A Mixed Factors Model for Dimension Reduction and Extraction of a Group Structure in Gene Expression Data

Ryo Yoshida
The Graduate University for Advanced Studies,
4-6-7 Minami-Azabu, Minato-ku, Tokyo, 103-8569, Japan
yoshidar@ism.ac.jp

Tomoyuki Higuchi
Institute of Statistical Mathematics
4-6-7 Minami-Azabu, Minato-ku, Tokyo, 103-8569, Japan
higuchi@ism.ac.jp

Seiya Imoto
Human Genome Center, Institute of Medical Science, University of Tokyo,
4-6-1 Shirokanedai, Minato-ku, Tokyo, 108-8639, Japan
imoto@ims.u-tokyo.ac.jp

Abstract

When we cluster tissue samples on the basis of genes, the number of observations to be grouped is much smaller than the dimension of feature vector. In such a case, the applicability of conventional model-based clustering is limited since the high dimensionality of feature vector leads to overfitting during the density estimation process. To overcome such difficulty, we attempt a methodological extension of the factor analysis. Our approach enables us not only to prevent from the occurrence of overfitting, but also to handle the issues of clustering, data compression and extracting a set of genes to be relevant to explain the group structure. The potential usefulness are demonstrated with the application to the leukemia dataset.

1. Introduction

Cluster analysis of microarray gene expression data plays an important role in the automated search and the validation for the various classes in either tissue samples [2, 5, 7] and genes [4, 19]. A distinction of microarray dataset is that the number of tissue samples is much smaller than the number of genes. Our work in this study focuses on clustering of tissue samples.

For this purpose, one of the widely used techniques has been hierarchical clustering. Although the method has con-tributed to find distinct subtypes of disease [2, 5], there is weakness as to systematic guidance for solving practical questions, e.g. what genes are relevant to explain the group structure and, how many clusters there are. Furthermore, it is not evident for this approach to classify a set of future samples since the decision boundary is not explicit.

The model-based clustering using the finite mixture model is a well-established technique for finding groups in multivariate dataset [10, 11, 17]. Most commonly, the attention has focused on the use of Gaussian mixture because of the computational convenience. However, in some microarray experiments, the applicability of this approach is limited. That is, when we wish to cluster tissue samples on the basis of genes, then the sample size is much smaller than the dimension of feature vector. In such a case, the finite mixture models often lead to overfitting during the density estimation process. Therefore, we need to reduce the dimensionality of data before proceeding to cluster analysis.

The principal component analysis (PCA, [3]) is a commonly used method for reducing the dimensionality of microarrays [11, 12]. In spite of its usefulness, PCA is not justified in clustering context since the projections of data corresponding to the dominant eigenvalues do not necessary reflect the presence of groups in dataset. Most such limitations are related to the fact that PCA only takes into consideration the second-order characteristic of data. Some authors gave the illustrations that PCA fails to reveal underlying groups [6, 17].

This article attempts a methodological extension of the factor analysis. In our model, so referred to as the *mixed factors model*, the factor variable plays a role to give a parsimonious description of clusters in the feature space. The model presents a parsimonious parameterization of Gaussian mixture. Consequently, our approach enables us to prevent from the occurrence of overfitting in the density estimation even when the dimension of data is about several thousand. The application of the mixed factors analysis covers the issues of clustering and dimension reduction. The reduced-dimensionality representation of original data is constructed as to be plausible estimate of signal revealing the existence of group structure. The proposed method also can extract genes to be relevant to explain the group structure. In this process, sets of genes that are functionally co-worked are automatically detected.

The mixture of factor analyzers (MFA, [16, 17]), which is an extension of the mixture of probabilistic principal components analysis (MPPCA, [21]), is closely related to our model. Our mixed factors model is distinguished from MFA and MPPCA in terms of the parameterization of Gaussian mixture. While the number of free parameters of MFA and MPPCA grows quickly as the number of clusters tending to large, our approach can mitigate such difficulty. Some researchers might be motivated by grouping the tissue samples into the large number of clusters across several thousand genes. In such situations the superiority of our method will be apparent.

The rest of this article is organized as follows. In Section 2, we will present the mixed factors model and outline some materials used in the following sections. Our model will be also discussed in relation to MFA and MPPCA. Section 3 covers the EM algorithm for the maximum likelihood estimation of our model. Section 4 will express the procedures of clustering, data compression, visualization and selection of genes to be relevant to explain the presence of groups. In Section 5, the potential usefulness of our approach will be demonstrated with the application to a well-known dataset, the leukemia data [12]. Finally, the concluding remarks are given in Section 6.

2. Mixed Factors Model

2.1. Probability Model

Let x be an observed variable distributed over $\mathbf{R}^d$. When we cluster the tissue samples across the genes, the dimensional of feature vector, d, corresponds to the number of genes which is typically ranging from 10^2 to 10^4. The basic idea underlying factor analysis [3] is to relate x to the factor variable $f \in \mathbf{R}^q$;

$$x = \Xi f + \epsilon. \tag{1}$$

Here $q < d$, and the ϵ is an observational noise to be Gaussian, $\epsilon \sim N(\mathbf{0}, \Lambda)$, with $\Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_d)$. The matrix of order $d \times q$, Ξ, contains the factor loadings and is referred to the *factor loading matrix*. The factor variable and the observational noise are conventionally assumed to be mutually independent random variables. In the following discussions, we assume that the noise covariance is to be isotropic, i.e. $\Lambda = \lambda I_d$, where I_d denotes d-dimensional identity matrix. Note that, for a given factor variable, the observed variable is distributed according to $x|f \sim N(\Xi f, \lambda I_d)$. A key motivation of usual factor analysis is that due to the diagonality of noise covariance matrix, all variables in $x = (x_1, \cdots, x_d)$ are conditionally independent for a given factor variable. Therefore, it can be considered as that the factor variable plays a role to give a parsimonious explanation for the dependencies in x.

A key idea of the proposed method is to describe clusters on $\mathbf{R}^d$ by using the factor variable. Suppose that the total population of x consists of G subpopulations, $\mathcal{P}_1, \cdots, \mathcal{P}_G$. Here we let $l^T = (l_1, \cdots, l_G)$ be a vector of unknown class labels to indicate the subpopulations by

$$l_g = \begin{cases} 1 & x \in \mathcal{P}_g \\ 0 & x \notin \mathcal{P}_g. \end{cases}$$

Consider that the l follows from multinomial distribution, $l \sim M_G(\alpha)$ with probabilities $\alpha^T = (\alpha_1, \cdots, \alpha_G)$, and that given $l_g = 1$, the factor variable is distributed to be Gaussian, $f|l_g = 1 \sim N(\mu_g, \Sigma_g)$. Then the unconditional distribution of f is given by the G-components Gaussian mixture with density

$$p(f) = \sum_{g=1}^{G} \alpha_g \phi(f; \mu_g, \Sigma_g). \tag{2}$$

Here the $\phi(f; \mu_g, \Sigma_g)$ denotes the normal density with mean μ_g and covariance matrix Σ_g. We refer to the observed system, composed of (1) and (2), the *mixed factors model*.

For this model, the complete data is taken to be $y^T = (x^T, f^T, l^T)$, where the missing variables correspond to f and l. The model of the complete data y is factorized as

$$p(y) = p(f|l)p(l)p(x|f), \tag{3}$$

where

$$
\begin{aligned}
p(x|f) &= \phi(x; \Xi f, \lambda I_d), \\
p(f|l) &= \prod_{g=1}^{G} \phi(f; \mu_g, \Sigma_g)^{l_g}, \\
p(l) &= \prod_{g=1}^{G} \alpha_g^{l_g}.
\end{aligned}
$$

Notice that $p(\boldsymbol{x}|\boldsymbol{f}, l) = p(\boldsymbol{x}|\boldsymbol{f})$ holds in the mixed factors model. This implies that given a factor variable, the class label has no effect to the conditional distribution of $\boldsymbol{x}$. The presence of clusters on $\mathbf{R}^d$ is completely explained by the factor variable.

Under this generative model, the observed variable is unconditionally distributed to be the G-components Gaussian mixture,

$$p(\boldsymbol{x}) = \sum_{g=1}^{G} \alpha_g \phi(\boldsymbol{x}; \boldsymbol{\Xi}\boldsymbol{\mu}_g, \boldsymbol{\Xi}\boldsymbol{\Sigma}_g\boldsymbol{\Xi}^T + \lambda\boldsymbol{I}_d). \qquad (4)$$

Thus, the feature variable is characterized by clusters centered at mean $\boldsymbol{\Xi}\boldsymbol{\mu}_g$ in which each group size is defined by the mixing proportion α_g, $g \in \{1 \cdots, G\}$. The covariance matrix formed in $\boldsymbol{\Xi}\boldsymbol{\Sigma}_g\boldsymbol{\Xi}^T + \lambda\boldsymbol{I}_d$ imposes a geometric feature on the gth cluster. For the Gaussian mixture with unrestricted covariance, there are $d(d+1)/2$ distinct parameters in each component covariance matrix. In cluster analysis of microarray data, the unrestricted covariance leads to overfitting in the density estimation process since the number of free parameters grows quickly as G tends to large. When we are in such situations, the mixed factors model gives a natural approach to the parsimonious parameterization for Gaussian mixture. The occurrence of overfitting can be avoided by choosing q as to be appropriate for a given dimensionality of data and the specified number of components.

