Vol.2014-ICS-176 No.2
2014/7/22

Bayesian multiple and co-clustering methods: Application to fMRI data

Tomoki Tokuda[1,a)] Junichiro Yoshimoto[1] Yu Shimizu[1] Kosuke Yoshida[2,1] Shigeru Toki[3]
Go Okada[3] Masahiro Takamura[3] Tetsuya Yamamoto[3] Shinpei Yoshimura[4]
Yasumasa Okamoto[3] Shigeto Yamawaki[3] Noriaki Yahata[5] Kenji Doya[1]

Abstract: We propose a novel approach for the dimension reduction of high dimensional data to make the data available for conventional statistical evaluations. Our method is based on nonparametric multiple Gaussian clustering, in which we assume that in each cluster block, the instances follow an independent and identically (i.i.d.) univariate Gaussian distribution. We show theoretically that our model can fit multivariate Gaussian distributions with exchangeable features. We further show how the clusters derived with this specific model can be used to effectively reduce the dimension of data taking into account associations between attributes. Finally, we demonstrate our approach in an application to resting state functional magnetic resonance imaging (fMRI) data , which implies subtypes of depression may be characterized by the treatment effect of antidepressant drug SSRI.

1. Introduction

With the advent of sophisticated data acquisition technologies and associated research questions, high-dimensional data are widely available in scientific research. In the field of neuroscience, high-dimensional brain imaging data such as magnetic resonance imaging (MRI) and positron emission tomography (PET) are now increasingly used for a better understanding of the brain. In these data, typically, voxel[*1]-wise analyses such as the two-sample *t*-test are used to find associations between phenotypes of subjects (e.g., major depressive disorder) and brain areas (see e.g. [10]). Such univariate approaches might overlook sets of voxels whose co-activation pattern differs between two phenotypes of subjects. However, application of multivariate analyses is challenging to such data, since the number of features (voxels) is usually much larger than the number of samples (subjects).

One possible solution for this problem is to reduce the dimension of features so that the number of features is smaller than that of the samples: In biology, typically, a factor analytical approach such as Principal Component Analysis (PCA) is applied [7]. However, interpretation of the derived principal components is difficult [13]. Moreover, since PCA focuses only on the variability of the data, there is a potential for the true data structure (such as a cluster structure) to become distorted [6].

In this report, we consider a novel approach of dimension reduction by means of co-clustering and multiple-clustering (more general class of co-clustering), which are an extension of nonparametric Gaussian models for feature space. Co-clustering is a specific class of clustering which simultaneously models partitions in both subjects and features, with the assumption of a subject-cluster (clusters of subjects) structure common to all feature-clusters (clusters of features) [7]. On the other hand, multiple clustering relaxes this assumption, allowing for the subject-clusters within one feature-cluster to be independent from subject-clusters in other feature clusters (see Figure 1 for these differences). Specifically, in our approach, we consider a model that assumes that the entries in each cluster block follow an i.i.d. univariate Gaussian distribution. This is a special case of the clustering method laid out in [4].

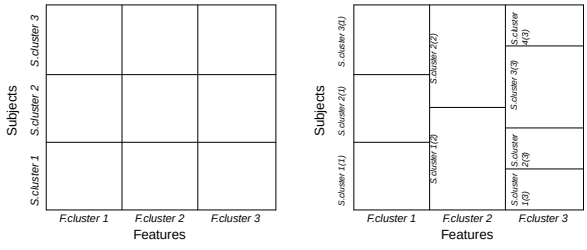

Fig. 1 *Illustration of co-clustering and multiple clustering for subject-times-feature data (the left and right panels, respectively). S.cluster and F.cluster denote subject- and feature-cluster, respectively. In the right panel, subject-clusters are dependent on the corresponding feature-cluster (the digit in parenthesis denotes the feature-cluster number); subjects are arranged differently in each feature-cluster.*

In both models, features with specific similar distribution patterns are allocated to the same feature-cluster. Further, using the

1 Okinawa Institute of Science and Technology Graduate University, Japan
2 Kyoto University, Japan
3 Hiroshima University, Japan
4 Otemon Gakuin University, Japan
5 Tokyo University, Japan
a) tomoki.tokuda@oist.jp

*1 voxel, a volume element in three dimensional space

derived feature- and subject-clusters in this model, we characterize each subject with the mean value of a Gaussian distribution fitted to the instances in each cluster block. Through this process of characterization of the data, the feature dimension is largely reduced, making it possible to apply a number of conventional statistical methods. In this report, we show that for brain imaging data, our method of dimension reduction based on the co-clustering model allows for easy statistical evaluation of brain activation, originally measured for thousands of voxels. As a consequence, we identify subgroups of depressive subjects, which are characterized by treatment effect of the antidepressant drug SSRI (Selective Serotonin Reuptake Inhibitors).

