International Journal of Pure and Applied Mathematics
Volume 46 No. 3 2008, 403-414

ON MIXTURE REGRESSION SHRINKAGE AND SELECTION VIA THE MR-LASSO

Ronghua Luo[1], Hansheng Wang[2], Chih-Ling Tsai[3] §
[1,2]Guanghua School of Management
Peking University
Beijing, 100871, P.R. CHINA
[1]e-mail: luoronghua@gsm.pku.edu.cn
[2]e-mail: hansheng@gsm.pku.edu.cn
[3]Graduate School of Management
University of California
Davis, CA 95616-8609, USA
e-mail: cltsai@ucdavis.edu

Abstract: In finite mixture regression models, we generalize the application of the *least absolute shrinkage and selection operator* (LASSO) to obtain MR-LASSO, which incorporates both mixture and regression penalties. Because MR-LASSO jointly penalizes both regression coefficients and mixture components, it enables simultaneous identification of significant variables and determination of important mixture components. Simulation studies indicate that MR-LASSO outperforms LASSO. Extensions to mixture non-Gaussian and mixture time series models are briefly described.

AMS Subject Classification: 62J07
Key Words: finite mixture model, LASSO, mixture penalty

1. Introduction

In the last two decades, finite mixture regression models have been extensively used in various areas (e.g., astronomy, biology, genetics, engineering, social science, and marketing). These models provide an approach for classifying objects

Received: June 2, 2008

§Correspondence author

into components (or clusters) and estimating regression models across different components simultaneously. Several useful reference books can be found in Titterington et al [11], McLachlan and Peel [8], and Wedel and Kamakura [15].

In practice, the number of components and the regression coefficients included in each component are often unknown. Hence, the determination of the number of components and variables is critical in fitting the data with a finite mixture regression model. It is natural to extend the classical variable selection criteria, e.g., Akaike information criterion (Akaike [1]) and Bayesian information criterion (Schwarz [9]), to mixture regression models so that they are able to jointly choose the number of components and variables (Desarbo and Corn [5]). However, the classical variable selection criteria can be unstable (Breiman [2]) and their computational costs are expensive. To overcome those limitations, various shrinkage methods have been proposed in the past decade. Those useful methods include but are not limited to: *least absolute shrinkage and selection operator* (Tibshirani [10], LASSO), *bridge regression* (Fu [7]), and *smoothly clipped absolute deviation* (Fan and Li [6]). It has been shown that those shrinkage method is able to yield more stable and interpretable models.

For illustration purpose, we only considered LASSO-type penalty in this article. Nevertheless, similar idea can be extended to other useful shrinkage method without much difficulty. To better understand the motivation of LASSO, we consider the single component regression model, $y_i = x_i'\beta + \varepsilon_i$, where y_i is the response, $x_i = (x_{i1}, \cdots, x_{ip})'$ is a p-dimensional covariate, $\beta = (b_1, \cdots, b_p)'$, and ε_i are independent identically random errors for $i = 1, \cdots, n$. Then, the LASSO estimate is $\tilde{\beta} = \arg_\beta \min\{\sum_i (y_i - \sum_j b_j x_{ij})^2 + \gamma \sum_j |b_j|\}$, where $\gamma \geq 0$. Lasso yields good estimators when the regression model has the single component.

In this article, we extend the application of LASSO from the classical regression model to the finite mixture regression model. We propose a mixture regression LASSO (MR-LASSO) approach to jointly take into account both the mixture and regression penalties. The mixture penalty (MP) penalizes the L_2 distance of the regression coefficients between components, while the regression penalty (RP) penalizes the non-zero regression coefficients within each component. In other words, MP encourages combining similar mixture components into one common component, thereby forcing them to share the same regression coefficients to produce sparse solutions between components. In contrast, RP produces sparse solutions within each component. As a result, the MR-LASSO simultaneously forms important mixture components by merging similar components together and selects significant regression variables within

each component.

The rest of the paper is organized as follows. In Section 2, we define MR-LASSO for mixture regression models and propose a modified EM algorithm to obtain MR-LASSO estimates. Section 3 presents Monte Carlo studies. Then, the article is concluded with a short discussion in Section 4.