2.2. Rotational Ambiguity and Parameter Restriction

The parameter vector $\boldsymbol{\theta}$ in the mixed factors model consists of all elements in $\boldsymbol{\Xi}$, λ and the component parameters α_g, $\boldsymbol{\mu}_g$, $\boldsymbol{\Sigma}_g$ for $g \in \{1, \cdots, G\}$. While our approach provides a parsimonious parameterization for Gaussian mixture, such modeling leads to the lack of identifiability of the parameter space.

Let $\boldsymbol{R}$ be any nonsingular matrix of order q. Note that each component mean and covariance in (4) are invariant under the manipulation by $\boldsymbol{R}$ as

$$\boldsymbol{\Xi}\boldsymbol{R}\boldsymbol{R}^{-1}\boldsymbol{\mu}_g, \quad \boldsymbol{\Xi}\boldsymbol{R}\boldsymbol{R}^{-1}\boldsymbol{\Sigma}_g\boldsymbol{R}^{-T}\boldsymbol{R}^T\boldsymbol{\Xi}^T + \lambda\boldsymbol{I}.$$

This will occur when we apply a nonsingular linear transformation so that $\boldsymbol{\Xi} \to \boldsymbol{\Xi}\boldsymbol{R}$, $\boldsymbol{f} \to \boldsymbol{R}^{-1}\boldsymbol{f}$. Therefore, there is an infinity of choices for $\boldsymbol{\mu}_g$, $\boldsymbol{\Sigma}_g$ and $\boldsymbol{\Xi}$

To avoid the nonidentifiability, we need to reduce the degree of freedom for the parameters by imposing q^2 constraints which corresponds to the order of $\boldsymbol{R}$. In this paper, we consider a natural approach as follows;

(a) $\boldsymbol{\Sigma}_g = \mathrm{diag}(\sigma_{1g} \cdots \sigma_{qg})$, $g \in \{1, \cdots, G\}$,

(b) $\boldsymbol{\Xi}^T\boldsymbol{\Xi} = \boldsymbol{I}_q$.

The conditions (a) and (b) impose $q(q-1)/2$ and $q(q+1)/2$

restrictions on the number of free parameters, respectively. Then, the total size of restrictions results in q^2 to be imposed. The condition (b) offers the orthonormality of q columns in the factor loading matrix, $\boldsymbol{\Xi} = (\boldsymbol{\xi}_1 \cdots \boldsymbol{\xi}_q)$, i.e. $\|\boldsymbol{\xi}_k\| = 1$, $k \in \{1, \cdots, q\}$, $\boldsymbol{\xi}_k^T\boldsymbol{\xi}_h = 0$ for all $h \neq k$.

Consequently, the mixed factors model has unknown parameters with the degree of freedom,

$$\mathcal{N}_{\boldsymbol{\theta}}(d, G) = (G-1) + dq + 1 + 2Gq - \left(\frac{q^2 + q}{2}\right). \quad (5)$$

Note that $\mathcal{N}_{\boldsymbol{\theta}}(d, G)$ grows with only proportional to q as d tends to large, and also that it grows with only proportional to $2q + 1$ as G tends to large. As will be remarked in next discussion, this property will be desirable in situations where researchers are motivated by grouping the tissue samples into the large number of clusters across genes.

2.3. Related Models

The mixture of factor analyzers (MFA, [16, 17]) presents the G-components Gaussian mixture as follows;

$$p(\boldsymbol{x}) = \sum_{g=1}^{G} \alpha_g \phi(\boldsymbol{x}; \boldsymbol{m}_g, \boldsymbol{W}_g\boldsymbol{W}_g^T + \boldsymbol{\Lambda}_g). \qquad (6)$$

Here the $\boldsymbol{W}_g$ is $d \times q$ matrix and $\boldsymbol{\Lambda}_g = \mathrm{diag}\{\lambda_{1g}, \cdots, \lambda_{qg}\}$. This is a generalization of the mixture of probabilistic PCA (MPPCA, [21]) which takes $\boldsymbol{\Lambda}_g$ to be isotropic, $\boldsymbol{\Lambda}_g = \lambda_g\boldsymbol{I}_d$. It was shown by [17, 21] that there implicitly exists a generative model behind this form of Gaussian mixture. Suppose that we have G submodels in which each gth model is given by

$$\boldsymbol{x} = \boldsymbol{m}_g + \boldsymbol{W}_g\boldsymbol{t} + \boldsymbol{\epsilon}_g,$$

for $g \in \{1, \cdots, G\}$. Here the q-dimensional latent variable $\boldsymbol{t}$ is defined to be Gaussian $N(\boldsymbol{0}, \boldsymbol{I}_q)$, and the noise model is $\boldsymbol{\epsilon}_g \sim N(\boldsymbol{0}, \boldsymbol{\Lambda}_g)$. Mixing these submodels with probability α_g $g \in \{1, \cdots, G\}$, then one can obtain the Gaussian mixture in the form of (6).

As well as our approach, MFA also presents a parsimonious parameterization of Gaussian mixture model. This model characterizes more flexible geometric feature of clusters than that of the mixed factors model. However, in cluster analysis to be faced in microarray studies, it may still suffer from the problem of overfitting. It follows from [17] that the number of free parameters of MPPCA is given by

$$\tilde{\mathcal{N}}_{\boldsymbol{\theta}}(d, G) = 2G - 1 + Gdq + Gd - G\left(\frac{q^2 - q}{2}\right).$$

Table 1 shows a comparison for the number of free parameters between the mixed factors model and MPPCA for varying G and d. For instance, consider that we have a

Mixed factors model ($q = 2$)

	$G = 2$	$G = 3$	$G = 4$	$G = 5$
$d = 50$	107	112	117	122
$d = 200$	407	412	417	422
$d = 1000$	2007	2012	2017	2022
$d = 6000$	12007	12012	12017	12022
$d = 20000$	40007	40012	40017	40022

MPPCA ($q = 2$)

	$G = 2$	$G = 3$	$G = 4$	$G = 5$
$d = 50$	301	452	603	754
$d = 200$	1201	1802	2403	3004
$d = 1000$	6001	9002	12003	15004
$d = 6000$	36001	54002	72003	90004
$d = 20000$	120001	180002	240003	300004

Table 1. Comparison for the number of free parameters ($q = 2$) between the mixed factors model (top) and the mixture of principal component analyzers (bottom), against some specified values of the number of groups G and the dimension of feature vector d

.

set of 1000-dimensional observations, and are motivated by grouping them into two clusters using 2-dimenional latent variable ($d = 1000$, $G = 2$, $q = 2$). This situation is considered to be typical in microarray studies. Then, the number of free parameters of MPPCA is $\tilde{\mathcal{N}}_{\boldsymbol{\theta}}(d, G) = 6001$, while our model gives $\mathcal{N}_{\boldsymbol{\theta}}(d, G) = 2007$. Further, if $d = 1000$, $G = 3$, $q = 2$, then, $\tilde{\mathcal{N}}_{\boldsymbol{\theta}}(d, G) = 9002$ and $\mathcal{N}_{\boldsymbol{\theta}}(d, G) = 2012$. While $\tilde{\mathcal{N}}_{\boldsymbol{\theta}}(d, G)$ becomes large quickly as G tends to large, our approach enables the increase in the number of parameters to be saved. In essence, the challenge faced by the biological scientists is to use the large-scaled dataset whereas the several organisms are composed of the large number of genes, e.g. the genome of *Saccharomyces cerevisiae* contains more than 6000 genes. In such situations, we can not expect some estimates obtained from MFA and MPPCA approaches to be reliable, and so the scope of their application is limited.

2.4. Posterior Distributions

An objective of the mixed factors analysis is to find the plausible values of $\boldsymbol{f}$ and l based on an observation $\boldsymbol{x}$. It is achieved by the suitable estimators to be close to the estimands on the average. Our analysis aims to reduce the dimensionality of data by estimating the factor variable and to divide the feature space, $\mathbf{R}^d$, by attributing the labels l for

all $\boldsymbol{x} \in \mathbf{R}^d$, simultaneously. Most common approach for estimating the latent random variables is based on Bayes' rule. We will revisit to these issues in the later sections. In this context, the posterior distributions of $\boldsymbol{f}$ and l play a key role in constructing the Bayes estimators. Hence, in sequel we will investigate the functional form of them.

