Available online at www.isas.org.in/jisas

JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 64(1) 2010 61-76

Lassoing Mixtures with Applications to Proteomic Mass Spectroscopy Analysis

Guan Xing[*1] and J. Sunil Rao[2]

[1]*Bristol-Myers Squibb*

[2]*Case Western Reserve University*

SUMMARY

We propose a new estimation method for finite mixture models. Important in this estimation process is the determination of the number of mixture components. Traditional methods either perform sequential hypothesis testing, or perform model selection based on some criteria such as AIC, BIC, and Kullback-Leibler (KL) distance. We treat the component densities as predictors and generate pseudo-response based on the CDF/PDF of a saturated mixture model. To get a sparse component representation, we use a variation of the LASSO – a L1-constraint optimization that produces many zero components weights. We then iterate between LASSO and EM steps to update the estimates of the component density parameters and component weights. Our approach is very general and can be extended naturally to handle finite multivariate mixtures and mixtures with non-normal components. A series of simulations illustrate the competitiveness of our approach. We then apply the methodology to a problem of classifying ovarian cancer patients based on protein mass spectroscopy data profiles.

Keywords : Mixture models, LASSO, Proteomic mass spectroscopy, Selection, Shrinkage.

1. INTRODUCTION

Finite mixture densities are mainly used in modelling heterogeneous data, where observations come from different populations but we do not know the exact source of each observation. Let X denote a random variable to be measured, and let $f(x|\theta_j)$ denote the density function for the jth population, and π_j denote the corresponding probability that X comes from this population. The finite mixture model has the following form

$$f(x|\Theta) = \sum_{j=1}^{k} \pi_j f(x|\theta_j) \qquad (1)$$

with $\theta_j \in \Theta$, $\sum_{j=1}^{k} \pi_j = 1$, $\pi_j \geq 0$. To avoid unidentifiability, we assume the location parameters are in a strictly increasing order. For an observed random sample $x_1, \ldots, x_n$ from equation 1, the loglikelihood function is

$$l_n(x|\underline{\pi}, \underline{\theta}) = \sum_{i=1}^{n} \log [\sum_{j=1}^{k} \log \pi_j f(x_i|\theta_j)] \qquad (2)$$

Finite mixture models have a large number of applications in statistical modelling. For example, Smith (1961) introduced the mixture model for genetic linkage heterogeneity based on the recombination fraction θ and the proportion of linked families. Other applications include fisheries (Macdonald *et al*. 1979), genetics (Ott 1999), physics (Tanner 1962), psychology (Broadben 1966), medicine (Clark *et al*. 1968), botany (Gordon and Prentice 1977), disease mapping (Schlattmann and Böhning 1993), and meta-analysis (Böhning 1999).

The component number estimation in finite mixture modelling is not a trivial question when the component number is unknown. It has long been

[*] *Corresponding author* : Guan Xing

E-mail address : guan.xing@bms.com

recognized that this problem is unidentifiable under the null hypothesis of homogeneity. The common likelihood ratio test (LRT) fails in the case of the finite mixture models since the LRT statistic goes to infinity with probability 1 (Hartigan 1985). Many new methods have been proposed to choose an appropriate component number, including the modified likelihood ratio test (MLRT) (Chen *et al*. 2001), D-test (Charnigo and Sun 2004), Kullback-Leibler (KL) distance (Sahu and Cheng 2002) and Bayes factor (Ishwaran *et al*. 2001). However, most methods are complex in computation and often lose the clear interpretation of LRT. Besides, many methods are designed to deal with simple cases such as a 2-component mixture versus the homogenous distribution, and as a result, they are not easily extended to more complicated situations.

We propose a new approach from a different point of view. With the constructed pseudo-response from a saturated model, we transform the mixture density estimation problem to a variable selection problem in linear regression. The weight vector of the mixture density is treated as the regression parameters conditional on the constructed response and design matrix, and calculated using a variant of the LASSO approach with special constraint that all regression parameters are non-negative and the summation is 1. EM algorithm is used to optimize the parameters of the component densities, and the whole process is repeated until the convergence of the final estimate. From our experience, the algorithm's converging speed is at the same order with EM.

The paper is organized as follows. Section 2 introduces our proposed method. Section 3 shows some theoretical results. Extensive simulation results are provided in Section 4. In Section 5, we apply our approach to a proteomic mass spectroscopy ovarian cancer data analysis. Section 6 includes a brief summary and some discussions.

2. PROPOSED METHOD

Suppose there is a random sample of n observations, $x_1, \ldots, x_n$, from a finite mixture distribution $\sum_{j=1}^{k} \pi_j f(x|\theta_j)$. The component number k, component weights $\underline{\pi}$, and the distribution parameters $\underline{\theta}$ are all unknown. We use Y to denote the mixture density. The true density values at the sample points are $y_i = \sum_{j=1}^{k} \pi_j f(x_i|\theta_j)$, $i = 1, \ldots, n$. Suppose we have fitted a saturated mixture model $\tilde{f}$ with component number m for the data, $m > k$, and use $\tilde{Y}$ to denote the density of the sample points estimated with $\tilde{f}$. The choice of the saturated model will be discussed in Section 2.1. Then we can approximate the relationship between Y and $\tilde{Y}$.

Let ε_i denote the difference between the estimate $\tilde{y}_i$ and the true value y_i,

$$\tilde{y}_i = y_i + \varepsilon_i = \sum_{j=1}^{k} \pi_j f(x_i|\theta_j) + \varepsilon_i,\ i = 1,...,n \qquad (3)$$

However, since we do not know the true value of k, we use m to approximate it as the start, and expect that there are $m - k$ components with weight 0.

$$\begin{aligned} \tilde{y}_i &= \sum_{j=1}^{k} \pi_j f(x_i|\theta_j) + \sum_{j=k+1}^{m} 0 \times f(x_i|\theta_j) + \varepsilon_i \\ &= \sum_{j=1}^{m} \pi_j f(x_i|\theta_j) + \varepsilon_i \text{ with } \sum_{j} I_{\{\pi_j = 0\}} = m - k \\ &\qquad i = 1, ..., n \end{aligned} \qquad (4)$$

If we make an assumption for the error term $\underline{\varepsilon}$ that $E(\underline{\varepsilon}) = \underline{0}$, $Var(\underline{\varepsilon}) = \Sigma$ for some well defined positive definite matrix, we can treat $\underline{\pi}$ as the regression parameters conditional on the response $\tilde{Y}$ and the design matrix $\mathbf{X}$ with the (i, j)th entry $f(x_i|\theta_j)$. In other words, if we assume that the component density parameters $\underline{\theta}$ are known temporarily, we can re-express Equation 4 as

$$\tilde{y}_i = \sum_{j=1}^{m} \pi_j X_{ij} + \varepsilon_i,\ i = 1, \ldots, n \qquad (5)$$

where the design matrix $\mathbf{X}$ is constructed using $f(x_i|\theta_j)$, and $m - k$ of the regression parameters π_j should have true value 0.

Note that the estimated density could be seen as the summation of the true density, some pseudo component densities with weight zero, and an error term. Thus we can estimate mixture weights as regression parameters due to the linear structure of the mixture model.

We can use either the cumulative distribution function (CDF) or the probability density function (PDF) to relate Y to $\tilde{Y}$. There are subtle differences between these two functions though. The correlation

between predictors ($f(x|\theta_j)$) is larger under the CDF case, while the design matrix is nearly orthogonal under the PDF case. The correlation between the columns of the design matrix affects the imposition of sparsity on the mixing weights. The estimates using PDF are usually larger than those using CDF, which will be shown in the simulation examples.

In our approximation, we know that in Equation 4, some true π_j values should be exactly zero. The ordinary least square (OLS) estimator of the mixing weights under the regression-type relation derived in Equation 5 does not help here because it gives m non-zero regression parameters. To impose sparsity, we will use a variant of Tibshirani's least absolute shrinkage and selection operator (LASSO) (Tibshirani 1996).

