Extending the Akaike Information Criterion to Mixture Regression Models

Prasad A. NAIK, Peide SHI, and Chih-Ling TSAI

We examine the problem of jointly selecting the number of components and variables in finite mixture regression models. We find that the Akaike information criterion is unsatisfactory for this purpose because it overestimates the number of components, which in turn results in incorrect variables being retained in the model. Therefore, we derive a new information criterion, the mixture regression criterion (MRC), that yields marked improvement in model selection due to what we call the "clustering penalty function." Moreover, we prove the asymptotic efficiency of the MRC. We show that it performs well in Monte Carlo studies for the same or different covariates across components with equal or unequal sample sizes. We also present an empirical example on sales territory management to illustrate the application and efficacy of the MRC. Finally, we generalize the MRC to mixture quasi-likelihood and mixture autoregressive models, thus extending its applicability to non-Gaussian models, discrete responses, and dependent data.

KEY WORDS: Akaike information criterion; Cluster analysis; EM algorithm; Mixture models; Model selection; Variable selection.

1. INTRODUCTION

In the life sciences, engineering, medical, and business disciplines, researchers encounter the need to group similar objects and separate dissimilar ones to better understand the substantive phenomenon of interest. Cluster analysis provides one way to group objects into various clusters that are maximally distant from each other (e.g., Hartigan 1975). Once this classification is done, researchers seek to understand the differential impact of explanatory variables on some phenomenon of interest across various clusters. Toward this end, they may estimate a regression model in each cluster, but the resulting estimated coefficients are seriously biased even if clusters are well separated (Bryant and Williamson 1978). On the other hand, finite mixture regression models (see McLachlan and Peel 2000) provide an approach to classifying objects into various components (or clusters) and estimating regression models across components simultaneously (e.g., DeSarbo and Corn 1988; Wedel and Kamakura 2000, chap. 7).

Presently, users can apply extant approaches (e.g., Biernacki, Celeux, and Govaert 2000; Tibshirani, Walther, and Hastie 2001; Tadesse, Sha, and Vannucci 2005) to determine the number of components (but not variables) or information criteria, such as the Akaike information criterion (AIC) (Akaike 1973) and Bayes information criterion (BIC) (Schwarz 1978), to select the variables (but not components). Recently, Raftery and Dean (2006) investigated the simultaneous selection of the variables to use for clustering and the number of clusters retained in the model; however, their model does not specify a family of multiple regression models to predict the response variable across components. In summary, there is no method available to aid the joint selection of components and variables in mixture regression models. Moreover, adapting the penalty term in the AIC or BIC to select both components and variables may not provide satisfactory performance, especially when the sample size is small or the number of variables is large. Specifically, these criteria fit too many components (i.e., overcluster) and retain too many variables (i.e., overfit). A serious consequence of overclustering is that it results in fitting *spurious* regressions in nonexistent components, whereas overfitting reduces the accuracy of estimated effects and lowers the precision of forecasts (Altham 1984).

The objective of this article is to develop a method for the simultaneous determination of the number of components and variables in finite mixture regression models. We first derive the *mixture regression criterion* (MRC) using the complete-data log-likelihood discrepancy between the true and candidate models. The MRC consists of three terms: the first measures the lack of fit, the second imposes a penalty for regression parameters, and the third is what we call the *clustering penalty* function. The second term is the multicomponent generalization of the penalty function given by Hurvich and Tsai (1989), whereas the third term penalizes the number of components to be retained. We prove that the MRC is an efficient criterion. Using Monte Carlo studies, we show that MRC performs satisfactorily because the clustering penalty function mitigates the problem of overclustering. Specifically, the MRC performs well when different components contain the same or different covariates with either equal or unequal sample sizes. It outperforms both the AIC and BIC in small-sample data and high-dimensional models. Using empirical data, we illustrate that the MRC yields meaningful results, whereas the AIC tends to overcluster and overfit the data. Finally, we generalize the MRC's applicability to mixture quasi-likelihood models (McCullagh and Nelder 1989) and mixture autoregressive time series models (Le, Martin, and Raftery 1996; Wong and Li 2000). Thus the MRC can be applied not only to discrete response or non-Gaussian data such as those arising from logistic or Poisson mixture regression models, but also to dependent and nonlinear stochastic processes that exhibit various phenomena, such as flat stretches, cycles, outliers, and conditional heteroscedasticity.