Let $\boldsymbol{u}(\boldsymbol{x}) = \boldsymbol{\Xi}^T \boldsymbol{x}$ and $\tilde{\boldsymbol{\epsilon}} = \boldsymbol{\Xi}^T \boldsymbol{\epsilon}$, that is, the orthogonal transformations of $\boldsymbol{x}$ and $\boldsymbol{\epsilon}$ onto $\mathbf{R}^q$. Then, the generative model (1) can be replaced by

$$\boldsymbol{u}(\boldsymbol{x}) = \boldsymbol{f} + \tilde{\boldsymbol{\epsilon}}, \tag{7}$$

where $\tilde{\boldsymbol{\epsilon}} \sim N(\boldsymbol{0}, \lambda \boldsymbol{I}_q)$. Given this formulation, the q-dimensional variable $\boldsymbol{u}(\boldsymbol{x})$ is distributed according to the G-component Gaussian mixture as

$$p(\boldsymbol{u}(\boldsymbol{x})) = \sum_{g=1}^{G} \alpha_g \phi(\boldsymbol{u}(\boldsymbol{x}); \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g + \lambda \boldsymbol{I}_q). \tag{8}$$

Here the gth component model is defined by $\boldsymbol{u}(\boldsymbol{x})|l_g = 1 \sim N(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g + \lambda \boldsymbol{I}_q)$. Notice that the density of observed variable can be rewritten as

$$p(\boldsymbol{x}) = (2\pi)^{-d+q} \lambda^{-d+q} \exp\left[-\frac{1}{2\lambda}(\|\boldsymbol{x}\|^2 - \|\boldsymbol{u}(\boldsymbol{x})\|^2) \right]$$
$$\times p(\boldsymbol{u}(\boldsymbol{x})), \tag{9}$$

if the parameters satisfy the restrictions (a) and (b).

Using Bayes theorem, the posterior probability of $l_g = 1$ can be assessed by

$$\Pr(\boldsymbol{x} \in \mathcal{P}_g) = \frac{\alpha_g \phi(\boldsymbol{u}(\boldsymbol{x}); \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g + \lambda \boldsymbol{I}_q)}{p(\boldsymbol{u}(\boldsymbol{x}))}, \tag{10}$$

for $g \in \{1, \cdots, G\}$. Each value assigns the membership probability that an observation $\boldsymbol{x}$ belongs to the gth subpopulation $\mathcal{P}_g$. Therefore, the posterior probability of l is given by multinomial distribution having the probability mass function $P(l|\boldsymbol{x}) = \prod_{g=1}^{G} \Pr(\boldsymbol{x} \in \mathcal{P}_g)^{l_g}$. Note that the probability of belonging is defined on the q-dimensional variable $\boldsymbol{u}(\boldsymbol{x})$, and so we can write $\Pr(\boldsymbol{x} \in \mathcal{P}_g) = \Pr(\boldsymbol{u}(\boldsymbol{x}) \in \mathcal{P}_g)$ without loss of generality.

Next let us consider the posterior distribution of the factor variable. Given $l_g = 1$, the $\boldsymbol{u}(\boldsymbol{x})$ and the $\boldsymbol{f}$ are unconditionally distributed to be $2q$-dimensional Gaussian with mean $(\boldsymbol{\mu}_g^T, \boldsymbol{\mu}_g^T)^T$ and covariance matrix

$$\left(\begin{array}{cc} \boldsymbol{\Sigma}_g + \lambda \boldsymbol{I}_q & \boldsymbol{\Sigma}_g \\ \boldsymbol{\Sigma}_g & \boldsymbol{\Sigma}_g \end{array} \right).$$

Then, it follows from standard property of the multivariate normal distribution that

$$p(\boldsymbol{f}|\boldsymbol{x}, l_g = 1) = \phi(\boldsymbol{f}; \boldsymbol{\psi}_g(\boldsymbol{x}), \boldsymbol{\Gamma}_g),$$

where the posterior mean and covariance are given by

$$
\begin{aligned}
\boldsymbol{\psi}_g(\boldsymbol{x}) &= \boldsymbol{\mu}_g + \boldsymbol{\Sigma}_g(\boldsymbol{\Sigma}_g + \lambda\boldsymbol{I}_q)^{-1}(\boldsymbol{u}(\boldsymbol{x}) - \boldsymbol{\mu}_g) \\
&= \boldsymbol{\mu}_g + \boldsymbol{Q}_g(\boldsymbol{u}(\boldsymbol{x}) - \boldsymbol{\mu}_g),
\end{aligned}
$$

and

$$
\begin{aligned}
\boldsymbol{\Gamma}_g &= \boldsymbol{\Sigma}_g - \boldsymbol{\Sigma}_g(\boldsymbol{\Sigma}_g + \lambda\boldsymbol{I}_q)^{-1}\boldsymbol{\Sigma}_g \\
&= \lambda\boldsymbol{Q}_g. \tag{11}
\end{aligned}
$$

Here the $q \times q$ matrix $\boldsymbol{Q}_g$ is diagonal so that each diagonal element consists of the signal to noise ratio $\sigma_{kg}/(\sigma_{kg} + \lambda)$, $k \in \{1, \cdots, q\}$ corresponding to the generative model (7) in which σ_{kg} denotes the kth diagonal element of $\boldsymbol{\Sigma}_g$. Hence, this gives the posterior distribution of $\boldsymbol{f}$ in the form of the Gaussian mixture;

$$
p(\boldsymbol{f}|\boldsymbol{x}) = \sum_{g=1}^{G} \Pr(\boldsymbol{x} \in \mathcal{P}_g)\phi(\boldsymbol{f}; \boldsymbol{\psi}_g(\boldsymbol{x}), \boldsymbol{\Gamma}_g). \tag{12}
$$

The mixing proportions correspond to the posterior probabilities of belonging. These posterior distributions are used to estimate $\boldsymbol{\theta}$ (Section 3) and to construct Bayes estimators for the latent variables, e.g. the posterior expectation, the maximum a posteriori (MAP) estimator.

Finally, note that we have no need to calculate d-dimensional density for assessing either (10) and (9). In microarray studies where the dimension of feature vector is ranging from 10^3 to 10^4, the direct calculation based on the form of $\Pr(\boldsymbol{x} \in \mathcal{P}_g) \propto \alpha_g\phi(\boldsymbol{x}; \boldsymbol{\Xi}\boldsymbol{\mu}_g, \boldsymbol{\Xi}\boldsymbol{\Sigma}_g\boldsymbol{\Xi}^T + \lambda\boldsymbol{I}_d)$ and (4) might fail since the high dimensional densities take extremely small values, and also are computationally very demanding. In contrast to such intractability, our approach of the parameter restriction provides a way of saving the computational resources and avoiding the overflow.

3. Model Estimation

3.1. EM algorithm

Suppose that we have a sample of size N, $\boldsymbol{X} = (\boldsymbol{x}_1 \cdots \boldsymbol{x}_N)$ where $\boldsymbol{X}$ denotes a data matrix of order $d \times N$. Given data, the mixed factors model can be fitted under the principle of the maximum likelihood although there exists no closed-form solution for the estimator. The EM algorithm is a general approach to maximum likelihood estimation for the problem in which a probability model includes one or more latent variables.

We let $\boldsymbol{f}_j^T = (f_{1j}, \cdots, f_{qj})$ and $\boldsymbol{l}_j^T = (l_{1j}, \cdots, l_{Gj})$ be the realizations of factor variable and class label vector corresponding to the jth observation $\boldsymbol{x}_j$. The complete data $\boldsymbol{y}_j^T = (\boldsymbol{x}_j^T, \boldsymbol{f}_j^T, \boldsymbol{l}_j^T)$ is assumed to be i.i.d. sample drawn from (3). Suppose that we are now having an estimate $\hat{\boldsymbol{\theta}}$ at

the current step of the EM algorithm. The method alternates between two steps, say E step and M step.

