A Penalized Likelihood Estimation on Transcriptional Module-based Clustering

Ryo Yoshida[1], Seiya Imoto[2] and Tomoyuki Higuchi[1]

[1] Institute of Statistical Mathematics 4-6-7 Minami-Azabu, Minato-ku, Tokyo, JPN
[2] Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo, JPN

Abstract. In this paper, we propose a new clustering procedure for high dimensional microarray data. Major difficulty in cluster analysis of microarray data is that the number of samples to be clustered is much smaller than the dimension of data which is equal to the number of genes used in an analysis. In such a case, the applicability of conventional model-based clustering is limited by the occurence of overlearning. A key idea of the proposed method is to seek a linear mapping of data onto the low-dimensional subspace before proceeding to cluster analysis. The linear mapping is constructed such that the transformed data successfully reveal clusters existed in the original data space. A clustering rule is applied to the transformed data rather than the original data. We also establish a link between this method and a probabilistic framework, that is, a penalized likelihood estimation of the mixed factors model. The effectiveness of the proposed method is demonstrated through the real application.

1 Introduction

Microarray dataset is a collection of microarray experiments, $\mathbf{x}_j \in \mathbb{R}^d$, $j \in \{1, \cdots, N\}$ in which each experiment represents the expression levels of d genes corresponding to the jth sample. Usually, microarray dataset has a fairly small sample size N, typically less than one hundred, whereas the number of genes involved is more than several thousands. Cluster analysis of microarray has been considered as a challenge to the automated search for molecular subtypes of desease. In view of statistics, major difficulty in this problem is that the number of samples to be clustered is much smaller than that of genes, i.e. $N << d$. This fact limits the applicability of conventional model-based (or distance-based) clustering by the occurence of overlearning. For instance, clustering based on the Gaussian mixture model, which also includes the K-means clustering as a special case, usually leads to the overfitting during the density estimation process with $N << d$. In this article, a new procedure is proposed to overcome such intractability inherent in microarray studies.

The goal of cluster analysis is to partition a set of N samples $\{\mathbf{x}_j\}_{j=1}^N$ into G-nonoverlapping clusters $\{\mathcal{P}_g\}_{g=1}^G$, such that those in a particular cluster are cohesive and separated from those in other clusters. This problem amounts to

estimating the vector of G-unknown class labels $\mathbf{c}(\mathbf{x}_j)^T = (c_1(\mathbf{x}_j), \cdots, c_G(\mathbf{x}_j))$, $j \in \{1, \cdots, N\}$:

$$c_g(\mathbf{x}_j) = \begin{cases} 1 & \text{if } \mathbf{x}_j \in \mathcal{P}_g \\ 0 & \text{otherwise.} \end{cases}$$

The estimation of $\{\mathbf{c}(\mathbf{x}_j)\}_{j=1}^N$ is achieved by constructing a suitable classifier $\hat{\mathbf{c}}(\mathbf{x}_j) = \{\hat{c}_g(\mathbf{x}_j)\}_g$ which declares the assignment of the jth sample to the gth cluster by $\hat{c}_g(\mathbf{x}_j) = 1$ and $\hat{c}_h(\mathbf{x}_j) = 0$ for $h \neq g$.

Unfortunately, constructing the clustering rule as defined over $\mathbb{R}^d$ is very hard with $N << d$ as the finite mixture model leads to the overfiting during the density estimation. Reducing the dimension of data, that is construction of a mapping of data onto the low-dimensional subspace, has been considered as a key issue in microarray study. In this article we consider to seek a linear mapping of data onto the low-dimensional subspace, $\mathbf{P}^T\mathbf{x}_j \in \mathbb{R}^q$ as with $q << d$ before proceeding to cluster analysis:

$$\mathbf{P}^T\mathbf{x}_j = \{\mathbf{p}_k^T\mathbf{x}_j\}_k.$$

Here, the $\mathbf{p}_k$ stands for the kth column of $\mathbf{P}$. Then, the corresponding classifier is defined over $\mathbb{R}^q$ rather than $\mathbb{R}^d$:

$$\hat{\mathbf{c}}(\mathbf{x}_j) \equiv \hat{\mathbf{c}}(\mathbf{P}^T\mathbf{x}_j), \quad \mathbf{x}_j \in \mathbb{R}^d.$$

Hereafter, we implicitly assume a correspondence between the q-mappings of data and the transcriptional module genes as each direction $\mathbf{p}_k$ plays a role to correct up the gene expression patterns in a transcriptional module. In this sense, we call the clustering system based on a linear mapping the transcriptional module-based clustering.

In clustering context, the q-directions $\{\mathbf{p}_k\}_{k=1}^q$ should be chosen such that the transformed data $\{\mathbf{P}^T\mathbf{x}_j\}_{j=1}^N$ successfully reveal the clusters existed in $\mathbb{R}^d$. Then we can identify the clusters based on the lower-dimensional dataset. These two tasks are formulated as the statistical estimation for $\{\mathbf{c}, \mathbf{P}\}$. This problem amounts to an optimization problem that minimizes a loss function $Q(\mathbf{c}, \mathbf{P})$ with respect to the unknown encoders function $\mathbf{c}(\mathbf{x})$ and the q-directions $\{\mathbf{p}_k\}_{k=1}^q$. One of the key results in this study is to establish a link these two processes, i.e. the dimension reduction of data and the clustering algorithm, and a probabilistic framework. In this context, the optimization for $\min_{c,P} Q(\mathbf{c}, \mathbf{P})$ is converted into a penalized likelihood estimation of a probability model of which we call the mixed factors model. Such formulation gives us a great deal of utilities, in either the computation for finding $\min_{c,P} Q(\mathbf{c}, \mathbf{P})$ and the determination of the number of clusters and the appropriate dimension of projected data space, $\{G, q\}$, respectively.

