Bayesian Independent Component Analysis under Hierarchical Model on Independent Components

Kai Asaba*, Shota Saito*, Shunsuke Horii*, and Toshiyasu Matsushima*
* Waseda University, Tokyo, Japan
E-mail: kaikai0426@akane.waseda.jp, shota@aoni.waseda.jp, s.horii@aoni.waseda.jp, toshimat@waseda.jp

***Abstract*—Independent component analysis (ICA) deals with the problem of estimating unknown latent variables (independent components) from observed data. One of the previous studies of ICA assumes a Laplace distribution on independent components. However, this assumption makes it difficult to calculate the posterior distribution of independent components. On the other hand, in the problem of sparse linear regression, several studies have approximately calculated the posterior distribution of parameters by assuming a hierarchical model expressing a Laplace distribution. This paper considers ICA in which a hierarchical model expressing a Laplace distribution is assumed on independent components. For this hierarchical model, we propose a method of calculating the approximate posterior distribution of independent components by using a variational Bayes method. Through some experiments, we show the effectiveness of our proposed method.**

I. Introduction

Independent component analysis (ICA) deals with the problem of estimating unknown latent variables (independent components) from observed data. It is applied to speech signal processing, time series analysis, image feature extraction, and so on (see, e.g., [6]). In ICA, independent components are assumed to be mutually independent and nongaussian. To give a model generality, many studies on ICA do not explicitly assume a nongaussian distribution on independent components.

On the other hand, when we have some prior knowledge of independent components, the improvement of the estimation accuracy is expected by using this prior knowledge. For this reason, several studies (e.g., [5], [8], [10], [11], [12]) have explicitly assumed a nongaussian distribution on independent components. Especially, in [5], a Laplace distribution is assumed for independent components. This is partly because a distribution of image data and speech signal is empirically known to have high kurtosis and a Laplace distribution is widely used to express such a distribution. However, one problem is that we have difficulty deriving a posterior distribution of independent components due to an absolute value in a probability density function of a Laplace distribution.

A similar problem also arises in the problem of sparse linear regression. That is, we cannot calculate a posterior distribution of a parameter analytically when we assume a Laplace distribution as a prior distribution of a parameter. Nevertheless, we can approximately calculate the posterior distribution using an EM algorithm, a variational Bayes method, and a Gibbs sampling by assuming a hierarchical model. This is because a Laplace distribution can be expressed as a mixture of Gaussian distributions whose variances obey exponential distributions (see, e.g., [2], [4], [7], [9]).

The preceding discussions are summarized as follows:

1) In ICA, the previous study [5] has assumed a Laplace distribution on independent components. However, this assumption makes it difficult to derive a posterior distribution of independent components.
2) In the problem of sparse linear regression, it is possible to calculate a posterior distribution of a parameter by expressing a Laplace distribution as a hierarchical model.

In view of 1) and 2), we consider ICA in which a hierarchical model expressing a Laplace distribution is assumed on independent components. For this hierarchical model, we derive an approximate posterior distribution of independent components by using a variational Bayes method. The ICA model that we consider can be seen as the problem of dictionary learning. Thus, we discuss a relationship between our study and the previous study [13] in which a hierarchical model is assumed for the problem of dictionary learning. Moreover, through some experiments, we show the effectiveness of our proposed method.

The organization of this paper is as follows. Section II formulates ICA model and describes a hierarchical prior distribution of independent components. Section III compares our study and the previous study [13]. Section IV derives an approximate posterior distribution of independent components by using a variational Bayes method. Section V describes our experiments and Section VI discusses the experimental results. Finally, Section VII concludes this paper.

II. Hierarchical model of independent components

Suppose we have N observed data $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$, where $\boldsymbol{x}_n = (x_{1n}, \ldots, x_{d_x n})^T \in \mathbb{R}^{d_x}$ for $n = 1, \ldots, N$. We assume that the observed data are expressed as a linear transformation of independent latent variables called independent components. That is, the observed data are expressed as

$$\boldsymbol{x}_n = W\boldsymbol{u}_n + \boldsymbol{\epsilon}_n, \quad (n = 1, \ldots, N) \tag{1}$$

where $W \in \mathbb{R}^{d_x \times d_u}$ is a matrix which expresses a linear transformation, $\boldsymbol{u}_n = (u_{1n}, \ldots, u_{d_u n})^T \in \mathbb{R}^{d_u}$ represents an independent component, and $\boldsymbol{\epsilon}_n = (\epsilon_{1n}, \ldots, \epsilon_{d_x n})^T \in \mathbb{R}^{d_x}$ is a noise vector whose component ϵ_{in} obeys a Gaussian distribution $\mathcal{N}(\epsilon_{in} \mid 0, \sigma^2)$.