For a linear model $\tilde{Y} = X\underline{\pi} + \varepsilon$, LASSO estimator minimizes the residual sum of squares (RSS) subject to the constraint $\sum_j |\pi_j| \le t$ for some $t > 0$.

$$\hat{\underline{\pi}}^{LASSO} = \underset{\underline{\pi}}{argmin} \sum_{i=1}^{n} (\tilde{y}_i - \sum_{j=1}^{p} x_{ij}\pi_j)^2$$

subject to $\sum_j |\pi_j| \le t$

which is equivalent to

$$\hat{\underline{\pi}}^{LASSO} = \underset{\underline{\pi}}{argmin} \left\{ \sum_{i=1}^{n} (\tilde{y}_i - \sum_{j=1}^{p} x_{ij}\pi_j)^2 + \lambda \sum_{j=1}^{p} |\pi_j| \right\}$$

with $\lambda > 0$. Because of the special geometry of LASSO constraint $\sum_j |\pi_j| \le t$ (the rotated diamond), the LASSO solution can produce exact 0 coefficients for some predictors. As a result, the LASSO does coefficient shrinkage and predictor selection simultaneously and continuously. It retains the good features of both subset selection and shrinkage re-gression while minimizing their shortcomings.

However, an additional special requirement is needed for the mixture problem as indicated in Equation 5. The mixing weights must also satisfy $\sum_j \pi_j = 1$ and $\pi_j \ge 0$, since $\underline{\pi}$ represents the weights of the component distributions in a finite mixture model. We will use a variant of the LASSO called the positive LASSO approach proposed by Efron *et al*. (2004), and find the solution with exact summation 1. For efficient computation, we will make use of the least angle regression (LARS) representation of the positive LASSO as in Efron et al's paper. The detailed algorithm will be described in Section 2.3.

2.1 Saturated Model Choice

To construct a saturated model, we need to use a component number m large enough. Lindsay (1983) proved that the upper bound of k is the number of distinct points in the sample. For many scientific questions, we can have a stricter upper bound d for the estimate of k (Ishwaran *et al*. 2001), and the model with component number larger than d is a saturated model, i.e., $m > d$ will suffice the requirement for a saturated model. While our inference is robust to the choice of the saturated model, we suggest using the most saturated model satisfying the constraint. For the normal mixture model with unconstrained means and variances, the log-likelihood function is unbounded for any number of entertained components (Eguchi and Yoshioka 2001). Thus we typically need to impose some structural constraints to the model to eliminate the unboundedness problem, one common option is the equal variance constraint that $\sigma_1^2 = ... = \sigma_k^2 = \sigma^2$ with the largest possible component number $m = n - 1$.

2.2 Connections with other Methods

The idea of using penalized optimization for mixture models has been suggested by other statisticians. In this section, we want to show the relationship between our approach and these methods.

The modified likelihood ratio test suggested by Chen *et al*. (2001) has been popular because it is distribution-free and asymptotically locally most powerful. It adds a penalty term on the mixture weights to the likelihood function, and solves the optimization of the penalized likelihood

$$(\underline{\pi}, \underline{\theta}) = \underset{\underline{\pi}, \underline{\theta}}{arg\,max} \{\log \sum_{i=1}^{n} f(x_i | \underline{\pi}, \underline{\theta}) + C \sum_{j=1}^{k} \log(\pi_j)\} \quad (6)$$

The LASSO solves the following constrained optimization problem

$$\underline{\hat{\pi}}^{LASSO} = \underset{\underline{\pi}}{argmin}\{\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\pi_j)^2 + \lambda \sum_{j=1}^{p}|\pi_j|\} \quad (7)$$

which is equivalent to Equation 6. As we all know, in the regression setting, the ordinary least square estimate is equivalent to the maximum likelihood estimate.

$$\hat{\beta}^{ols} = \underset{\beta}{argmin}\{\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2\}$$

$$= \underset{\beta}{argmax}\{\log f(Y|\mathbf{X}\beta)\} \quad (8)$$

Chen *et al*. (2001) also discussed the Bayesian view of their method, the penalty term is considered as the prior distribution of $\underline{\pi}$. This is also the same with the Bayesian explanation of LASSO, which has a double exponential prior distribution.

The general idea of generating the pseudo-response for non-regression situations and employing linear regression modelling is also suggested by other researchers. Zou *et al*. (2006) introduced a new method called sparse principal component analysis (SPCA). They contrived the ordinary principal component analysis (PCA) as the response and used the LASSO (elastic net) to produce the modified principal components with sparse loadings.

2.3 Lassoing Mixture Algorithm

Suppose we want to fit a finite normal mixture distribution to the data with *n* samples. The fitting algorithm is described in Algorithm 1.

We start with the most saturated model, where the number of components *m* is *n* – 1. The contrived response is the cumulative density or probability density of the fitted *m*-component normal mixture at the data points. We use the ordered data as the initial corresponding location parameters, and the sample variance divided by m as the initial common variance parameter. The entries of the design matrix **X** will just be the CDF/PDF of each component distribution at the data points. For example, X_{ij} will be the CDF/PDF for the j^{th} component distribution $N(x|\theta_j)$ at the i^{th} data point.

Algorithm 1. Lassoing Mixture

Fit a m component saturated mixture model $\tilde{f}$ and let $\tilde{y}_i = \tilde{f}(x_i)$.

With initial values $\underline{\theta}^0$, construct the design matrix $\mathbf{X}_{ij}^0 = N(x_i|\theta_j^0)$.

Repeat

$\underline{\hat{\pi}} = \underset{\underline{\pi}}{argmin}\sum_{i=1}^{n}(\tilde{y}_i - \sum_{j=1}^{m}\mathbf{X}_{ij}\pi_j)^2$ with the constraint $\sum \pi_j = 1$ and $\pi_j \geq 0$.

Delete the redundant components with weight zero.

Update the component parameters using EM with fixed weights.

$$\mathbf{X}_{ij}^{r+1} = N(x_i|\theta_j^{r+1})$$

until the convergence of the parameters.

A final full EM fit is suggested to optimize the location and scale parameters.

Using the ordered data as the initial values for the corresponding location parameters of an *m*-component normal mixture is a reasonable choice, which implies each data point is from its own component distribution. However, the initial value of the common variance can not be easily determined. In practice, we find that the large initial value tends to produce the solutions too sparse, while the initial value too small gives a model with redundant components. Our choice of the sample variance divided by *m* is a heuristic attempt based on our experience. Besides, we find that our model selection result is very robust for the choice of the saturated model, a mixture model with the component number large enough can be used as the saturated model.

We use two simulated data sets from the simulation design 2 presented later in the paper to show the parameter traces of our algorithm. Each data consists of 100 observations from a 2-component

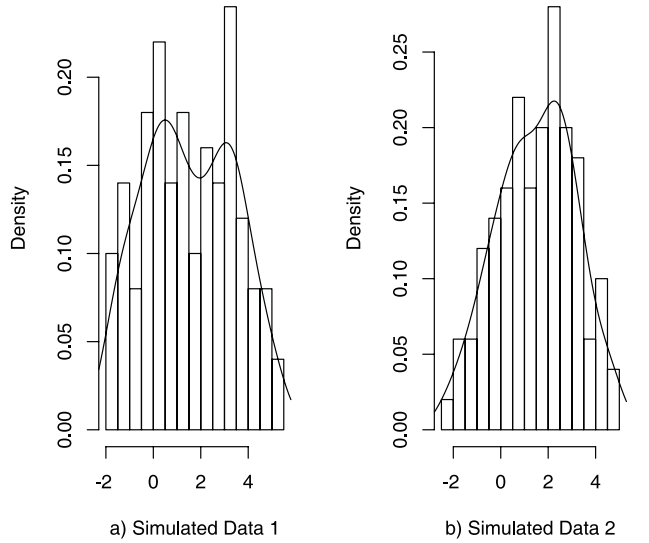

Fig. 1. (a) is the histogram plot of the simulated data set 1 with the kernel density curve. (b) is the histogram plot of the simulated data set 2 with the kernel density curve.

normal mixture with homogeneous unit variance, equal weight, and location parameter μ = (0, 3). Fig. 1 shows the histogram plots with kernel density curves of these two data. Fig. 2 and Fig. 3 are the parameter trace plots. Both analysis start with a 99-component normal mixture with the ordered data as the initial mean vector and $\sigma_{data}/\sqrt{100}$ as the initial common standard deviation. Because the component number drops sharply from 99 to a small number at the first iteration, we only plot the trace from the second step until convergence.