2. MR-LASSO and Modified EM Algorithm

2.1. MR-LASSO

Consider a finite mixture regression model whose density function is

$$f(Y;\phi) = \prod_{i=1}^{n} \sum_{k=1}^{K} \alpha_k f_k(y_i; x_i'\beta_k, \sigma_k), \tag{2.1}$$

where $Y = (y_1, \cdots, y_n)'$, $\phi = \{(\alpha_k, \beta_k, \sigma_k^2) : k = 1, \cdots, K\}$, $0 \le \alpha_k \le 1$, $\sum_{k=1}^{K} \alpha_k = 1$, $f_k(y_i; x_i'\beta_k, \sigma_k)$ is the normal density with mean $x_i'\beta_k$ and standard deviation σ_k, and $\beta_k = (b_{k1}, \cdots, b_{kp})'$. Without loss of generality, we consider $p \ge \max\{p_k; k = 1, \cdots, K\}$, where p_k is the number of covariates in the k-th component (e.g., p is the number of different covariates that occured in all K clusters). We further assume that there is K_0 so that $\alpha_k > 0$ for $k \le K^0 \le K$ and $\alpha_k = 0$ for $K^0 < k \le K$. In other words, we assume that only K_0 of K candidate mixture components truly exist.

To estimate unknown parameters ϕ in the mixture regression model (2.1), we can follow McLachlan and Peel's [8] direct approach to find parameter estimators which maximize the loglikelihood function of ϕ. However, this approach would preclude a straightforward extension of Tibshirani's [10] LASSO to the finite mixture regression model. Hence, we consider the complete-data loglikelihood function (see McLachlan and Peel [8]),

$$\begin{aligned} &-\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\left\{\log f(y_i; x_i'\beta_k, \sigma_k) + \log(\alpha_k)\right\} \\ &\quad = \sum_{k=1}^{K}\Big[\frac{1}{n}\sum_{i=1}^{n} z_{ik}\left\{(y_i - x_i'\beta_k)^2/(2\sigma_k^2) - \log(2\pi\sigma_k)/2 - \log(\alpha_k)\right\}\Big], \end{aligned} \tag{2.2}$$

where $z_{ik} = 1$ when the i-th observation arises from the k-th mixture component and $z_{ik} = 0$ otherwise. For the known z_{ik}, the parameter estimate of β_k can be

obtained by simply minimizing the following least squares function,

$$\sum_{k=1}^{K}\left\{\frac{1}{2n}\sum_{i=1}^{n} z_{ik}(y_i - x_i'\beta_k)^2\right\}. \tag{2.3}$$

Based on (2.3), we are able to directly adapt Tibshirani's [10] approach to shrink unnecessary coefficients to zero in finite mixture models, which yields the following LASSO criterion,

$$\sum_{k=1}^{K}\left\{\frac{1}{2n}\sum_{i=1}^{n} z_{ik}(y_i - x_i'\beta_k)^2 + \sum_{j=1}^{p}\gamma|b_{kj}|\right\}.$$

Because LASSO employs a common tuning parameter for all regression coefficients, it may yield a large bias (see Fan and Li [6]). Therefore, we follow the idea of Zou [18], Zhang and Lu [17], Wang et al [13] and propose the following adaptive LASSO criterion, aLASSO, which allows for different tuning parameters for different coefficients:

$$\sum_{k=1}^{K}\left\{\frac{1}{2n}\sum_{i=1}^{n} z_{ik}(y_i - x_i'\beta_k)^2 + \sum_{j=1}^{p}\gamma_{kj}|b_{kj}|\right\},$$

where γ_{kj} are the non-negative tuning parameters. In practice, however, the number of true components is often unknown. Hence, we further generalize the idea of LASSO to finite mixture regression models by proposing the following MR-LASSO criterion,

$$\sum_{k=1}^{K}\left\{\frac{1}{2n}\sum_{i=1}^{n} z_{ik}(y_i - x_i'\beta_k)^2 + \frac{\lambda}{2}\sum_{l\neq k}^{K}||\beta_k - \beta_l|| + \sum_{j=1}^{p}\gamma_{kj}|b_{kj}|\right\}, \tag{2.4}$$

where $||\cdot||$ stands for a usual L_2 norm, and λ is the non-negative tuning parameter for mixture components. This criterion includes two LASSO-type penalties. The first is the second term of equation (2.4), which penalizes the L_2 distance of the regression coefficients between the mixture components. Hence, we refer to it as the mixture penalty (MP). The second is the last term of equation (2.4), which penalizes the regression coefficients within each component, and which we term the regression penalty (RP).