Define $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \in \mathbb{R}^{d_x \times N}$, $U = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N) \in \mathbb{R}^{d_u \times N}$, and $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_N) \in \mathbb{R}^{d_x \times N}$. Then, the model (1) is expressed as

$$X = WU + \boldsymbol{\epsilon}. \tag{2}$$

In this study, we assume the following three-layer hierarchical prior distribution on independent components. In the first layer, we place a Gaussian distribution

$$p(\boldsymbol{u}_n \mid \boldsymbol{\lambda}) = \prod_{i=1}^{d_u} p(u_{in} \mid \lambda_i) \tag{3}$$

$$= \prod_{i=1}^{d_u} \mathcal{N}(u_{in} \mid 0, \lambda_i), \tag{4}$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_{d_u})^T \in \mathbb{R}^{d_u}$. Next, in the second layer, $\boldsymbol{\lambda}$ is assigned an exponential distribution

$$p(\boldsymbol{\lambda} \mid \boldsymbol{\alpha}) = \prod_{i=1}^{d_u} p(\lambda_i \mid \alpha_i) \tag{5}$$

$$= \prod_{i=1}^{d_u} \mathrm{Exp}(\lambda_i \mid \alpha_i), \tag{6}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_{d_u})^T \in \mathbb{R}^{d_u}$. Lastly, in the third layer, $\boldsymbol{\alpha}$ is assigned a Gamma distribution

$$p(\boldsymbol{\alpha}) = \prod_{i=1}^{d_u} p(\alpha_i) \tag{7}$$

$$= \prod_{i=1}^{d_u} \mathrm{Gam}(\alpha_i; a, b), \tag{8}$$

where a and b are parameters of a Gamma distribution. Integrating out λ_i, we obtain a Laplace distribution, i.e.,

$$p(u_{in} \mid \alpha_i) = \int p(u_{in} \mid \lambda_i) p(\lambda_i \mid \alpha_i) d\lambda_i \tag{9}$$

is a Laplace distribution (see, e.g., [1]).

We assume that each component of the matrix $W \in \mathbb{R}^{d_x \times d_u}$ obeys a Gaussian distribution $\mathcal{N}(w_{ij} \mid \mu, \tau^2)$.

III. Relationship with the previous study in the problem of dictionary learning

The model (2) can be seen as the problem of dictionary learning in the sense that we decompose the given data X into the matrix W (which expresses the basis) and the sparse matrix U (which expresses the coefficient). In view of this observation, this section describes the relationship between our study and the previous study [13] in which a hierarchical model is assumed for the problem of dictionary learning.

In [13], the model (2) is assumed and the following two-layer hierarchical prior distribution is considered. In the first layer, a Gaussian distribution is placed on U, i.e.,

$$p(U \mid \boldsymbol{\beta}) = \prod_{i=1}^{d_u} \prod_{n=1}^{N} p(u_{in} \mid \beta_{in}) \tag{10}$$

$$= \prod_{i=1}^{d_u} \prod_{n=1}^{N} \mathcal{N}\left(u_{in} \;\middle|\; 0, \frac{1}{\beta_{in}}\right), \tag{11}$$

where $\boldsymbol{\beta}$ is a vector which has a component β_{in} ($i = 1, \ldots, d_u, n = 1, \ldots, N$). In the second layer, $\boldsymbol{\beta}$ is assigned a Gamma distribution

$$p(\boldsymbol{\beta}) = \prod_{i=1}^{d_u} \prod_{n=1}^{N} p(\beta_{in}) \tag{12}$$

$$= \prod_{i=1}^{d_u} \prod_{n=1}^{N} \mathrm{Gam}(\beta_{in}; c, d), \tag{13}$$

where c and d are parameters of a Gamma distribution. Integrating out β_{in}, we obtain a student-t distribution, i.e.,

$$p(u_{in}) = \int p(u_{in} \mid \beta_{in}) p(\beta_{in}) d\beta_{in} \tag{14}$$

is a student-t distribution.

Thus, the differences between the previous study [13] and our study are summarized as follows:

- In [13], the two-layer hierarchical model is assumed. A variance parameter $1/\beta_{in}$ of the matrix U is different from each component u_{in} as shown in (11). Further, $p(u_{in})$ is a student-t distribution as shown in (14).
- In our study, the three-layer hierarchical model is assumed. A variance parameter λ_i of the matrix U is different from each row as shown in (4). Moreover, $p(u_{in} \mid \alpha_i)$ is a Laplace distribution as shown in (9).

