Learning bases from Natural Sounds

by

Hsing-Lung Chou

Advisor

Ray-Bing Chen

Institute of Statistics,
National University of Kaohsiung
Kaohsiung, Taiwan 811 R.O.C.

June 2005

Contents

1 Introduction **1**

2 Null-space algorithm **2**

3 Experimental Results **5**

3.1 GGD with r=1 6

3.2 GGD with r=0.5 13

4 Discussion **16**

References **18**

Learning bases from Natural Sounds

Advisor: Dr. Ray-Bing Chen
Institute of Statistics
National University of Kaohsiung

Student: Hsing-Lung Chou
Institute of Statistics
National University of Kaohsiung

ABSTRACT

In this thesis, we are interested in learning the bases from natural sounds. The null-space algorithm, proposed by Chen and Wu (2002) and Chen (2003), is applied here for solving thus problem. Based on two different source assumptions, we demonstrate what bases we learn from monkey sounds. Finally we compare these bases according to the different source assumptions.

Keywords : independent component analysis, overcomplete representation, null-space algorithm, noiseless model.

從自然界聲音中找尋有用之聲音基底

指導教授: 陳瑞彬 博士
國立高雄大學統計學研究所

學生: 周興隆
國立高雄大學統計學研究所

摘要

在這篇論文裡，我們有興趣在自然界聲音中找尋有用的聲音基底。陳和吳(2002)以及陳(2003)提出了零空間演算法，可用以解決這一個問題。基於兩種不同的來源之模型假定，我們展示了從猴子聲音所找到的基底。最後我們將根據不同的來源假設來比較這些基底之差異。

關鍵字 :獨立因子分析, 過度完全表示法, 零空間演算法, 無誤差模型.

1 Introduction

Natural signals have been characterized statistical dependencies across space and time. One viewpoint of sensory systems is that they need to uncover these dependencies by processing them with filters whose form depends on the characteristic statistics of the ensemble of signals to which they are exposed (Barlow, 1989, and Atick and Redlich, 1990). This is called the reduce redundancy principle (Barlow, 1961 and 1989). Under this principle, the decompositions of the observations should be as independent as possible. In order to reduce the dependence among the observations, *Independent Component Analysis* (ICA, e.g., Comon, 1994, and Bell and Sejnowski, 1995) is one of the efficient algorithms to elucidate the higher-order statistics of natural signals. In ICA, the observations are assumed to be the linear mixtures of the independent sources. Let m be the number of observations, and M be the number of the independent sources. Then the observations is represented as

$$x_i = a_{i1}s_1 + \cdots + a_{iM}s_M, i = 1, \ldots, m,$$

where x_i is the i^{th} observation; s_j is the j^{th} source, $j = 1, \cdots, M$, and a_{ij} is the corresponding coefficient of the j^{th} source to the ith observation. Using the vector-matrix notation, this learn model is written as

$$\mathbf{x} = A\mathbf{s}, \tag{1}$$

where $\mathbf{x} = (x_1, \cdots, x_m)'$ is an m-dimensional observation vector, $\mathbf{s} = (s_1, \cdots, s_M)'$ is an M-dimensional source vector, and A is an $m \times M$ unknown basis matrix, and the Equation (1) is called independent component analysis (or ICA) model. In ICA algorithm, the mixing matrix A is assumed to be an invertible square matrix. That is the number of observations is the same as the number of sources, i.e. $M = m$. Hence the goal of ICA is to find an invertible square matrix W that makes outputs as independent as possible, i.e.,

$$\mathbf{u} = W\mathbf{x} = WA\mathbf{s},$$

where $\mathbf{u}$ is considered as an estimate of the source vector. The sources could be exactly recovered when W is the inverse of A up to a permutation and scale change, i.e.

$$WA = RS,$$

where R is a permutation matrix, and S is the scaling matrix.

Besides ICA algorithm, Olshausen and Field (1996) proposed a similar learning algorithm, Sparse Coding, for recovering the independent sources based on the noise model, i.e.

$$\mathbf{x} = A\mathbf{s} + \epsilon \tag{2}$$

where $\mathbf{x} = (x_1, \cdots, x_m)'$ is an m-dimensional observation vector; $\mathbf{s} = (s_1, \cdots, s_M)'$ is an M-dimensional source vector; A is an $m \times M$ basis matrix, and $\epsilon = (\epsilon_1, \cdots, \epsilon_m)'$ is Gaussian additive noise. Since in Sparse Coding, there is no restriction on the number of observations and sources, i.e. A may not be the square matrix, Olshausen and Field (1997) studied the sparse coding for the overcomplete situation. That is the A is a rectangular matrix with more columns than rows ($M > m$). Based on the same noise model, Equation (2), Lewicki and Olshausen (1999), and Lewicki and Sejnowski (2000) proposed the overcomplete representation by approximating the likelihood function with a Gaussian mound around the posterior estimation and can be considered as the extension of complete ICA. However, the original ICA model is a noiseless model. For this "noiseless" model, Chen and Wu (2002) and Chen (2003) proposed the null-space algorithm for the overcomplete independent component analysis. In their work, the model is

$$\mathbf{x} = A\mathbf{s}$$

where $\mathbf{x} = (x_1, \cdots, x_m)'$ is an m-dimensional observation vector, $\mathbf{s} = (s_1, \cdots, s_M)'$ is an M-dimensional source vector, A is an $m \times M$ rectangular basis matrix with $M > m$. Except the null-space algorithm, ICA, sparse coding and overcomplete representation have been applied successfully in learning the bases from image patches and natural sounds. For example: Bell and Sejnowski (1996) applied ICA algorithm to learn the higher order structures of natural sounds, and Lewicki and Sejnowski (2000) studied the human speech signals by overcomplete representation. In this work, we are interested in learning the bases from natural sounds by the null-space algorithm, i.e. estimate the unknown matrix A.

This thesis is organized as the follows. In Section 2, the null-space algorithm will be introduced. Then we apply this algorithm to learn the bases from natural sounds, and the results are shown in Section 3. Finally a discussion is given in Section 4.