The rest of this article is organized as follows. In section 2 we introduce two criteria to be minimized in the construction of linear mapping. In section 3, we will define a generalized loss function that links the two criteria introduced in section 2. Section 4 presents a probabilistic formulation of this approach. Section

5 present an optimization algorithm for minimizing the proposed generalized loss function. Section 6 contains the determination of the number of clusters and some another parameters. In section 7, the effectiveness of our method will be demonstrated thorough the application to a well-known microarray data, the small round cell tumors of childhood. Finally, the concluding remarks are give in Section 8.

2 Clustering based on Linear Mapping

2.1 Principal Component Analysis

Principal component analysis (PCA, [1]) is one of the most commonly used techniques for constructing a linear mapping of data in statistical data analysis including bioinformatics ([4],[5]). PCA determines the q-directions $\{\mathbf{p}_k\}_{k=1}^q$ to minimize the negative variance of $\{\mathbf{P}^T\mathbf{x}_j\}_{j=1}^N$ with taking account $||\mathbf{p}_k||^2 = 1$, $k \in \{1, \cdots, q\}$. Thus, the objective function to be minimized is

$$Q_A(\mathbf{P}) := -\frac{1}{N}\sum_{k=1}^{q}\sum_{j=1}^{N}\mathbf{p}_k^T\mathbf{x}_j\mathbf{x}_j^T\mathbf{p}_k + \sum_{k=1}^{q}\lambda_k(||\mathbf{p}_k||^2 - 1).$$

where the $\{\lambda_k\}_{k=1}^q$ denote the Lagrange multipliers to impose $||\mathbf{p}_k||^2 = 1$, $k \in \{1, \cdots, q\}$. Here, we assume that the origin of $\{\mathbf{x}_j\}_{j=1}^N$ has been shifted to zero by subtracting the sample mean from all samples.

The optimal q-directions $\{\hat{\mathbf{p}}\}_{k=1}^q$ are equal to the q-principal axes of the sample covariance matrix corresponding to the dominant eigenvalues $\{\hat{\lambda_k}\}_{k=1}^q$. However, as was remarked by some literatures, PCA sometimes fails to reveal the presence of clusters shown by the original data [2, 7]. For instance, when the within-cluster variance on a particular cluster largely dominates the between-clusters variance, a direction tends to the principal axis corresponding to one clusters [2, 7]. Most such limitation are related to the fact that PCA only takes into consideration the second order characteristic of data.

2.2 Within-Cluster Variances

Alternatively, consider to seek a linear mapping to minimize the overlaps of clusters revealed onto $\mathbb{R}^q$. Let us define a loss function to be the Euclid distance between $\mathbf{P}^T\mathbf{x}$ and the unknown centroids $\{\mu_g\}_{g=1}^G$ of G-clusters:

$$L(\mathbf{P}, \mu; \mathbf{x}) := \sum_{g=1}^{G} c_g(\mathbf{x})||\mathbf{P}^T\mathbf{x} - \mu_g||^2, \quad \mathbf{x} \in \mathbb{R}^d. \tag{1}$$

Hereafter we stand for the true distribution of data by $f(\mathbf{x})$. Besides, we also denote the conditional distribution of $\mathbf{c}(\mathbf{x})$ by

$$f(\mathbf{c}(\mathbf{x})|\mathbf{x}) := \prod_{g=1}^{G} w_g(\mathbf{x})^{c_g(\mathbf{x})},$$

where the unknown functionals $\mathbf{w}(\mathbf{x}) = \{w_g(\mathbf{x})\}_g$ satisfy

$$\{w_g(\mathbf{x}) \geq 0\}_{g=1}^{G}, \quad \sum_{g=1}^{G} w_g(\mathbf{x}) = 1, \quad \mathbf{x} \in \mathbb{R}^d.$$

Taking the expectation of (1) with respect to $f(\mathbf{x}, \mathbf{c}) = f(\mathbf{x})f(\mathbf{c}(\mathbf{x})|\mathbf{x})$ defines a risk function to be minimized in the construction of estimators for $\{\mathbf{c}(\mathbf{x}), \mathbf{P}\}$ although the true distribution of $\{\mathbf{x}, \mathbf{c}(\mathbf{x})\}$ is unknown. Instead, replacing $f(\mathbf{x})$ by the empirical distribution $\hat{f}(\mathbf{x})$, we can obtaine an empirical loss function

$$\begin{aligned} Q_B(\mathbf{w}, \mathbf{P}, \mu) :&= E_f L(\mathbf{P}, \mu; \mathbf{x}) \\ &= \frac{1}{N} \sum_{g=1}^{G} \sum_{j=1}^{N} w_g(\mathbf{x}_j) ||\mathbf{P}^T \mathbf{x}_j - \mu_g||^2 - \sum_{k=1}^{q} \lambda_k (||\mathbf{p}_k||^2 - 1). \quad (2) \end{aligned}$$

Here the $\{\lambda_k\}_{k=1}^{q}$ denote the Lagrange multiplier. The first term in (2) presents just the within-cluster variances of $\{\mathbf{p}_k^T \mathbf{x}_j\}_{j=1}^{N}$, $k \in \{1, \cdots, q\}$. An optimal $\hat{\mathbf{P}}$ minimizes the overlap of G-clusters revealed onto the $\mathbb{R}^q$ although the conditional distributions $\{\mathbf{w}(\mathbf{x}_j)\}_{j=1}^{N}$ and the G-centroids $\{\mu_g\}_{g=1}^{G}$ remain to be unknown. The optimization method will be described in later under more general setting.