2. Clustering model

As has been seen in Figure 1, the co-clustering model can be considered as a special case of the multiple-clustering. Hence, for model description, we focus mainly on the multiple clustering model, for which several variants have so far been proposed (see review in [7]). Among these, we focus on the nonparametric Bayesian model for Gaussian distributions [4]. Specifically, our model assumes that univariate Gaussian distribution fits each cluster block, while the model in [4] considers general covariance structure. In a nutshell, our model is much simpler than that of [4]. However, in case of high-dimensional case such as fMRI data, our model has a computational advantage.

2.1 Model

For i.i.d. d-dimensional random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, we consider a $d \times V$ feature-partition matrix $\boldsymbol{Y}$ in which $Y_{j,v} = 1$ if the jth feature belongs to the vth feature-cluster. Similarly, we consider a $V \times n \times K$ subject-partition (3rd-order) tensor $\boldsymbol{Z}$ in which $Z_{v,i,k} = 1$ if the ith subject belongs to the kth subject-cluster in the vth feature-cluster (V and K are sufficiently large). Further, we denote the parameters for each cluster block (v, k) as $\boldsymbol{\theta}_{v,k}$. We assume that the observation vector in a cluster block follows an i.i.d. Gaussian distribution. Hence, $\boldsymbol{\theta}_{v,k}$ can be decomposed into a mean vector $\boldsymbol{\mu}_{v,k}$ and a covariance matrix $\boldsymbol{\Sigma}_{v,k}$. With this notation, the conditional probability of $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^t$ on $\boldsymbol{Y}$, $\boldsymbol{Z}$, and $\boldsymbol{\theta}_{v,k}$ is given by

$$P(\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\Theta}) = \prod_{v=1}^{V} \prod_{i=1}^{n} \mathrm{Gauss}(\boldsymbol{\Xi}_{v,i}|\boldsymbol{\mu}_{v,k(v,i)}, \boldsymbol{\Sigma}_{v,k(v,i)}), \qquad (1)$$

where $\boldsymbol{\Theta}$ denotes $\{\boldsymbol{\theta}_{v,k}\}$; $\boldsymbol{\Xi}_{v,i}$ is the vector composed of $X_{i,j}$ ($j = 1, \ldots, d$) such that $Y_{j,v} = 1$; $\mathrm{Gauss}(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$; $k(v, i)$ the subject-cluster index in the vth feature-cluster to which the ith subject belongs (i.e., $Z_{v,i,k} = 1$).

To estimate the number of feature-clusters and the number of subject-clusters (conditional on a particular feature-cluster), we further assume the following nonparametric prior distributions for $\boldsymbol{Y}$ and $\boldsymbol{Z}$ (via Dirichlet process).

$$\begin{aligned}
w_s &\sim \mathrm{Beta}(\cdot|1, \alpha),\ s = 1, 2, \ldots \\
\pi_v &= w_v \prod_{s=1}^{v-1} (1 - w_s),\ v = 1, 2, \ldots \\
\boldsymbol{Y}_j &\sim \mathrm{Mul}(\cdot|\boldsymbol{\pi}) \\
u_{l,v} &\sim \mathrm{Beta}(\cdot|1, \beta),\ l = 1, 2, \ldots,\ v = 1, 2, \ldots \\
\eta_{k,v} &= u_{k,v} \prod_{l=1}^{k-1} (1 - u_{l,v}),\ v = 1, 2, \ldots,\ k = 1, 2, \ldots \\
\boldsymbol{Z}_{v,i} &\sim \mathrm{Mul}(\cdot|\boldsymbol{\eta}_v),
\end{aligned}$$

where $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots)$, $\boldsymbol{\eta}_v = (\eta_{1,v}, \eta_{2,v}, \ldots)$, $\boldsymbol{Y}_j = (Y_{j,1}, Y_{j,2}, \ldots)$, and $\boldsymbol{Z}_{v,i} = (Z_{v,i,1}, Z_{v,i,2}, \ldots)$. $\mathrm{Beta}(\cdot|a, b)$ is a Beta distribution with prior sample sizes (a, b). $\mathrm{Mul}(\cdot|\boldsymbol{\eta})$ is a multinomial distribution of one sample size and probability $\boldsymbol{\eta}$. We set the hyperparameters α and β (concentration parameters in Dirichlet process) to 1.