2 Null-space algorithm

Consider the model

$$\mathbf{x}_t = A\mathbf{s}_t, t = 1, \cdots, T, \tag{3}$$

where $\mathbf{x}_t = (x_{1t}, \cdots, x_{mt})'$ collects the observation vector at time t, $\mathbf{s}_t = (s_{1t}, \cdots, s_{Mt})'$ collects the source vector at time t, and A is an $m \times M$ basis matrix with $M > m$. Since A is a rectangular matrix, the matrix A can be decomposed by singular value decomposition (SVD), i.e.

$$A = U \left(D \quad \mathbf{0} \right) V',$$

where U is an $m \times m$ orthogonal matrix; V is an $M \times M$ orthogonal matrix; $\mathbf{0}$ is an $m \times (M-m)$ zero matrix, and D is an $m \times m$ diagonal matrix with the real diagonal elements, d_i, such that

$$d_1 \geq d_2 \cdots \geq d_m \geq 0.$$

Hence, the solutions of Equation (3) can be represented as

$$\begin{aligned} \mathbf{s}_t &= A^{-}\mathbf{x}_t + V_2 c_t, \\ &= V_1 D^{-1} U' \mathbf{x}_t + V_2 c_t \\ &= V \left[\begin{pmatrix} D^{-1} \\ \mathbf{0} \end{pmatrix} U'\mathbf{x}_t + \begin{pmatrix} \mathbf{0} \\ I_{M-m} \end{pmatrix} c_t \right], \end{aligned} \tag{4}$$

where $A^{-} = V_1 D^{-1} U'$ is a generalized inverse of A; $V_1 = (v_1, \cdots, v_m)$ is the bases of the row space of A; $V_2 = (v_{m+1}, \cdots, v_M)$ is the bases of the null space of A; v_i is the i^{th} column of V, and c_t is a vector of coordinates in the null space. Then Equation (4) is called the null-space representation.

Assume the joint distribution of sources $\mathbf{s}_1, \cdots, \mathbf{s}_T$ to be $P(\mathbf{s}_1, \cdots, \mathbf{s}_T)$. With the null space representation, we have the joint pdf of $\mathbf{x}_1, \cdots, \mathbf{x}_T$ and $\mathbf{c}_1, \cdots, \mathbf{c}_T$,

$$P(c_1, \cdots, c_T, \mathbf{x}_1, \cdots, \mathbf{x}_T | A) = P(\mathbf{s}_1, \cdots, \mathbf{s}_T)|D|^{-T}.$$

Here $\mathbf{x}_1, \cdots, \mathbf{x}_T$ are observations; $c_1, \cdots, c_T$ are latent variables, and $A = U \begin{pmatrix} D & \mathbf{0} \end{pmatrix} V'$ is the unknown parameter. Based on Bayesian framework, we put uniform priors on U, V and $\log(D)$. The reason for why we can work on $\log(D)$ is that D is a scaling matrix. Then the posterior distribution of A is

$$\begin{aligned} & P(U, V, \log(D) | c_1, \cdots, c_T, \mathbf{x}_1, \cdots, \mathbf{x}_T) \\ \propto \; & P(U, V, \log(D)) P(c_1, \cdots, c_T, \mathbf{x}_1, \cdots, \mathbf{x}_T | U, V, \log(D)). \end{aligned} \tag{5}$$

According to this posterior, the estimation of the unknown parameter A can be found by the data augmentation algorithm of Tanner and Wong (1987). In fact, the data augmentation algorithm is a stochastic version of the EM algorithm (Dempster, Laird, and Rubin, 1997), which iterates of the two steps :

Step 1. Recovering s_t by sampling from $P(c_1, \cdots, c_T | \mathbf{x}_1, \cdots, \mathbf{x}_T, A)$, it means the distribution of the missing data given the observed data and the parameter.

Step 2. Estimating A by sampling from $P(U, V, \log(D) | c_1, \cdots, c_T, \mathbf{x}_1, \cdots, \mathbf{x}_T)$, it means complete data posterior distribution.

Here $P(c_1, \cdots, c_T|\mathbf{x}_1, \cdots, \mathbf{x}_T, A)$ and $P(U, V, \log(D)|c_1, \cdots, c_T, \mathbf{x}_1, \cdots, \mathbf{x}_T)$ are all proportional to the joint posterior, Equation (5). The difference of them is which is fixed and which is random.

To perform these two steps, Chen and Wu (2002) and Chen (2003) proposed the null-space algorithm. There are two parts in this algorithm. One is inhibition algorithm, and the other one is Givens sampler. We would describe these two algorithms in the following :

Part 1. Inhibition algorithm: This algorithm is to sample c_t from their conditional distribution $P(c_1, \cdots, c_T|\mathbf{x}_1, \cdots, \mathbf{x}_T, A)$ by Langevin-Eular moves. Assume the target distribution is $\pi(\mathbf{c}) \propto \exp(-H(\mathbf{c}))$. Then

$$\mathbf{c}(\tau + 1) = \mathbf{c}(\tau) - \frac{h}{2}\frac{\partial H(\mathbf{c})}{\partial \mathbf{c}}|_{\mathbf{c}=\mathbf{c}(\tau)} + \sqrt{h}Z_\tau, \tag{6}$$

where $\mathbf{c}(\tau)$ is the value of $\mathbf{c}$ at τ^{th} iteration of the Langevin-Euler process, i.e. $\mathbf{c}(\tau)$ collects the value of $(c_1, \cdots, c_T)$ at the τ^{th} iteration of Langevin-Euler move, Z_τ is white noise vector, and $h > 0$ is a suitable constant.

Part 2. Given sampler: This method is used to sample the orthogonal matrices U and V, and to estimate the diagonal matrix D from their posteriors. As mentioned before, we work on the log scale with $w_i = \log(d_i)$, and $w_1, \cdots, w_m$ satisfy $w_1 > \cdots > w_m$. The prior of $w_1, \cdots, w_m$ is assumed to be uniform with the order constraint. Hence the posterior distribution of $w_1, \cdots, w_m$ is

$$\begin{aligned} & P(w_1, \cdots, w_m|\mathbf{x}_1, \cdots, \mathbf{x}_T, c_1, \cdots, c_T, U, V) \\ \propto\ & P(\mathbf{s}_1, \cdots, \mathbf{s}_T) \exp(-T\sum_{i=1}^{m} w_i)). \end{aligned}$$

Therefore the maximum a posterior (MAP) estimation of $\mathbf{w} = (w_1, \cdots, w_m)'$ can be found by solving

$$\frac{\partial \log P(\mathbf{s}_1, \cdots, \mathbf{s}_T)}{\partial \mathbf{w}} = T\frac{\partial \sum_{i=1}^{m} w_i}{\partial \mathbf{w}}. \tag{7}$$

Now we consider how to sample U and V from their posterior distributions. The columns of U and V are orthogonal to each other, and we must maintain the orthogonality when we update U and V. Hence it is not good to sample U and V directly. So we accomplish this work with the following procedure. Suppose we want to update U. We randomly select two columns u_i, and u_j of U, and rotate them by an angle θ_{ij} on the plane spanned by the two vectors, i.e,

$$\begin{aligned} u_i &\leftarrow u_i \cos\theta_{ij} + u_j \sin\theta_{ij}, \\ u_j &\leftarrow -u_i \sin\theta_{ij} + u_j \cos\theta_{ij}. \end{aligned}$$

The distribution of θ_{ij} can be easily derived from the posterior distribution of U given everything else, i.e.,

$$P(\theta_{ij}) \propto P(c_1, \cdots, c_T, \mathbf{x}_1, \cdots, \mathbf{x}_T | A = U(i, j, \theta_{ij}) \, (\, D \quad \mathbf{0} \,) \, V) \tag{8}$$

and θ_{ij} can be drawn from $P(\theta_{ij})$ by the inversion method. Based on this procedure, we can maintain the orthogonality of all the column vectors, and the updating procedure is the same for V.

