Kernel Mixture Survival Models for Identifying Cancer Subtypes, Predicting Patient’s Cancer Types and Survival Probabilities

Tomohiro Ando  Seiya Imoto  Satoru Miyano
ando@ims.u-tokyo.ac.jp  imoto@ims.u-tokyo.ac.jp  miyano@ims.u-tokyo.ac.jp

Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan

Abstract

One important application of microarray gene expression data is to study the relationship between the clinical phenotype of cancer patients and gene expression profiles on the whole-genome scale. The clinical phenotype includes several different types of cancers, survival times, relapse times, drug responses and so on. Under the situation that the subtypes of cancer have not been previously identified or known to exist, we develop a new kernel mixture modeling method that performs simultaneously identification of the subtype of cancer, prediction of the probabilities of both cancer type and patient’s survival, and detection of a set of marker genes on which to base a diagnosis. The proposed method is successfully performed on real data analysis and simulation studies.

Keywords: classification, gene selection, kernel method, mixture model, survival analysis

1 Introduction

The DNA microarray technology, to produce expression data on a large number of genes in a parallel fashion, is increasingly being performed in cancer researches [2, 12]. Recently, there is an increasing need to link the expression profiles with various phenotypic data obtained from cancer patients for predicting clinical outcomes; such as cancer types, survival times and drug responses. Under the situation that all different cancer types in a given patient are known in advance, a number of successful applications of statistical learning methods have been reported [14, 15, 17, 22, 23].

Unfortunately, if the cancer subtypes have not been previously identified, the analysis result may lead to the wrong diagnosis and clinical treatment, since most of statistical learning methods are only applicable under the situation that all cancer types are known completely. For example, it is well known that the diffuse large B-cell lymphoma is clinically heterogeneous; patients differently respond to the same treatment and have different survival rates [2, 25, 27].

If different subtypes of cancer are known to exist, the most popular method in relating gene expression profiles to survival phenotypic data is the two stage approach. It first apply the clustering method to group cancer samples into several clusters based on gene expression profiles across many genes. Clustering method has successfully identified clinically relevant cancer subtypes. Based on the clustering result, as the second step, one can apply some statistical learning methods to each cluster [2, 11, 27, 28]. It is obvious that the clustering results depend on the choice of clustering method, and further depends on the choice of the number of clusters. Additionally, they ignore the survival information completely. Thus they may identify cancer subtypes that are uncorrelated with the clinical outcome and not useful for diagnostic purposes.

The main purpose of this paper is to develop a new statistical learning method that simultaneously identifies cancer subtypes and predicts both patient’s cancer type and survival probability in the
contest the mixtures of experts framework [13]. Attention is also focused on the detection of a set of marker genes on which to base a diagnosis. The proposed method jointly utilizes both gene expression profiles and survival phenotypic information and therefore the selected genes are helpful to predict both cancer type and survival time of cancer patient more accurately.

The remainder of this paper is organized as follows. First, we present the details on the proposed statistical learning method. Then we investigate the performance of the proposed method through Monte Carlo simulations. Finally, we discuss the experimental results of the proposed method using diffuse large B-cell lymphoma data [2] and give a conclusion.

2 Method

2.1 A Mixture of Weibull Experts Model

Consider an individual exposed to $G$ cancer types, where $G$ is assumed to be unknown. If all different cancer types are already known in a given patient, $G$ would be a fixed value. The probabilistic event of interest is the random pair $(g, t)$, where $g$ takes a value from the set $\{1, \ldots, G\}$ to indicate the type of cancer and $t$ is a non-negative random variable representing the individual’s survival or censored time.

We assume that each individual belongs to one of the clinically relevant $G$ type of cancer, and their cancer types $g$ and survival or censored times $t$ are characterized by the $p$-dimensional gene expression profiles $x = (x_1, \ldots, x_p)'$, which is the subset of original high-dimensional gene expression data.