For our model, we consider a more restrictive structure for $\boldsymbol{\mu}_{v,k}$ and $\boldsymbol{\Sigma}_{v,k}$ than originally defined in [4]: We assume that $\boldsymbol{\mu}_{v,k}$ is a multiple of the unit vector $\mathbf{1}$ and that $\boldsymbol{\Sigma}_{v,k}$ is a multiple of the identity matrix $\boldsymbol{I}$. With these restrictions, the conjugate prior distributions for the Gaussian parameters can be reduced to a univariate distribution as follows:

$$\begin{aligned}
\boldsymbol{\Sigma}_{v,k} &= \sigma_{v,k}^2 \boldsymbol{I} \\
\sigma_{v,k}^{-2} &\sim \mathrm{Ga}(\cdot|\sigma_0^{-2}/2, \gamma_0 \sigma_0^{-2}/2) \\
\boldsymbol{\mu}_{v,k} &= \lambda_{v,k} \mathbf{1} \\
\lambda_{v,k} &\sim \mathrm{Gauss}(\cdot|\lambda_0, \gamma_1 \sigma_{v,k}^2),
\end{aligned}$$

where $\mathrm{Ga}(\cdot|a, b)$ denotes the Gamma distribution with shape and rate parameters (a, b). In the present paper, we set $\sigma_0^{-2} = 10^{-4}$, $\gamma_0 = 1$, $\gamma_1 = 10^4$, and $\lambda_0 = 10^{-4}$ so that the prior distributions are nearly non-informative (though they may be set arbitrarily).

For the co-clustering model, we restrict that subject-cluster is common to all feature-clusters. Thus, the model in (1) is simply replaced by the following equation:

$$P(\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\Theta}) = \prod_{v=1}^{V} \prod_{i=1}^{n} \mathrm{Gauss}(\boldsymbol{\Xi}_{v,i}|\boldsymbol{\mu}_{v,k(i)}, \boldsymbol{\Sigma}_{v,k(i)}),$$

where $\boldsymbol{Z}$ is a $n \times K$ matrix in which $Z_{i,k} = 1$ if the ith subject belongs to the kth subject-cluster. The remainders of equations regarding its priors are changed accordingly.

Since the estimation of these parameters from the data in closed form is intractable, we rely on variational inferences. The variational method allows for approximation of the maximum a posteriori (MAP) estimate for each parameter in an iterative way. Finally, we estimate the posterior membership probabilities of each feature and subject (i.e., in terms of $\boldsymbol{Y}$ and $\boldsymbol{Z}$), which leads to a multiple clustering solution (see [4] for details).

2.2 Factor-analytical representation of model

Although our multiple clustering model is simply built based on an ensemble of univariate Gaussian distributions, it has some interesting properties for its model representation. As a general property of mixture models, our model can flexibly fit different types of underlying generative distributions of the data. In particular, in the present section, we explore factor-analytical representation of the model. To clarify, we focus on a single vector

$\Xi_{v,i}$ for a specific pair of (v, i) and for simplicity denote it as $\Xi = (\Xi_1, \ldots, \Xi_{d'})$ (i.e., we focus on a particular feature-cluster, while considiering mixture structures in subjects). Hewitt-Savage's theorem (generalization of the deFinetti theorem [3]) states that a finite set of infinite random sequences, $\Xi_1, \ldots, \Xi_{d'}$, is exchangeable, if and only if

$$P(\Xi) = P(\Xi_1, \ldots, \Xi_{d'}) = \int \prod_{j=1}^{d'} P(\Xi_j|\phi) m(\phi),$$

where m is a probability measure of ϕ on a σ-field [5]. Our model includes a special case of this theorem where $P(\cdot|\phi) = \text{Gauss}(\cdot|\phi, (1-\rho)c)$ and $\phi \sim \text{Gauss}(\lambda_0, \rho c)$ with $0 \leq \rho < 1$ and $c > 0$, resulting in

$$P(\Xi) = \text{Gauss}(\Xi|\lambda_0 \mathbf{1}, c(\rho \mathbf{1}\mathbf{1}^T + (1-\rho)\boldsymbol{I})). \quad (2)$$