To better understand the motivation underlying MR-LASSO, we consider a case with two clusters ($K = 2$) and one predictor within each cluster ($p = 1$). For the purpose of illustration, we also assume that $\lambda = \gamma_{kj} = 1.0$. Then, the resulting MR-LASSO penalties becomes

$$|\beta_1 - \beta_2| + |\beta_1| + |\beta_2|.$$

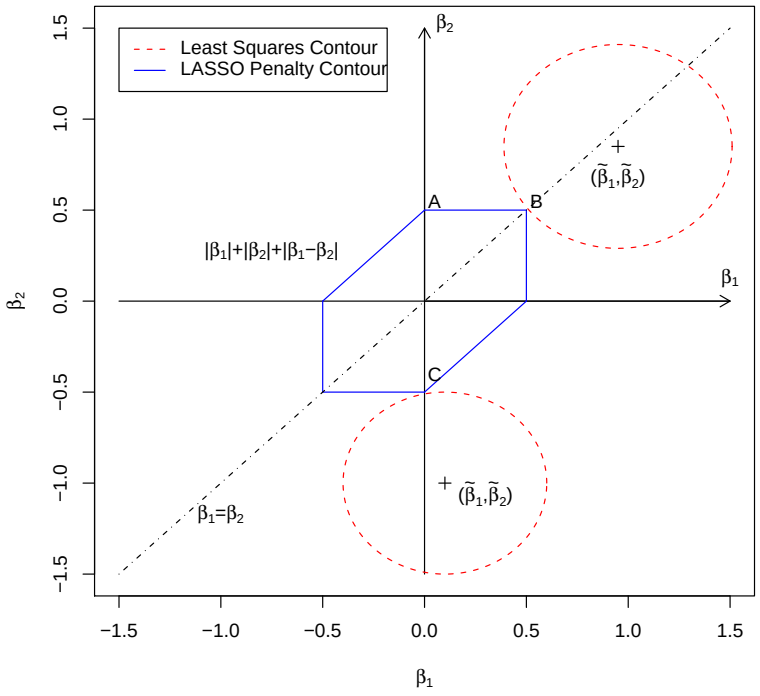

Figure 1: Graphical display of the MR-LASSO Penalty

Figure 1 depicts the six-sided polygon (see the solid line) of the two LASSO-typed penalties and the circle contours (see the dashed-line) of the non-constrained parameters centered with their respective least squares estimators, $(\tilde{\beta}_1, \tilde{\beta}_2)$. It shows that if the circle contour hits the polygon at points "A" and "C", then the sparse solutions ($\beta_1 = 0$, $\beta_2 > 0$) and ($\beta_1 = 0$, $\beta_2 < 0$) are obtained, respectively. In contrast, if the circle contour hits the polygon at point "B", then the sparse solution ($\beta_1 = \beta_2$) is produced. Therefore, the proposed penalty functions produce sparse solutions for both components and regression coefficients.

2.2. A Modified EM Algorithm

In practice, the z_{ik} in (2.2) is unobserved missing data. To obtain MR-LASSO estimates, we adapt McLachlan and Peel's [8] approach to replace z_{ik} by its expected value via the expectation-maximization (EM) algorithm (Dempster et al [4]). Specifically, in the E-step, we replace z_{ik} by

$$\tau_{ik}^{(m)} = \tau_{ik}^{(m)}(\phi^{(m)}) = \frac{\alpha_k^{(m)} f(y_i; x_i'\beta_k^{(m)}, \sigma_k^{(m)})}{\sum_{l=1}^K \alpha_l^{(m)} f(y_i; x_i'\beta_l^{(m)}, \sigma_l^{(m)})},$$

where

$$\phi^{(m)} = \{(\alpha_k^{(m)}, \beta_k^{(m)}, \sigma_k^{2(m)}) : k = 1, \cdots, K\}$$

denotes the provisional estimates at the m-th iteration.