In the context of mixtures of expert framework [13], the cancer type $g$ is chosen from a multinomial distribution with probability $\pi_j(x)$ such that $\sum_{j=1}^{G} \pi_j(x) = 1$, and for the cancer type $j$, the survival time $t$ is generated with the survival probability function $S_j(t|x) = \Pr(T > t|x, g = j)$. The total survival probability function of $t$, conditioned on $x$, is given by the mixture density

$$S(t|x) = \sum_{j=1}^{G} \pi_j(x)S_j(t|x).$$

A variety of survival probability functions can be used for $S_j(t|x)$, for, e.g., exponential, exponential Weibull, Weibull, gamma, log-normal and so on. One major choice would be the exponential model, however, the assumption of a constant hazard function is too restrictive. In this paper, we shall use Weibull model, which is more suitable for representing the survival times and is widely used in medical studies.

One of the difficulties in dealing with gene expression data is that the dimension of gene expression profiles are usually much larger than the sample size. To deal with the problem of high-dimensionality, several researchers constructed the linear combinations of the original variables; for example, principal component analysis [5], partial least squares [22, 24] and so on. However, even if it is little, informational loss cannot be denied in both approaches.

Recently the kernel method has been shown to be an effective tool to treat high-dimensional data without introducing any computational difficulty [8, 29]. A number of successful applications of the kernel method to gene expression data analysis have been reported [17, 19]. By introducing the kernel method, we express the survival probability function $S_j(t|x)$ and $\pi_j(x)$ by Weibull survival probability function and multiple logit function:

$$S_j(t|x; w_j, \zeta_j) = \exp \left[ -t^\gamma \exp \left\{ w_j^\prime \phi(x) \right\} \right], \quad (\zeta_j > 0),$$

$$\pi_j(x; \gamma) = \frac{\exp \{ \gamma_j^\prime \phi(x) \}}{\sum_{k=1}^{G} \exp \{ \gamma_k^\prime \phi(x) \}},$$

where $\phi(x) = (1, \phi_1(x), \ldots, \phi_m(x))'$ is the unknown kernel function vector, $w_j = (w_{j0}, w_{j1}, \ldots, w_{jm})'$ and $\gamma = (\gamma_1, \ldots, \gamma_{G-1})'$, $\gamma_j = (\gamma_{j0}, \gamma_{j1}, \ldots, \gamma_{jm})'$ are unknown parameter vector and $\gamma_G = 0$ for model identifiability. For the kernel functions $\phi_k(x)$, we use Gaussian kernel $\phi_k(x) = \exp\{-||x-\mu_k||^2/(2\sigma^2)\}$, where $\mu_k$ is the center and $\sigma^2$ is the spread parameter.

The problem is how to construct Gaussian kernel functions and how to estimate the unknown parameters $\Psi = (w_1', \ldots, w_G', \zeta_1, \ldots, \zeta_G, \gamma_1', \ldots, \gamma_{G-1}')'$ included in the model.
2.2 Estimation

Let \( \{(x_\alpha, t_\alpha, z_\alpha); \alpha = 1, \ldots, n\} \) be a set of independent observations, where \( x_\alpha \) is a vector of gene expression profiles associated with the \( \alpha \)-th individual, \( t_\alpha \) is the failure time or censoring time for the \( \alpha \)-th individual, \( z_\alpha = 1 \) indicates that the \( \alpha \)-th individual fails and \( z_\alpha = 0 \) represents a censoring.

The process of estimating the mixture density (1) consists of two steps. In the first step (Step 1), we construct a set of \( m \) Gaussian kernel functions. In the second step (Step 2), the unknown parameters \( \Psi \) are estimated by the maximum likelihood method. One approach for constructing the model is the full-supervised learning that estimates \( \mu_k, \sigma^2 \) and \( \Psi \) simultaneously. It does not guarantee the basis functions localized well [21] and causes the identification problem, however. From a statistical perspective, the serious problem frequently happens. If the centers and the spread parameters are regarded as parameters to be optimized by supervised learning, then the number of parameters exceeds the sample size, which we call over-parameterization.