Here we summarize the null-space algorithm. The null-space algorithm iterates the following two steps :

Recovering source vector $\mathbf{s}_t$:

We use the Langevin-Euler moves to sample the null-space coefficient $c = (c_1, \cdots, c_T)$ from its condition distribution, and then update the sources according to the null-space representation.

Estimating basis matrix A :

Based on the updating sources, we can get the estimation of D by solving Equation (7) and update two columns of U or V by sampling the corresponding angle from Equation (8).

3 Experimental Results

In this thesis, we want to learn the bases from the natural sounds by null-space algorithm, and to see if there exist some special local structures in the bases. The observations we use here is the monkey's sound with 120 samples. The unrecovered sources are assumed to be independent at time t, and to come from a generalized gaussian distribution (GGD). Hence, the joint pdf of $\mathbf{s}_1, \cdots, \mathbf{s}_T$ is

$$\begin{aligned} P(\mathbf{s}_1, \cdots, \mathbf{s}_T) &= \prod_{t=1}^{T} P(\mathbf{s}_t) \\ &= \prod_{t=1}^{T} [\prod_{i=1}^{M} P(\mathbf{s}_{it})] \\ &= \prod_{t=1}^{T} \prod_{i=1}^{M} \frac{\lambda}{2} \exp\{-\lambda |\mathbf{s}_{it}|^r\}, \end{aligned} \tag{9}$$

where r denotes the shape parameter, and λ is related to the variance of the distribution. Here, the bases are learned by the null-space algorithm based on different shape parameters. Before running the null-space algorithm, we center the observations and do a whitening transformation on the observations to make the problem simpler.

3.1 GGD with r=1

When $r = 1$, the generalized gaussian distribution is the double exponential distribution. Then the joint distribution of $s_1, \cdots, s_T$ is

$$\begin{aligned} P(\mathbf{s}_1, \cdots, \mathbf{s}_T) &= \prod_{t=1}^{T} P(\mathbf{s}_t) \\ &= \prod_{t=1}^{T} [\prod_{i=1}^{M} P(\mathbf{s}_{it})] \\ &= \prod_{t=1}^{T} \prod_{i=1}^{M} \frac{\lambda}{2} \exp\{-\lambda |\mathbf{s}_{it}|\} \\ &= (\frac{\lambda}{2})^{MT} \exp\{-\lambda \sum_{t=1}^{T} \sum_{i=1}^{M} |\mathbf{s}_{it}|\} \end{aligned} \tag{10}$$

Based on this source assumption, we consider different sample sizes of observations to see if there exists any special pattern in the bases.

Case 1. In this case, for each time t, each observation vector $\mathbf{x_t}$ is 30×1, and the source vector $\mathbf{s_t}$ is assumed to be 40×1 . Then the basis matrix A a is 30×40 matrix. After 40000 iterations, the basis vectors are shown in Figure 1, and the distributions of recovered sources with the corresponding kurtoses are shown in Figure 2.

Case 2. In this case, for each time t, each observation vector $\mathbf{x_t}$ is 40×1, and the source vector $\mathbf{s_t}$ is assumed to be 50×1. Then the basis matrix A a is 40×50 matrix. After 60000 iterations, the basis vectors are shown in Figure 3, and the distributions of recovered sources with the corresponding kurtoses are shown in Figure 4.

Case 3. In this case, for each time t, each observation vector $\mathbf{x_t}$ is 50×1, and the source vector $\mathbf{s_t}$ is assumed to be 60×1. Then the basis matrix A is a 50×60 matrix. After 90000 iterations, the basis vectors are shown in Figure 5, and the distributions of recovered sources with the corresponding kurtoses are shown in Figure 6.

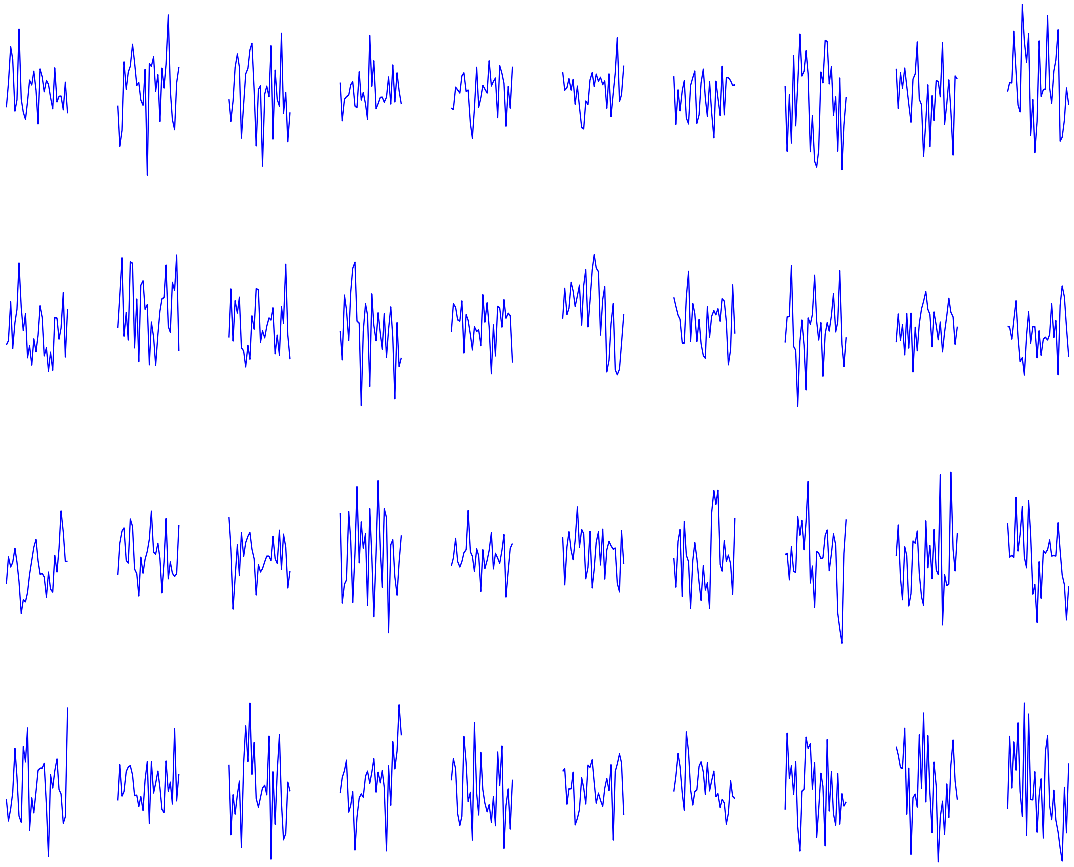

Figure 1: The patterns of the learned bases when the size of the mixing matrix A is 30×40.