3 Generalized Criterion

While the minimum within-cluster variances are a suitable criterion in the construction of linear mapping to reflect the group structure of original dataset, its applicability might be limited due to the dimensionality of the data. Most limitations are related to the occurence of overlearning. Such unsuitableness occurs due to the fact that the N data points are sparsely distributed on $\mathbb{R}^d$. Then, the degree of freedom in the determination of $\{\mathbf{p}_k\}_{k=1}^{q}$ is extreamly large. Accordingly, the compressed samples $\{\mathbf{P}^T \mathbf{x}_j\}_{j=1}^{N}$ might improperly exhibit the clusters despite no clusters on $\mathbb{R}^d$.

To overcome such limitation, we propose a criterion for estimating parameters by combining the score functions $Q_A(\mathbf{P})$ and $Q_B(\mathbf{w}, \mathbf{P}, \mu)$ of the form

$$\begin{aligned} Q_\alpha(\mathbf{w}, \mathbf{P}, \mu) = &-\frac{1}{N} \sum_{j=1}^{N} ||\mathbf{P}^T \mathbf{x}_j||^2 + \frac{\alpha}{N} \sum_{g=1}^{G} \sum_{j=1}^{N} w_g(\mathbf{x}_j) ||\mathbf{P}^T \mathbf{x}_j - \mu_g||^2 \\ &+ \sum_{k=1}^{q} \lambda_k (||\mathbf{p}_k||^2 - 1), \end{aligned} \quad (3)$$

where $\alpha \in [0, 1]$ is a mixing rate that controls the trade-off between the total variance and the between-clusters variance. Here the $\{\lambda_k\}_{k=1}^{q}$ denote the Lagrange multipliers for taking account $\{||\mathbf{p}_k||^2 = 1\}_{k=1}^{q}$. Notice that for any $\{\mathbf{w}(\mathbf{x}_j), \mathbf{P}\}$,

the minimization of (3) with respect to the G-centroids is accomplished by the weighted average of $\{\mathbf{P}^T\mathbf{x}_j\}_{j=1}^N$:

$$\hat{\mu}_g = \frac{1}{N\bar{w}_g}\sum_{j=1}^{N} w_g(\mathbf{x}_j)\mathbf{P}^T\mathbf{x}_j, \tag{4}$$

where $\bar{w}_g = (1/N)\sum_{j=1}^{N} w_g(\mathbf{x}_j)$. Equating $\mu_g = \hat{\mu}_g$ for $g \in \{1, \cdots, G\}$, the first two terms in (3) can be rewritten as

$$-\frac{1}{N}\sum_{j=1}^{N}||\mathbf{P}^T\mathbf{x}_j||^2 + \frac{\alpha}{N}\sum_{g=1}^{G}\sum_{j=1}^{N} w_g(\mathbf{x}_j)||\mathbf{P}^T(\mathbf{x}_j - \bar{\mathbf{x}}_g)||^2. \tag{5}$$

where the $\{\bar{\mathbf{x}}_g\}_{g=1}^G$ denote the group means corresponding to $\{\mathbf{x}_j\}$,

$$\bar{\mathbf{x}}_g = \frac{1}{N\bar{w}_g}\sum_{j=1}^{N} w_g(\mathbf{x}_j)\mathbf{x}_j, \quad g \in \{1, \cdots, G\}. \tag{6}$$

As $\alpha \to 0$, the quantity (5) tends to the variance of $\{\mathbf{P}^T\mathbf{x}_j\}_{j=1}^N$, and then, the optimal $\{\hat{\mathbf{p}}_k\}_{k=1}^q$ tends to the principal axes. To the contrary, as $\alpha \to 1$, the (5) tends to the negative between-clusters variance of $\{\mathbf{P}^T\mathbf{x}_j\}_{j=1}^N$:

$$-\text{trace}\left(\mathbf{P}^T\sum_{g=1}^{G}\bar{\mathbf{x}}_g\bar{\mathbf{x}}_g^T\mathbf{P}\right) = -\text{trace}\sum_{g=1}^{G}\hat{\mu}_g\hat{\mu}_g^T.$$

Then, a linear mapping of data with the optimal q-directions tends to separate the G-centroids.

Next, consider a computational aspect in the construction of the optimal q-directions. Differentiating (3) with respect to $\mathbf{p}_k$ with equating $\mu_g = \hat{\mu}_g$ leads to an equation to be solved,

$$\left[\frac{1}{N}\sum_{j=1}^{N}\mathbf{x}_j\mathbf{x}_j^T - \frac{\alpha}{N}\sum_{g=1}^{G}\sum_{j=1}^{N} w_g(\mathbf{x}_j)(\mathbf{x}_j - \bar{\mathbf{x}}_g)(\mathbf{x}_j - \bar{\mathbf{x}}_g)^T - \lambda_k\mathbf{I}\right]\mathbf{p}_k = \mathbf{0}. \tag{7}$$

Obviously, the solutions can be given by the corresponding eigenvalues which satisfy

$$\hat{\lambda}_k = \mathbf{p}_k^T\left[\frac{1}{N}\sum_{j=1}^{N}\mathbf{x}_j\mathbf{x}_j^T - \frac{\alpha}{N}\sum_{g=1}^{G}\sum_{j=1}^{N} w_g(\mathbf{x}_j)(\mathbf{x}_j - \bar{\mathbf{x}}_g)(\mathbf{x}_j - \bar{\mathbf{x}}_g)^T\right]\mathbf{p}_k. \tag{8}$$

Thus, the $\min Q_\alpha(\mathbf{w}, \mathbf{P}, \mu)$ with any fixed third arguments $\{\mathbf{w}\}$ can be attained at $Q^* = -\sum_{k=1}^{q}\hat{\lambda}_k$ with a series of the dominant eigenvalues, $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq, \cdots, \geq \hat{\lambda}_q$: the optimal $\hat{\mathbf{p}}_1$ corresponds to the largest $\hat{\lambda}_1$, and the rest of directions $\{\hat{\mathbf{p}}_k\}_{k=2}^q$ are orthogonal to all of the preceding ones.