**Step 1:** The centers \( \mu_k \) in Gaussian kernel are determined by using a k-means clustering algorithm, which divides the data set \( \{x_\alpha; \alpha = 1, \ldots, n\} \) into \( m \) clusters \( A_1, \ldots, A_m \) that correspond to the number of the kernel functions. The centers are then determined by \( \mu_k = \frac{1}{n_k} \sum_{\alpha \in A_k} x_\alpha \), where \( n_k \) is the number of the observations which belong to the \( k \)-th cluster \( A_k \). Gaussian kernels are then constructed by substituting the estimated centers and an appropriate value of \( \sigma^2 \). An appropriate choice of \( \sigma^2 \) is discussed in the later section.

**Step 2:** We next estimate the unknown parameters \( \Psi \). The log-likelihood function of the mixture density (1) with components (2) and (3) is given by

\[
\log L(\Psi) = \sum_{\alpha=1}^{n} \log \left\{ \sum_{j=1}^{G} \pi_j(x_\alpha; \gamma) \ell_j(t_\alpha|x_\alpha; w_j, \zeta_j) \right\},
\]

where \( \ell_j(t_\alpha|x_\alpha; w_j, \zeta_j) \) is the \( j \)-th cancer type likelihood for the \( \alpha \)-th individual failed or censored at time \( t_\alpha \), which is given by \(-dS_j(t_\alpha|x_\alpha; w_j, \zeta_j)/dt\) if \( z_\alpha = 1 \) and \( S_j(t_\alpha|x_\alpha; w_j, \zeta_j) \) otherwise. Although the maximum likelihood estimate \( \hat{\Psi} \) can be obtained by maximizing \( \log L(\Psi) \) directly, an EM algorithm [9] allows us to obtain the solution \( \hat{\Psi} \) by iteratively solving the simpler problem.

To solve the simpler problem, an unobservable random vector \( y_\alpha = (y_{ja}; a = 1, \ldots, n) \) of zero-one indicator variables is introduced for each observation \( \alpha \). Here \( y_{ja} = 1 \) or \( y_{ja} = 0 \) according as the \( \alpha \)-th individual belongs to \( j \)-th cancer type or not. Then the complete-data log-likelihood is given by

\[
\log L_c(\Psi) = \sum_{\alpha=1}^{n} \sum_{j=1}^{G} y_{ja} \log \left\{ \pi_j(x_\alpha; \gamma) \ell_j(t_\alpha|x_\alpha; w_j, \zeta_j) \right\}.
\]

It follows on application of the EM algorithm that at the \( k \)-th iteration of the E-step, we calculate the Q-function, which is the expectation of \( \log L_c(\Psi) \) conditional on the current estimate of the parameter \( \Psi^{(k)} \) and the observed data

\[
Q(\Psi; \Psi^{(k)}) = \sum_{\alpha=1}^{n} \sum_{j=1}^{G} \tau_{ja}^{(k)} \log \left\{ \pi_j(x_\alpha; \gamma) \ell_j(t_\alpha|x_\alpha; w_j, \zeta_j) \right\},
\]

where \( \tau_{ja}^{(k)} = E \left( y_{ja}|\Psi^{(k)} \right) = \frac{\pi_j(x_\alpha; \gamma) \ell_j(t_\alpha|x_\alpha; w_j, \zeta_j)}{\sum_{k=1}^{G} \pi_k(x_\alpha; \gamma) \ell_k(t_\alpha|x_\alpha; w_k, \zeta_k)} \bigg| \psi = \Psi^{(k)} \).

The M-step provides the updated estimate \( \Psi^{(k+1)} \) that maximizes \( Q(\Psi; \Psi^{(k)}) \). The updated estimate \( \Psi^{(k+1)} \) is given as the solution of \( \partial Q(\Psi; \Psi^{(k)})/\partial \Psi = 0 \), and is obtained by employing a nonlinear optimization scheme. Using the first and second derivatives \( g(\Psi) = \partial Q(\Psi; \Psi^{(k)})/\partial \Psi \) and \( H(\Psi) = \partial^2 Q(\Psi; \Psi^{(k)})/\partial \Psi \partial \Psi' \), the updated estimate \( \Psi^{(k+1)} \) is obtained by repeating Newton-Raphson update: \( \Psi^{new} = \Psi^{old} - H(\Psi^{old})^{-1}g(\Psi^{old}) \). Until a suitable condition is satisfied, the EM algorithm updates the current parameter \( \Psi^{(k)} \) to \( \Psi^{(k+1)} \) by maximizing \( Q(\Psi; \Psi^{(k)}) \).
2.3 The Number of Mixture Components and Marker Gene Selection