This form is a typical instance when $\Xi_1, \ldots, \Xi_{d'}$ are exchangeable, implying that our model fits data such that variables within the same feature-cluster are exchangeable Gaussian. Furthermore, the combination of Eqs. (2) and (1) implies that $\boldsymbol{X}_i$ follows a multivariate Gaussian distribution of [*2]

$$\begin{aligned} P(\boldsymbol{X}_i|\boldsymbol{Y}) &= \prod_{v=1}^{V} \text{Gauss}(\Xi_v|\lambda_{0,v}\mathbf{1}, c_v(\rho_v \mathbf{1}\mathbf{1}^T + (1-\rho_v)\boldsymbol{I})) \\ &= \text{Gauss}(\boldsymbol{X}_i|\boldsymbol{\mu}, \boldsymbol{W}\boldsymbol{W}^T + \boldsymbol{\Psi}), \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{\mu} &= (\lambda_{0,1}\mathbf{1}^T, \ldots, \lambda_{0,V}\mathbf{1}^T)^T \\ \boldsymbol{W} &= \begin{pmatrix} \sqrt{c_1\rho_1}\mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sqrt{c_V\rho_V}\mathbf{1} \end{pmatrix} \\ \boldsymbol{\Psi} &= \begin{pmatrix} c_1(1-\rho_1)\boldsymbol{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & c_V(1-\rho_V)\boldsymbol{I} \end{pmatrix}, \end{aligned} \quad (3)$$

where $\boldsymbol{\mu}$ is a $d \times 1$ vector; $\boldsymbol{W}$ a $d \times V$ matrix; $\boldsymbol{\Psi}$ a $d \times d$ matrix.

Equivalently, $\boldsymbol{X}_i$, conditional on the feature-cluster matrix $\boldsymbol{Y}$, is given using V-dimensional latent variables $\boldsymbol{Z}_i$:

$$\boldsymbol{X}_i|\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{Z}_i + \boldsymbol{\mu} + \boldsymbol{\epsilon}_i,$$

where $\boldsymbol{Z}_i \sim \text{Gauss}(\cdot|\mathbf{0}, \boldsymbol{I})$ and $\boldsymbol{\epsilon}_i \sim \text{Gauss}(\cdot|\mathbf{0}, \boldsymbol{\Psi})$.

This model has the following restrictions: 1) features within the same feature-cluster share the same mean; 2) the covariance of features within the same feature-cluster is represented as $c(\rho\mathbf{1}\mathbf{1}^t + (1-\rho)\boldsymbol{I})$; 3) features belonging to different feature-clusters are uncorrelated. Restrictions 2) and 3) imply that this model has the effect of a factor analyzer similar to the probabilistic PCA [8] which is modeled as (using our notation)

$$P(\boldsymbol{X}_i) = \text{Gauss}(\boldsymbol{X}_i|\boldsymbol{\mu}, \boldsymbol{W}\boldsymbol{W}^T + \boldsymbol{\Psi}),$$

where $\boldsymbol{\Psi} = \sigma^2\boldsymbol{I}$.

More specifically, our model has the effect of sparse PCA because of zero elements in the off-diagonal blocks of $\boldsymbol{W}$ in Eqs. (3). In particular, our model makes $\boldsymbol{W}$ sparse in such a way that a single latent variable represents a specific set of features.

[*2] Note that the suffix v denotes that the corresponding parameters are specific to the vth feature-cluster; we sort features according to their membership of the feature-clusters, but the order of features within the same feature-cluster is arbitrary.

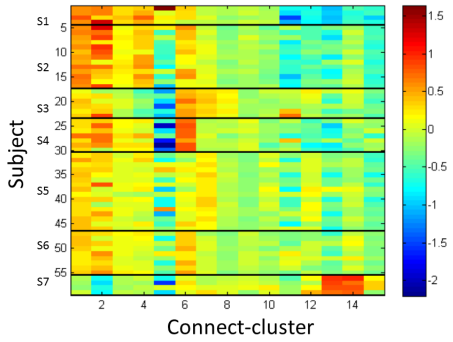

Fig. 2 *Heatmap of the reduced data. Subject-clusters are sorted by the proportion of remission (that is based on HRSD scores); the lines in the heatmap show boundaries of the subject-clusters ; the subject-clusters are indexed as $S1, \ldots, S7$; connect-clusters are sorted by their correlations with the proportion of remission.*