Given $\{\mathbf{P}, \mu\}$, the estimation of $\{\mathbf{w}(\mathbf{x}_j)\}_{j=1}^N$ can be accomplished by the K-means-like rule: the solution will put unit value on $w_g(\mathbf{x}_j)$ with a smallest distance between $\hat{\mathbf{P}}^T\mathbf{x}_j$ and $\{\hat{\mu}_g\}_{g=1}^G$. In this article, we generalized this type of clustering, i.e. the hard clustering, to the soft clustering. This goal can be accomplished by imposing a smoothness on the $\{w_g(\mathbf{x})\}_g$. One such penalization is the negative entropy of $\{w_g(\mathbf{x})\}_g$:

$$H(\mathbf{w}(\mathbf{x})) := \sum_{g=1}^{G} w_g(\mathbf{x}) \log w_g(\mathbf{x}).$$

This function achieves the minimum value for equal values $w_g(\mathbf{x}) = 1/G$, $g \in \{1, \cdots, G\}$, and is correspondingly larger as the $\{w_g(\mathbf{x})\}_{g=1}^G$ tends to more unequal. Consequently, the modified criterion becomes

$$Q_\alpha(\mathbf{w}, \mathbf{P}, \mu) - \beta \sum_{j=1}^{N} H(\mathbf{w}(\mathbf{x}_j)). \tag{9}$$

The quantity $\beta \geq 0$ controls the strength of penalty that tunes the trade-off between soft and hard clustering. The optimal $\{\hat{w}_g(\mathbf{x}_j)\}_{g=1}^G$ for this objective function turns to

$$\hat{w}_g(\mathbf{x}_j) \propto \exp\Big(-\frac{\alpha}{\beta} ||\mathbf{P}^T\mathbf{x}_j - \mu_g||^2 \Big). \tag{10}$$

for all $j \in \{1, \cdots, N\}$. This solution puts the increased weight on a particular group to be the smallest $||\mathbf{P}^T\mathbf{x}_j - \mu_g||^2$ as β tends to small, and setting $\beta \to \infty$ places the equal weights on all groups. Correspondingly, our approach alternates between two steps, solving (7) and the grouping (10) with an initial starting value until a series of the corresponding Q_α is in convergence.

We will revist the computational aspect of this method in section 5. The proposed optimization algorithm can be implemented without solving the eigenvalues equation that might be computationaly very demanding in microarray study. The remined tasks are the determination of smoothness $\{\alpha, \beta\}$ and a suitable q on which data are mapped. Moreover, the number of clusters G must often be deduced from data. As will be shown in section 6, these tasks can be converted into the statistical model selection throught the probabilistic formulation of the method.

4 Probabilistic Formulation

We now discuss the proposed clustering method within a probabilistic framework. Let $\mathbf{f}_j \in \mathbb{R}^q$ be a latent random variable corresponding to the jth sample where q is much smaller than d. Then, suppose that a set $\{\mathbf{x}_j, \mathbf{f}_j\}_{j=1}^N$ is independently distributed according to

$$\mathbf{x}_j = \mathbf{P}\mathbf{f}_j + \epsilon_j, \tag{11}$$

$$\mathbf{f}_j|c_g(\mathbf{x}_j)=1 \sim N(\mu_g, \sigma\mathbf{I}), \quad g \in \{1,\cdots,G\}, \tag{12}$$

where the ϵ_j is assumed to be Gaussian noise with $N(0,\gamma\mathbf{I})$ and to be independent to $\mathbf{f}_j$. The observational equation (11) states that for a given $\mathbf{f}_j$, the $\mathbf{x}_j$ is distributed to be $N(\mathbf{P}\mathbf{f}_j, \gamma\mathbf{I})$. Accordingly, this generative model also states the distribution of data conditional on the class label by

$$\mathbf{x}_j|c_g(\mathbf{x}_j)=1 \sim N(\mathbf{P}\mu_g, \sigma\mathbf{P}\mathbf{P}^T+\gamma\mathbf{I}), \quad g \in \{1,\cdots,G\}.$$

Thus, the distributional aspect of data is characterized by G-clusters centered at $\{\mathbf{P}\mu_g\}_{g=1}^G$. This model, called the mixed factors model, was originally proposed by Yoshida et al. [8] to intend a parsimonious parameterization of the Gaussian mixture.

As the preceding method imposes the orthogonality on the q-directions, we now assume the orthogonality of q-columns in the loading matrix $\mathbf{P}=\{\mathbf{p}_k\}_k$. Then the logged-density of $\mathbf{x}$ can be written as

$$\log P(\mathbf{x}|\mathbf{c}(\mathbf{x})) = const. - \frac{1}{\gamma}(||\mathbf{x}||^2 - ||\mathbf{P}^T\mathbf{x}||^2) - \frac{1}{\gamma+\sigma}\sum_{g=1}^{G} c_g(\mathbf{x})||\mathbf{P}^T\mathbf{x}-\mu_g||^2 \tag{13}$$

Taking the expectation of (13) with respect to the empirical distribution $\hat{f}(\mathbf{x})$ and the conditional distribution of unknown class labels $f(\mathbf{c}(\mathbf{x})|\mathbf{x})$ leads to the log-likelihood function of unknown parameters $\{\mathbf{w},\mathbf{P},\mu\}$ after multiplying (13) by γ:

$$L(\mathbf{w},\mathbf{P},\mu) = const. + \frac{1}{N}\sum_{j=1}^{N}||\mathbf{P}^T\mathbf{x}_j||^2 - \frac{\alpha}{N}\sum_{g=1}^{G}\sum_{j=1}^{N} w_g(\mathbf{x}_j)||\mathbf{P}^T\mathbf{x}_j-\mu_g)||^2,$$

where $\alpha=\gamma/(\gamma+\sigma)$. Adding both the regularization term $\beta\sum_{j=1}^{N}H(\mathbf{x}_j)$ and the Lagrange terms $-\sum_{k=1}^{q}\lambda_k(||\mathbf{p}_k||-1)$ to this function gives a criterion equivalent to the negative of (9). This implies that the problem to be solved in our method turns to a penalized likelihood estimation of the mixed factors model.