From the trace plots, we see that the parameter $\underline{\pi}$ and $\underline{\mu}$ jump randomly at the beginning. With the degeneration of the component number k, other parameter estimates converge to the stationary values

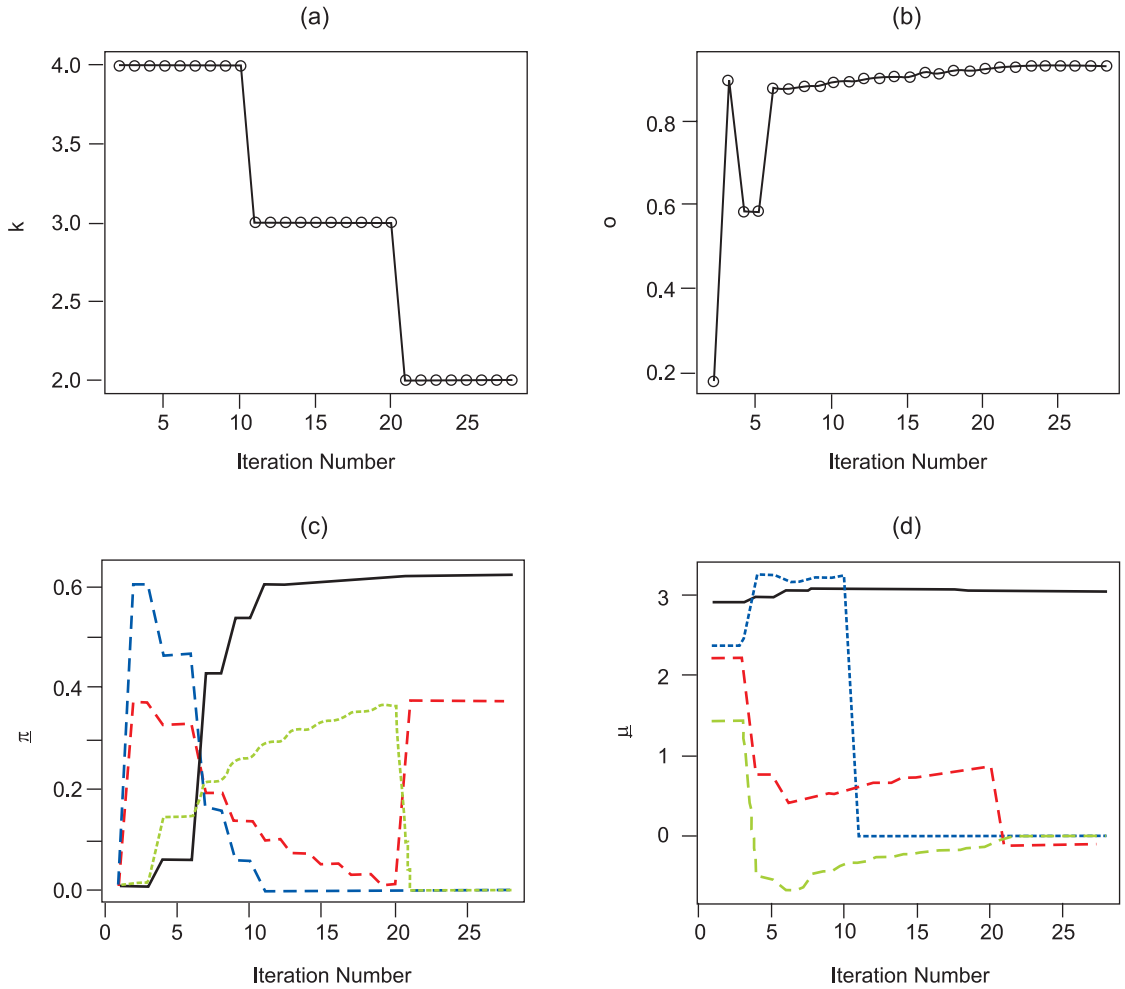

Fig. 2. Trace plot of the simulated data set 1. (*a*) is the trace of the component number *k*. (*b*) is the trace of the common standard deviation σ · (*c*) is the trace of the weight vector $\underline{\pi}$ · (*d*) is the trace of the mean vector $\underline{\mu}$.

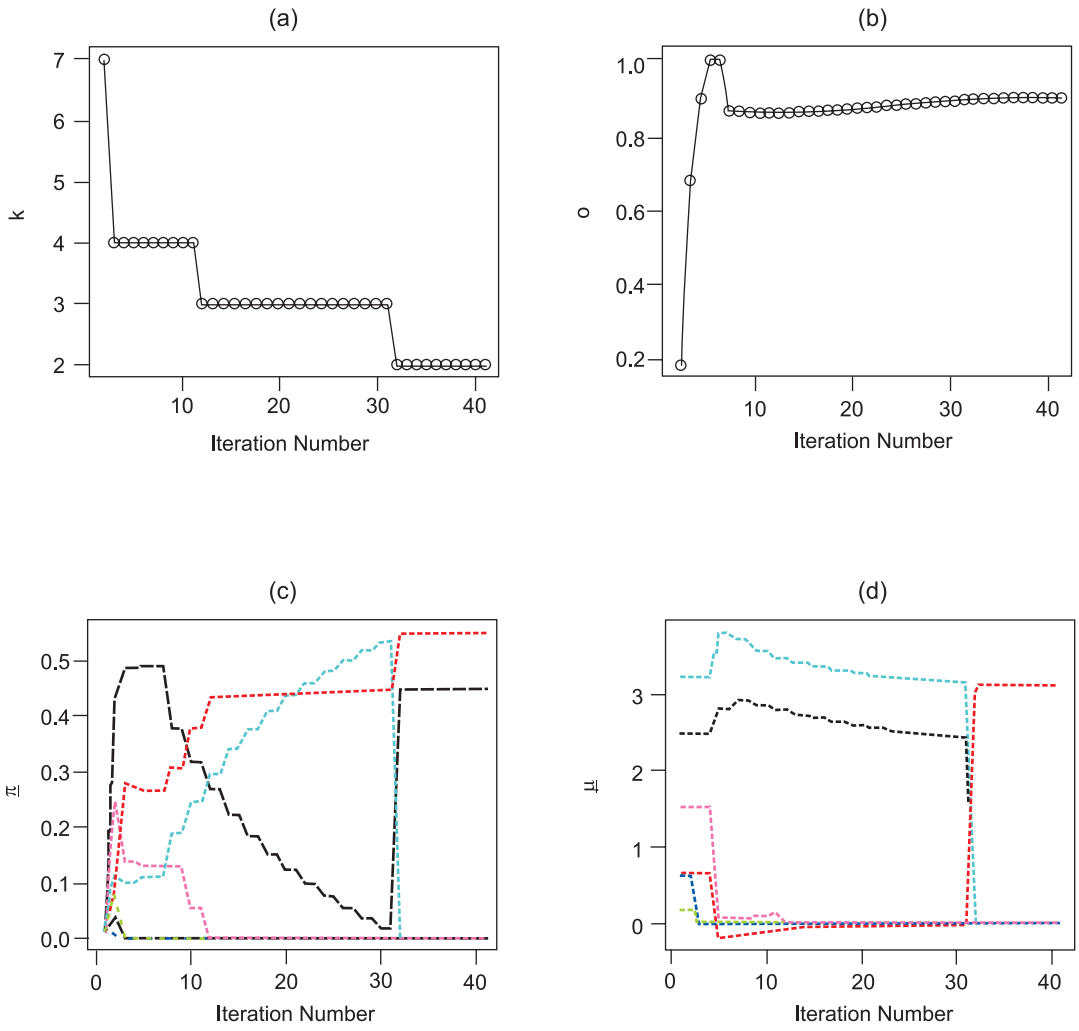

Fig. 3. Trace plot of the simulated data set 2. (*a*) is the trace of the component number *k*. (*b*) is the trace of the common standard deviation σ. (*c*) is the trace of the weight vector $\underline{\pi}$. (*d*) is the trace of the mean vector $\underline{\mu}$.

gradually. The final model selected has 2 components. It is interesting that components with larger weights at early iterations do not necessarily survive in the end.