Firstly, consider to update the hth column of the factor loading matrix, $\boldsymbol{\xi}_h$, while all of parameters except to $\boldsymbol{\xi}_h$ are fixed at the obtained values. The evaluation of $\boldsymbol{\xi}_h$ must take account of the constraints that $\boldsymbol{\xi}_h^T\hat{\boldsymbol{\xi}}_k = 0$ for all $k \neq h$. This can be achieved by the use of Lagrange multiplier γ_k for $k \neq h$. Then, the complete data log-likelihood with the Lagrange terms is given by

$$
\begin{aligned}
\mathcal{D}_\mathcal{C} &= \sum_{j=1}^{N} \log\phi(\boldsymbol{x}_j; f_{hj}\boldsymbol{\xi}_h + \sum_{k \neq h} f_{kj}\hat{\boldsymbol{\xi}}_k, \hat{\lambda}\boldsymbol{I}_d) \\
&\quad + \sum_{k \neq h} \gamma_k\boldsymbol{\xi}_h^T\hat{\boldsymbol{\xi}}_k, \tag{13}
\end{aligned}
$$

where $\hat{\boldsymbol{\xi}}_k$ denotes the fixed values and we have omitted the terms independent of $\boldsymbol{\xi}_h$. Here, we have no concern to the imposition on the norm of $\boldsymbol{\xi}_h$ since this can be scaled a posteriori as will be seen in later. But it is assumed that all of $\hat{\boldsymbol{\xi}}_k$ for $k \neq h$ have been normalized such that $\|\hat{\boldsymbol{\xi}}_k\| = 1$.

By taking derivative of (13) with respect to $\boldsymbol{\xi}_h$ and multiplying it by some constants, one can obtain the gradient as follow;

$$
\begin{aligned}
\boldsymbol{g}(\boldsymbol{\xi}_h) &= \boldsymbol{\xi}_h\sum_j f_{hj}^2 + \sum_{k \neq h}\hat{\boldsymbol{\xi}}_k\sum_j f_{hj}f_{kj} \\
&\quad - \sum_j \boldsymbol{x}_j f_{hj} + \lambda\sum_{k \neq h}\gamma_k\hat{\boldsymbol{\xi}}_k.
\end{aligned}
$$

Here the quantities $(\sum_j f_{hj}f_{1j}, \cdots, \sum_j f_{hj}f_{qj})^T$ and $\sum_j \boldsymbol{x}_j f_{hj}$ are summarized in the hth column of the matrices of the sufficient statistics,

$$
\boldsymbol{T}_{ff} = \sum_j \boldsymbol{f}_j\boldsymbol{f}_j^T, \quad \boldsymbol{T}_{xf} = \sum_j \boldsymbol{x}_j\boldsymbol{f}_j^T. \tag{14}
$$

In the E step, we replace $(\sum_j f_{hj}f_{1j}, \cdots, \sum_j f_{hj}f_{qj})^T$ and $\sum_j \boldsymbol{x}_j f_{hj}$ by the conditional expectations with respect to $\hat{p}(\boldsymbol{f}|\boldsymbol{x})$ (12) where the posterior distribution $\hat{p}(\boldsymbol{f}|\boldsymbol{x})$ is assessed by the current values of parameters.

These are given by the hth column of

$$
\langle\boldsymbol{T}_{ff}\rangle = \sum_j\sum_g \hat{\Pr}(\boldsymbol{x}_j \in \mathcal{P}_g)\left(\hat{\boldsymbol{\Gamma}}_g + \hat{\boldsymbol{\psi}}_g(\boldsymbol{x}_j)\hat{\boldsymbol{\psi}}_g(\boldsymbol{x}_j)^T\right), \tag{15}
$$

$$
\langle\boldsymbol{T}_{xf}\rangle = \sum_j \boldsymbol{x}_j\sum_g \hat{\Pr}(\boldsymbol{x}_j \in \mathcal{P}_g)\hat{\boldsymbol{\psi}}_g(\boldsymbol{x}_j), \tag{16}
$$

where again, all posterior quantities are evaluated by the current values of parameters. Then, these give the estimating equation for $\boldsymbol{\xi}_h$ as follow;

$$
\begin{aligned}
\boldsymbol{\xi}_h\langle\sum_j f_{hj}^2\rangle &+ \sum_{k \neq h}\hat{\boldsymbol{\xi}}_k\langle\sum_j f_{hj}f_{kj}\rangle - \langle\sum_j \boldsymbol{x}_j f_{hj}\rangle \\
&+ \lambda\sum_{k \neq h}\gamma_k\hat{\boldsymbol{\xi}}_k = \boldsymbol{0}. \tag{17}
\end{aligned}
$$

To find a solution for γ_k, consider to multiply (17) by $\hat{\boldsymbol{\xi}}_k^T$. Then, this leads to

$$\hat{\gamma}_k = \frac{1}{\lambda}\left(\hat{\boldsymbol{\xi}}_k^T \langle\sum_j \boldsymbol{x}_j f_{hj}\rangle - \langle\sum_j f_{hj}f_{kj}\rangle\right),$$

for all $k \neq h$. Further, substituting $\hat{\gamma}_k$ to (17) give a solution as

$$\hat{\boldsymbol{\xi}}_h = \frac{1}{\langle\sum_j f_{hj}^2\rangle}\left\{\langle\sum_j \boldsymbol{x}_j f_{hj}\rangle - \sum_{k\neq h}\hat{\boldsymbol{\xi}}_k^T\langle\sum_j \boldsymbol{x}_j f_{hj}\rangle\hat{\boldsymbol{\xi}}_k\right\}. \quad (18)$$

Converting this new values to the old one yields $\hat{\boldsymbol{\Xi}}$.

Here, notice that since $\|\hat{\boldsymbol{\xi}}_h\| \neq 1$, we need to normalize it. Let $\boldsymbol{U}_h$ be a diagonal matrix of order q such that the hth diagonal element is $\|\hat{\boldsymbol{\xi}}_h\|^{1/2}$, otherwise 1. Since $\boldsymbol{U}_h\hat{\boldsymbol{\Sigma}}_g\boldsymbol{U}_h^T$ remains diagonal, the probabilistic nature of (1) are still preserved even if $\hat{\boldsymbol{\Xi}} \to \hat{\boldsymbol{\Xi}}\boldsymbol{U}_h^{-1}$ and $\boldsymbol{f} \to \boldsymbol{U}_h\boldsymbol{f}$. Thus, without loss of generality, we can rescale the estimates of parameters by

$$\hat{\boldsymbol{\Xi}}\boldsymbol{U}_h^{-1} \to \hat{\boldsymbol{\Xi}}, \;\; \boldsymbol{U}_h\hat{\boldsymbol{\mu}}_g \to \hat{\boldsymbol{\mu}}_g, \;\; \boldsymbol{U}_h\hat{\boldsymbol{\Sigma}}_g\boldsymbol{U}_h^T \to \hat{\boldsymbol{\Sigma}}_g. \quad (19)$$

Repeating these processes, (18) and (19), for $h = 1, \cdots, q$, we would have an estimate of factor loading matrix.

Next, differentiating the complete data log-likelihood with respect to the remaining parameters and setting all derivatives to zero, one can obtain the following equations after some manipulations,

$$\text{tr}\left(N\lambda\boldsymbol{I}_d - \boldsymbol{T}_{xx} - \hat{\boldsymbol{\Xi}}\boldsymbol{T}_{ff}\hat{\boldsymbol{\Xi}}^T + \hat{\boldsymbol{\Xi}}\boldsymbol{T}_{xf}^T + \boldsymbol{T}_{xf}\hat{\boldsymbol{\Xi}}\right) = \boldsymbol{0},$$

and for $g \in \{1, \cdots, G\}$,

$$\boldsymbol{T}_{fl_g} - \boldsymbol{\mu}_g T_{l_g} = \boldsymbol{0},$$
$$T_{l_g}\boldsymbol{\Sigma}_g - \boldsymbol{T}_{ffl_g} + T_{l_g}\boldsymbol{\mu}_g\boldsymbol{\mu}_g^T = \boldsymbol{O},$$
$$\frac{1}{\alpha_g}T_{l_g} + \kappa = 0.$$

Here, the sufficient statistics is comprised of (14) and

$$\boldsymbol{T}_{xx} = \sum_j \boldsymbol{x}_j\boldsymbol{x}_j^T, \;\; \boldsymbol{T}_{fl_g} = \sum_j l_{gj}\boldsymbol{f}_j,$$
$$\boldsymbol{T}_{ffl_g} = \sum_j l_{gj}\boldsymbol{f}_j\boldsymbol{f}_j^T, \;\; T_{l_g} = \sum_j l_{gj}.$$

The E step requires the expectation of sufficient statistics conditional on $\boldsymbol{X}$ and the current values of parameters. It follows from the results presented in Section 2 that

$$\langle\boldsymbol{T}_{xx}\rangle = \sum_j \boldsymbol{x}_j\boldsymbol{x}_j^T,$$
$$\langle\boldsymbol{T}_{fl_g}\rangle = \sum_j \hat{\text{Pr}}\,(\boldsymbol{x}_j \in \mathcal{P}_g)\hat{\boldsymbol{\psi}}_g(\boldsymbol{x}_j),$$

$$\langle\boldsymbol{T}_{ffl_g}\rangle = \sum_j \hat{\text{Pr}}(\boldsymbol{x}_j \in \mathcal{P}_g)(\hat{\boldsymbol{\Gamma}}_g + \hat{\boldsymbol{\psi}}_g(\boldsymbol{x}_j)\hat{\boldsymbol{\psi}}_g(\boldsymbol{x}_j)^T),$$
$$\langle T_{l_g}\rangle = \sum_j \hat{\text{Pr}}\,(\boldsymbol{x}_j \in \mathcal{P}_g), \quad (20)$$

and (15), (16).