The article is organized as follows. In Section 2 we derive the MRC criterion for mixture regression models and prove its asymptotic efficiency. In Section 3 we report Monte Carlo studies and present an empirical example. In the concluding Section 4 we extend the MRC to non-Gaussian and mixture time series models and also suggest three avenues for further research.

Prasad A. Naik is Professor, Graduate School of Management, University of California, Davis, CA 95616 (E-mail: *panaik@ucdavis.edu*). Peide Shi is Senior Analyst, Nuclear Safety Solutions Ltd., Toronto, Ontario, Canada, M5G 1X6 (E-mail: *pdshi@yahoo.ca*). Chih-Ling Tsai is Professor, Graduate School of Management, University of California, Davis, CA 95616, and Guanghua School of Management, Peking University, Beijing, 100871, People's Republic of China (E-mail: *cltsai@ucdavis.edu*). The authors thank the joint editor, associate editor, and referees for their helpful suggestions and comments.

Journal of the American Statistical Association
March 2007, Vol. 102, No. 477, Theory and Methods
DOI 10.1198/016214506000000861

2. INFORMATION CRITERION FOR MIXTURE REGRESSION MODELS

In this section we describe the structure and estimation of finite-mixture regression models, derive the model selection criterion, and prove its asymptotic efficiency.

2.1 Model Structure

Consider a candidate model with density function

$$f(y; \mathbf{x}, \phi) = \sum_{k=1}^{K} \alpha_k f_k(y; \mathbf{x}, \boldsymbol{\beta}_k, \sigma_k), \tag{1}$$

where $0 < \alpha_k \leq 1$ and $\sum_{k=1}^{K} \alpha_k = 1$; $f_k(y; \mathbf{x}, \boldsymbol{\beta}_k, \sigma_k)$ is the normal density with mean $\mathbf{x}'\boldsymbol{\beta}_k$ and variance σ_k^2; $\mathbf{x}$ is a $p \times 1$ vector of explanatory variables, which are given and considered fixed (i.e., nonrandom); $\boldsymbol{\beta}_k$ is a conformable parameter vector; and $\phi = \{(\alpha_k, \boldsymbol{\beta}_k, \sigma_k) : k = 1, \ldots, K\}$. Let $\mathbf{Z}$ be an $n \times K$ indicator matrix with jkth element, z_{jk}, equalling unity when y_j arises from the kth component of the mixture and 0 otherwise. Then, for the given data $\{(y_j, \mathbf{x}_j, z_j) : j = 1, \ldots, n\}$, the complete-data log-likelihood function (see Titterington, Smith, and Markov 1985, p. 84; McLachlan and Peel 2000, p. 48) of the candidate model is

$$L(\phi; \mathbf{Z}, \mathbf{Y}, \mathbf{X}) = \sum_{k=1}^{K} \sum_{j=1}^{n} z_{jk}\{\log \alpha_k + \log f_k(y_j; \mathbf{x}_j, \boldsymbol{\beta}_k, \sigma_k)\}, \tag{2}$$

where $\mathbf{Y} = (y_1, \ldots, y_n)'$, and the explanatory vectors $\mathbf{x}_j$ $(j = 1, \ldots, n)$ are stored in an $n \times p$ matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$. Similarly, the complete-data log-likelihood function of the true model is obtained by replacing L, f_k, ϕ, $\mathbf{X}$, $\mathbf{x}_j$, $\mathbf{Z}$, z_{jk}, α_k, $\boldsymbol{\beta}_k$, σ_k, p, and K in (2) with L^0, f_k^0, ϕ^0, $\mathbf{X}_0$, $\mathbf{x}_j^0$, $\mathbf{Z}^0$, z_{jk}^0, α_k^0, $\boldsymbol{\beta}_k^0$, σ_k^0, p^0, and K^0.