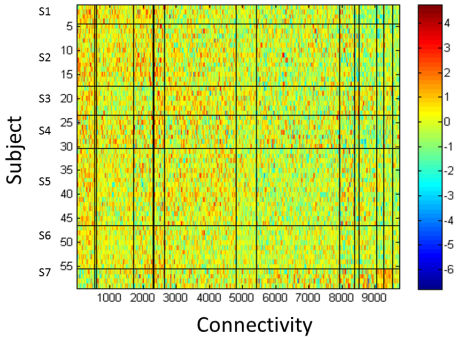

Fig. 3 *Heatmap of the original data. The subject- and connectivity-clusters are sorted as in Figure 2.*

2.3 Dimension reduction: characterization of feature-cluster

Making use of the exchangeability, we can characterize a subject by assigning the mean value of the corresponding cluster univariate Gaussian distribution, mapping the $n \times d$ matrix $\boldsymbol{X}$ to a $n \times V$ matrix $\boldsymbol{S}$,

$$f: X_{i,j} \to S_{i,v(j)} \text{ with } S_{i,v(j)} = \lambda_{v(j),k(v(j),i)}, \quad (4)$$

where $v(j)$ denotes the index of the feature-cluster to which the jth feature belongs. Alternatively, we can characterize a subject by assigning the feature-cluster entry mean of this subject:

$$f: X_{i,j} \to S_{i,v(j)} \text{ with } S_{i,v(j)} = \sum_{j' \in v(j)} X_{i,j'}/|v(j)|, \quad (5)$$

where $|v(j)|$ is the cluster size of the feature-cluster $v(j)$. Unlike Eqs. (4), the dimension reduction method in Eqs. (5) retains the subject variability in each feature-cluster. These mappings can substantially reduce data size and enable further statistical analysis. We see an example in the following section.

3. Application to real data

In this section, by means of the co-clustering method, we analyze resting state functional MRI (rsfMRI) data, which has recently become a common source of investigation in the study of brain activity related to mental disorders. Our objective is to identify subtypes of depressive subjects in an unsupervised manner.

The sample size of our data is 125, in which 66 were obtained

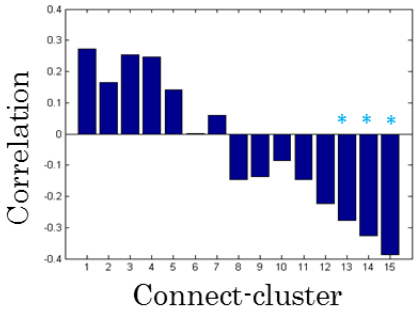

Fig. 4 *Correlations between HRSD scores after 6 weeks of the dosage of SSRI, and mean values of each connect-cluster; An asterisk * denotes that the corresponding correlation is significant at level of* $\alpha = 0.05$

from healthy subjects (control) and 59 from major depressive disorder subjects (depression). For the depressive subjects, the brain scan of rsfMRI was taken before the start of the treatment by the antidepressant drug SSRI (Selective Serotonin Reuptake Inhibitors). Features of this dataset are composed by 9730 functional connectivity (FC) between 140 pre-specified brain areas. As pre-processing of data, we standardized each feature (FC) using the mean and the standard deviation of the control subjects, which yielded the data of the depressive subjects with size 59×9739.

First, we applied the co-clustering method to this data. Due to computational costs[*3] and later inference difficulties, we restricted the maximum number of feature-clusters to 15 (in our context, we call them connect-clusters). The results of dimension reduction by means of the co-clustering method are displayed in Figure 2 (compare these results with those of the original data displayed in Figure 3). As can be seen in Figure 2, the co-clustering method reveals possible cluster-structures both in subjects and features.

Second, as a next step, we further analyzed the reduced data to characterize the subject-clusters by the treatment effect of antidepressant drug SSRI. We evaluated the treatment effect of the drug on depression in terms of HRSD (Hamilton Rating Scale for Depression) score after six weeks of the dosage of SSRI. HRSD score is based on dozens clinical questions to measure severity of depression: the larger the score is, the more severe the depressive disorder. We evaluated correlations between HRSD scores and the mean values of each connect-cluster. The results are displayed in Figure 4 in which significant correlations are shown with an asterisk. As can be seen in Figure 4, the correlations for connect-clusters 13, 14 and 15 are significant at level $\alpha = 0.05$. This suggests that the subject-clusters of depression may be characterized by the SSRI treatment effect and that the relevant brain regions for the treatment effect may be included in connect-clusters 13, 14, and 15. For these connect-clusters, the most relevant brain areas in terms of frequencies of connectivity to the other brain areas are identified as follows: ParaCENT.Lob.R, HIPP.L, ParaCENT.Lob.L, and PtCENT.R for connect-cluster 13; TEMP.Mid.R 20, and TEMP.Inf.R for connect-cluster 14; ParaCENT.Lob.R, and FRNT.Med.Orb.R for connect-cluster 15.