In the M step, all sufficient statistics are replaced by these posterior quantities. Then the M step is given by

$$\hat{\lambda} = \frac{1}{dN}\text{tr}\left(\langle\boldsymbol{T}_{xx}\rangle + \hat{\boldsymbol{\Xi}}\langle\boldsymbol{T}_{ff}\rangle\hat{\boldsymbol{\Xi}}^T - 2\langle\boldsymbol{T}_{xf}\rangle\hat{\boldsymbol{\Xi}}\right), \quad (21)$$

and for $g \in \{1, \cdots, G\}$

$$\hat{\alpha}_g = \frac{1}{n}\langle T_{l_g}\rangle, \quad (22)$$
$$\hat{\boldsymbol{\mu}}_g = \frac{1}{\langle T_{l_g}\rangle}\langle\boldsymbol{T}_{fl_g}\rangle, \quad (23)$$
$$\hat{\boldsymbol{\Sigma}}_g = \frac{1}{\langle T_{l_g}\rangle}\langle\boldsymbol{T}_{ffl_g}\rangle - \hat{\boldsymbol{\mu}}_g\hat{\boldsymbol{\mu}}_g^T. \quad (24)$$

Thus, the EM algorithm is summarized as follows:

1. Specify initial values of parameters, $\hat{\boldsymbol{\theta}}$.

2. (Factor loadings) Repeat the following steps for $h = 1, \cdots, q$.

   (a) Evaluate the hth column of $\langle\boldsymbol{T}_{ff}\rangle$ and $\langle\boldsymbol{T}_{xf}\rangle$ based on $\hat{\boldsymbol{\theta}}$.

   (b) Compute $\hat{\boldsymbol{\xi}}_h$ by equation (18).

   (c) Normalize $\hat{\boldsymbol{\xi}}_h$ and rescale $\hat{\boldsymbol{\mu}}_g$, $\hat{\boldsymbol{\Sigma}}_g$, $g \in \{1, \cdots, G\}$.

   (d) Set these parameters to $\hat{\boldsymbol{\theta}}$.

3. (Remaining parameters)

   (a) Evaluate $\langle\boldsymbol{T}_{xx}\rangle$, $\langle\boldsymbol{T}_{xf}\rangle$, $\langle\boldsymbol{T}_{ff}\rangle$, $\langle\boldsymbol{T}_{fl_g}\rangle$, $\langle\boldsymbol{T}_{ffl_g}\rangle$, $\langle T_{l_g}\rangle$ based on $\hat{\boldsymbol{\theta}}$.

   (b) Update λ, α_g, $\boldsymbol{\mu}_g$, and $\boldsymbol{\Sigma}_g$ by (21),(22),(23),(24).

   (c) Set these parameters to $\hat{\boldsymbol{\theta}}$.

4. If the sequence of parameters and log-likelihood is judged to have converged, the iteration is stopped, otherwise go to 2.

The factor loading matrix are updated iteratively according to (18) and the rescaling (19) in which one stage is composed of q cycles. Then, the equations from (21) to (24) follow. The EM algorithm alternates between these processes until a sequence of the produced parameters and the values of likelihood is judged to have converged. Under mild conditions, the method is guaranteed to find a local maximum of the log-likelihood.

3.2. Implementations

As well as another mixture models, the likelihood equations of the mixed factor model have the multiple roots, and so the EM algorithm should be started from a wide choice of starting values to search for all local maxima. An obvious choice for the roots is the one corresponding to the largest of the local maxima. In below, we will describe an ad hoc way of specifying initial parameters.

Firstly, consider to determine each element in the factor loading matrix, denoted by $(\Xi)_{hk}$, $h \in \{1, \cdots, d\}$ and $k \in \{1, \cdots, q\}$ and the noise variance λ. The main difficulty is that the factor loadings must take account the condition $\Xi^T \Xi = I_q$. Let $\bar{x}_k = (1/N) \sum_j x_{kj}$ and $s_k = (1/N) \sum_j (x_{kj} - \bar{x}_k)^2$, that is, sample mean and variance of kth variable in x. Then our approach to the initialization of these parameters is summarized as follows:

1. Set $\lambda = (1/d) \sum_{k=1}^{d} s_k$.

2. Generate the factor loadings by $(\Xi)_{hk} \sim N(0, 1)$, for all h, k.

3. Implement the cholesky decomposition, $\Xi^T \Xi = C^T C$ where C is the lower triangle matrix of order q, and recompute the factor loadings by

$$\Xi C^{-1} \to \Xi.$$

Subsequently, these parameters give a set of orthogonal transformation of original data $u(x_j) = \Xi^T x_j$, $j \in \{1, \cdots, N\}$ and also the mean and variance, $\bar{u} = (1/N) \sum_j u(x_j)$ and $S_u = (1/N) \text{diag} \sum_j (u(x_j) - \bar{u})(u(x_j) - \bar{u})^T$ which will be useful for specifying the component parameters of $p(f)$. An outright way is that $\alpha_g = 1/G$, $\mu_g \sim N(\bar{u}, S_u)$ and $\Sigma_g = S_u$, $g \in \{1, \cdots, G\}$.

For some gene expression datasets, each array contains some genes with fluorescence intensity measurements that were flagged by the experimenter and recorded as missing data points. For instance, [9] reported that the mean percentage of missing data points per array is 6.6% for the lymphoma datasets [2] and 3.3% for the NCI 60 datasets [19]. Our modeling offers a natural approach to the analysis in the cases where some values in $x_j^T = (x_{1j} \cdots x_{dj})$ exhibit one or more missing values. The presence of missing values requires a minor change in the EM algorithm in which the missing values are considered as to be a set of latent variables. However, we are omitted to discuss this issue more deeply in this study.

4. Mixed Factors Analysis

Once the model has been fitted to a dataset, the mixed factors analysis offers the applications to clustering, dimension reduction and data visualization. These can be addressed by finding the plausible values of latent variables f_j, l_j. The analysis also covers the method of extracting sets of genes to be relevant to explain the presence of groups. In this process, the sets of genes that are considered to be functionally co-worked is automatically detected. In sequel, we will discuss these methods.

4.1 Model Selection

The basic issues arising in the mixed factors analysis are determination of the number of clusters and the dimension of factor variable. In statistics, these issues can be converted into the problem of model selection among all possible models being in consideration, $\mathcal{M}_k$, $k = 1, \cdots, K$.

A commonly used approach to this problem is based on the Bayesian Information Criterion (BIC, [20]);

$$BIC(k) = -2\hat{\mathcal{D}}_k + \nu_k \log N. \tag{25}$$

Here the $\hat{\mathcal{D}}_k$ is a selected local maxima of log-likelihood of $\mathcal{M}_k$, and ν_k denotes the number of free parameter given by (5). We should choose a model to be most likely in sense of the minimum BIC. Unfortunately, finite mixture models do not satisfy the regularity conditions that underlies the published proofs of (25), but several results suggest its appropriateness and good performance in a range of applications of model-based clustering [17].

There also exists another possible approaches to be applicable in this context, e.g. the classical hypothesis testing, as the likelihood ratio test [15], Akaike Information Criterion (AIC, [1]) and the B-fold cross validation ([1, 18]). However, the validity of these methods also depends on the same regularity conditions needed for the asymptotic expansions in the derivation of BIC. These conditions break down due to the lack of identifiability in the mixture model.

4.2. Cluster Analysis

The goal of cluster analysis is to classify data x_j, $j \in \{1, \cdots, N\}$ into nonoverlapping G groupings. In terms of model-based clustering, it is converted into the problem to infer the l_j on the basis of the feature data x_j or to divide the feature space $\mathbf{R}^d$ by attributing labels l_g, $g \in \{1, \cdots, G\}$ for all $x \in \mathbf{R}^d$. Our analysis facilitates the clustering based on the estimated posterior probability that x_j belongs to $\mathcal{P}_g$, $g \in \{1, \cdots, G\}$. The most common classifier is to assign x_j to a class with the highest posterior probability of belonging;

$$\hat{l}_g(x_j) = \begin{cases} 1 & \text{if } \hat{\Pr}(x_j \in \mathcal{P}_g) = \max_{h \in G} \hat{\Pr}(x_j \in \mathcal{P}_h), \\ 0 & \text{otherwise.} \end{cases}$$

If the estimate $\hat{\Pr}(x_j \in \mathcal{P}_g)$ were true, this classification rule would be the Bayes rule which minimizes the overall

misclassification rate [18]. This clustering is made based on $\hat{u}(x_j)$ and the estimated q-dimensional Gaussian mixture $\hat{p}(\hat{u}(x))$. If the specified dimension of $\hat{u}(x_j)$ is less or equal to two, it is possible to visualize the decision boundary.