IV. Posterior distribution based on a Variational Bayes method

This study derives an approximate posterior distribution of a true posterior distributuion $p(W, U, \boldsymbol{\lambda}, \boldsymbol{\alpha} \mid X)$ by using a variational Bayes method [3]. In the approximation by the variational Bayes method, the objective is to obtain an approximate posterior distribution $q^\star(W, U, \boldsymbol{\lambda}, \boldsymbol{\alpha})$ which minimizes Kullback-Leibler (KL) divergence between the approximate posterior distribution and the true posterior distribution. That is, we aim to obtain

$$q^\star(W, U, \boldsymbol{\lambda}, \boldsymbol{\alpha})$$

$$= \operatorname*{arg\,min}_{q(W,U,\boldsymbol{\lambda},\boldsymbol{\alpha})} \int q(W, U, \boldsymbol{\lambda}, \boldsymbol{\alpha}) \log \frac{q(W, U, \boldsymbol{\lambda}, \boldsymbol{\alpha})}{p(W, U, \boldsymbol{\lambda}, \boldsymbol{\alpha} \mid X)}. \tag{15}$$

However, it is difficult to carry out the minimization in (15) for arbitrary probability distribution. Thus, we restrict the optimization distribution to $q(W, U, \boldsymbol{\lambda}, \boldsymbol{\alpha})$ which can be factrized as

$$q(W, U, \boldsymbol{\lambda}, \boldsymbol{\alpha}) = q(W) q(U) q(\boldsymbol{\lambda}) q(\boldsymbol{\alpha}). \tag{16}$$

Using this factorization, we can calculate the minimization in (15) by updating $q(\cdot)$ iteratively. We denote by $q_t(W)$, $q_t(U)$, $q_t(\boldsymbol{\lambda})$, and $q_t(\boldsymbol{\alpha})$ an approximate distribution of W, U, $\boldsymbol{\lambda}$, and $\boldsymbol{\alpha}$ at t-th iteration, respectively. Then, the updates of posterior distributions are given as follows [3]:

$$\ln q_{t+1}(W) \propto \mathbb{E}_{q_t(U) q_t(\boldsymbol{\lambda}) q_t(\boldsymbol{\alpha})}[\ln p(X, W, U, \boldsymbol{\lambda}, \boldsymbol{\alpha})], \tag{17}$$

$$\ln q_{t+1}(U) \propto \mathbb{E}_{q_t(W) q_t(\boldsymbol{\lambda}) q_t(\boldsymbol{\alpha})}[\ln p(X, W, U, \boldsymbol{\lambda}, \boldsymbol{\alpha})], \tag{18}$$

$$\ln q_{t+1}(\boldsymbol{\lambda}) \propto \mathbb{E}_{q_t(W) q_t(U) q_t(\boldsymbol{\alpha})}[\ln p(X, W, U, \boldsymbol{\lambda}, \boldsymbol{\alpha})], \tag{19}$$

$$\ln q_{t+1}(\boldsymbol{\alpha}) \propto \mathbb{E}_{q_t(W) q_t(U) q_t(\boldsymbol{\lambda})}[\ln p(X, W, U, \boldsymbol{\lambda}, \boldsymbol{\alpha})]. \tag{20}$$

In the following, we describe the concrete update equations of $q(W)$, $q(U)$, $q(\boldsymbol{\lambda})$, and $q(\boldsymbol{\alpha})$.

A. Update of $q(W)$

We denote by $\boldsymbol{w}_{j\cdot}$ the j-th row of $W \in \mathbb{R}^{d_x \times d_u}$. From (17), the update equation of $q(W)$ is given by

$$q_{t+1}(W) = \prod_{j=1}^{d_x} \mathcal{N}(\boldsymbol{w}_{j\cdot} | \boldsymbol{b}_{j\cdot}^{(t)} A^{(t)}, A^{(t)}), \tag{21}$$

where

$$A^{(t)} = \left(\frac{1}{\sigma^2} \mathbb{E}_{q_t(U)}[UU^T] + \frac{1}{\tau^2} \mathbf{I} \right)^{-1} \tag{22}$$

and $\mathbf{I} \in R^{d_u \times d_u}$ is the identity metrix; $\boldsymbol{b}_{j\cdot}^{(t)}$ denotes the j-th row of

$$B^{(t)} = \frac{1}{\sigma^2} X (\mathbb{E}_{q_t(U)}[U])^T + \frac{1}{\tau^2} M \tag{23}$$

and $M \in R^{d_x \times d_u}$ is the matrix whose components are all μ.