3. UNDERSTANDING THE ESTIMATOR

Firstly we constrain the problem to the situation where the mixture component number *k* is known. The common approach is fitting the model with EM to get the maximum likelihood estimates (MLE) of the other parameters which maximize function 2. We want to show that we can also get MLE by the following Algorithm 2, which iteratively uses the OLS estimator and EM with the fixed weights.

The following Lemma 1, Lemma 2, and Theorem 1 show that the likelihood is increased at each step of the Algorithm 2. Thus, the final estimates are MLE.

Lemma 1. Suppose in the finite mixture model estimation, the number of components *k* is known and $\hat{\underline{\pi}}$ is the EM estimate for the weight vector $\underline{\pi}$, then $E(\hat{\underline{\pi}}) = \underline{\pi}$.

Lemma 2. Suppose in the finite mixture model estimation, the number of components *k* is known, the component parameters $\underline{\theta}$ are known, and the weight vector $\underline{\pi}$ is unknown. Construct the design matrix **X** and response vector $\tilde{Y}$ as in Algorithms 1 and 2, and

Algorithm 2. OLS + EM

Fit a saturated mixture model $\tilde{f}$ and let $\tilde{y}_i = \tilde{f}(x_i)$. With initial values $\underline{\theta}^0$, construct the design matrix $\mathbf{X}_{ij}^0 = N(x_i \mid \theta_j^0)$.

Repeat

$$\underline{\hat{\pi}} = \underset{\underline{\pi}}{argmin} \sum_{i=1}^{n} (\tilde{y}_i - \sum_{j=1}^{k} \mathbf{X}_{ij} \pi_j)^2$$

Update the component parameters using EM with the fixed weights.

$$\mathbf{X}_{ij}^{r+1} = N(x_i \mid \theta_j^{r+1})$$

until the convergence of the parameters.

estimate $\underline{\pi}$ with $\underline{\hat{\pi}}^{ols} = \underset{\underline{\pi}}{arg\ \min} (\hat{Y} - \mathbf{X}\underline{\pi})^2$. Then for the class of the unbiased estimates of $\underline{\pi}$, $\underline{\hat{\pi}}^{ols}$ has the largest loglikelihood. $\underline{\hat{\pi}}^{ols} = \underset{\underline{\pi}}{argmax}\ l_n(x|\underline{\pi}, \underline{\theta})$.

Theorem 1. Suppose in the finite mixture model estimation, the component number k is known. With some initial parameter values, the estimates by Algorithm 2 are

$$(\underline{\hat{\pi}}, \underline{\hat{\theta}}) = \underset{\underline{\pi}, \underline{\theta}}{argmax}\ l_n(x|\underline{\pi}, \underline{\theta})$$

When the mixture component number k is unknown, a general extension of Algorithm 2 is Algorithm 1, in which we replace the OLS estimator with the positive LASSO and start from the $k = m$ model, which is the largest possible saturated model. Setting $k = m$ allows the lassoing mixture approach to be cast as a *penalized MLE* (PMLE) with penalty imposed on the mixture weights by the L-1 parameter constraints of the (positive) LASSO. Due to the sparsity property of the LASSO, redundant mixture component weights may be estimated to be 0. Thus the lassoing mixtures approach can be used in the general case when the mixture component number is unknown.

Next, we will show that the LASSOed mixture estimate from Algorithm 1 is indeed consistent. Additionally, consistency of the estimate of the number of components k can be obtained by making use of the new adaptive LASSO algorithm of Zou (2005). Note that the general strategy of using a saturated model as a starting point to fit the mixture model has also been suggested in sorts as far back as Laird (1978), although not in a regularized fashion as what is presented here.

The following Theorem 2 shows the consistency of the final lassoed mixture estimate.

Theorem 2. Let f denote the true mixture density and $\hat{f}$ denote our estimate, then $\hat{f} - f \xrightarrow{p} 0$.

Remarks. Clearly order selection of the finite mixtures can be cast as a variable selection problem. However, the variable selection with LASSO is not always consistent. Meinshausen and Bühlmann (2006) and Ishwaran and Rao (2005) showed the confiict between prediction and variable selection, the optimal λ for prediction gives inconsistent variable selection results. Meinshausen and Bühlmann (2006) derived some sufficient conditions for the consistency of LASSO variable selection, which are not satisfied always. To obtain the general variable selection consistency, Zou (2005) proposed an adaptive LASSO approach. Under the model of Equation 5, he defined data-dependent weights w_j, $j = 1, \ldots, n$ to different coefficients. The adaptive LASSO estimate is

$$\underline{\hat{\pi}}^{Ada-LASSO} = \underset{\underline{\pi}}{argmin} \{\sum_{i=1}^{n} (\tilde{y}_i - \sum_{j=1}^{p} X_{ij} \pi_j)^2 + \lambda \sum_{j=1}^{p} w_j \mid \pi_j \mid\}$$

Then if $\hat{w} = 1/\mid \underline{\hat{\pi}}^{ols} \mid^{\gamma}$, where $\gamma > 0$, the estimate $\underline{\hat{\pi}}^{Ada-LASSO}$ will be consistent and have the optimal estimation rate.

Ishwaran and Rao (2005) also suggested that we could use a hard thresholding procedure for the LASSO estimates where the thresholding parameter must be a function of the sample size to achieve the consistency of LASSO variable selection.

4. SIMULATION STUDIES

To illustrate how our method performs, we make use of the simulation design from Ishwaran *et al*. (2001). They developed a weighted Bayes factor method for estimating the finite mixture models by implemented an generalized weighted Chinese restaurant (GWCR) Monte Carlo algorithm, and did the performance comparison on the simulated data sets with the AIC and BIC approaches. There are 10 different simulations in which data are independently drawn from a finite location normal mixture with unit variance. All experiments except Experiment 1 have uniform weights for the components. In Experiment 1, $\underline{\pi}$ = (1/3, 2/3). Experiments 1-3 have two components and $\underline{\mu}$ = (0, 3), (0, 3), (0, 1.8) respectively. Experiments 4-6 have four components and $\underline{\mu}$ = (0,3, 6, 9), (0, 1.5, 3, 4. 5), (0, 1.5, 3, 6) respectively. Experiments 7-10 have seven components and $\underline{\mu}$ = (0, 3, 6, 9, 12, 15, 18), (0, 1.5, 3, 4.5, 6, 7.5, 9), (0, 1.5, 3, 4.5, 6, 9.5, 12.5), (0, 1.5, 3, 4.5, 9, 10.5, 12) respectively. Experiments 1-6 have sample size n = 100, while Experiments 7-10 have sample size n = 400. Each simulation is repeated 500 times. We fix the same random seed for all experiments to avoid the random fiuctuation due to changing random seeds. Fig. 4 shows the true mixture densities in each simulation. For Experiments 3, 5, 6, 8-10, the mode number is less than the component number because of the close distance between components. We apply our lassoing mixture method to each setting and compare the selection performance to the results published in Ishwaran *et al*. (2001).