In the M-step, we first replace z_{ik} in equation (2.2) by $\tau_{ik}^{(m)}$ and then maximize it with respect to $(\alpha_k, \beta_k, \sigma_k)$ which satisfies the constraint $\sum_{k=1}^K \alpha_k = 1$. This maximization yields the closed-form estimates of α_k and σ_k^2 at the $(m+1)$-th iteration:

$$\alpha_k^{(m+1)} = \frac{1}{n}\sum_{i=1}^n \tau_{ik}^{(m)}, \text{ and } \sigma_k^{2(m+1)} = Y_k^{(m)'}(I - H_k^{(m)})Y_k^{(m)}/\mathrm{tr}(W_k^{(m)}),$$

where

$$H_k^{(m)} = X_k^{(m)}(X_k^{(m)'}X_k^{(m)})^{-1}X_k^{(m)'}, \qquad X_k^{(m)} = W_k^{(m)1/2}X,$$

$$X = (x_1, \cdots, x_n)', \qquad Y_k^{(m)} = W_k^{(m)1/2}Y, \qquad W_k^{(m)} = \mathrm{diag}(\tau_k^{(m)}),$$

and $\tau_k^{(m)} = (\tau_{1k}^{(m)}, \cdots, \tau_{nk}^{(m)})'$.

To obtain the closed form of $\beta_k^{(m+1)}$, we next adapt the local quadratic function (Tibshirani [10], Fan and Li [6]) to approximate the penalty functions of (2.4) that result in the approximate penalized least squares function

$$\sum_{k=1}^K \left\{ \frac{1}{2n}\sum_{i=1}^n \tau_{ik}^{(m)} \left(y_i - x_i'\beta_k^{(m+1)}\right)^2 + \frac{\lambda||\beta_k^{(m)} - \beta_l^{(m)}||}{||\beta_k^{(m+1)} - \beta_l^{(m)}||^2} + \sum_{j=1}^p \frac{\gamma_{kj}}{|b_{kj}^{(m)}|}|b_{kj}^{(m+1)}|^2 \right\}.$$

As $\beta_k^{(m)}$, $\tau_{ik}^{(m)}$ and $\sigma_k^{2(m)}$ converge, the above approximation and equation (2.4) (replacing its unobserved missing data and unknown parameters by their corresponding estimates) yield the same minimum. Moreover, this approximation leads to a closed form of the regression parameter estimate

$$\beta_k^{(m+1)} = (X_k^{(m)'}X_k^{(m)} + D_k^{(m)})^{-1}X_k^{(m)'}(Y_k^{(m)} + V_k^{(m)}),$$

where $D_k^{(m)}$ is a $p \times p$ diagonal matrix with the j-th component

$$2\frac{\gamma_{kj}}{|b_{kj}^{(m)}|} + \sum_{l \neq k} \frac{\lambda}{||\beta_k^{(m)} - \beta_l^{(m)}||} \quad \text{and} \quad V_k^{(m)} = \sum_{l \neq k} \frac{\lambda}{||\beta_k^{(m)} - \beta_l^{(m)}||} \beta_l^{(m)} .$$

Finally, the computation of $\beta_k^{(m+1)}$ requires the tuning parameters γ_{kj}. Because there are many tuning parameters to be calculated, the commonly used GCV approach cannot effectively find solutions. Hence, we propose a simple procedure to dynamically update the tuning parameters within each iteration as follows:

$$\gamma_{kj}^{(m+1)} = \frac{\log(n)}{n|b_{kj}^{(m)}|} .$$

The above tuning parameters shrink to 0 faster for significant variables, and more slowly for non-significant variables. Extensive simulations (not presented here) demonstrate that the proposed approach performs well in finite samples. We remark that such a simple procedure is computationally fast. If one do wish to select tuning parameters in a completely data driven manner, we can replace the factor $\log(n)$ by a hyper tuning parameter (e.g., κ_0). Then, the value of κ_0 can be selected by the BIC-type criterion as suggested by Wang et al [14] and Wang and Leng [12].