5 Maximization-Maximization Algorithm

Here, we present an optimization algorithm to maximize the penalized likelihood of the mixed factors model, that is equivalent to find the minimizer of (9). This can be achieved by the EM algorithm (Dempster et al.[3]). The EM algorithm takes $\{\mathbf{x}_j,\mathbf{f}_j\}_{j=1}^N$ as a complete data set and then alternates between the two steps: the expectation of the complete data likelihood with respect to the posterior distribution of the unknown factors $\{\mathbf{f}_j\}_{j=1}^N$,

$$P(\mathbf{f}_j|c_g(\mathbf{x}_j),\mathbf{x}_j) = \phi(\mathbf{f}_j;\alpha\mu_g+(1-\alpha)\mathbf{P}^T\mathbf{x}_j, \lambda\alpha\mathbf{I}),$$

and the maximization of the expected complete data likelihood. Hereafter, we let $\phi(\ ;\mathbf{a},\mathbf{B})$ be the Gaussian density with the mean $\mathbf{a}$ and the covariance matrix $\mathbf{B}$.

Consider now to update the hth direction $\mathbf{p}_h$ whereas the another parameters are fixed at the current values $\{\hat{\mathbf{p}}_k\}_{k\neq h}$, $\{\hat{\mu}_g\}_g$ and $\{\hat{\mathbf{w}}(\mathbf{x}_j)\}_j$. By the definition, the complete data log-likelihood of the mixed factors model, $L_c = (1/N)\sum_j \log P(\mathbf{x}_j, \mathbf{f}_j|\mathbf{c}(\mathbf{x}_j))$, can be explicitly represented by

$$L_c(\mathbf{w}, \mathbf{P}, \mu) := \frac{1}{N}\sum_{j=1}^{N} \phi(\mathbf{x}_j; f_{hj}\mathbf{p}_h + \sum_{h\neq k} f_{kj}\hat{\mathbf{p}}_k, \gamma\mathbf{I}) + \frac{1}{N}\sum_{g=1}^{G}\sum_{j=1}^{N} \hat{w}_g(\mathbf{x}_j)\phi(\mathbf{f}_j; \hat{\mu}_g, \sigma\mathbf{I}), \quad (14)$$

where f_{hj} is the hth element of $\mathbf{f}_j$. Note that the hth direction depends only the first term in (14) which corresponds to the observational equation (11).

Let $\langle L_c \rangle$ be the conditional expectation of (14) where the expectation is taken with respect to the $P(\mathbf{f}_j|c_g(\mathbf{x}_j), \mathbf{x}_j)$ evaluated with the current parameters $\{\hat{\mathbf{w}}, \hat{\mathbf{P}}, \hat{\mu}\}$. Then the objective function to be maximized at this step is

$$\langle L_c \rangle + \eta_h(||\mathbf{p}_h||^2 - 1) + \sum_{k\neq h} \eta_k \hat{\mathbf{p}}_k^T \mathbf{p}_h.$$

Here, the $\{\eta_k\}_{k=1}^q$ are the Lagrange multipliers to impose the orthogonality on the q-directions. Solving this gives the optimal $\hat{\mathbf{p}}_h$ as

$$\hat{\mathbf{p}}_h = \frac{1}{S}\Big[\sum_{j=1}^{N}\langle f_{hj}\rangle \mathbf{x}_j - \sum_{k\neq h}\hat{\mathbf{p}}_k^T \sum_{j=1}^{N}\langle f_{hj}\rangle \mathbf{x}_j \hat{\mathbf{p}}_k\Big],$$

where the S denotes the normalizing constant to satisfy $||\hat{\mathbf{p}}_h||^2 = 1$, and the conditional expectation of the latent variables, $\langle f_{hj}\rangle$, is equal to the hth element of

$$\langle \mathbf{f}_j \rangle = \alpha \sum_{g=1}^{G} \mu_g + (1-\alpha)\mathbf{P}^T \mathbf{x}_j.$$

Repeating this process for $h \in \{1, \cdots, q\}$, we would have a series of q-directions $\{\hat{\mathbf{p}}_k\}_{k=1}^q$.