Tables 1-3 present the results of our method using CDF/PDF and the results from the AIC, BIC, GWCR algorithms (Ishwaran *et al*. 2001). In Experiments 1-2, Lassoing recognizes the 2-component mixture, though not as consistent as AIC, BIC and GWCR. In Experiment 3, Lassoing discovers the true 2 components, while the other methods tend to recognize 1 component. In Experiment 4, AIC and Lassoing PDF are the winner, and Lassoing CDF method uncover the true model about 20% of the time. In Experiments 5-6, Lassoing methods tend to have higher frequencies of finding the correct number of components as compared with other methods. In Experiments 7-10, Lassoing methods are better than BIC.

Table 1. Results of Simulations 1-3: Samples size is 100 and all distributions has two components. Entries in the last five columns are the percentage of times out of the 500 samples for which the component number estimate equals a candidate dimension value k. Percentages highlighted by boxes indicate highest value and thus represent the best model for a specific procedure.

Exp	# modes	k	AIC	BIC	GWCR	Lassoing	
						CDF	PDF
1	2	1	0.018	0.150	0.018	0.050	0.074
		2	0.896	0.838	0.920	0.584	0.432
		3	0.062	0.012	0.058	0.186	0.220
		4	0.024	0.000	0.004	0.068	0.136
		5	0.000	0.000	0.000	0.046	0.062
		≥ 6	0.000	0.000	0.000	0.066	0.076
2	2	1	0.022	0.212	0.030	0.022	0.016
		2	0.900	0.780	0.916	0.510	0.386
		3	0.050	0.006	0.054	0.232	0.238
		4	0.028	0.002	0.000	0.132	0.172
		5	0.000	0.000	0.000	0.026	0.084
		≥ 6	0.000	0.000	0.000	0.078	0.104
3	1	1	0.702	0.968	0.868	0.106	0.194
		2	0.264	0.030	0.130	0.572	0.416
		3	0.024	0.002	0.002	0.186	0.184
		4	0.000	0.000	0.000	0.044	0.096
		5	0.000	0.000	0.000	0.016	0.048
		≥ 6	0.000	0.000	0.000	0.076	0.062

A proportion of Lassoing results have large estimates for the component number k. One possible reason is that we start from the largest possible model $m = n - 1$. Lassoing PDF tends to have larger estimates than that of Lassoing CDF, one of the possible reasons is the different correlations between the columns of the design matrices of them, which is mentioned in Section 2.

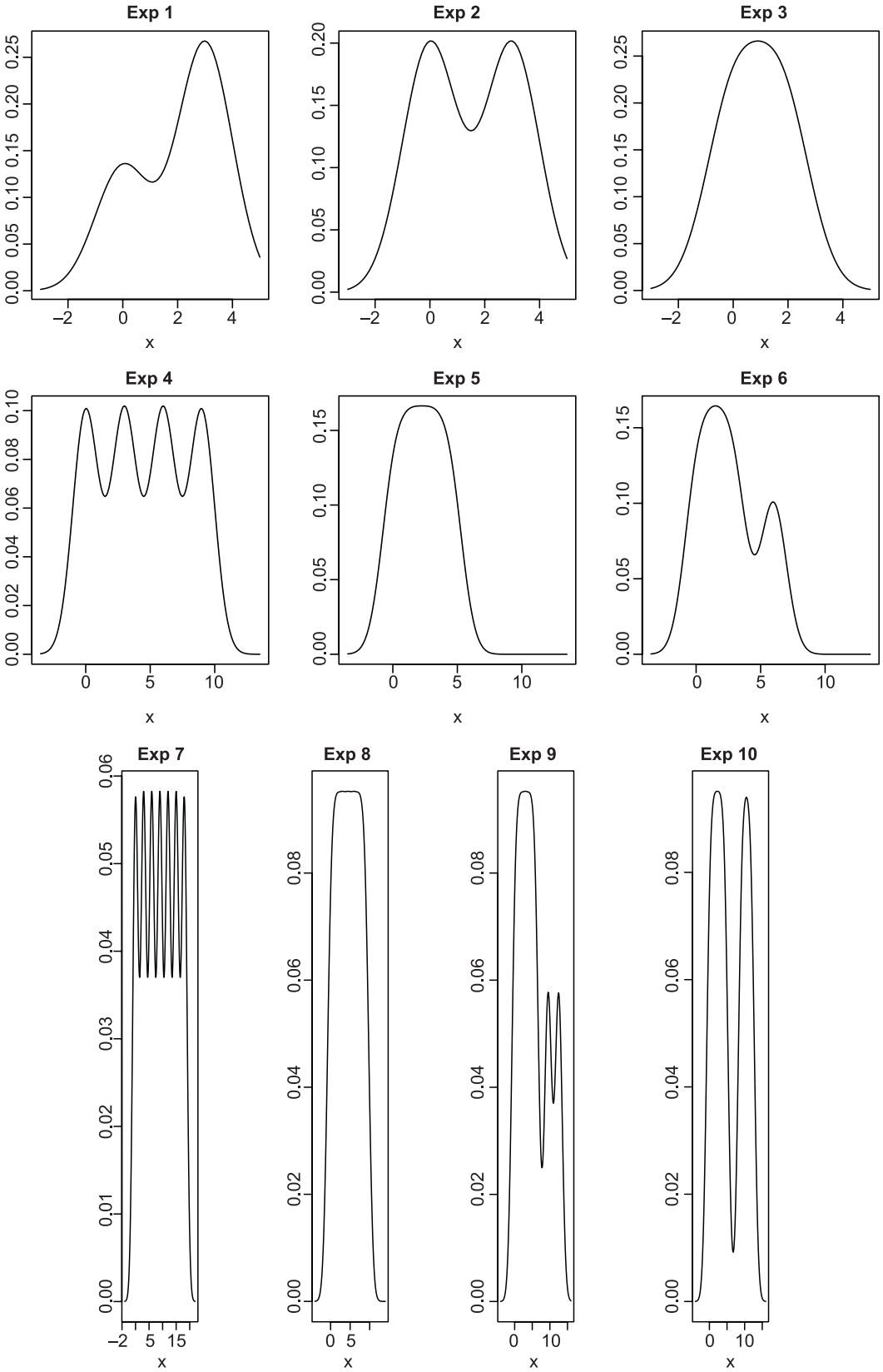

Fig. 4. True mixture densities used in simulation Experiments 1-10.

Table 2. Results of Simulations 4-6: Samples size is 100 and all distributions has four components. Format and methods used are similar to that described in Table 1.

Exp	# modes	*k*	AIC	BIC	GWCR	Lassoing	
						CDF	PDF
4	4	1	0.000	0.110	0.000	0.014	0.034
		2	0.178	0.596	0.102	0.348	0.080
		3	0.110	0.110	0.554	0.198	0.090
		4	0.674	0.182	0.306	0.194	0.268
		5	0.038	0.002	0.038	0.078	0.182
		6	0.000	0.000	0.000	0.050	0.108
		≥ 7	0.000	0.000	0.000	0.118	0.238
5	1	1	0.244	0.748	0.144	0.028	0.060
		2	0.556	0.246	0.818	0.494	0.312
		3	0.142	0.004	0.032	0.230	0.274
		4	0.044	0.002	0.006	0.122	0.142
		5	0.014	0.000	0.000	0.040	0.086
		6	0.000	0.000	0.000	0.018	0.060
		≥ 7	0.000	0.000	0.000	0.068	0.066
6	2	1	0.016	0.188	0.000	0.022	0.060
		2	0.474	0.698	0.612	0.476	0.216
		3	0.392	0.106	0.368	0.208	0.280
		4	0.102	0.008	0.020	0.106	0.184
		5	0.014	0.000	0.000	0.054	0.120
		6	0.000	0.000	0.000	0.036	0.062
		≥ 7	0.002	0.000	0.000	0.098	0.078

5. ANALYSIS OF PROTEOMIC MASS SPECTROSCOPY DATA FOR OVARIAN CANCER CLASSIFICATION

Microarray technology is widely used because it can provide the expression of thousands of genes of the samples at the same time. However, it is argued that proteins are closer to actual biologic functions of cells than mRNAs or DNAs, protein biomarkers of a disease should offer more information about disease than the genetic biomarkers (Wu *et al*. 2003). Protein mass

Table 3. Results of Simulations 7-10: Samples size is 400 and all distributions has seven components. Format and methods used are similar to that described in Table 1.

