ICA with a small allowable noise level and a small degree of sparsity captures independent components mainly for textures. The acquired Gabor functions with small aspect ratios are suitable for the representation of textures because of the efficiency. Previous studies on the ICA of natural images concentrated on this case. On the other hand, ICA with a large allowable noise level and a large degree of sparsity captures independent components for contours, where textures are discarded as noises. The contours prefer the elongate bases.

The features of the independent components generally become more complex

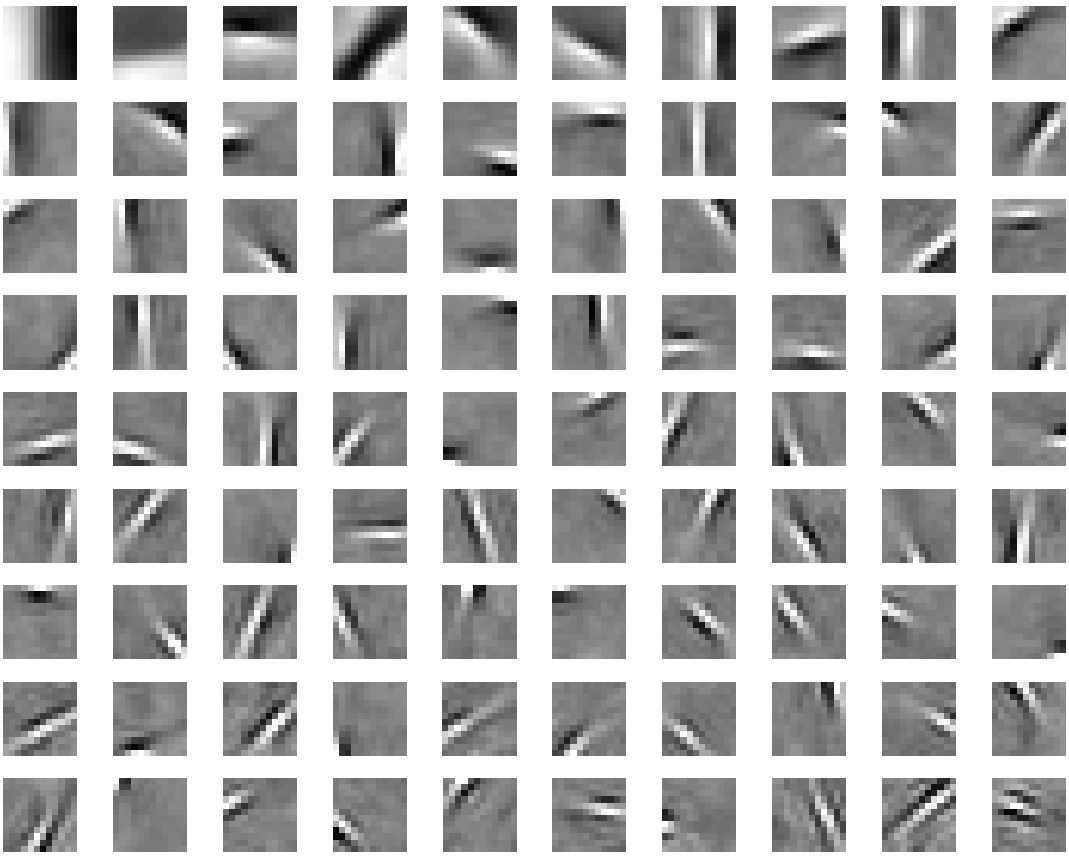

Figure 3: Learned bases at $t = 500$ in the warming session.

and macroscopic, as the size of the basis dictionary increases with the degree of sparsity. In the extreme case when the dictionary size equals to the data size ($m = n$), each data point can become the basis of an individual independent component, and this sparsest code attains the maximum complexity of the features. In fact, such a trivial case does not be considered, since the data should be regarded as infinite relative to the dictionary. Furthermore, the introduction of large allowable noises makes the model focus on components with high intensity by discarding weak components as noises. This increases the size of the potential basis dictionary, and thus the complexity of the features grows.