Given an estimate of q-directions, the G-centroids of clusters and the conditional distribution of class labels, $\{w_g(\mathbf{x}_j)\}_g$, for $j \in \{1, \cdots, N\}$ are estimated by (6) and (10), respectively. Thus, we just compute the simple recursive formulas until the sequence of estimates and the corresponding penalized likelihood are judged to be converged. Such sequence of parameters yields a non-decreasing sequence of the penalized likelihood of the mixed factors model. To sum up, we summarize this algorithm in below:

1. Set the initial values for $\{\hat{\mathbf{w}}, \hat{\mathbf{P}}, \hat{\mu}\}$ and $\{G, q, \alpha, \beta\}$. Then repeat the step 2 to 4 until the sequence of either parameters and the corresponding penalized likelihood will be converged:

2. (q-directions)
   for $h = 1$ to q, update $\hat{\mathbf{p}}_h$ by

$$\hat{\mathbf{p}}_h = \frac{1}{S}\Big[\sum_{j=1}^{N}\langle f_{hj}\rangle \mathbf{x}_j - \sum_{k\neq h}\hat{\mathbf{p}_k}^T \sum_{j=1}^{N}\langle f_{hj}\rangle \mathbf{x}_j\hat{\mathbf{p}}_k\Big].$$

3. (G-centroids) for $g = 1$ to G, update $\hat{\mu}_g$ by

$$\hat{\mu}_g = \frac{1}{N\bar{w}_g}\sum_{j=1}^{N}\hat{w}_g(\mathbf{x}_j)\hat{\mathbf{P}}^T\mathbf{x}_j.$$

4. (Grouping function) for $g = 1$ to G and $j = 1$ to M, update $\hat{w}_g(\mathbf{x}_j)$ by

$$\hat{w}_g(\mathbf{x}_j) \propto \exp\Big(-\frac{\alpha}{\beta}||\hat{\mathbf{P}}^T\mathbf{x}_j - \hat{\mu}_g||^2\Big).$$

Notice that we have no need to evaluate the noise variances $\{\gamma, \sigma\}$ of the mixed factors model during this procudure. However, the model selection method described in next section requires the evaluation of these parameters. It follows from $\alpha = \gamma/(\gamma + \sigma)$ that the σ can be estimated by $\hat{\sigma} = (1 - \alpha)\hat{\gamma}/\alpha$ with a given estimate $\hat{\gamma}$. By the simple calculation, it can be seen that an optimal $\hat{\gamma}$ necessarily satisfies

$$\hat{\gamma} = \frac{1}{(d-q)N}\sum_{j=1}^{N}\Big(||\mathbf{x}_j||^2 - ||\hat{\mathbf{P}}^T\mathbf{x}_j||^2\Big).$$

6 Penalized Mixed Factors Analysis

A basic issue arising in this method is the determination of the number of clusters G, the suitable dimension of the linear mapping q and the strength of penalties $\{\alpha, \beta\}$. Within statistical framework, this issue can be converted into the model selection problem that chooses a suitable set $\{G^*, q^*, \alpha^*, \beta^*\}$ among the possible combinations. In this article, we address this problem by selecting a particular combination to show the best predictability.

Consider to split $\{\mathbf{x}_j\}_{j=1}^{N}$ into the two disjoint subsets, a training sample set $\{\mathbf{x}_j^e\}_{j=1}^{N_e}$ used in the estimation of parameters and a set of the blinded test sample $\{\mathbf{x}_j^b\}_{j=1}^{N_b}$. Let $\{\hat{\mathbf{w}}^e, \hat{\mathbf{P}}^e, \hat{\mu}^e\}$ be a set of parameters estimated by the training samples with a particular $\{G, q, \alpha, \beta\}$. One possible approach is to select a combination $\{G^*, q^*, \alpha^*, \beta^*\}$ to minimize the prediction error

$$C(\{G, q, \alpha, \beta\}) := -\frac{1}{N_b}\sum_{j=1}^{N_b}\log P(\mathbf{x}_j^b; \{\hat{\mathbf{w}}^e, \hat{\mathbf{P}}^e, \hat{\mu}^e\}),$$

where $P(\mathbf{x}; \{\mathbf{w}, \mathbf{P}, \mu\})$ is the unconditional density of data to be the Gaussian mixture as

$$P(\mathbf{x}; \{\hat{\mathbf{w}}, \hat{\mathbf{P}}, \hat{\mu}\}) = \sum_{g=1}^{G} \frac{1}{G} \phi(\mathbf{x}; \hat{\mathbf{P}}\hat{\mu}_g, \hat{\sigma}\hat{\mathbf{P}}\hat{\mathbf{P}}^T + \hat{\gamma}\mathbf{I}). \tag{15}$$

Given a set $\{G^*, q^*, \alpha^*, \beta^*\}$, our method calibrates G-clusters based on the estimated conditional distribution of class labels $\{w_g(\mathbf{x}_j)\}_g$. The most common classifier is to assign $\mathbf{x}_j$ to a cluster with the highest posterior probability of belonging:

$$\hat{c}_g(\hat{\mathbf{P}}^T\mathbf{x}_j) = \begin{cases} 1 \text{ if } w_g(\mathbf{x}_j) = \max\limits_{h \in \{1,\cdots,G\}} w_h(\mathbf{x}_j), \\ 0 \qquad\qquad \text{otherwise.} \end{cases}$$

Biological interpretation of q-coordinates corresponding to $\{\hat{\mathbf{P}}^T\mathbf{x}_j\}_{j=1}^N$ is important for real data analysis. This can be achieved by investigating the values in q-directions $\{\hat{\mathbf{p}}_k\}_k$. Obviously, a particular element of $\{\mathbf{x}_j\}_{j=1}^N$ had a large contribution in the calibration of clusters if the corresponding element of $|\hat{\mathbf{p}}_k|$ takes a large value. To the contrary, if an element of $|\hat{\mathbf{p}}_k|$ takes a value close to zero, the kth coordinate is not affected by the corresponding gene. In this way, by investigating all values in $\hat{\mathbf{P}}$, each of q-directions can be understood. In practice, it will be helpful to list the top L of genes to give the highest positive values in $\hat{\mathbf{p}}_k$ at Ω_+^k and to give the highest negative values in $\hat{\mathbf{p}}_k$ at Ω_-^k for each $k \in \{1, \cdots, q\}$. As will be demonstrated in next section, for the gene expression analysis, these $2q$ sets can be useful either to find the biologically meaningful groups of genes and to elucidate a causal link from the calibrated clusters to the biological knowledge.