4.3. Dimension Reduction

The mixed factors analysis gives a scope to transform the data into some reduced-dimensionality representation. This can be carried out by the posterior mean of factor variable,

$$\hat{f}(x_j) = \sum_{g=1}^{G} \hat{\Pr}\left(x_j \in \mathcal{P}_g\right)\hat{\psi}_g(x_j),$$

under the Bayes rule [14]. An alternative way of constructing a reduced-rank data may be the orthogonal mapping of feature data onto $\mathbf{R}^q$,

$$E(f + \tilde{\epsilon}|x) = \hat{u}(x_j),$$

although we have no clarity to do so. When the specified dimension of factor is less or equal to three, the estimated factor variables are useful for data visualization. Even when $q > 3$, the data visualization is possible by using some techniques, e.g. the scatter plot.

Alternative method of data visualization is to select some axes to be plotted among $\{1, \cdots, q\}$ such that set of the mappings reveals the presence of group structure. Consider now to select two axes. Let $\hat{u}_{kh}(x) = (\hat{u}_k(x), \hat{u}_h(x))^T$. We consider that the degree of separation given by the set, $\hat{u}_{kh}(x_j)$, $j \in \{1, \cdots, N\}$ is measured by using the minus entropy of the component memberships;

$$E_{kh} = \sum_{j=1}^{N}\sum_{g=1}^{G}\hat{\Pr}(\hat{u}_{kh}(x_j) \in \mathcal{P}_g)\log\hat{\Pr}\left(\hat{u}_{kh}(x_j) \in \mathcal{P}_g\right),$$

where

$$\hat{\Pr}\left(\hat{u}_{kh}(x_j) \in \mathcal{P}_g\right) \propto \alpha_g \phi(\hat{u}_{kh}(x_j); \hat{\mu}_g^{kh}, \hat{\Sigma}_g^{kh} + \lambda I_2).$$

The $\hat{\mu}_g^{kh}$ consists of the kth and hth elements of $\hat{\mu}_g$, and the $\hat{\Sigma}_g^{kh}$ is diagonal matrix such that the elements are given by $\hat{\Sigma}_g$. Here it is assumed that $\hat{\Pr}\left(\hat{u}_{kh}(x_j) \in \mathcal{P}_g\right)\log\hat{\Pr}(\hat{u}_{kh}(x_j) \in \mathcal{P}_g) = 0$ if $\hat{\Pr}(\hat{u}_{kh}(x_j) \in \mathcal{P}_g) = 0$.

Note that the E_{kh} is minimized if and only if $\hat{\Pr}\left(\hat{u}_{kh}(x_j) \in \mathcal{P}_g\right) = 1/G$ for all $g \in \{1, \cdots, G\}$ and $j \in \{1, \cdots, N\}$. This implies the poor separation of $\hat{u}_{kh}(x_j)$, $j \in \{1, \cdots, N\}$ in sense of that all of $\hat{u}_{kh}(x_j)$ can not be assigned to a particular group. In contrast to this, the E_{kh} is maximized if and only if $\hat{\Pr}(\hat{u}_{kh}(x_j) \in \mathcal{P}_g) = 1$ for certain g and $\hat{Pr}(\hat{u}_{kh}(x_j) \in \mathcal{P}_m) = 0$ for $m \neq g$, $j \in \{1, \cdots, N\}$. This means that each observation is completely classified into a particular group with probability one. In this sense, the minus entropy of the component memberships can be interpreted as a quantity for measuring the degree of separation exhibited by $\hat{u}_{kh}(x_j)$, $k, h \in \{1, \cdots, q\}$. Thus we select $\hat{f}_{k'}$ and $\hat{f}_{h'}$ to be visualized such that $(k', h') = \mathrm{argmax}_{k,h\in\{1,\cdots,q\}}E_{kh}$.

The axes selection can also be achieved by using the another possible quantities for measuring the degree of separation, e.g. the between group variance under the estimated class labels. However, we have no idea to select one among possible approaches.

4.4 Interpretation of Mixed Factors

The researchers might often desire to obtain a biological interpretation of the estimated mixed factors, and also to extract variables from $x^T = (x_1, \cdots, x_d)$ so as to contribute the presence of groups on the feature space or to exclude ones not to do so. In our context, this can be achieved by assessing the dependency between $x_1, \cdots, x_d$ and $f_1, \cdots, f_q$. A natural measure to summarize the dependency in these variables might be covariance

$$Cov(x, f) = \hat{\Xi}\left(\sum_{g=1}^{G}\hat{\alpha}_g\hat{\Sigma}_g\right). \tag{26}$$

It may be more convenient to use the correlation matrix $\rho(x, f)$ which is of that each (h, k)th element of (26) are divided by the square of $Var(f_k) = \sum_{g=1}^{G}\hat{\alpha}_g\hat{\sigma}_{kg}$ and the hth diagonal element of $Var(x) = \hat{\Xi}(\sum_{g=1}^{G}\hat{\alpha}_g\hat{\Sigma}_g)\hat{\Xi}^T + \hat{\lambda}I_d$. By investigating the values in $\rho(x, f)$ or $\mathrm{Cov}(x, f)$, each of q-coordinates can be understood. If the hth gene, that is, x_h, is highly correlated with f_k, then it is considered to be relevant to explain the grouping shown in kth coordinate.

In practice, it will be helpful to list some genes to give the highly positive correlation with f_k at Ω_+^k and to give the highly negative correlation with f_k at Ω_-^k for $k \in \{1, \cdots, q\}$. As will be demonstrated in next section, in context of gene expression analysis, these $2q$ sets can be useful to find the biologically meaningful groups of genes to be functionally co-worked and also to explain the existence of group structure.

5. Real Data Analysis

In this section we will give an illustration of the mixed factors analysis with a well-known dataset, the leukemia data of Goloub *et al*. [12]. This dataset is available at http://www.broad.mit.edu/cancer/. Our objective is to cluster the leukemia tissues on the basis of genes. Although the class labels have been available, we applied the mixed factors analysis to the dataset without this
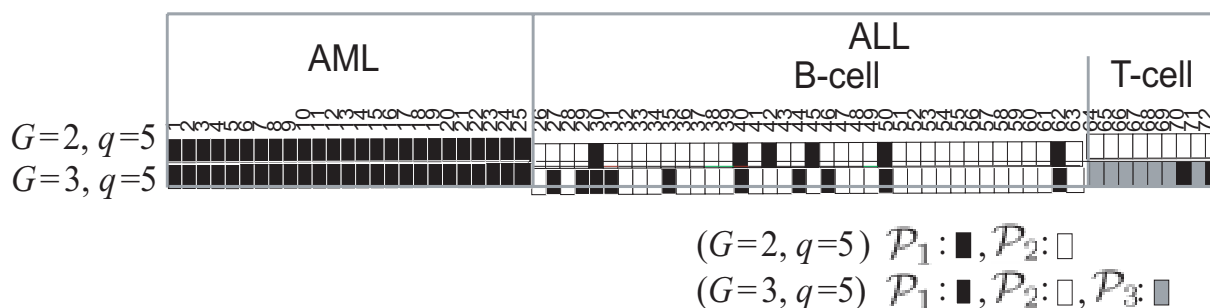
Figure 1. Results of clustering. The order of tissues was rearranged so that the cases of AML are labeled by $1-25$ and the ALL are labeled by $26-72$ (B-cell ALL, $26-63$ and T-cell ALL, $64-72$). For clustering given by the model of $q=5$ and $G=2$, the black and white sites correspond to the tissues grouped into $\mathcal{P}_1$ and $\mathcal{P}_2$, respectively. For clustering given by the model of $q=5$ and $G=3$, the black, white and gray sites indicate the tissues grouped into $\mathcal{P}_1$, $\mathcal{P}_2$ and $\mathcal{P}_3$, respectively.
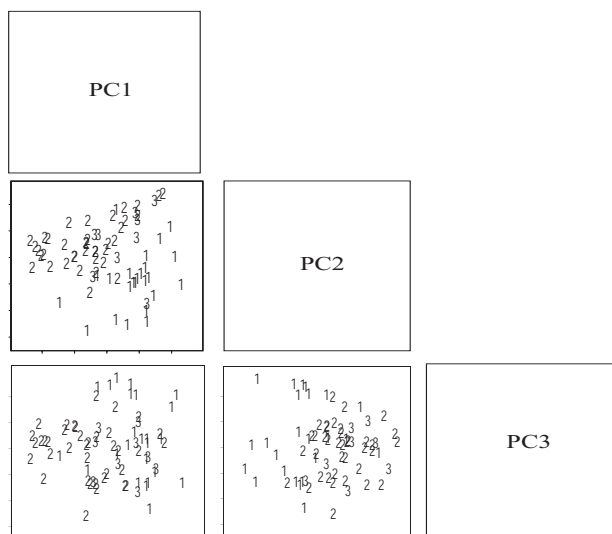
Figure 2. Scatter plots of the first three of principal components. The true groups are labeled by AML $\to 1$, B-cell ALL $\to 2$ and T-cell ALL $\to 3$.
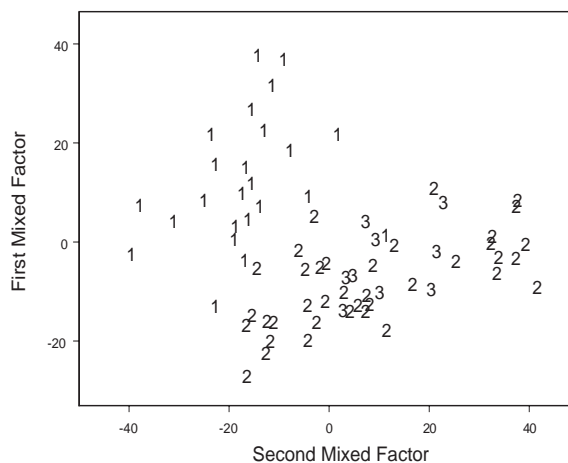