2.2 Model Estimation

To estimate mixture regression models, we apply the expectation–maximization (EM) algorithm (Dempster, Laird, and Rubin 1977). Let $\phi^{(m)} = \{(\alpha_k^{(m)}, \boldsymbol{\beta}_k^{(m)}, \sigma_k^{(m)}) : k = 1, \ldots, K\}$ denote the provisional estimates at the mth iteration and define $Q(\phi; \phi^{(m)}) = E[L|\mathbf{Y}, \mathbf{X}, \phi^{(m)}]$. In the E-step, we obtain $Q(\phi; \phi^{(m)})$ by replacing z_{jk} in (2) with the expected value $\tau_{jk} = E[z_{jk}|y_j]$, which is given by

$$\tau_{jk}^{(m)} = \frac{\alpha_k^{(m)} f_k(y_j; \mathbf{x}_j, \boldsymbol{\beta}_k^{(m)}, \sigma_k^{(m)})}{\sum_{k=1}^{K} \alpha_k^{(m)} f_k(y_j; \mathbf{x}_j, \boldsymbol{\beta}_k^{(m)}, \sigma_k^{(m)})}.$$

In the M-step, we maximize $Q(\phi; \phi^{(m)})$ with respect to $(\alpha_k, \boldsymbol{\beta}_k, \sigma_k)$. This maximization yields closed-form estimates for the $(m+1)$th iteration,

$$\alpha_k^{(m+1)} = \sum_{j=1}^{n} \frac{\tau_{jk}^{(m)}}{n},$$

$$\boldsymbol{\beta}_k^{(m+1)} = \left(\tilde{\mathbf{X}}_k^{(m)\prime} \tilde{\mathbf{X}}_k^{(m)}\right)^{-1} \tilde{\mathbf{X}}_k^{(m)\prime} \tilde{\mathbf{Y}}_k^{(m)},$$

and

$$\sigma_k^{2(m+1)} = \frac{\tilde{\mathbf{Y}}_k^{(m)\prime} (\mathbf{I} - \tilde{\mathbf{H}}_k^{(m)}) \tilde{\mathbf{Y}}_k^{(m)}}{\operatorname{tr}(\mathbf{W}_k^{(m)})},$$

for $k = 1, \ldots, K$, where $\mathbf{W}_k^{(m)} = \operatorname{diag}(\boldsymbol{\tau}_k^{(m)})$, $\boldsymbol{\tau}_k^{(m)} = (\tau_{1k}^{(m)}, \ldots, \tau_{nk}^{(m)})'$, $\tilde{\mathbf{X}}_k^{(m)} = \mathbf{W}_k^{(m)1/2}\mathbf{X}$, $\tilde{\mathbf{Y}}_k^{(m)} = \mathbf{W}_k^{(m)1/2}\mathbf{Y}$, and $\tilde{\mathbf{H}}_k^{(m)} = \tilde{\mathbf{X}}_k^{(m)}(\tilde{\mathbf{X}}_k^{(m)\prime}\tilde{\mathbf{X}}_k^{(m)})^{-1}\tilde{\mathbf{X}}_k^{(m)\prime}$.