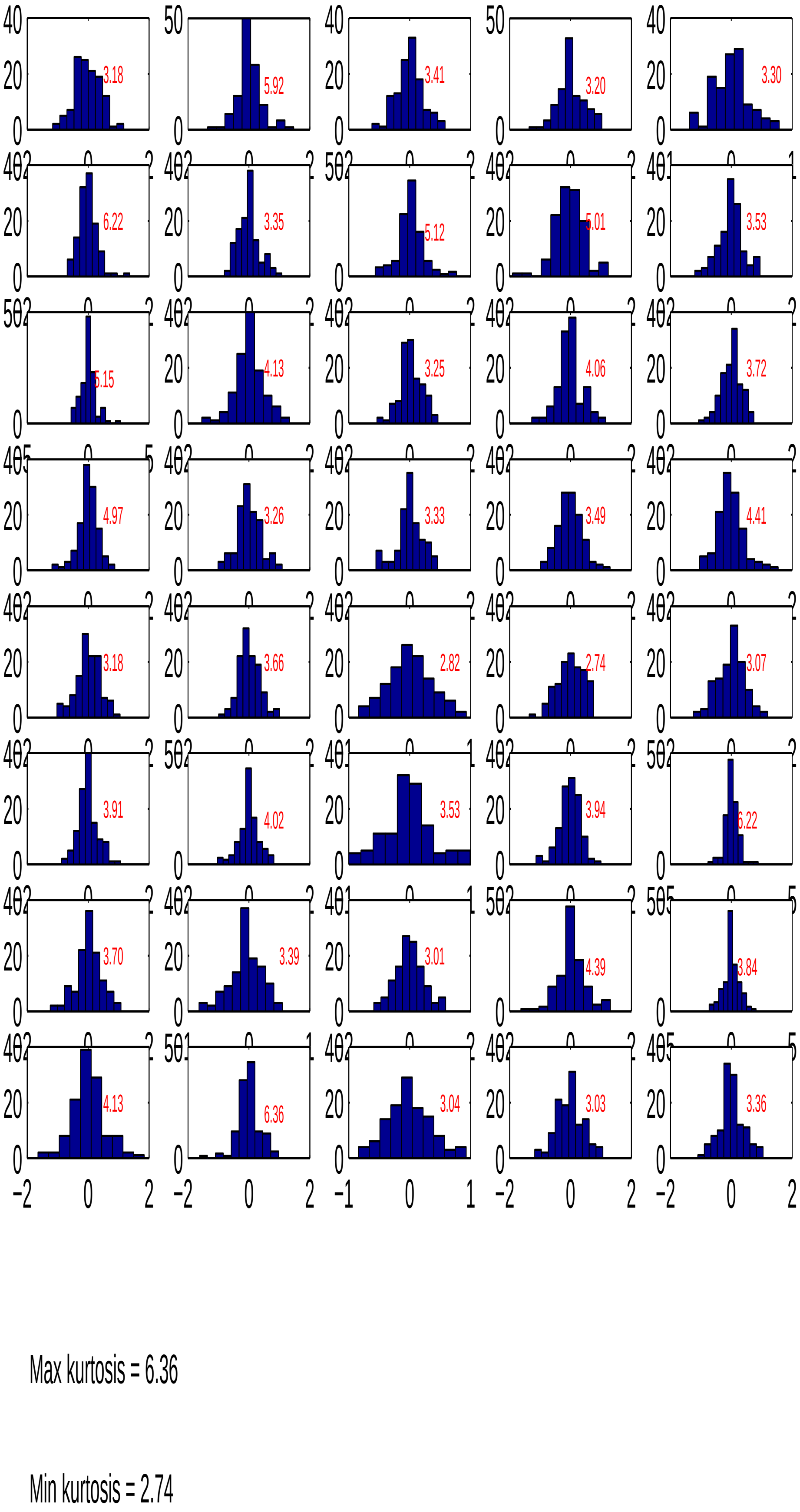

Figure 2: Distributions of learned sources and their corresponding kurtoses.