7 Real Application

Khan et al. [6] classified the small round blue cell tumors (SRBCT's) of childhood into the four diagnostic categories, neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin's lymphoma (NHL) and the Ewing family of tumors (EWS) using cDNA gene expression profiles. The dataset is available at the website `http://www.nhgri.nih.gov/DIR/Microarray/Supplement/`. For each of the 83 SRBCT samples, the expression levels of $2,308$ genes were measured. Khan et al. [6] splited the data into two parts; the training set comprising 63 cases (NB, 12; RMS, 20; BL, 8; EWS, 23) and the test set, 20 cases (NB, 6; RMS, 5; BL, 3; EWS, 6) where Burkitt's lymphoma (BL) is a subset of NHL. All samples are summarized in Figure 1. Note that the name of sample specifies the cancer type suffixed with -T for a tumor biopsy material and -C for a cell line. Khan et al. [6] successfully classified the tumor types into the four categories using artificial neural networks. Unlike this, the purpose of our study is to identify the clusters of these SRBCT's in the unsupervaised manner, and then, to look at the association between the caribrated clusters and some medical outcome, that is, the unsupervised learning.

For the preprocessing, we removed genes whose range of expression values across 83 samples is less than 3.0, 680 genes then remain to be analysed. We then adjusted the columns of 680×88 data matrix to have mean zero after centering the rows. To find an optimal $\{\hat{\mathbf{w}}, \hat{\mathbf{P}}, \hat{\mu}\}$, we used the 63 training samples including tumors and cell lines, 13 EWS-T, 10 EWS-C, 12 NB, 8 BL, 10 RWS-T and 10 RWS-C. The 20 blinded samples were used to select $\{G^*, q^*, \alpha^*, \beta^*\}$ and also to assess the predictability of the resulting clusters.

We candidated a set of the number of clusters ranging from $G = 4, 5, 6$ and the dimension onto which the data are mapped varying $q = 1$ to 11. We also candidated a set of possible combinations of the regularization parameters as $\{\alpha, \beta\} \in \{0.1, 0, 2, \cdots, 0.9\} \times \{0.8, 0.9, 1.0, 1.1, 1.2\}$. The smallest local minimum of the generalized criterion corresponding to $q = 8$ gave the minimum scores of $C(\{G, q, \alpha, \beta\})$ for all G in which the most suitable smoothing parameters $\{\alpha^*, \beta^*\}$ were given. Figure 1 shows the groupings given by $G = 4, 5, 6$ fixed at $q = 8$. The calibrated model with $G = 4$ correctly grouped all samples into the diagnostic categories, i.e. EWS, NB, BL and RWS. It also could be seen from $G = 5$ in Figure 1 that the RWS samples were divided into the two subgroups as corresponding to the heterogeneity between RWS-Ts and RWS-Cs. Moreover, the clustering given by $G = 6$ yielded a partition as reflecting the molecular dissimilarity between the tumor samples and cell lines on the EWSs. Indeed, the model selection based on $C(\{G, q, \alpha, \beta\})$ showed the evidence of molecular subtypes on either EMS and RWS as the model of $G = 6$ was judged to be optimal. We also tested the capability of the caribrated clustering rule using the 20 blinded samples (see Figure 1). When these samples were assigned into a particular cluster using the resulting classifiers for each G, we obtained the plausible grouping as likely to reflect the diagnostic categories of cancer types, for all G. For instance, of the 20 blinded samples, TEST20-EWS-T were misclassified into the RMS related category for $G = 4$. In addition, for $G = 6$, 19 of the 20 test samples were correctly grouped into the related diagnostic categories in which TEST19-EWS-T was misclassified into the EWS-C related cluster. From this analysis, the predictability confirm us the effectiveness of the estimated grouping.

A causal link from the clusters to the biological knowledge can be elucidated thorough the inspection of relevant genes. Figure 1 illustrates the expression patterns of 16 set of relevant genes selected by $G = 6$ and $q = 8$. For instance, the genes in Ω^1_+ are good discriminators on the basis of the lack of expression in BL and RMS, and the high expression in EWS and NB. Note also that the genes in Ω^1_- showed the opposite expression patterns to Ω^1_+. The relevant genes in the some sets were expressed in one or two of the six molecular categories as the Ω^5_+, Ω^7_- and Ω^8_+ are specific to NB, EWS and RMS, respectively. Of interest is that the genes in Ω^2_+ are specifically expressed in the tumor samples as EWS-T, RWS-T and and not expressed in the cell lines. The genes in Ω^2_- shows the opposite patterns to Ω^2_+ as the lack of expression in the tumor samples and the high-expression in the cell lines. This fact validates the presence of heterogeneity corresponding to the molecular types within a cancer type.

8 Concluding Remarks

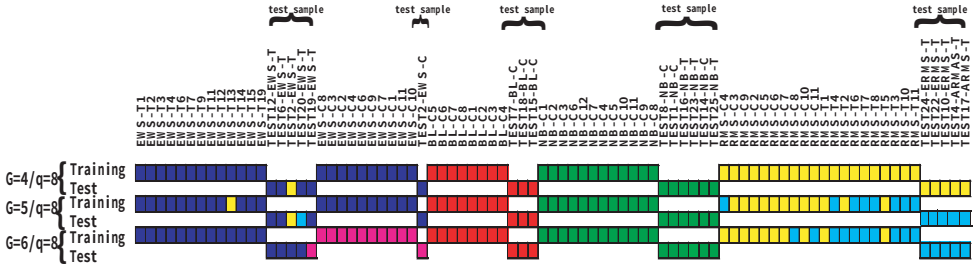
In this study, we proposed a method of clustering for the high-dimensional microarray dataset. A distinction of our method is that the clustering rule is applied to the linear mapping of data onto the low-dimensional subspace, rather than the original dataset. In the construction of linear mapping, the directions are chosen as to minimize a criterion that links the variance and the within-cluster variances of the compressed data. We also established an optimization algorithm to find such directions and a suitable clustering rule.