Can we obtain unlimitedly complex independent components by increasing the dictionary size and the patch size? This is actually difficult in simulation. However, the most serious problem in implementation is the explosion of the number of components covering huge variation of such complex features. A natural solution to this problem is to use *an activity pattern over units*, rather than *the activity of a*

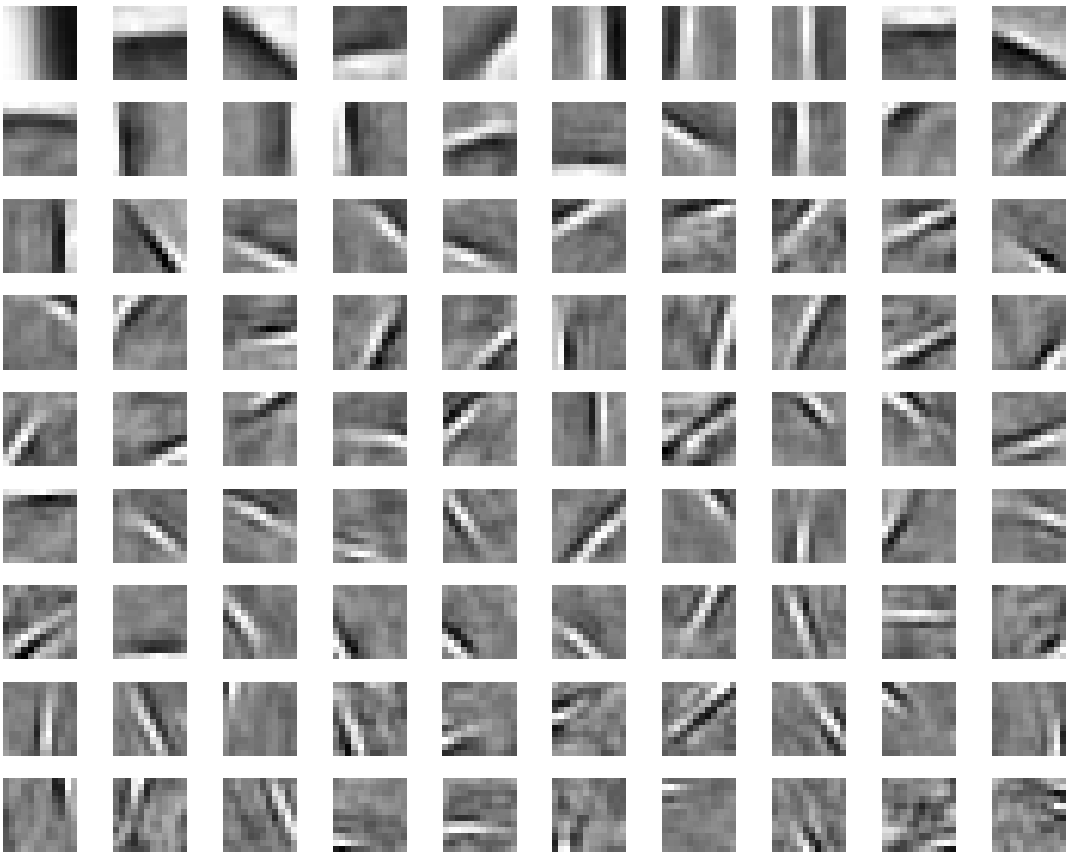

Figure 4: Learned bases at $t = 1000$ in the warming session (the sparsest condition).

single unit, as a component. Activity patterns over m units constitute a bounded domain S in $\boldsymbol{R}^m$. Each point in S corresponds to an independent component. (Alternatively, some points in S should be considered as equivalent.) The basis of the independent component is given by summing up the pattern vectors of the units weighted by their activities. Thus, the set of the bases is regarded as a smooth function over S. This smoothness may enable learning from finite data. The generative topographic mapping (GTM) [5][6][20] has similar structure, though it assumes that each data point is generated from a smooth function of a single point on a topological space. On the other hand, our model assumes that each data point is generated from the sum of multiple outputs of a smooth function, whose inputs are evoked independently over an internal space. The activation probabilities of the independent components can be given by a Markov random field (MRF), that is, an undirected graphical network over the units. The MRF is specified by a relatively small number of parameters representing the interaction among the units.

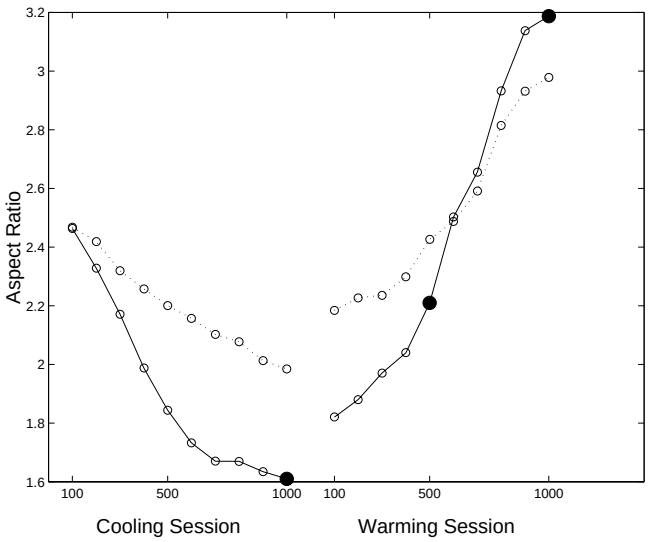

Figure 5: The variation of mean aspect ratio of the learned bases.