In previous section, we implicitly assumed that the number of mixture component $G$ and a set of "marker" genes $x$ are known in advance. In practical situations, however, they are unknown and to be identified. In addition, we have to choose the number of kernel functions $m$ and the values of the spread parameter $\sigma^2$ in Gaussian kernel.

In this paper, we shall use Akaike[1]'s information criterion since it has some attractive properties in practical applications [6]:

$$\text{AIC}(G, \sigma, m, x) = -2 \log L(\hat{\Psi}) + 2p,$$

where $L(\hat{\Psi})$ is the likelihood evaluated at the maximum likelihood estimate $\hat{\Psi}$ and $p = (2G - 1)(m + 1) + G$ is the number of free parameters included in the model. We choose the values $G$, $m$ and $\sigma^2$ and a set of marker genes which minimize AIC score.

Under the fixed values of $G$, $m$ and $\sigma^2$, the best set of marker genes can be obtained by searching all possible gene combinations. However, it is a time-consuming task to consider all possible gene combinations. For example, consider the situation that there are 100 genes, then the number of all possible gene combinations is larger than $10^{30}$. Therefore, to reduce the searching space, we shall use the forward stepwise selection method.

The forward stepwise selection algorithm begins with no genes in the model (1). At the forward step, AIC score is calculated for each gene and the best gene is added into the model. The forward step calculates AIC score again for the genes still remaining outside the model and adds the best gene into the model. This evaluation process is repeated until the AIC score is not improved. Next, at the backward elimination step, the genes already in the model are examined for removal according to AIC score. Because we would like to eliminate a low contributory gene already in the model, the backward step would be useful. The forward step and the backward elimination step are repeated until none of genes outside the model has an entry AIC score and every gene in the model is the best to stay. This procedure is carried out for various combinations of $G$, $m$ and $\sigma^2$. Finally, we choose the best one that minimizes AIC among the candidates.

2.4 Bayes Classification Rule and Survival Probability Prediction

After the number of mixture components and a set of marker genes are determined, and the maximum likelihood estimate $\hat{\Psi}$ is obtained, a Bayes classification rule assigns the $\alpha$-th individual to cancer type $j$, if the posterior probability of cancer type $j$ is maximum over the groups:

$$P(g = j|x_\alpha; \hat{\Psi}) = \frac{\pi_j(x_\alpha; \gamma)\ell_j(t_\alpha|x_\alpha; w_j, \zeta_j)}{\sum_{k=1}^{G} \pi_k(x_\alpha; \gamma)\ell_k(t_\alpha|x_\alpha; w_k, \zeta_k)}_{\Psi=\hat{\Psi}}.$$  

Inference on the survival probability for $\alpha$-th individual is done by

$$S(t|x_\alpha) = \sum_{j=1}^{G} \pi_j(x_\alpha; \hat{\gamma})S_j(t|x_\alpha; \hat{w}_j, \hat{\zeta}_j).$$

In general, the future observation will only have the gene expression profiles. In such a case, we first estimate the posterior probability that an individual belongs to the $j$-th cancer type based on $\Pr(g = j|x; \hat{\Psi}) = \pi_j(x; \hat{\gamma})$, and assign to the $j$-th cancer type, as the maximizer of posterior probability over the groups. Estimating the posterior probability $\Pr(g = j|x; \hat{\Psi})$ in (7) by $\pi_j(x; \hat{\gamma})$, we can predict the survival function for the future observation.