Exp	# modes	*k*	AIC	BIC	GWCR	Lassoing	
						CDF	PDF
7	7	1	0.004	0.816	0.000	0.008	0.026
		2	0.000	0.000	0.000	0.260	0.218
		3	0.000	0.000	0.010	0.228	0.232
		4	0.302	0.168	0.188	0.234	0.206
		5	0.212	0.016	0.424	0.160	0.086
		6	0.098	0.000	0.178	0.062	0.054
		7	0.326	0.000	0.114	0.032	0.116
		8	0.036	0.000	0.056	0.006	0.042
		9	0.022	0.000	0.030	0.010	0.012
		≥10	0.000	0.000	0.000	0.000	0.008
8	1	1	0.030	0.538	0.000	0.002	0.014
		2	0.684	0.462	0.078	0.354	0.282
		3	0.000	0.000	0.590	0.252	0.326
		4	0.248	0.000	0.272	0.234	0.246
		5	0.000	0.000	0.048	0.096	0.094
		6	0.012	0.000	0.008	0.036	0.026
		7	0.024	0.000	0.004	0.016	0.010
		8	0.002	0.000	0.000	0.010	0.002
9	2+1	1	0.002	0.458	0.000	0.014	0.101
		2	0.000	0.000	0.002	0.384	0.334
		3	0.144	0.398	0.120	0.220	0.268
		4	0.460	0.138	0.408	0.196	0.170
		5	0.308	0.006	0.312	0.106	0.138
		6	0.048	0.000	0.128	0.046	0.058
		7	0.016	0.000	0.024	0.020	0.014
		8	0.022	0.000	0.006	0.010	0.006
		9	0.000	0.000	0.000	0.004	0.002
10	1+1	1	0.000	0.000	0.000	0.010	0.014
		2	0.496	0.992	0.020	0.292	0.232
		3	0.000	0.000	0.370	0.220	0.310
		4	0.302	0.006	0.466	0.256	0.242
		5	0.118	0.002	0.128	0.112	0.150
		6	0.064	0.000	0.010	0.060	0.026
		7	0.016	0.000	0.006	0.034	0.010
		8	0.004	0.000	0.006	0.008	0.008
		9	0.000	0.000	0.000	0.008	0.002
		≥10	0.000	0.000	0.000	0.000	0.006

spectrometry is the new technique to analyze protein expression. It produces the mass/charge ratio (m/z) spectra of the interested proteins with high definition, as is shown in Fig. 5. The x axis of the spectra is the protein mass divided by the number of charges introduced by ionization, and the y axis is the protein intensity of the corresponding x value. This analysis can be conducted on thousands of proteins over a large number of samples simultaneously and can be used to detect the quantitative or qualitative changes between samples. An important application in the early cancer detection is to classify and predict cancer on the basis of protein spectra.

The ovarian cancer data set was produced by the Keck Laboratory at Yale University. It consists of MALDI-MS spectra of 47 patients and 42 normal persons. Each spectra includes 91360 measurements spaced 0.019 dalton (Da) apart, where dalton is the atomic mass unit. It has been analyzed by Wu *et al*. (2003) and Tibshirani *et al*. (2004).

Firstly, since each spectrum consists of intensity measurements at the ordered grid points, it is natural to use a density curve to smooth the data and construct the classifier. Finite normal mixture distribution is a convenient and robust option for the spectra density fitting. Because of the wide span of the measurements (800 Da-3500 Da), it is inappropriate to use only one density curve to fit the data. We split the data into M pieces in order with the width L Da for each piece, and construct the classifier for each piece. Let $G_m(x)$ denote the classifier for the mth piece, $m = 1, 2, \ldots, M$, $C \in (-1, 1)$ denote the patient and control classes, the predictions from all of the classifiers are then combined through a majority vote to produce the final prediction,

$$G(x) = sign\left(\sum_{m=1}^{M} G_m(x)\right) \tag{9}$$

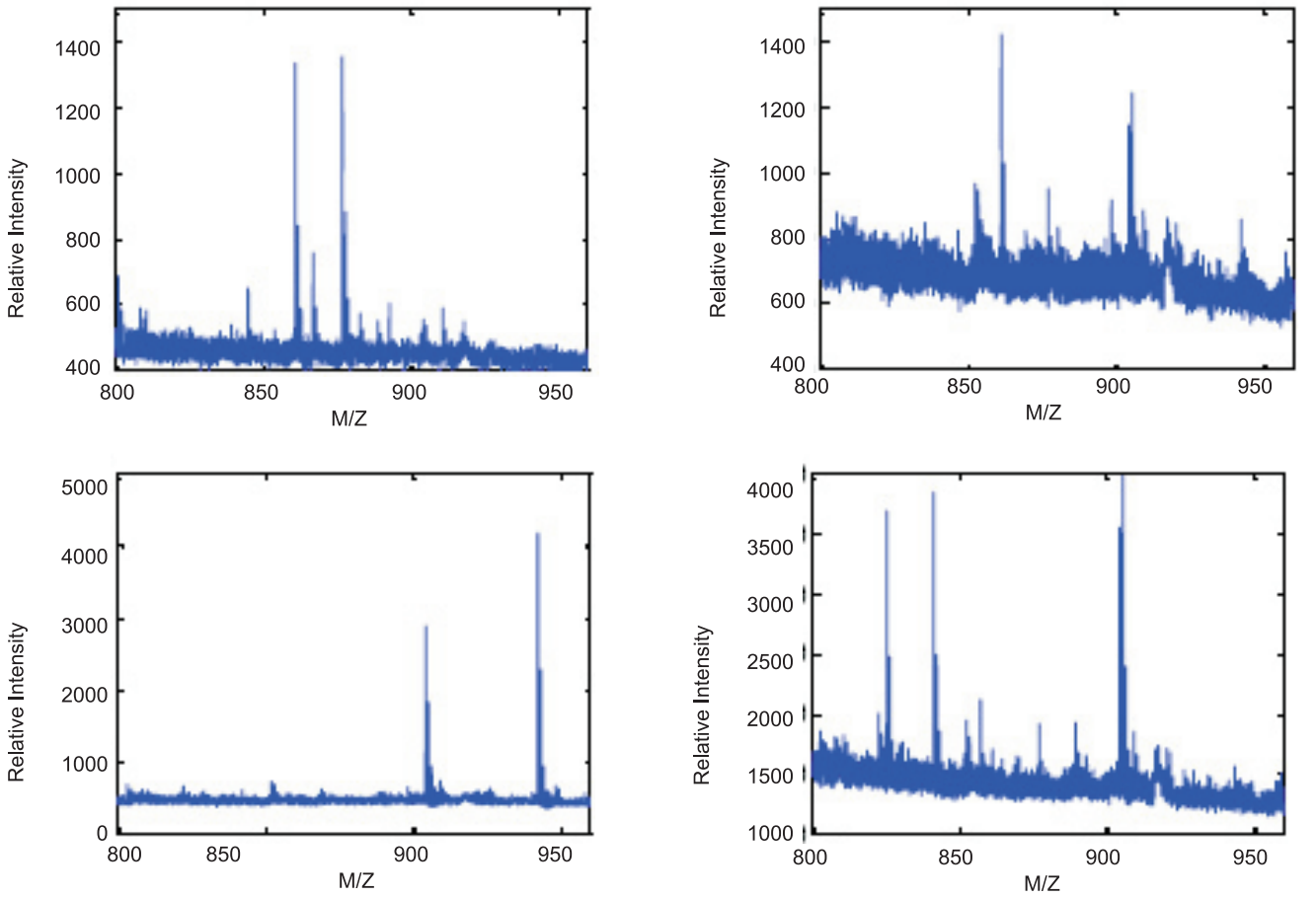

Fig. 5. Protein mass spectrometry sample plots. The top two spectra are from the cancer samples and the bottom two spectra are from the control samples.