Figure 3. Plot of $\hat{f}_{k'}(x_j)$, $\hat{f}_{h'}(x_j)$ given by the mixed factors model with $G=2$ and $q=5$. The true groups are labeled by AML $\to 1$, B-cell ALL $\to 2$ and T-cell ALL $\to 3$.

knowledge in order to demonstrate its potential usefulness as a method of unsupervised learning.

5.1. Leukemia Data and Preprocessing

The leukemia data were studied by Goloub *et al.* [12]. Originally it was reported that the leukemia data contains two types of acute leukemias: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). The 7129 gene expression levels were measured using

Affymetrix high density oligonucleotide arrays on 72 patients, consisting of 47 cases of ALL and 25 cases of AML (38 B-cell ALL and 9 T-cell ALL). As shown in Figure 1, the order of tissues was rearranged so that the cases of AML are labeled by $1-25$ and the ALL are labeled by $26-72$ (B-cell ALL, $26-63$ and T-cell ALL, $64-72$). Following [9, 12, 16], three preprocessing steps were applied to the normalized matrix of intensity values available on the website: (a) thresholding, floor of 100 and ceiling of $16,000$; (b) filtering, exclusion of genes with $\max / \min \leq 5$ or $(\max - \min) \leq 500$ where $\max$ and $\min$ refer to the maximum and minimum intensities for a particular gene across the 72 samples; (c) the natural logarithm of the expression
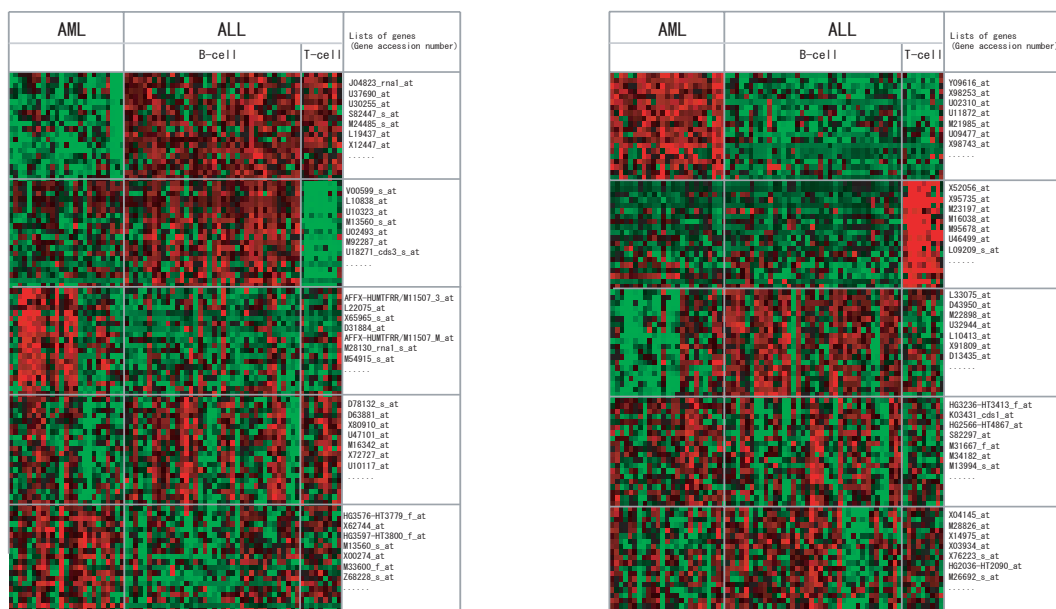
Figure 4. Heat map for the expression levels of genes judged to be relevant to the group structure. These genes are selected by the mixed factors model with $G = 3$, $q = 5$. Each of 5 rows shows the expression levels corresponding to 20 genes in Ω_+^k (left) and Ω_-^k (right), for $k = 1, \cdots, 5$. The 20 genes in Ω_+^k were selected such that these are top 20 out of $x_1 \cdots, x_d$ to give the highest positive correlation with f_k. The 20 genes in Ω_-^k are top 20 out of $x_1 \cdots, x_d$ to give the highest negative correlation with f_k. The first 25 samples refer to the AML cases, and the next 47 samples, the ALL $26 - 72$ cases (B-cell ALL, $26 - 63$ and T-cell ALL, $64 - 72$). The name of gene follows from the gene accession number to be available at the website.

levels was taken. After preprocessing, the 3571 genes remain and so produce the data matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{72})$ of order 3571×72.

The retained dataset was firstly standardized so that each columns in the matrix of the logged microarray data have mean 0 and variance 1, and then we standardized the rows of $\boldsymbol{X}$ to have mean zero and unit variance.

5.2. Clustering and Selecting Relevant Genes

We firstly applied PCA to this retained dataset based on the correlation matrix. Figure 2 displays the first three of principal components. These projections slightly revealed clusters implying the existence of two classes, ALL and AML. However these provided no evidence for the presence of subclasses, AML, B-cell ALL and T-cell ALL. If proceeding to clustering of these projections using some techniques, e.g. k-means, Gaussian mixture clustering, we would obtain the large number of misallocations.

Next we considered clustering of the tissue samples on the basis of the retained 3571 genes using the mixed factors model with $G = 2$. After fitting the models ranging from $q = 2$ to $q = 7$ with 40 starting values of parameters, the

model of $q = 5$ was chosen to be best in the sense of the minimum BIC. This gave the following groups;

$$
\begin{aligned}
\mathcal{P}_1 &= \{1 - 25, 30, 40, 42, 45, 50, 62\}, \\
\mathcal{P}_2 &= \{26 - 29, 31 - 39, 41, 43, 44, 46 - 49, \\
&\quad 51 - 61, 63 - 72\}. \quad (27)
\end{aligned}
$$

This grouping is also summarized in Figeure 1. It can be found from (27) and Figure 1 that the two clusters reflect the partition corresponding to the ALL and AML leukemia, and that most part of members in $\mathcal{P}_1$ correspond to the AML leukemia tissues. Particularly, the AML and T-cell ALL were completely classified. The misallocations were equal to $\{30, 40, 42, 45, 50, 62\}$, and so, the error rate is about 8%. All of misclassifications corresponds to B-cell ALL. We confirmed that the mixed factors analysis could provide a meaningful grouping despite of the high-dimensionality of dataset.

Figure 3 displays $\hat{f}_{k'}(\boldsymbol{x}_j), \hat{f}_{h'}(\boldsymbol{x}_j), j \in \{1, \cdots, N\}$ obtained from the selected model, where the coordinates k' and h' were chosen according to the minus entropy of the component memberships. Each data points is labeled by the true class. This plot is helpful to understand the group structure, visually. These projections present the existence

of two clusters. All of the 6 projections corresponding to misallocations were located near the boundary.