3 Results

3.1 Simulated Data Examples

To demonstrate the proposed mixture modeling methods, we first present the results of simulation experiments. Attention is focused on both the identification of true number of mixture components and on the detection of the marker genes that characterize the different cancer types and their survival probability functions.
Kernel Mixture Survival Models

We generated a set of data \( \{(x_\alpha, t_\alpha, z_\alpha); \alpha = 1, \ldots, n\} \) from the three component mixture model. The 3,000 dimensional gene expression values \( x = (x_1, \ldots, x_{3000})' \) were generated as standard normal with a few exceptions. The expression values of genes 1 and 2 \( (x_1, x_2)' \) were generated independently from the normal with means \( c_1 = (1, -0.8)' \) in cancer type 1, \( c_2 = (-1, 0)' \) in cancer type 2, and \( c_3 = (0, 0.8)' \) in cancer type 3 with variance \( \sigma^2 \). Based on the gene expression values of genes 1 and 2, the survival times were generated from the specified survival probability functions. Thus, genes 1 and 2 are a set of marker genes to be specified.

We consider the two situations. In the first case (C1), we assume the exponential survival probability functions for each component \( S_j(t|x_1, x_2) = \exp[-t \exp(-h_j(x_1, x_2))] \), where the true functions \( h_j(x_1, x_2) \) are respectively given by

\[
\begin{align*}
  h_1(x_1, x_2) &= 0.3x_1 + 0.4x_2 - 0.1 \quad \text{(type 1)} \\
  h_2(x_1, x_2) &= 0.5(x_1 - x_2)^2 + 1 \quad \text{(type 2)} \\
  h_3(x_1, x_2) &= 0.06x_1^3 - 0.2x_2 \quad \text{(type 3)}
\end{align*}
\]

For the second case (C2), the exponential survival probability functions of each component is replaced by Weibull survival probability functions \( S_j(t|x_1, x_2; \zeta_j) = \exp[-t^{\zeta_j} \exp(-h_j(x_1, x_2))] \) with \( h_j(x_1, x_2) \), \( j = 1, 2 \) and 3 are given in (8), \( \zeta_1 = 0.4, \zeta_2 = 1.8 \) and \( \zeta_3 = 1 \), respectively. Thus \( S_3(t|x_1, x_2; \zeta_3) \) is equivalent to the exponential survival probability function.

Table 1: Percentage of times the information criterion AIC estimates \( G \) equal to 1, 2, 3, 4 or 5. Percentages of times that the marker genes \( x_1 \) and \( x_2 \) appeared in the selected genes are also reported. The value \( \max_{j \neq 1,2} \{x_j\} \) indicates the maximum value of percentage of times the genes \( x_3 \) to \( x_{3000} \) appeared in the selected genes. The means of the number of selected genes \( \bar{p} \) are also shown.

<table>
<thead>
<tr>
<th>Case</th>
<th>( (n, \sigma^2) )</th>
<th>Frequency of ( \hat{G} )</th>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( \max_{j \neq 1,2} {x_j} )</th>
<th>( \bar{p} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n )</td>
<td>( \sigma^2 )</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>C1</td>
<td>(50,0.3)</td>
<td>14</td>
<td>27</td>
<td>46</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(50,0.6)</td>
<td>16</td>
<td>26</td>
<td>42</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(100,0.3)</td>
<td>10</td>
<td>18</td>
<td>53</td>
<td>19</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(100,0.6)</td>
<td>9</td>
<td>21</td>
<td>48</td>
<td>22</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(150,0.3)</td>
<td>5</td>
<td>12</td>
<td>60</td>
<td>21</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>(150,0.6)</td>
<td>4</td>
<td>11</td>
<td>56</td>
<td>25</td>
<td>4</td>
</tr>
<tr>
<td>C2</td>
<td>(50,0.3)</td>
<td>12</td>
<td>25</td>
<td>45</td>
<td>18</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(50,0.6)</td>
<td>13</td>
<td>23</td>
<td>44</td>
<td>20</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(100,0.3)</td>
<td>7</td>
<td>15</td>
<td>51</td>
<td>25</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>(100,0.6)</td>
<td>5</td>
<td>18</td>
<td>49</td>
<td>25</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>(150,0.3)</td>
<td>2</td>
<td>10</td>
<td>58</td>
<td>24</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>(150,0.6)</td>
<td>3</td>
<td>9</td>
<td>55</td>
<td>26</td>
<td>7</td>
</tr>
</tbody>
</table>

Given that a sample belongs to the \( j \)-th cancer type, a failure time was generated from the \( j \)-th survival probability function according to the inverse transform method. The censoring time was also generated from a uniform distribution \( U(20, 50) \) for each observation. If the failure time were greater than the censoring time, it was taken to be censored at this censoring time. For each simulation set, we generated \( n \) independent samples with prior probabilities for each class \( p = (0.3, 0.4, 0.3) \).