For the given mixture tuning parameter, λ, the E- and M-steps are alternated until the difference of $||\phi^{(m)} - \phi^{(m+1)}||_1 = \sum_{k=1}^{K} \sum_{j=1}^{p} |\beta_{kj}^{(m)} - \beta_{kj}^{(m+1)}| + \sum_{k=1}^{K} |\alpha_k^{(m)} - \alpha_k^{(m+1)}| + \sum_{k=1}^{K} |\sigma_k^{2(m)} - \sigma_k^{2(m+1)}|$ decreases below a preset tolerance, where $||\cdot||_1$ is L_1 norm. Let the resulting estimator be $\hat{\phi}(\lambda)$. We then apply a known selection criterion $BIC = \log\{f(Y; \hat{\phi}(\lambda))\} + \log(n)\{\hat{K}(\lambda) + \sum_{k=1}^{\hat{K}(\lambda)} \hat{p}_k(\lambda)\}$ (see Schwarz [9]) to choose the tuning parameter $\hat{\lambda}$ that minimizes BIC, where $\hat{K}(\lambda)$ and $\hat{p}_k(\lambda)$ are the estimators of K and p_k, respectively, for the given λ. Consequently, the $\hat{\lambda}$ together with the above EM algorithm yields the MR-LASSO estimator $\hat{\phi}$.

3. Simulation Results

We present two simulation examples to assess the performance of the LASSO, LASSO*, and MR-LASSO via 1,000 realizations. The number of true components are $K^0 = 1, 3$, while the number of candidate components are $K = 5$. The total number of sample sizes are $n = 100$, 150 and 250. For each realization,

			Mixture Component			Regression Variables			Correct
σ	n	Method	Underfit	Correct	Overfit	Underfit	Correct	Overfit	Model
1.5	100	LASSO	0.000	0.031	0.969	0.000	0.030	0.970	0.000
		aLASSO	0.000	0.192	0.808	0.203	0.797	0.000	0.183
		MR-LASSO	0.000	0.929	0.071	0.016	0.984	0.000	0.914
	150	LASSO	0.000	0.323	0.677	0.000	0.056	0.944	0.015
		aLASSO	0.000	0.300	0.700	0.035	0.965	0.000	0.299
		MR-LASSO	0.000	0.969	0.031	0.000	1.000	0.000	0.969
	250	LASSO	0.000	0.736	0.264	0.000	0.089	0.911	0.059
		aLASSO	0.000	0.504	0.496	0.001	0.999	0.000	0.504
		MR-LASSO	0.000	0.992	0.008	0.000	1.000	0.000	0.992
0.5	100	LASSO	0.000	0.675	0.325	0.000	0.245	0.755	0.171
		aLASSO	0.000	0.832	0.168	0.008	0.992	0.000	0.829
		MR-LASSO	0.000	0.983	0.017	0.000	1.000	0.000	0.983
	150	LASSO	0.000	0.876	0.124	0.000	0.318	0.682	0.282
		aLASSO	0.000	0.880	0.120	0.000	1.000	0.000	0.880
		MR-LASSO	0.000	0.996	0.004	0.000	1.000	0.000	0.996
	250	LASSO	0.000	0.981	0.019	0.000	0.336	0.664	0.329
		aLASSO	0.000	0.953	0.047	0.000	1.000	0.000	0.953
		MR-LASSO	0.000	0.998	0.002	0.000	1.000	0.000	0.998

Table 1: Simulation results for Example 1 ($K^0 = 1$)

the following linear models are used to generate the data

$$\begin{aligned} &\text{Component 1: } y = x_1 + x_2 + x_3 + x_4 + 0x_5 + 0x_6 + 0x_7 + \sigma\epsilon\,, \\ &\text{Component 2: } y = x_1 + 2x_2 + 3x_3 + 4x_4 + 0x_5 + 0x_6 + 0x_7 + \sigma\epsilon\,, \\ &\text{Component 3: } y = 5x_1 + 6x_2 + 7x_3 + 8x_4 + 0x_5 + 0x_6 + 0x_7 + \sigma\epsilon\,, \end{aligned} \tag{3.1}$$

where $x_i, i = 1, \cdots, 7$ are independently generated from a uniform distribution with support $[0, \sqrt{12}]$, and ϵ is independent standard normal random variable with the standard deviation σ, and $\sigma = 0.5$ and 1.5. In sum, we consider the following two examples.