The 47 ALL tissues consists of 9 T-cell and 38 B-cell types. Given the existence of three subclasses, 25 AML, 38 B-cell ALL and 9 T-cell ALL, Chow *et al.* [7] and McLachlan *et al.* [16] attributed these samples into three groups. We also considered clustering this dataset into three groups by using the mixed factors model of $G = 3$. Given the models ranging from $q = 2$ to $q = 7$ with 40 starting values, the scores of BIC produced by each local maxima decided the goodness of $q = 5$. This gave the following clusters (see also Figeure 1),

$$
\begin{aligned}
\mathcal{P}_1 &= \{1 - 25, 27, 29 - 31, 35, 40, 44, 46, 50, 62, 70, 72\}, \\
\mathcal{P}_2 &= \{26, 28, 32 - 34, 36 - 39, 41 - 43, 45, 47 - 49, \\
&\quad\quad 51 - 61, 63\}, \\
\mathcal{P}_3 &= \{64 - 69, 71\}.
\end{aligned}
$$

This split implies the weak association of three groups $\mathcal{P}_g$, $g = 1, 2, 3$, with AML, B-cell ALL and T-cell ALL, respectively. The first cluster $\mathcal{P}_1$ consisted of 25 AML, 10 B-cell ALL plus 2 T-cell ALL. The $\mathcal{P}_2$ consisted of the remaining B-cell cases. All of the members belonging to $\mathcal{P}_3$ were T-cell cases. All of AML cases were completely classified as well as clustering by $G = 2$. Most part of misallocations correspond to B-cell ALL cases and were classified into $\mathcal{P}_1$.

Figure 4 displays the heat map of the expression levels given by genes in Ω_+^k and Ω_-^k, $k \in \{1, \cdots, 5\}$. These 10 sets of 20 genes showed the 10 expression patterns. All genes in a set exhibit the similar expression pattern. These genes are considered to be functionally co-worked. In addition, notice that a pair of Ω_+^k and Ω_-^k shows the opposite expression patterns. We can interpret that all genes in Ω_+^k are expressed in combination with ones in Ω_-^k. The two sets are negatively correlated each other. Figure 5 displays $\hat{f}_{k'}(\boldsymbol{x}_j)$, $\hat{f}_{h'}(\boldsymbol{x}_j)$, $j \in \{1, \cdots, N\}$ which were chosen according to the minus entropy of the component memberships. The expression patterns given by the genes in Ω_+^k and Ω_-^k, $k \in \{1, 2\}$ (top two of five sets in Figure 4) are compressed into $\hat{f}_{k'}(\boldsymbol{x}_j)$, $\hat{f}_{h'}(\boldsymbol{x}_j)$, $j \in \{1, \cdots, N\}$. This plot provides the evidence for the presence of three subclasses in the leukemia tissues.

6. Concluding Remarks

When we cluster tissue samples on the basis of genes, the number of observations to be grouped is typically much smaller than the dimension of feature vector. In such a case, the applicability of the conventional mixture model-based approach is limited. In this paper, we have shown the method of the mixed factors analysis. The mixed factors model presents a parsimonious parameterization of Gaussian mixture. Consequently, our approach enables us to
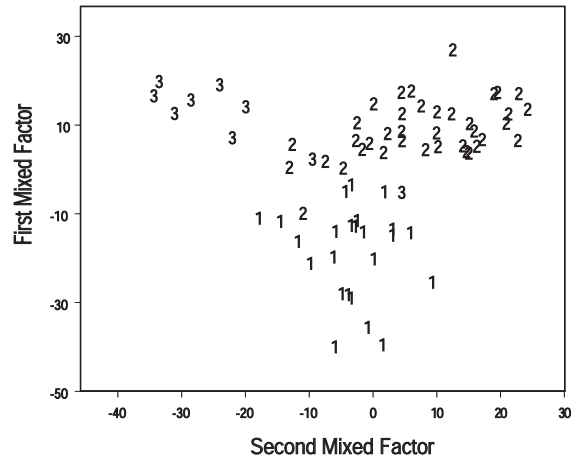
Figure 5. Plot of $\hat{f}_{k'}(\boldsymbol{x}_j)$, $\hat{f}_{h'}(\boldsymbol{x}_j)$ given by the mixed factors model with $G = 3$ and $q = 5$. The true groups are labeled as AML $\rightarrow 1$, B-cell ALL $\rightarrow 2$ and T-cell ALL $\rightarrow 3$.

prevent from the occurrence of overfitting during the density estimation process. In the process of clustering, the method automatically reduce the dimensionality of feature data, to extract genes to be relevant to explain the presence of groups and also to detect genes that are expressed in combination. The mixed factors analysis was applied to a well-known dataset, the leukemia data for highlighting its usefulness. The density estimation succeeded in spite of that we used the 3571-dimensional feature data and the clustering produced the biologically meaningful groups of leukemia tissues. The results showed the potential usefulness and the role of our method in the application to microarray datasets. We expect that the mixed factors analysis will contribute to find molecular subtypes of disease.

Acknowledgements

References

[1] H. Akaike, A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, 19(30A), pp. 9–14, 1974.

[2] A.A. Alizadeh, M.B. Eisen, R.E. Davis, C.Ma, I.S. Lossos, A. Rosenwald, J.C. Boldrick, H. Sabet, T. Tran, X. Yu, J.I. Powell, L. Yang, G.E. Marti, T. Moore, J. Hudson, L. Lu, R. Lewis, D.B Tibshirani, G. Sherlock, W.C. Chan, T.C. Greiner, D.D. Weisenberger, J.O. Armitage, R. Warnke, and

L.M. Staudt, Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature*, 403, pp. 503–511, 2000.

[3] T.W. Anderson, *An Introduction to multivariate statistical analysis*, Wiley, New York, 1984.

[4] Y. Barash, and N. Friedman, Context-specific Bayesian clustering for gene expression data, *Proceedings of the Fifth Annual International Conference on Computational Biology*, pp. 22–25, 2001.

[5] M. Bittener, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, N. Sampas, E. Dougherty, E. Wang, F. Marincola, C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, J. Carpten, E. Gillanders, D. Leja, K. Dietrich, C. Beaudry, M. Berens, D. Alberts, and V. Sondak, Molecular classification of cutaneous malignant melanoma by gene expression profiling, *Nature*, 406, pp. 536–540, 2000.

[6] W.C. Chang, On using principal components before separating a mixture of two multivariate normal distributions, *Applied Statistics*, 32, pp. 267–275, 1983.

[7] M.L. Chow, E.J. Moler, and I.S. Mian, Identifying marker genes in transcription profiling data using a mixture of feature relevance experts, *Physiol. Genomics*, 5, pp. 99–111, 2001.

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

[9] S. Dudoit, J. Fridlyand, and T.P. Speed, Comparison of discrimination methods for the classification of tumors using gene expression data, *Journal of the American Statistical Association*, 97(457), pp. 77–87, 2002.

[10] C. Fraley and A.E. Raftery, Model-based clustering, discriminant analysis, and density estimation, *Journal of the American Statistical Association*, 97(458), pp. 611–631, 2002.

[11] D. Ghosh and A.M. Chinnaiyan, Mixture modeling of gene expression data from microarray experiments, *Bioinformatics*, 18(2), pp. 275–286, 2002.

[12] T.R. Goloub, D.K. Slonim, P. Tamayo, C. Huard, M. Gassenbeck, J.P. Mersirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, 286, pp. 531–537, 1999.

[13] I. Holmes and W.J. Bruno, Finding regulatory elements using joint likelihoods for sequence and expression profile data, *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pp. 19–23, 2000.

[14] E.L. Lehmann, *Theory of point estimation*, Wiley, New York, 1983.

[15] E. L. Lehmann, *Testing statistical hypotheses*, Wiley, New York, 1986.

[16] G.J. McLachlan, R.W. Bean, and D. Peel, A mixture model-based approach to the clustering of microarray expression data, *Bioinformatics*, 18(3), pp. 413–422, 2002.

[17] G.J. McLachlan and D. Peel, *Finite Mixture Models*, Wiley, New York, 1997.

[18] B.D. Ripley, *Pattern recognition and neural networks*, Cambridge University Press, Cambridge, 1996.

[19] D.T. Ross, U. Scherf, M.B. Eisen, C.M. Perou, P. Spellman, V. Iyer, S.S. Jeffrey, M.V. deRijin, M. Waltham, A. Pergamenshikov, C.F. Lee, D. Lashkari, D. Shalon, T.G. Myers, J.N. Weinstein, D. Botstein, and P.O. Brown, Systematic variation in gene expression patterns in human cancer cell lines, *Nature Genetics*, 24, pp. 227–234, 2000.

[20] G. Schwartz, Estimating the dimension of a model, *Annals of Statistics*, 6, pp. 461–464, 1978.

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