The estimated number of mixture components and the summaries of selected genes for various settings are reported in Table 1. The simulation results were obtained by 100 repeated Monte Carlo trials. From Table 1, as the number of samples becomes large, the proposed method identified the true number of mixture component more accuracy. Table 1 also indicates that the proposed method frequently identified the marker genes. Even when the error variance is large \( \sigma^2 = 0.6 \), the forward stepwise selection algorithm can select the marker genes effectively if we have enough samples. We also
Ando et al. checked the percentage of times a set of non informative genes $x_3, ..., x_{3000}$ appeared in the selected genes. Table 1 indicates that the largest value was 5%. From these results, we can expect that the proposed method can work effectively in the real data analysis.

### 3.2 Application to Diffused Large B-Cell Lymphoma Data

The lymphoma data set [2] consists of gene expression levels from cDNA experiments involving three prevalent adult lymphoid malignancies. These three malignancies are the diffuse large B-cell lymphoma, the B-cell chronic lymphocytic leukemia and the follicular lymphoma, respectively. It is well known that the diffuse large B-cell lymphoma is clinically heterogeneous: less than half patients respond well to current clinical therapy, whereas the remainder does not [2]. [2] identified two molecularly distinct diffuse large B-cell lymphoma; germinal centre B-like (GCB-like) and activated B-like (AB-like). Among 42 diffuse large B-cell lymphoma patients, 40 of them had survival information, including 22 death with death time and 18 being still alive with the follow-up times. All patients have not been treated before obtaining the biopsy sample. After biopsy, the patients were received standard multi-agent chemotherapy regimens. Deleting the subjects with any missing covariates, we use the 2041 gene expression profiles.

It would be suitable to reduce the sampling effects, since we have only 40 observations. To reduce the sampling effects, we performed a full leave one out cross validation procedure. Excluding one sample, we construct a mixture of Weibull experts model (1) by using remained $n-1$ samples. We choose the number of mixture components $G$ for which AIC score is minimized over a set of competing models. By cycling through all observations, we obtained $n$ estimates of the number of mixture components. We found that the three components models are selected most frequently (62.5%). It indicates that three subtypes of diffuse large B-cell lymphoma are included in the data.

Table 2 shows the 20 genes, which are frequently included (more than 30%) in the models designed in the full leave one out cross validation procedure. We investigated these genes and found that the proposed method successfully identified many important genes. For example, [7] reported that IgG Fc receptors can be targets for deregulation through chromosomal translocation in lymphoma. [20] investigated whether interleukin-10 (IL-10) genetic polymorphisms influence the cytokine production as well as the incidence and outcome of diffuse large B-cell lymphoma. They concluded that IL-10 production contributes to the clinical course of diffuse large B-Cell lymphoma and that this phenomenon involves a substantial genetic component. [30] reported that the gene tyk2 appears to be expressed more frequently in cancer tissues. The gene IP-10 plays an important role in the pathogenesis of tissue necrosis and vascular damage associated with certain EBV-positive lymphoproliferative processes [10]. Casein kinase II alpha chain is an useful marker gene to distinguish between GCB-like and AB-like [2]. In the breast cancer study, [16] reported that a more effective strategy to control breast cancer is to target AIB1-mediated and ovarian hormone-initiated pathways. The gene expression of CD70 was used for identifying a novel subpopulation of germinal center B cells [18].