Example 1. Component 1 in (3.1) is used to generate the data so that there is only one true component (i.e., $K^0 = 1$) with four nonzero coefficients.

Example 2. Components 1, 2 and 3 in (3.1) are used to generate the data so that there are three true components (i.e., $K^0 = 3$), and each component contains four non-zero coefficients. In addition, the mixing probabilities of Components 1, 2 and 3 are 0.50, 0.30 and 0.20, respectively.

In these two examples, we apply the modified EM algorithm to compute parameter estimates. The algorithm is first started by setting equal weights on each candidate component and then randomly assigning each observation into

			Mixture Component			Regression Variables			Correct
σ	n	Method	Underfit	Correct	Overfit	Underfit	Correct	Overfit	Model
1.5	100	LASSO	0.006	0.291	0.703	0.000	0.002	0.998	0.001
		aLASSO	0.003	0.247	0.750	0.003	0.614	0.383	0.076
		MR-LASSO	0.006	0.766	0.228	0.007	0.818	0.175	0.313
	150	LASSO	0.001	0.491	0.508	0.000	0.003	0.997	0.002
		aLASSO	0.000	0.470	0.530	0.000	0.953	0.047	0.317
		MR-LASSO	0.000	0.885	0.115	0.000	0.959	0.041	0.719
	250	LASSO	0.000	0.688	0.312	0.000	0.004	0.996	0.004
		aLASSO	0.000	0.510	0.490	0.000	0.999	0.001	0.484
		MR-LASSO	0.000	0.965	0.035	0.000	0.976	0.024	0.930
0.5	100	LASSO	0.004	0.756	0.240	0.000	0.009	0.991	0.007
		aLASSO	0.002	0.700	0.298	0.001	0.563	0.436	0.280
		MR-LASSO	0.005	0.913	0.082	0.001	0.694	0.305	0.591
	150	LASSO	0.002	0.938	0.060	0.000	0.017	0.983	0.017
		aLASSO	0.002	0.912	0.086	0.000	0.776	0.224	0.654
		MR-LASSO	0.002	0.989	0.009	0.000	0.785	0.215	0.773
	250	LASSO	0.000	0.988	0.012	0.000	0.030	0.970	0.030
		aLASSO	0.000	0.935	0.065	0.000	0.918	0.082	0.851
		MR-LASSO	0.000	0.995	0.005	0.000	0.918	0.082	0.913

Table 2: Simulation results for Example 2 ($K^0 = 3$)

each component. Next, the algorithm is iterated without the MR penalty until the difference between two consecutive estimators falls below an initial critical value of 10^{-3}. Finally, the algorithm is fully iterated with the MR penalty until the convergence criterion of 10^{-6} is satisfied. By the time convergence is reached, the estimate of α_k, b_{kj}, or $||\beta_j - \beta_l||/\sqrt{p}$ has been shrunk to 0 if its corresponding absolute value is less than 0.01.

To compare LASSO, aLASSO, and MR-LASSO, Tables 1 and 2 present the percentage of correctly (under, over) estimated numbers of components; the percentage of correctly (under, over) estimated numbers of regression coefficients, and the percentage of both mixture components and significant variables being correctly identified. Apparently, Lasso performs poorly with severe overfittings. This is because LASSO's tuning parameter is fixed, and therefore is not able to effectively shrink non-significant coefficients to zero. In contrast, aLASSO, with the dynamic tuning parameter is able to compress a greater number of unnecessary coefficients to zero. Consequently, aLASSO performs better than LASSO in identifying the correct mixture models, especially when the sample size is large or $K^0 = 3$. However, aLASSO often lacks the ability to merge redundant components (e.g., see $\sigma = 1.5$). This finding is not surprising since

aLASSO does not have the MP penalty to prevent overfitting of components.