The E- and M-steps are alternated until the value of $\log\{f(\mathbf{Y}; \mathbf{X}, \phi^{(m+1)})/f(\mathbf{Y}; \mathbf{X}, \phi^{(m)})\}$ decreases below a preset tolerance. We initialize the algorithm and obtain $\tau_{jk}^{(0)}$ by partitioning $\mathbf{X}$ into K clusters either randomly or rationally through the K-means clustering method (MacQueen 1967). The resulting estimates eventually converge to the maximum likelihood estimators, $\hat{\phi} = \{(\hat{\alpha}_k, \hat{\boldsymbol{\beta}}_k, \hat{\sigma}_k) : k = 1, \ldots, K\}$, which have the desirable asymptotic properties (Redner and Walker 1984; Wu 1983). In addition, we follow the suggestion of Hathaway (1985) to set the lower bound such that $\min_{k,l}(\hat{\sigma}_k/\hat{\sigma}_l) \geq c$, $c \in (0, 1]$, to eliminate singularities in the likelihood function and reduce the likelihood of spurious local maxima. Hathaway (1985, p. 799) suggested that a good value of c can be determined by varying it dynamically over the unit interval. To choose K components and p variables for retention in a mixture regression model, we next propose the MRC model selection criterion.

2.3 Derivation of the Mixture Regression Criterion

The main goal of model selection is to approximate the true model using candidate models with different combinations of (K, p) and then retain the model that entails a minimum loss of information. The true model is a mixture model of the form (1) for some integer K^0 of component models $(1 < K^0 < K)$ and some subvector of dimension p^0 of the vector of explanatory variables $(1 < p^0 < p)$. Consequently, the columns of $\mathbf{X}$ can be rearranged so that $\mathbf{X}^0\boldsymbol{\beta}_k^0 = \mathbf{X}\boldsymbol{\beta}_k^*$, where $\boldsymbol{\beta}_k^* = (\boldsymbol{\beta}_k^{0\prime}, \boldsymbol{\beta}_k^{1\prime})'$ and $\boldsymbol{\beta}_k^1$ is a $(p - p^0) \times 1$ vector of 0's for $k = 1, \ldots, K$ (see Hurvich and Tsai 1989).

Next, consider fitting a candidate model using the observed sample $\mathbf{Y}$. We apply the EM algorithm in Section 2.2 to obtain $\hat{\phi}$. More specifically, the resulting estimate $\hat{\alpha}_k$ is an average of $\hat{\tau}_{jk}$ over $j = 1, \ldots, n$; $\hat{\boldsymbol{\beta}}_k$ depends on $\hat{\tau}_{jk}$ through $\hat{\mathbf{W}}_k = \operatorname{diag}(\hat{\boldsymbol{\tau}}_k)$, and $\hat{\sigma}_k^2$ depends on $\hat{\tau}_{jk}$ through both $\hat{\mathbf{H}}_k = \hat{\mathbf{X}}_k(\hat{\mathbf{X}}_k'\hat{\mathbf{X}}_k)^{-1}\hat{\mathbf{X}}_k'$ and $\operatorname{tr}(\hat{\mathbf{W}}_k)$, where $\hat{\mathbf{X}}_k = \hat{\mathbf{W}}_k^{1/2}\mathbf{X}$. Thus, for each $k = 1, \ldots, K$, $(\hat{\alpha}_k, \hat{\boldsymbol{\beta}}_k, \hat{\sigma}_k^2)$ depends *implicitly* on $\hat{\boldsymbol{\tau}}_k = (\hat{\tau}_{1k}, \ldots, \hat{\tau}_{nk})'$. In other words, the estimated parameters $\hat{\boldsymbol{\theta}} = (\hat{\phi}, \hat{\boldsymbol{\tau}})$ are functions of the observed sample $\mathbf{Y}$, where $\hat{\boldsymbol{\tau}} = (\hat{\tau}_1, \ldots, \hat{\tau}_K)$. How well does this fitted model predict future samples $\mathbf{Y}^* = (y_1^*, \ldots, y_n^*)'$ that are generated from the true model and independent of $\mathbf{Y}$? Conceptually, $\mathbf{Y}^*$ serves as the holdout sample for assessing the quality of fitted models. We can assess the prediction quality using the metric $\Delta(\hat{\boldsymbol{\theta}}) = E_{\mathbf{Y}^*}\{L^0(\phi^0; \mathbf{Z}^*, \mathbf{Y}^*, \mathbf{X}) - L(\hat{\phi}; \hat{\boldsymbol{\tau}}, \mathbf{Y}^*, \mathbf{X})\}$, where $\mathbf{Z}^*$ is an $n \times K^0$ matrix whose jkth element, z_{jk}^*, equals unity when y_j^* arises from the kth component of the mixture and is 0 otherwise. Because $\hat{\boldsymbol{\theta}}$ depends on $\mathbf{Y}$, we adopt Akaike's (1985) predictive approach to eliminating this dependence on the particular sample by averaging $\Delta(\hat{\boldsymbol{\theta}}(\mathbf{Y}))$ across different independent samples $\mathbf{Y}$ drawn from the true model. The resulting complete-data log-likelihood (CL) discrepancy is