The effectiveness of the proposed method was demonstrated through the application to a well-known gene expression data, the small round blue cell tumors of childhood (SRBCTs). The clustering system could find the biologically meaningful groups of SRBCTs as we confirmed a plausible correspondence between the calibrated clusters and the diagnostic categories. Besides, the method identified sets of relevant genes associated with the calibrated clusters. These sets might be helpful to elucidate the causal link between the obtained grouping and the existing knowledge on biology.

References

1. Anderson, T.W.: An Introduction to multivariate statistical analysis. Wiley, New York, (1984)
2. Chang, W.C.: On using principal components before separating a mixture of two multivariate normal distributions. Applied Statistics **32** (1983) 267–275
3. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm (with discussion), J. Royal Stat. Soc. B. **39** (1977) 1–38
4. Ghosh, D., Chinnaiyan, A.M.: Mixture modeling of gene expression data from microarray experiments. Bioinformatics **18(2)** (2002) 275–286
5. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gassenbeck, M., Mersirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science **286** (1999) 531–537
6. Khan, J., Wei, J.S., Ringner, M., Saal, L.H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Atonescu, C.R., Peterson, C., Meltzer, P.S.: Classification and Diagnostic Prediction of Cancers using Gene Expression Profiling and Artificial Neural Networks Nature Medicine **7** (2001) 673–679
7. McLachlan, G.J., Peel, D.: Finite mixture models. Wiley New York (1997)
8. Yoshida, R., Higuchi, T., Imoto, S.: A mixed factors model for dimension reduction and extraction of a group structure in gene expression data. Proc. 3rd Computational Systems Bioinformatics (2004) 161-172

A. Clustering (G=4,5,6, q=8)



B. Relevant Genes

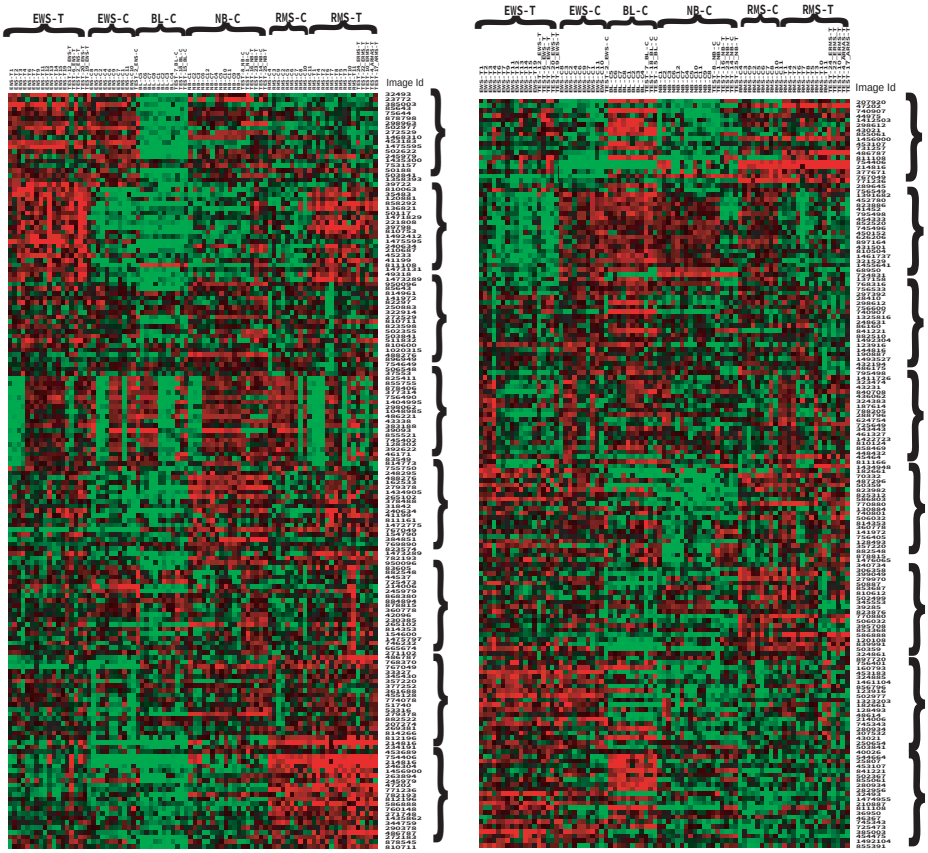


Fig. 1. Caribrated Clusters and relevant genes. **A**. Clustering result. The 63 SRBCTs samples (training samples) were used for finding $\{\hat{\mathbf{w}}, \hat{\mathbf{P}}, \hat{\mu}\}$, and then, the training set and the 20 test samples were grouped into clusters base on the calibrated classifiers for each combination $\{G, q, \alpha, \beta\}$. Shown here are clusters caribrated by the smallest local minima of the generalized criterion corresponding to $q = 8$ for $G = 4, 5, 6$ where the smoothness paramters were induced from the test samples. The resulting groups are depicted by the colors. **B**. Relevant genes selected by the optimal model, $G = 6$, $q = 8$. Shown in the left panel are the expression patterns of the 8 sets of 20 genes listed at Ω_+^k, $k \in \{1, \cdots, 8\}$. The expression patterns of genes listed at Ω_-^k, $k \in \{1, \cdots, 8\}$, are also shown in the right panel.