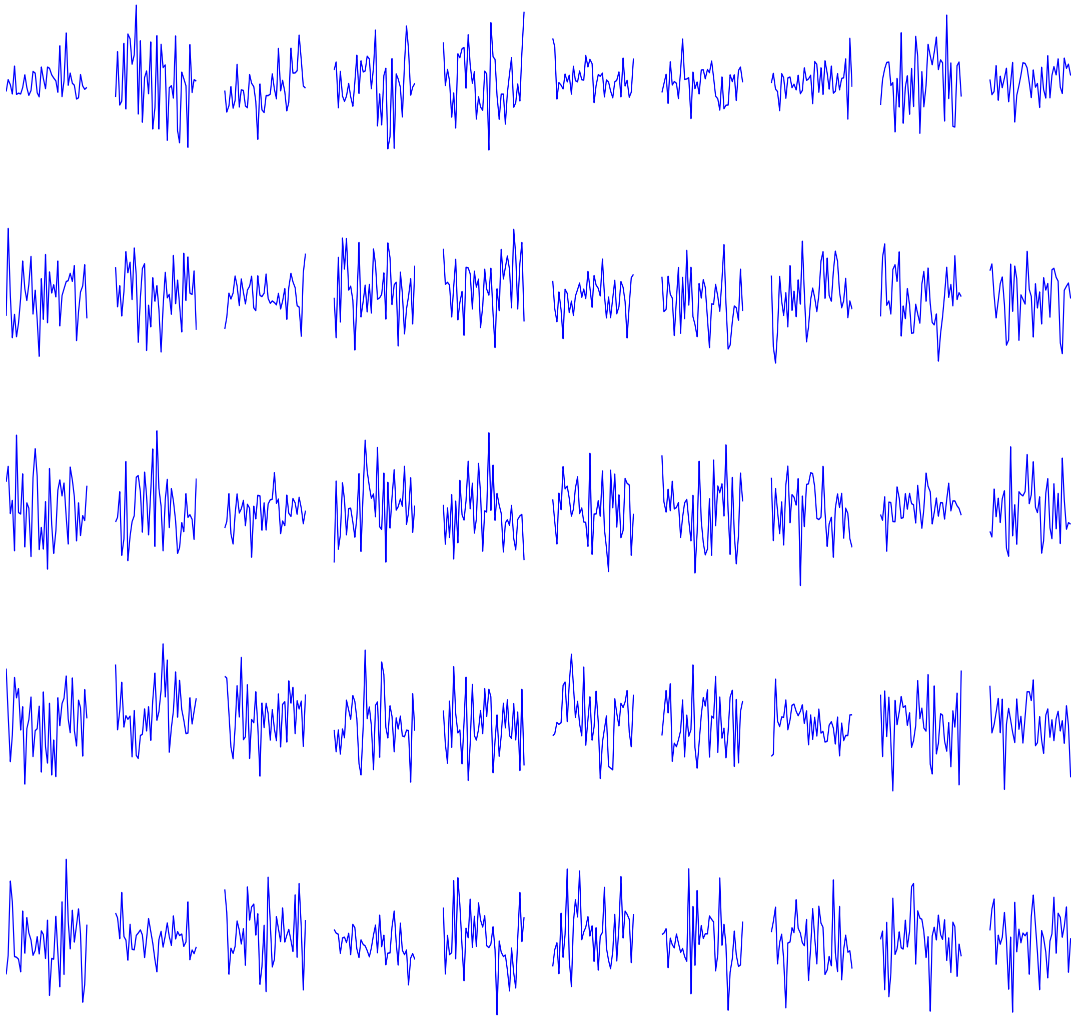

Figure 3: The patterns of the learned bases when the size of the mixing matrix A is 40×50.

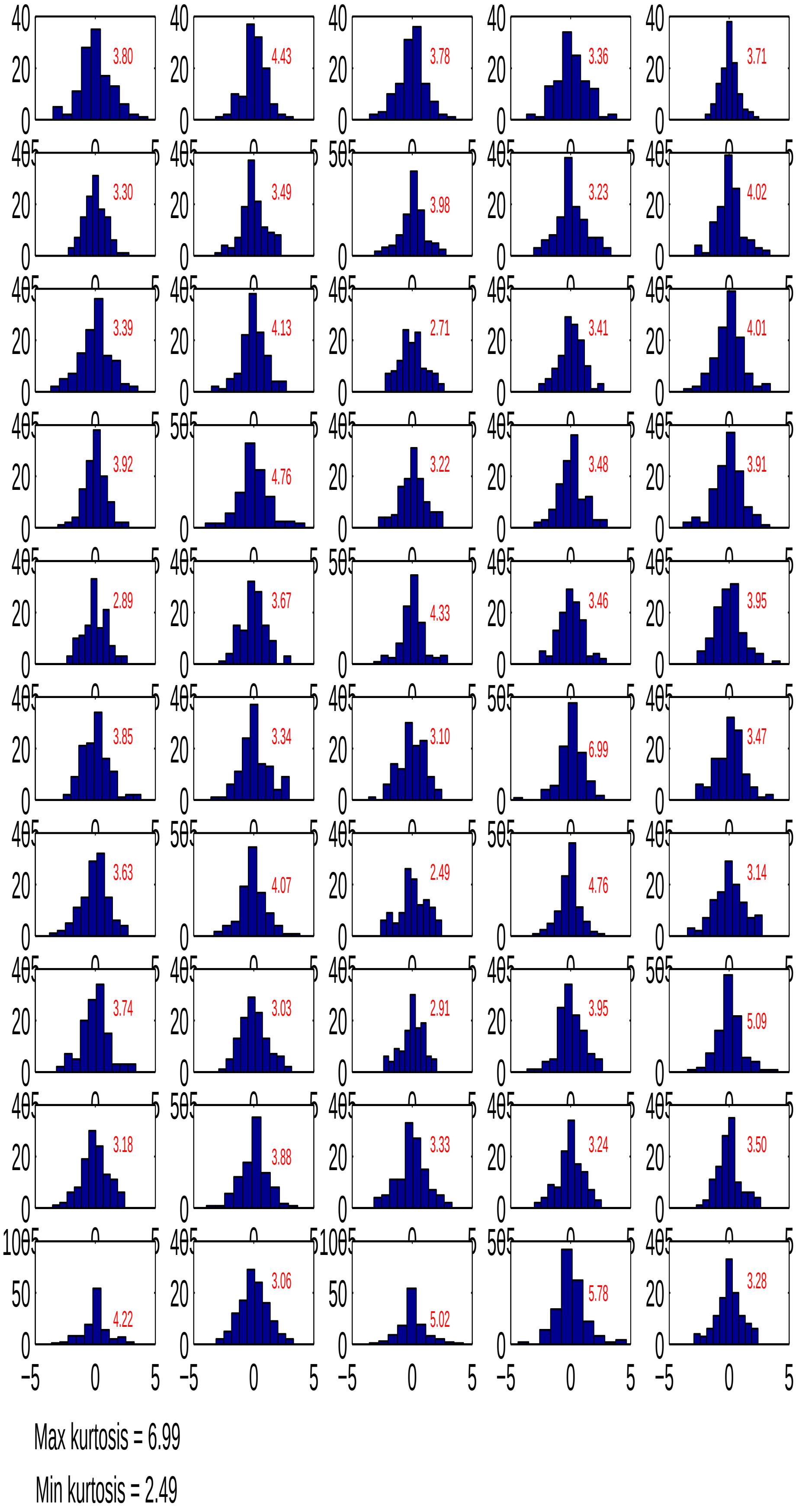

Figure 4: Distributions of learned sources and their corresponding kurtoses.