There are also many unknown genes in Table 2. We further investigated these unknown genes and found that the previously unknown gene (YORF=18585, Clone=1355675) was AIB1. Estrogen signaling plays an important role in a number of normal physiological processes and has important implications in the treatment of breast cancer. AIB1 is frequently amplified and overexpressed in human breast cancer. It has been shown to enhance estrogen-dependent transactivation [3]. Additionally, overexpression of this gene has been also observed in primary gastric tumors, head and neck cancers [4, 26]. Thus we suspect that AIB1 is an important gene for the treatment of diffuse large B-Cell lymphoma.

Based these 20 gene expression profiles, we next considered to fit the three components mixture model. Table 3 (a) summarizes the number of samples belongs to each mixture component. From Table 3, the cancer type 1 and cancer type 2 seem to be corresponding to the GCB-like and AB-like diffuse large B-cell lymphoma, respectively. [2] reported that the higher expression of BCL-6 is related to GCB-like subtypes. The higher expression of CD44 is related to AB-like subtypes [25]. We found
that the higher expression of BCL-6 is related to the cancer type 1 and that the higher expression of CD44 is related to the cancer type 2. Other marker genes for distinguishing with AB-like and GCB-like found in [2] and [25], for e.g., OP-1, CD21, CD22, JNK6 and IL-16, also indicated that the cancer type 1 and 2 is corresponding to the GCB-like and AB-like diffuse large B-cell lymphoma. Recently, three subtypes of diffuse large B-cell lymphoma; GCB-like, AB-like and new type 3 diffuse large B-cell lymphomas were identified by [25]. We found that the gene expressions JNK6 and CD44 in the cancer type 3 were both low level, which is related to new type 3 diffuse large B-cell lymphoma identified by [25]. Thus we suspect that the cancer type 3 may be corresponding new type 3 diffuse large B-cell lymphoma.

Table 2: List of a set of 20 marker genes based on the proposed method.

<table>
<thead>
<tr>
<th>YORF</th>
<th>Frequency</th>
<th>Gene Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>13338</td>
<td>0.650</td>
<td>Unknown UG Hs.123377 EST; Clone=1333741</td>
</tr>
<tr>
<td>17828</td>
<td>0.625</td>
<td>Low-affinity IgG Fc receptor II-B and C isoforms; Clone=292524</td>
</tr>
<tr>
<td>18585</td>
<td>0.625</td>
<td>Unknown; Clone=1355675</td>
</tr>
<tr>
<td>21012</td>
<td>0.625</td>
<td>Unknown UG Hs.36672 ESTs; Clone=1269317</td>
</tr>
<tr>
<td>16521</td>
<td>0.600</td>
<td>IL-10 receptor beta=Cytokine receptor family II, member 4; Clone=202498</td>
</tr>
<tr>
<td>21201</td>
<td>0.550</td>
<td>Calcium/calmodulin-dependent protein kinase II delta; Clone=1301950</td>
</tr>
<tr>
<td>14782</td>
<td>0.550</td>
<td>Unknown; Clone=1356601</td>
</tr>
<tr>
<td>17179</td>
<td>0.550</td>
<td>LPTPase=inducible lymphoid-specific, protein tyrosine phosphatase; Clone=686081</td>
</tr>
<tr>
<td>20359</td>
<td>0.525</td>
<td>Restin; Clone=713341</td>
</tr>
<tr>
<td>16806</td>
<td>0.450</td>
<td>Immunoglobulin-related 14.1; Clone=344134</td>
</tr>
<tr>
<td>18249</td>
<td>0.450</td>
<td>lamin B1; Clone=1184257</td>
</tr>
<tr>
<td>21512</td>
<td>0.425</td>
<td>Unknown; Clone=1352894</td>
</tr>
<tr>
<td>18348</td>
<td>0.425</td>
<td>tyk2=non-receptor protein tyrosine kinase; Clone=1271565</td>
</tr>
<tr>
<td>16447</td>
<td>0.375</td>
<td>branched chain aminotransferase precursor (BCATm); Clone=43773</td>
</tr>
<tr>
<td>20981</td>
<td>0.375</td>
<td>BAP31; Clone=1242061</td>
</tr>
<tr>
<td>16989</td>
<td>0.375</td>
<td>IP-10; Clone=491243</td>
</tr>
<tr>
<td>18386</td>
<td>0.375</td>
<td>Casein kinase II alpha chain; Clone=1286561</td>
</tr>
<tr>
<td>16509</td>
<td>0.350</td>
<td>AIB1=Amplified in Breast Cancer; Clone=197520</td>
</tr>
<tr>
<td>17541</td>
<td>0.325</td>
<td>CD70 antigen=CD27 ligand; Clone=1409032</td>
</tr>
<tr>
<td>15859</td>
<td>0.325</td>
<td>Phospholipid hydroperoxide glutathione peroxidase; Clone=809981</td>
</tr>
</tbody>
</table>