B. Update of $q(U)$

From (18), the update equation of $q(U)$ is given by

$$q_{t+1}(U) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{u}_n | \boldsymbol{\mu}_n^{(t)}, \Sigma_n^{(t)}), \tag{24}$$

where

$$\boldsymbol{\mu}_n^{(t)} = \frac{1}{\sigma^2} \Sigma_n^{(t)} (\mathbb{E}_{q_t(W)}[W])^T \boldsymbol{x}_n \tag{25}$$

and

$$\Sigma_n^{(t)} = \left(\frac{1}{\sigma^2} \mathbb{E}_{q_t(W)}[W^T W] \right. \tag{26}$$

$$\left. + \mathrm{diag} \left(\mathbb{E}_{q_t(\lambda)} \left[\frac{1}{\lambda_1} \right], \ldots, \mathbb{E}_{q_t(\lambda)} \left[\frac{1}{\lambda_{d_u}} \right] \right) \right)^{-1}. \tag{27}$$

C. Update of $q(\boldsymbol{\lambda})$

From (19), the update equation of $q(\boldsymbol{\lambda})$ is given by

$$q_{t+1}(\boldsymbol{\lambda}) = \prod_{i=1}^{d_u} \mathrm{GIG}(\lambda_i \mid a_{\lambda_i}^{(t)}, b_{\lambda_i}^{(t)}, p_{\lambda_i}^{(t)}), \tag{28}$$

where $\mathrm{GIG}(x \mid a, b, p)$ denotes a generalized inverse Gaussian (GIG) distribution

$$\mathrm{GIG}(x \mid a, b, p) = \frac{(a/b)^{p/2}}{2K_p(\sqrt{ab})} x^{p-1} \exp\left\{ -\frac{1}{2}(ax + \frac{b}{x}) \right\} \tag{29}$$

and $K_p(\cdot)$ is a modified Bessel function of the second kind; $a_{\lambda_i}^{(t)}$, $b_{\lambda_i}^{(t)}$, and $p_{\lambda_i}^{(t)}$ are given by

$$a_{\lambda_i}^{(t)} = 2\mathbb{E}_{q_t(\alpha_i)} \left[\frac{1}{\alpha_i} \right], \tag{30}$$

$$b_{\lambda_i}^{(t)} = \mathbb{E}_{q_t(u_{in})} \left[\sum_{n=1}^{N} u_{in}^2 \right], \tag{31}$$

$$p_{\lambda_i}^{(t)} = 1 - \frac{N}{2}. \tag{32}$$

D. Update of $q(\boldsymbol{\alpha})$

From (20), the update equation of $q(\boldsymbol{\alpha})$ is given by

$$q_{t+1}(\boldsymbol{\alpha}) = \prod_{i=1}^{d_u} \mathrm{GIG}(\alpha_i \mid 2b, 2\mathbb{E}_{q_t(\lambda_i)}[\lambda_i], a-1). \tag{33}$$

V. Experiments

To confirm the effectiveness of our proposed method, we evaluated the estimation accuracy of independent components for synthetic data. In the experiments, we set $a = 0.5$, $b = 0.01$, $N = 50$ and set the dimension of the observed data and independent components as the following three conditions:

- **Condition 1**: $d_x = 10$, $d_u = 10$,
- **Condition 2**: $d_x = 30$, $d_u = 30$,
- **Condition 3**: $d_x = 50$, $d_u = 50$.

In this experiment, independent components U^* were generated as follows: first, we generated parameters α_i $(i = 1, 2, \ldots, d_u)$ according to the gamma distribution $\mathrm{Gam}(\alpha_i; 0.5, 0.01)$. Next using these parameters α_i, we generated parameters λ_i $(i = 1, 2, \ldots, d_u)$ according to the exponential distribution $\mathrm{Exp}(\lambda_i \mid \alpha_i)$. Then, using these parameters λ_i, we generated independent components u_{in} $(i = 1, 2, \ldots, d_u, n = 1, 2, \ldots, N)$ according to the Gaussian distribution $\mathcal{N}(u_{in} \mid 0, \lambda_i)$.

After setting independent components U^* as above, we generated each component w_{ij} of the linear transformation matrix W^* according to the Gaussian distribution $\mathcal{N}(w_{ij}|0, 1)$. Using U^* and W^*, we generated the observed data X according to (2). In the experiment, we generated 100 observed data $X_1, X_2, \ldots, X_{100}$ according to $X_1 = W^*U^* + \boldsymbol{\epsilon}_1, X_2 = W^*U^* + \boldsymbol{\epsilon}_2, \ldots, X_{100} = W^*U^* + \boldsymbol{\epsilon}_{100}$, where $\boldsymbol{\epsilon}_i$ $(i = 1, 2, \ldots, 100)$ were generated independently according to the standard Gaussian distribution.