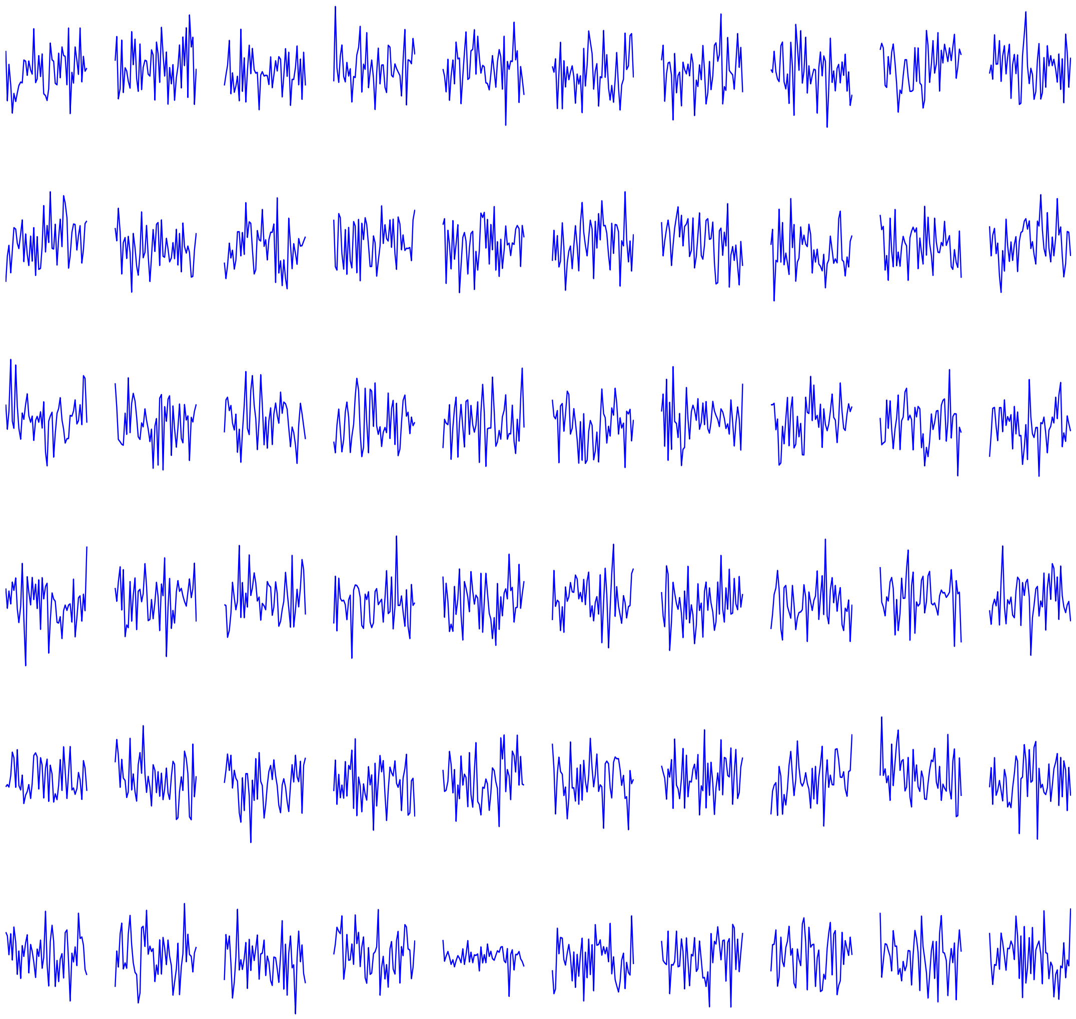

Figure 5: The patterns of the learned bases when the size of the mixing matrix A is 50×60.

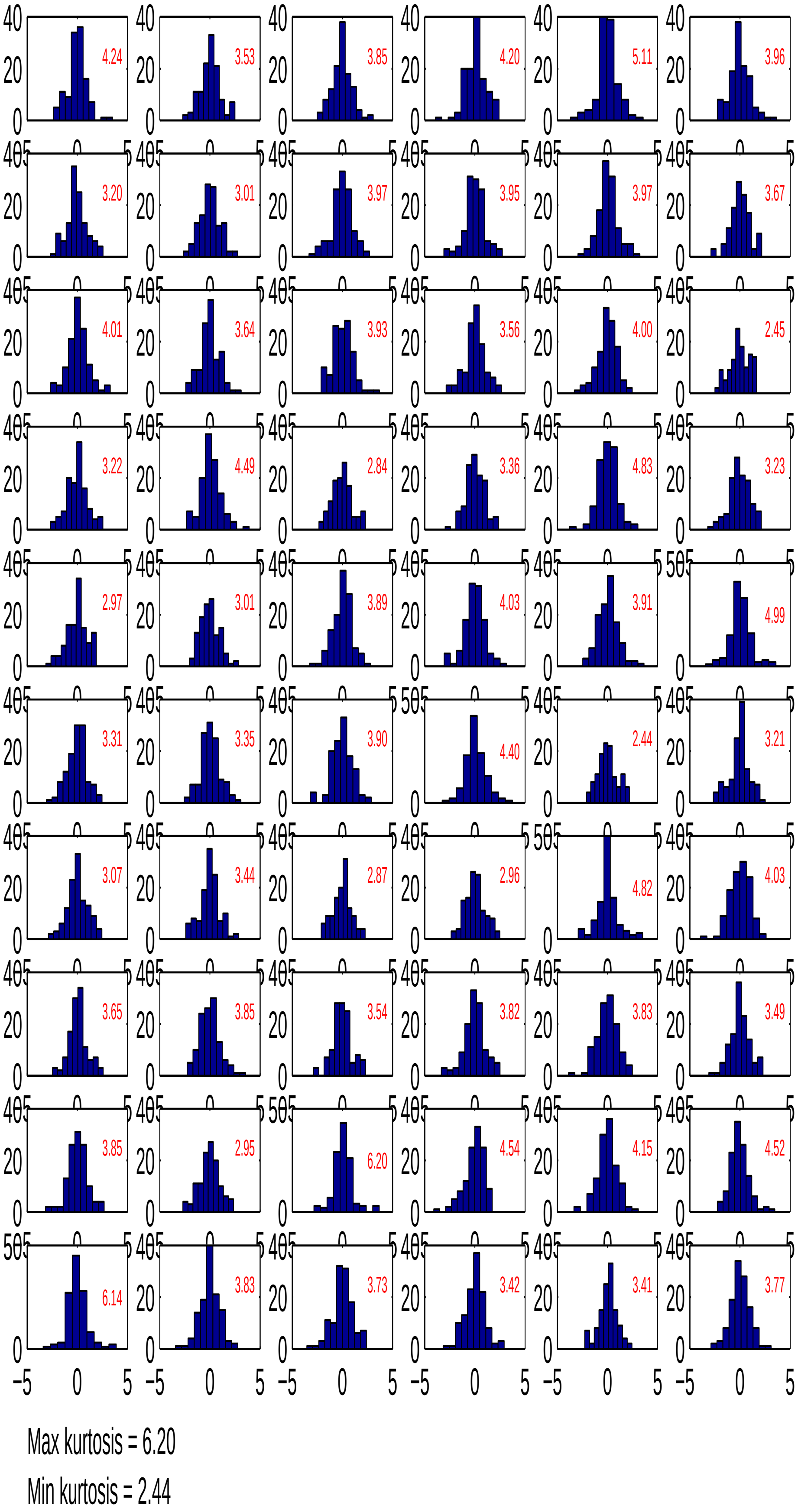

Figure 6: Distributions of learned sources and their corresponding kurtoses.