For each piece, we fit finite normal mixture densities to both patient data and control data. An observation will be assigned to the class with less distance.

$$G_m(x) = \begin{cases} -1, & \text{if } \sum (Y_m - \hat{F}_{m,p}(x))^2 \\ & \leq \sum (Y_m - \hat{F}_{m,c}(x))^2 \\ 1, & \text{otherwise} \end{cases} \quad (10)$$

Here Y_m is the normalized intensity (density) vector of the mth piece, $\hat{F}_{m,p}(x)$ and $\hat{F}_{m,c}(x)$ are the predicted densities for the m^{th} piece with the mixture models for the patients and controls correspondingly.

To fit the finite normal mixture models, we use the data points in the window of piece width $\pm m/2$ Da. The original intensities are normalized to make them a good approximation of the densities at the m/z values. In other words, $\int Y_m(x)\text{dx} \cong \sum Y_m(x_i)(x_{i+1} - x_i) = 1$. The common standard deviation of the component densities is fixed as 0.2 through some preliminary studies to make the mixture component number in the interval 1 – 10. The mixture component number and the location parameters of the component densities are selected by our lassoing approach. Due to the optimization complexity, we will not update our estimates by iteration. Instead, we use a sequence of fine grided location candidates to reduce the potential bias. We construct the candidate component densities with ordered location parameter spanned from the beginning to the end of the corresponding piece with the spaced interval 0.01, so the precision of the location parameter estimates is 0.01.

We use 5-fold cross validation to choose the piece width L, the result is summarized in Table 4. It seems that window width 5 Da has the best performance with cross validation error rate 26/89. Further possible improvement in the prediction accuracy can be achieved by biomarker selection and background noise detection.

Table 4. Results for 5-fold cross validation with window width 2-6 Da

Width (Da)	2	3	4	5	6
Cross Validation Errors/89	35	31	35	26	31

The peak probability contrasts (PPC) approach was proposed by Tibshirani *et al*. (2004). It extracts the common set of peaks from individual spectra, and apply the nearest shrunken centroid classifier to the set of extracted features. They used the same ovarian cancer data as the example, and showed that the tenfold cross-validated misclassification rate is 23/89-30/89 with different options in the algorithm. These results are comparable with that of our method.

Instead of using the extracted peak measurements, we use all information to construct the classification model. The reason is that a biomarker that has large relative intensity measurement does not always act differently in patients and controls, it may have large measurements in all spectra hence not a useful feature for classification. Our method uses mixture models, so it could provide pattern features that differentiate patients from controls. The common peaks extracted by the PPC method represent a discrete re-expression of the pattern features we find, which does not contain the discriminative information in other parts of the spectra. Interestingly, if we plot the mixture component locations versus the corresponding probabilities as in Fig. 6, we have a plot similar to the peak centroid plot in Tibshirani *et al*.'s paper. Each component of the mixture model represents a "peak", which does not have one-to-one correspondence with the real peaks in

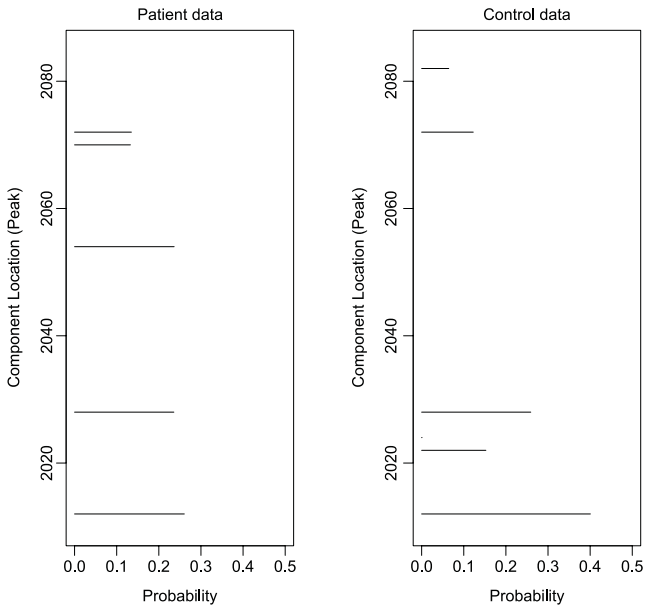

Fig. 6. Mixture component locations vs. probability plot.

spectra. Usually there are more components ("peaks") than the number of peaks extracted by Tibshirani's PPC method, since we need to use a few mixture components to capture the peak feature. We see that the patient model and control model share some components (peaks) though with different weights, and there are some components (peaks) distinguish these two groups.

6. DISCUSSION

In this paper we propose a new method to do mixture model estimation, especially for the situation when the number of components in unknown. By generating pseudo-response and candidate design matrix, we treat the mixture model estimation problem as a variable selection problem in linear regression. We show that our lassoing method is competitive with other methods such as GWCR, AIC, and BIC. Our method has a computation efficiency advantage since we do not do sequential model construction but instead soft thresholding using a variant of the LARS algorithm to select an appropriate model, which has the same computation order as the OLS estimation.

Multivariate mixture models are usually difficult for other methods to handle because of the model complexity and the heavy computation load when calculating the supremum of a multivariate density or drawing posterior samples from a multivariate parameter space. As we noted, only few paper such as Stephens (2000) included multivariate data application examples. Our approach can be extended easily to multivariate cases since the linear relationship of density is the same as that in univariate mixture models. The only additional computation is the EM algorithm for multivariate density.

ACKNOWLEDGEMENTS

Guan Xing is a Sr. Reserach Biostatistician at Bristol-Myers Squibb, and this work comprised a portion of his dissertation. J. Sunil Rao is Professor of Biostatistics and Genetic Epidemiology at Case Western Reserve University. His research is partially supported by NSF grant DMS-0203724. The authors would like to thank Hemant Ishwaran for helpful conversations while conducting this research.

REFERENCES

Böhning, D. (1999). *Computer-Assisted Analysis of Mixtures and Applications* : *Meta-Analysis, Disease Mapping and Others. Boca Raton*, FL, Chapman & Hall/CRC.

Broadbent, D.E. (1966). A difficulty in assessing bimodality in certain distributions. *British J. Math. Statist. Psych*., **19**, 125-126.

Charnigo, R. and Sun, J. (2004). Testing homogeneity in a mixture distribution via the L^2 distance between competing models. *J. Amer*. *Statist*. *Assoc*., **99**, 488-498.

Chen, H., Chen, J. and Kalbeisch, J.D. (2001). A modified likelihood ratio test for homogeneity in finite mixture models. *J. Roy. Statist. Soc. B,* **63**, 19-29.

Clark, V.A., Chapman, J.M., Coulson, A.H. and Hasselblad, V. (1968). Dividing the blood pressures from the Los Angeles heart study into two normal distributions. *Johns Hopkins Med. J*., **122**, 77-83.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann. Statist*., **32**, 407-499.

Eguchi, S. and Yoshioka, K. (2001). Maximum penalized likelihood estimation of finite mixtures with a structural model. *ISM Research Memo*, 809.

Gordon, A.D. and Prentice, I.C. (1977). Numerical methods in quaternary palaeoecology. IV. Separating mixtures of morphologically similar pollen taxa. *Rev. Palaeobotany Palynology*, **23**, 359-372.

Hartigan, J.A. (1985). A failure of likelihood asymptotics for normal mixtures. *Proc. of the Berkeley Conference in Honor of J. Neyman and J. Kiefer*, Vol. II, pp. 806-810.

Ishwaran, H., James, L.F., and Sun, J. (2001). Bayesian model selection in finite mixtures by marginal density decomposition. *J. Amer. Statist. Assoc*., **96**, 1316-1332.

Ishwaran, H. and Rao, J.S. (2005). Total risk for model selection. *Technical Report*. Department of Epidemiology and Biostatistics, Case Western Reserve University.