Given the observed data, we repeated our proposed algorithm until it converges. Then, we set the mean of $q(U)$ (it is denoted as $\hat{U}$) as the estimate of the true independent components U^*. We measured the estimation accuracy by using the mean squared error (MSE) $\frac{1}{Nd_u}||U^* - \hat{U}||_F^2$, where the notation $||\cdot||_F$ denotes the Frobenius norm. In the experiment, we derived the estimates $\hat{U}_1, \hat{U}_2, \ldots, \hat{U}_{100}$ from the observed data $X_1, X_2, \ldots, X_{100}$. Then, we calculated the average of the 100 MSE $\frac{1}{Nd_u}||U^* - \hat{U}_1||_F^2, \frac{1}{Nd_u}||U^* - \hat{U}_2||_F^2, \ldots, \frac{1}{Nd_u}||U^* - \hat{U}_{100}||_F^2$.

For comparison, we also calculated the MSE for the FastICA algorithm [6], which is one of the major algorithms for estimating independent components.

Table 1 shows the MSE of the previous method (FastICA) and our proposed method.

VI. Discussion

Compared with the FastICA algorithm, the MSE of the proposed method is smaller in all conditions. In this experiment, independent components were generated by the prior distribution that the proposed method assumes. On the other hand, the FastICA algorithm does not explicitly assume a

TABLE I
MSE OF THE PREVIOUS METHOD AND THE PROPOSED METHOD

	Proposed method	FastICA
Condition 1	4.85×10^{-8}	2.57×10^{-4}
Condition 2	5.01×10^{-7}	2.03×10^{-4}
Condition 3	3.53×10^{-7}	2.02×10^{-4}

nongaussian distribution on independent components. This is one of the reasons why the MSE of the proposed method is smaller than the FastICA algorithm.

VII. CONCLUSION

We have discussed the problem of ICA in which the hierarchical prior distribution on independent components is assumed. Due to this hierarchical prior distribution, the approximate posterior distributions can be calculated by using the variational Bayes method. Experiments results showed the effectiveness of our proposed method.

ACKNOWLEDGMENT

This work was supported in part by JSPS KAKENHI Grant Numbers JP16K00195, JP16K00417, JP17K00316, and JP17K06446.

REFERENCES

[1] D. F. Andrews and C. L. Mallows, "Scale mixtures of normal distributions," Journal of the Royal Statistical Society. Series B (Methodological), vol.36, no.1,pp.99–102, 1974.
[2] S. D. Babacan, S. Nakajima, and M. N. Do, "Bayesian group-sparse modeling and variational inference," IEEE Transactions on Signal Processing, vol. 62, no.11, pp.2906–2921, 2014.
[3] C. M. Bishop, Pattern recognition and machine learning, Springer, 2006.
[4] M. AT. Figueiredo, "Adaptive sparseness for supervised learning," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no.9, pp.1150–1159, 2003.
[5] A. Hyvärinen and K. Raju, "Imposing sparsity on the mixing matrix in independent component analysis," Neurocomputing, vol. 49, pp.151–162, 2002.
[6] A. Hyvärinen, J. Karhunen, and E. Oja, Independent component analysis, John Wiley & Sons, 2001.
[7] M. Kyung, J. Gill, M. Ghosh, and G. Casella, "Penalized regression, standard errors, and Bayesian lassos," Bayesian Analysis, vol. 5, no.2, pp.369–411, 2010.
[8] N. D. Lawrence and C. M. Bishop, "Variational Bayesian Independent Component Analysis," Tech. Rep., Computer Laboratory, University of Cambridge, 2000.
[9] T. Park and G. Casella, "The Bayesian lasso," Journal of the American Statistical Association, vol. 103, no.482, pp.681–686, 2008.
[10] S. Roberts and R. Choudrey, "Bayesian independent component analysis with prior constraints: an application in biosignal analysis, " Machine Learning Workshop, pp. 159–179, 2005.
[11] E. Roussos, S. Roberts, and I. Daubechies, "Variational Bayesian learning for wavelet independent component analysis,"Bayesian Inference and Maximum Entropy Methods in Science and Engineering, American Institute of Physics, pp.274–281, 2005.
[12] O. Winther and K. B. Petersen, "Bayesian independent component analysis: variational methods and non-negative decompositions," Digital Signal Processing, vol. 17, no. 5, pp. 858–872, 2007.
[13] L. Yang, J. Fang, H. Cheng, and H. Li, "Sparse Bayesian dictionary learning with a gaussian hierarchical model," Signal Processing, vol. 130, pp. 93–104, 2017.