3.2 GGD with r=0.5

After the experiments with the double exponential distributed assumption, we consider another generalize gaussian distribution with shape parameter $r = 0.5$. The joint distribution of $\mathbf{s}_1, \cdots, \mathbf{s}_T$ is

$$\begin{aligned} P(\mathbf{s}_1, \cdots, \mathbf{s}_T) &= \prod_{t=1}^{T} P(\mathbf{s}_t) \\ &= \prod_{t=1}^{T} [\prod_{i=1}^{M} P(\mathbf{s}_{it})] \\ &= \prod_{t=1}^{T} \prod_{i=1}^{M} \frac{\lambda}{2} \exp\{-\lambda |\mathbf{s}_{it}|^{0.5}\} \\ &= (\frac{\lambda}{2})^{MT} \exp\{-\lambda \sum_{t=1}^{T} \sum_{i=1}^{M} |\mathbf{s}_{it}|^{0.5}\} \end{aligned} \tag{11}$$

In this case, for each time t, each observation vector $\mathbf{x_t}$ is 30×1, and the source vector $\mathbf{s_t}$ is assumed to be 40×1. Then basis matrix A is a 30×40 matrix. After 40000 iterations, the basis vectors are shown in Figure 7, and the distributions of recovered sources with the corresponding kurtoses are shown in Figure 8.

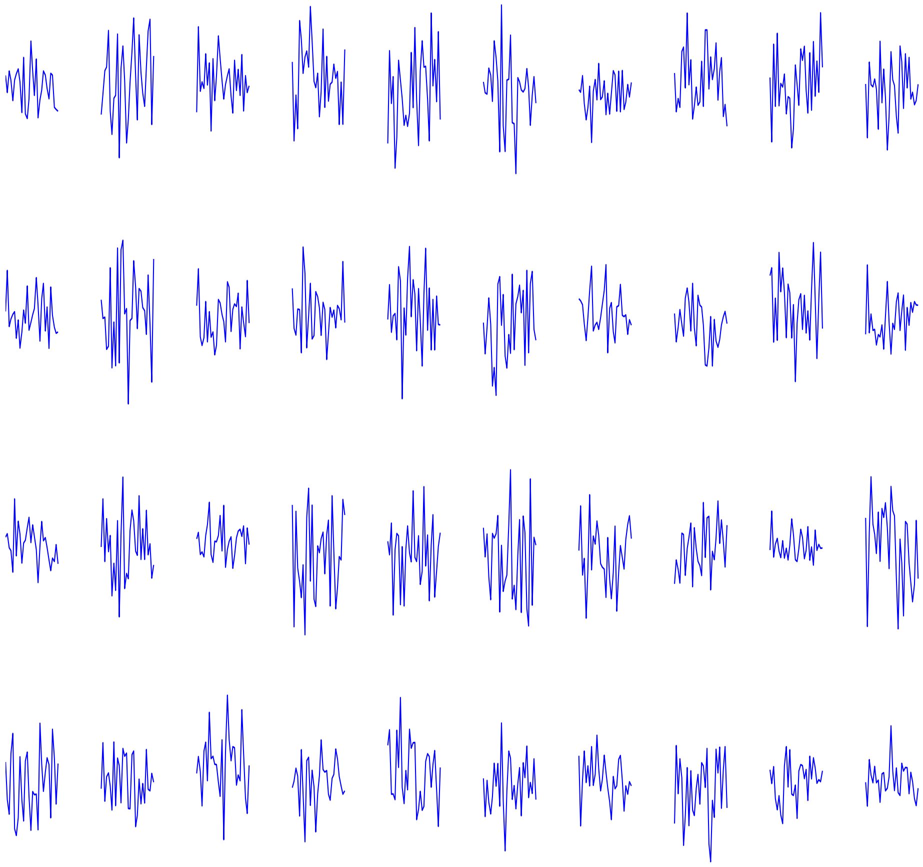

Figure 7: The patterns of the learned bases when the size of the mixing matrix A is 40×40.