[*3] The present simulation took five hours using two Intel Xeion E5649 CPUs with 6 cores each and a total of 48 GB main memory.

4. Conclusion

In this report, we introduced a specific class of multiple clustering model based on nonparametric Bayesian mixture models which assumes conditional independence of features within a cluster block, while considering a co-clustering model as a special case of the multiple clustering model. Using this model, we proposed a method to reduce data dimension without distorting correlated feature structures.

Theoretically, it was shown that this model can have an effect of (sparse) factor analyzer, capturing a specific type of correlations between features. In neurological data, we can assume that this type of co-activation of neurons is often the case, but they are overlooked or not modeled by other clustering methods. Moreover, our model evaluates partitions of samples and attributes simultaneously, which enables effective fitting of a given multiple clustering structure.

In an application to brain imaging data, our model provided a means to find subtypes of depressive subjects and simultaneously reduce data dimensions. Using the reduced dimension, it was implied that some brain regions are relevant to the treatment effect of antidepressant drug SSRI.

Acknowledgment

This study is the results of 'Integrated Research on Neuropsychiatric Disorders' and 'Brain Machine Interface Development' carried out under the Strategic Research Program for Brain Sciences by the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

[1] Attias, H.: Indeendent Factor Analysis, *Neural Computation II* (1999).
[2] Chan, K., Lee, T.-W. and Sejnowski, T. J.: Variational Bayesian learning of ICA with missing data, *Neural Computation*, Vol. 15, No. 8, pp. 1991–2011 (2003).
[3] de Finetti, B.: Funzione caratteristica di un fenomeno aleatorio, *Atti della R. Academia Nazionale dei Lincei, Serie 6.Memorie, Classe di Scienze Fisiche, Mathematice e Naturale*, Vol. 4, pp. 251–299 (1931).
[4] Guan, Y., Dy, J., Niu, D. and Ghahramani, Z.: Variational inference for nonparametric multiple clustering, *MultiClust: 1st International Workshop on Discovering, Summarizing and Using Multiple Clustering held at the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2012).
[5] Hewitt, E. and Savage, L.: Symmetric measures on Cartesian products, *Transactions of the American Mathematical Society*, Vol. 80, pp. 470–501 (1955).
[6] Law, M., Figueiredo, M. and Jain, A.: Simultaneous feature selection and clustering using mixture models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 26, No. 9, pp. 1154–1166 (2004).
[7] Murphy, K.: *Machine Learning: A Probabilistic Perspective*, Cambridge, Massachusetts: MIT Press (2012).
[8] Tipping, M. and Bishop, C.: Probabilistic principal component analysis, *Journal of the Royal Statistical Society Series B*, Vol. 61, No. 3, pp. 611–622 (2002).
[9] Xiao, Y., Zhu, Z., Wei, S. and Zhao, Y.: Discriminative ICA Model with Reconstruction Constraint for Image Classification, *MMf12, October 29?November 2, 2012, Nara* (2012).
[10] Yao, Z., Wang, L., Lu, Q., Liu, H. and Teng, G.: Regional homogeneity in depression and its relationship with separate depressive symptom clusters:a resting fMRI study, *Journal of Affective Disorders*, Vol. 115, No. 3, pp. 430–438 (2009).
[11] Yoshida, R., Higuchi, T. and Imoto, S.: A mixed factor model for dimension reduction and extraction of a group structure in gene expression data, *Proceedings of the 2004 IEEE Computational Systems* (2004).

[12] Zhang, K. and Chan, L.-W.: ICA with sparse connections, *Intelligent Data Engineering and Automated Learning–IDEAL 2006*, Springer, pp. 530–537 (2006).
[13] Zou, H., Hastie, T. and Tibshirani, R.: Sparse principal component analysis, *Journal of Computational and Graphical Statistics*, Vol. 15, No. 2, pp. 265–286 (2006).