After accounting for both the MP and RP penalties in parameter estimations, Tables 1 and 2 show that MR-LASSO clearly outperforms LASSO. As compared with aLASSO, MR-LASSO displays a similar ability to shrink non-significant coefficients to zero. However, MR-LASSO also enables the identification of a greater number of correct components than aLASSO across various sample sizes (n) and standard errors (σ). Hence, MR-LASSO is superior to aLASSO in overall correct model identification. It is worth noting that the MP plays a more important role in Example 1 than in Example 2, as there are more redundant components in Example 1 that need to be shrunk.

4. Discussion

In finite mixture regression models, we generalized Tibshirani's [10] LASSO to simultaneously compress regression coefficients to zero, while penalizing the difference of regression coefficients between mixture components. The resulting MR-LASSO enables us to not only identify significant variables but also to determine important mixture components. Because mixture models have been widely applied in practice, we can further extend MR-LASSO to the contexts of mixture Kalman filters (Chen and Liu [3]), mixture generalized linear models (see Wedel and Kamakura [15]; McLachlan and Peel [8]), and mixture autoregressive models (Wong and Li [16]). We believe these efforts would enhance the usefulness of MR-LASSO in mixture models and dimension reduction.

Acknowledgements

Ronghua Luo and Hansheng Wang's research are partially supported by NSFC Grant (10771006).

References

[1] Akaike, Information theory and an extension of the maximum likelihood principle, In: *2-nd International Symposium on Information Theory* (Ed-s: B.N. Petrov, F. Csaki), Budapest, Akademia Kiado (1973), 267-281.

[2] L. Breiman, Heuristics of instability and stabilization in model selection, *The Annals of Statistics*, **24** (1996), 2350-2383.

[3] R. Chen, J.S. Liu, Mixture Kalman filters, *Journal of Royal Statistical Society, Series B*, **62** (2000), 493-508.

[4] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the em algorithm (with discussion), *Journal of the Royal Statistical Society, Series B*, **39** (1977), 1-38.

[5] W.S. DeSarbo, W.L. Corn, A maximum likelihood method for clusterwise linear regression, *Journal of Classification*, **5** (1988), 249-282.

[6] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association*, **96** (2001), 1348-1360.

[7] W.J. Fu, Penalized regression: the bridge versus the LASSO, *Journal of Computational and Graphical Statistics*, **7** (1998), 397-416.

[8] G.J. McLachlan, D. Peel, *Finite Mixture Models*, John Wiley, New York (2000).

[9] G. Schwarz, Estimating the dimension of a model, *The Annals of Statistics*, **6** (1978), 461-464.

[10] R.J. Tibshirani, Regression shrinkage and selection via the LASSO, *Journal of the Royal Statistical Society, Series B*, **58** (1996), 267-288.

[11] D.M. Titterington, A.F.M. Smith, U.E. Makov, *Statistical Analysis of Finite Mixture Distributions*, John Wiley, New York (1985).

[12] H. Wang, C. Leng, Unified LASSO estimation via least squares approximation, *Journal of the American Statistical Association*, **101** (2007), 1418-1429.

[13] H. Wang, G. Li, C.L. Tsai, Regression coefficient and autoregressive order shrinkage and selection via LASSO, *Journal of Royal Statistical Society, Series B*, **69** (2007), 63-78.

[14] H. Wang, R. Li, C.L. Tsai, On the consistency of scad tuning parameter selector, *Biometrika*, **94** (2007), 553-558.

[15] M. Wedel, W. Kamakura, *Market Segmentation: Conceptual and Methodological Foundations*, Mass, Kluwer Academic (2000).

[16] C.S. Wong, W.K. Li, On a mixture autoregressive model, *Journal of Royal Statistical Society, Series B*, **62** (2000), 95-115.

[17] H.H. Zhang, W. Lu, Adaptive LASSO for Cox's proportional hazard model, *Biometrika* (2007), 691-703.

[18] H. Zou, The adaptive LASSO and its oracle properties, *Journal of the American Statistical Association*, **101** (2006), 1418-1429.