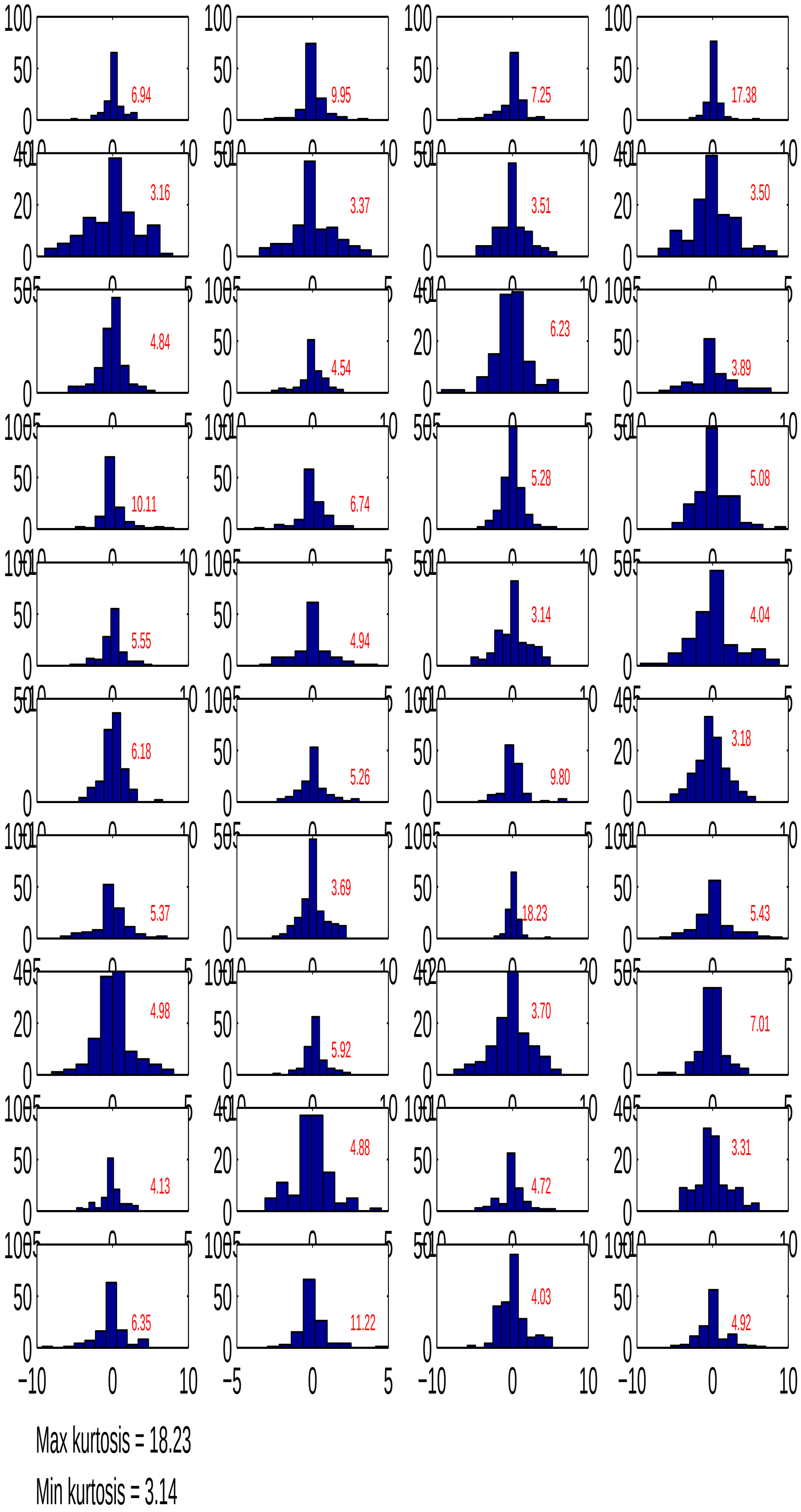

Figure 8: Distributions of learned sources and their corresponding kurtoses.

4 Discussion

In our work, the bases of natural sounds are learned by the null-space algorithm with the different source assumptions. Since all we have is the observations, how to get good results depends on the assumptions of sources. Here the marginal distributions of sources are assumed to be the generalized gaussian distributions with shape parameter $r = 1$ and $r = 0.5$, i.e.

$$P(\mathbf{s}_1, \cdots, \mathbf{s}_T) \quad = (\frac{\lambda}{2})^{MT} \exp\{-\lambda \sum_{t=1}^{T} \sum_{i=1}^{M} |\mathbf{s}_{it}|\}$$

and

$$P(\mathbf{s}_1, \cdots, \mathbf{s}_T) \quad = (\frac{\lambda}{2})^{MT} \exp\{-\lambda \sum_{t=1}^{T} \sum_{i=1}^{M} |\mathbf{s}_{it}|^{0.5}\}.$$

According to the results in Lewicki and Sejnowski (2000), what we want to do is to see if there exist localized structures in the bases. With the same observations, we compare the patterns of learned bases in Figure 1 and Figure 7. It seems that the bases in Figure 7 contain more localized and special structures than the bases in Figure 1. The distributions of the learned sources must have the sharp peak at zero and heavy tails, because the sources are assumed to follow the sparse distributions. Hence, we plot the histograms of distributions of sources to see the sparseness of sources. Clearly, most source distributions of GGD assumption with $r = 0.5$ have higher sharp peak at zero than those of double exponential assumption, and they also have the higher kurtoses. Therefore, we prefer the generalized gaussian distribution with shape parameter $r = 0.5$ to be the marginal distribution of our sources in our experiment.

In this thesis, we assume sound signals, s_{ij}, follow an independently identical distribution. However, the sound signals should not be independent in the real world. Lewicki (2002) assumed the signal observation $x(t)$ in a time window of length N to learn bases. That is

$$s_i(t) = \sum_{\tau=0}^{N} x(\tau) a_i(t - \tau)$$

s_i(t) is the i^{th} source, $i = 1, \cdots, M$, and a_i(t) is the bases of natural sounds at time t. Hence, the sources assumption is not time independent, and we may apply this idea with the null-space algorithm to learn bases from natural sounds.

Here we only consider the size of basis matrix A being 30×40; 40×50 or 50×60. These size may be too small to catch the localized structures. Hence we suggest to use bigger basis matrix. However, the bigger matrix is, the more computing time we need to learn. Therefore, we should improve our programming ability to have more efficient code. Besides learning the bases from natural sounds, Olshausen and Field (1996), Lewicki and Olshausen (1999), and Olshausen and Millman (2000) also considered the problems on learning the bases from image patches . Hence it could be interested to apply the null-space algorithm into small image patches to see what bases we would learn from images. As we mentioned before, we need to have more efficient coding ability because for images, the size of A should be larger and larger. For example A is assumed to be 64×128 in Olshausen and Millman (2000).

References

[1] Atick, J. J. and Redlich, A. N. (1990). Towards a theory of early visual processing. *Neural Computation*, **2**, 308-320.

[2] Barlow, H. B. (1961). Possible principles underlying the transformations of sensory messages. In: *Sensory Communication*, Rosenbluth, W. A. ed., MIT Press, 217-234.

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

[4] Bell, A. J. and Sejnowsi, T. J. (1995). An information-maximizaition approach to blind separation and blind deconvolution, *Neural Computation*, **7**(6), 1129-1159.

[5] Bell, A. J. and Sejnowsi, T. J. (1996). Learning the higher-order structure of a natural sound. *Network: Computation in Neural Systems*, **7**, 261-266.

[6] Bell, A. J. and Sejnowsi, T. J. (1997). The 'Independent components' of natural scences are edge filters. *Vision Research*, **37**, 3327-3338.

[7] Comon, P. (1994). Independent component analysis, a new concept? *Singal Processing*, **36**, 287-314.

[8] Chen, R.-B. and Wu, Y.-N. (2002). A null-space Representation for Overcomplete Independent Component Analysis. *2002 Proceedings of American Statistical Association, Statistical Computing Section [CD-ROM]. Alexandria, VA: American Statistical Association.*

[9] Chen, R.-B. (2003). A null-space algorithm for overcomplete independent component analysis. Ph.D. dissertation, Dept. of Statistics, University of California at Los Angeles.

[10] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via EM algorithm. *Journal of the Royal Statistical Society B*, **39**, 1-38.

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

[12] Field, D. J. and Olshausen, B. A. (1996). Emergence of simple-cell receptive-field properties by learning a sparse coding for natural images. *Natural*, **381**, 607-609.

[13] Field, D. J. and Olshausen, B. A. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, **11**, 3311-3325.

[14] Lewicki, M. S. and Olshausen, B. A. (1999). A probabilistic framwork for the adaption and comparison of image codes. *J. Opt. Soc. Am. A: Optics, Image Science, and Vision*, **16**(7), 1587-1601.

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

[16] Lewicki, M. S. (2002). Efficient coding of natural sounds. *Natural Neurosci*, **5**(4), 356-363.

[17] Millman, K. J. and Olshausen, B. A. (2000). Learning Sparse Codes with a Mixture-of-Gaussians Prior. *Advances in Neural Information Processing Systems*, **12**, 841-847.

[18] Tanner, M. and Wang, W. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, **82**, 528-550.