There are some neural network models for contour integration and contour classification using dynamic interaction among local edge detectors [13][23]. In these models, detectors that can constitute a long and smooth contour together have positive mutual interactions. Furthermore, the models have feature-binding mechanisms to differentiate independent contours in the activities of neural units. Although such models are not generative ones, their interaction structure over the units is available for the MRF.

We can introduce a global parameter controlling the strength of the mutual interaction. As the interaction strengthens, longer contour pieces have larger relative probabilities in the MRF. This corresponds to the elongation of the bases. Thus, this global parameter is related to the sparsity of the internal code. We can also consider the global parameter as the level of a hierarchical visual system. A dense code at the earliest stage provides basis representation for the various following stages, such as contour identification and uniform-texture-domain segmentation, where their proper sparser codes are constructed.

5 Conclusion

A generalized ICA model with variable sparsity was applied to natural image data. In the experiment, we observed the emergence of the independent components that resemble Gabor functions. This is consistent with the many studies on the ICA of natural images. Furthermore, we observed that the aspect ratios of the optimal

bases increase with the noise level and the degree of sparsity. This phenomenon was interpreted as the change of dominant visual features from textures to contours.

6 Acknowledgement

The author would like to acknowledge the helpful comments of the anonymous reviewers of this manuscript.

References

[1] J. J. Atick and A. N. Redlich, What does the retina know about natural scenes? *Neural Computation* **4** (1992) 196–210.

[2] H. Attias, Independent factor analysis, *Neural Computation* **11** (1999) 803–852.

[3] H. B. Barlow, Unsupervised learning, *Neural Computation* **1** (1989) 295–311.

[4] A. J. Bell and T. J. Sejnowski, The 'independent components' of natural scenes are edge filters, *Vision Res.* **37** (1997) 3327–3338.

[5] C. M. Bishop, M. Svensén and C. K. I. Williams, GTM: the generative topographic mapping, *Neural Computation* **10** (1998) 215–234.

[6] C. M. Bishop, M. Svensén and C. K. I. Williams, Developments of the generative topographic mapping, *Neurocomputing* **21** (1998) 203–224.

[7] M. Clyde, G. Parmigiani and B. Vidakovic, Multiple shrinkage and subset selection in wavelets, *Biometrika* **85** (1998) 391–402.

[8] D. J. Field, What is the goal of sensory coding? *Neural Computation* **6** (1994) 559–601.

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

[10] A. Hyvärinen and P. Pajunen, Nonlinear independent component analysis: Existence and uniqueness results, *Neural Networks* **12** (1999) 429–439.

[11] M. S. Lewicki and B. A. Olshausen, Probabilistic framework for the adaptation and comparison of image codes, *J. Opt. Soc. of Am. A* **16** (1999) 1587–1601.

[12] M. S. Lewicki and T. J. Sejnowski, Learning overcomplete representations, *Neural Computation* **12** (2000) 337–365.

[13] Z. Li, A neural model of contour integration in primary visual cortex, *Neural Computation* **10** (1998) 903–940.

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

[15] 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, eds., *Advances in Neural Information Processing Systems 12* (MIT press, Cambridge, 2000) 841–847.

[16] K. R. Rao and P. Yip, *Discrete Cosine Transform* (Academic Press, Boston, 1990).

[17] A. Taleb and C. Jutten, Source separation in post-nonlinear mixtures, *IEEE Trans. on Signal Processing* **47** (1999) 2807–2820.

[18] Y. Tan, J. Wang and J. M. Zurada, Nonlinear blind source separation using a radial basis function networks, *IEEE Trans. on Neural Networks* **12** (2001) 124–134.

[19] M. E. Tipping and C. M. Bishop, Mixtures of probabilistic principal component analyzers, *Neural Computation* **11** (1999) 443–482.

[20] A. Utsugi, Bayesian sampling and ensemble learning in generative topographic mapping, *Neural Processing Letters* **12** (2000) 277–290.

[21] A. Utsugi, Ensemble of independent factor analyzers with application to natural image analysis, *Neural Processing Letters* **14** (2001) 49–60.

[22] 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** (1998) 359–366.

[23] H. Wersing, J. J. Steil and H. Ritter, A competitive-layer model for feature binding and sensory segmentation, *Neural Computation* **13** (2001) 357–387.