Laird, N.M. (1978). Nonparametric maximum likelihood estiamtion of a mixing distribution. *J. Amer. Statist. Assoc*., **73**, 805-811.

Macdonald, P.D.M. and Pitcher, T.J. (1979). Age gropus from size-frequency data: A versatile and efficient method of

analyzing distribution mixtures. *J. Fish. Res. Board Can*., **36**, 987-1001.

Meinshausen, N. and Böuhlmann, P. (2006). High dimensional graphs and variable selection with the Lasso. *Ann. Statist*., **34**, 1436-1462.

Ott, J. (1999). *Analysis of Human Genetic Linkage*. 2nd ed., The Johns Hopkins University Press, Baltimore and London.

Sahu, S. and Cheng, R. (2002). A fast distance-based approach for determining the number of components in mixtures. *Can. J. Statist*., **31**, 3-22.

Schlattmann, P. and Böhning, D. (1993). Mixture models and disease mapping. *Statist. Med*., **12**, 943-950.

Smith, C.A.B. (1961). Homogeneity tests for linkage data. *Proc. Sec. Int. Congr*. *Hum. Genet*., **1**, 212-213.

Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components - An alternative to reversible jump methods. *Ann. Statist*., **28**, 40-74.

Tanner, W.F. (1962). Components of the hypsometic curve of the earth. *J. Geophy. Res.,* **67**, 2841-2843.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. B,* **58**, 267-288.

Tibshirani, R., Hastie, H., Narasimhan, B., Soltys, S., Koong A., and Le, Q. (2004). Sample classification from protein mass spectroscopy, by "peak probability contrasts". *Bioinformatics*, **20**, 3034-3044.

Wu, B., Abbott, T., Fishman, D., McMurray, W., Mor, G., Stone, K., Ward, D., Williams, K., and Zhao, H. (2003). Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, **19**, 1636-1643.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *J. Amer. Statist. Assoc*., **101**, 1418-1429.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Statist*., **15**, 265-286.

APPENDIX

Proof of Lemma 1.

Suppose we want to maximize the mixture likelihood $\prod_{i=1}^{n}[\sum_{j=1}^{k}\pi_j f(x_i|\theta_j)]$, where k is known. The complete likelihood is as follows:

$$L = \prod_{i=1}^{n}\sum_{j=1}^{k}[\pi_j f(x_i|\theta_j)]^{Zij}$$

where Z_{ij} is the indicator variable whether the ith observation comes from the jth component. The weight estimates are $\hat{\pi}_j^{s+1} = \frac{\sum_{i=1}^{n} Z_{ij}^{s}}{n}$, and $\hat{Z}_{ij}^{s+1} = \frac{\pi_j^s f(x_i|\theta_j)}{\sum_{j=1}^{k}\pi_j^s f(x_i|\theta_j)}$ where s is the iteration number of EM.

$Z_{1j}, \ldots, Z_{nj}$ are *i.i.d*, so we have

$$\begin{aligned} E(\hat{\pi}_j) &= \frac{\sum_{i=1}^{n} E(Z_{ij})}{n} \\ &= E(Z_{i=1,\,j}) \\ &= \frac{\pi_j E f(x_1|\theta_j)}{\sum_{j=1}^{k}\pi_j E f(x_1|\theta_j)} \\ &= \frac{\pi_j}{\sum_{j=1}^{k}\pi_j} = \pi_j \end{aligned}$$

The estimate of $\underline{\pi}$ is unbiased.

Proof of Lemma 2.

Let $y_1, y_2, \ldots, y_n$ denote the density estimates of a saturated model at the observations. Suppose there are another two density estimates of $X\pi_1$ and $X\pi_2$, for which we use $U = \{u_i, i = 1, \ldots, n\}$ and $V = \{v_i, i = 1, \ldots, n\}$ to denote, $U = X\pi_1$ and $V = X\pi_2$.

We want to show that if

$$\sum_{i=1}^{n}(y_i - u_i)^2 \le \sum_{i=1}^{n}(y_i - v_i)^2$$

then $\prod_{i=1}^{n} u_i \ge \prod_{i=1}^{n} v_i$

We need to restrict the choice of $\underline{\pi}$ in the class of unbiasedness. In other words, $E\underline{\pi}_1 = E\underline{\pi}_2$, which is a reasonable assumption based Lemma 1. It's equivalent to show that if $\sum_{i=1}^{n}(y_i - u_i)^2 \le \sum_{i=1}^{n}(y_i - v_i)^2$, then $\sum_{i=1}^{n} \ln u_i \ge \sum_{i=1}^{n} \ln v_i$. Using the Weak Law of Large Number, we change the problem to that if $E(Y - U)^2 \le E(Y - V)^2$, then $ElnU > ElnV$.

$$E(Y - U)^2 \le E(Y - V)^2$$

$$\Rightarrow \quad EU^2 - 2EY\,EU \le EV^2 - 2EY\,EV$$

Because that $E\underline{\pi}_1 = E\underline{\pi}_2$, $EU = EV$, we get $EU^2 \le EV^2$.

Using Taylor expansion $lnx = (x - 1) - \frac{1}{2}(x - 1)^2 + O(x^3)$ and only keep the first two terms, $E \ln U - E \ln V \approx E(2U - 2V + \frac{1}{2}V^2 - \frac{1}{2}U^2) \ge 0$. Because $\underline{\hat{\pi}}^{ols} = \underset{\underline{\pi}}{argmin}\,(Y - X\underline{\pi})^2$, $\underline{\hat{\pi}}^{ols}$ gives the largest likelihood for all unbiased $\underline{\hat{\pi}}$.

Proof of Theorem 1.

Let Y denote the pseudo-response from a saturated mixture model. When we use Algorithm 2, at the step of OLS, for fixed X, $X\underline{\hat{\pi}}^{ols}$, is closer to Y than any other $X\underline{\hat{\pi}}$, the likelihood is increased according to Lemma 2. At the step of Fixed-weight-EM, the likelihood is also increasing. So that our final estimates are MLE.

Proof of Theorem 2.

Suppose our algorithm stops at m steps. Let $\tilde{f}$ denote the random variable corresponding to the saturated distribution, $f_1, \ldots, f_m$ denote the corresponding random variables with corresponding

mixture distribution at the mth step, and $\hat{f}$ denote the random variable of the final fitted mixture distribution. $f_1, \ldots, f_m$ have different parameter sizes and $\hat{f}$ is derived from f_m by EM. We assume that all random variables are uniformly bounded. Then $f - c \xrightarrow{p} 0$ is equivalent to $E(f - c)^2 \xrightarrow{p} 0$. We will discuss three situations: $m = 1$, $m = o_p(n)$ and $m \asymp n$.

If $m = 1$ then

$$\begin{aligned} E(\tilde{f} - \hat{f})^2 &= \mathrm{E}(\tilde{f} - f_1 + f_1 - \hat{f})^2 \\ &= \mathrm{E}(\tilde{f} - f_1)^2 + E(f_1 - \hat{f})^2 \\ &\quad + 2E(\tilde{f} - f_1)(f_1 - \hat{f}) \end{aligned}$$

From the consistency of LASSO and EM, we have $f_1 - \tilde{f} \xrightarrow{p} 0$ and $\hat{f} - f_1 \xrightarrow{p} 0$. So the crossover term will be zero and $E(\tilde{f} - \hat{f})^2 \xrightarrow{p} 0$, $\hat{f} - \tilde{f} \xrightarrow{p} 0$. Because $\tilde{f}$ is consistent, $\tilde{f} - f \xrightarrow{p} 0$, our estimate $\hat{f}$ will also be consistent.

If $m \neq 1$ but $m = o_p(n)$, we will have a finite sum of $o_p(1)$, which will still be $o_p(1)$. Our estimate will still be consistent.

If $m \asymp n$, we could not show the consistency of our estimate. The simulation study shows that our algorithm converges very fast. We might not worry about this situation.