Figure 1 (a) shows the averaged survival functions \( \hat{S}_j(t) = \sum_{a,c,j} S_j(t_{a}\mid x_a; \tilde{w}_j, \tilde{z}_j)/n_j \) for three different groups, where \( n_j \) is the number of the observations which belong to the \( j \)-th cancer type. There is a remarkable difference between the survival functions; patients with cancer type 1 have a significantly better survival probability than those with cancer type 2. If we consider the mean survival functions of cancer type 1 and cancer type 2 as the survival function of GCB-like and AB-like, this estimated results are well corresponding to the result of [2]. [25] reported that the survival curve of new type 3 diffuse large B-cell lymphoma has higher survival rate than that of GCB-like while lower survival rate than that of AB-like. Considering the gene expressions of JNK6 and CD44 in the cancer type 3, it is natural to consider the cancer type 3 in Table 2 is corresponding to new type 3 diffuse large B-cell lymphoma.

We compared the above results made by our proposed methodology with those based on the two stage approach. It first divides observations into several clusters based only on gene expression profiles and then applies some statistical survival methods to each cluster. We applied the \( k \)-mean clustering method to group cancer samples into three clusters. Table 3 (b) shows the number of samples belongs to each cluster. It indicates that both cancer type 1 and 3 are corresponding to the AB-like diffuse large
B-cell lymphoma and the cancer type 2 corresponds to the GCB-like diffuse large B-cell lymphoma. Based on this clustering result, Kaplan-Meier survival curves are estimated for each cluster. Figure 1 (b) shows the fitted results. In contrast to our results, the survival curves of cancer type 1 and 3 are similar to each other. Since this two stage approach ignores survival phenotypic information completely, it is difficult to identify cancer subtypes that are correlated with both gene expression profiles and survival probabilities.

Table 3: Class distribution matrix.

<table>
<thead>
<tr>
<th></th>
<th>(a): Proposed method</th>
<th>(b): Two stage approach</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GCB-like</td>
<td>AB-like</td>
</tr>
<tr>
<td>Type 1</td>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td>Type 2</td>
<td>6</td>
<td>14</td>
</tr>
<tr>
<td>Type 3</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

Figure 1: (a): Mean survival probability functions $\tilde{S}_j(t)$ and (b): Kaplan-Meier survival curves in year for the cancer type 1 (---), the cancer type 2 (---) and the cancer type 3 (-----), respectively.

4 Conclusion

This article proposed the kernel mixture modeling method that jointly identifies the number of cancer subtypes and a set of marker genes, and predicts both cancer type and survival probability. In the context of mixture experts framework, we introduced the kernel multiple logit model and the kernel Weibull survival model for relating gene expression profiles to cancer type and survival probabilities. One advantage of the proposed method is that there is no computational or methodological limitation in term of the number of genes that can be used in the prediction of cancer type and patient’s survival time. The proposed method is also applicable even when different subtypes of cancer are not known to exist. Real data analysis and Monte Carlo simulations show that the proposed method can be useful for identifying the number of cancer subtypes and for predicting patients’ cancer type and survival probabilities.

The proposed methodology can be extended in various ways. For example, we used the common spread parameter $\sigma$ for Gaussian kernel. We expect that the performance of our proposed model could be improved if we introduce Gaussian kernel which has its own spread parameter. We would like to apply some extended methods to a larger real data set in our future research.
Acknowledgments

The authors would like to thank the anonymous reviewers for constructive and helpful comments that improved the quality of the paper.

References