$$\begin{aligned} d_{\mathrm{CL}} &= E_{\mathbf{Y}}\left\{\Delta(\hat{\boldsymbol{\theta}}(\mathbf{Y}))\right\} \\ &= 2E_{\mathbf{Y}}\left[E_{\mathbf{Y}^*}\{L^0(\phi^0; \mathbf{Z}^*, \mathbf{Y}^*, \mathbf{X}) - L(\hat{\phi}; \hat{\boldsymbol{\tau}}, \mathbf{Y}^*, \mathbf{X})\}\right], \end{aligned} \tag{3}$$

where both expectations are evaluated with respect to the true model. The double expectation in (3) has been commonly used to obtain model selection criteria (see Burnham and Anderson 2002, pp. 60, 443). For example, in single-component regression, Burnham and Anderson (2002, chap. 7.4) applied this approach to obtain a second-order improvement to the AIC and recommended its extension to finite-mixture models (p. 344). It is noteworthy that d_{CL} is based on Akaike's (1985) concept of predictive likelihood, which can be interpreted as cross-validation of the future samples $\mathbf{Y}^*$ using original $\mathbf{X}$ and estimated parameters (see Burnham and Anderson 2002, p. 365).

Given a collection of competing fitted candidate models, the one that minimizes d_{CL} is preferred. Appendix A shows that an estimator of d_{CL} is

$$\begin{aligned}\mathrm{MRC}^* = &\sum_{k=1}^{K} \operatorname{tr}(\hat{\mathbf{W}}_k)\log(\hat{\sigma}_k^2) \\ &+ \sum_{k=1}^{K} \frac{\operatorname{tr}(\hat{\mathbf{W}}_k)(\operatorname{tr}(\hat{\mathbf{W}}_k)+\hat{\nu}_{k1})(\hat{\delta}_{k1}/\hat{\delta}_{k2})}{\hat{\delta}_{k1}^2/\hat{\delta}_{k2}-2} \\ &- 2\sum_{k=1}^{K} \operatorname{tr}(\hat{\mathbf{W}}_k)\log(\hat{\alpha}_k). \qquad (4)\end{aligned}$$

In (4), $\hat{\delta}_{k1} = \operatorname{tr}\{(\mathbf{I}-\hat{\mathbf{H}}_k)\hat{\mathbf{W}}_k\}$, $\hat{\delta}_{k2} = \operatorname{tr}[\{(\mathbf{I}-\hat{\mathbf{H}}_k)\hat{\mathbf{W}}_k\}^2]$, and $\hat{\nu}_{k1} = \operatorname{tr}(\hat{\mathbf{H}}_k\hat{\mathbf{W}}_k)$, where $\hat{\mathbf{H}}_k = \hat{\mathbf{X}}_k(\hat{\mathbf{X}}_k'\hat{\mathbf{X}}_k)^{-1}\hat{\mathbf{X}}_k'$.

To further simplify the MRC*, we consider the true model in which clusters are well separated and observations are perfectly classified into only one cluster (say, the kth) so that the resulting diagonal elements of $\mathbf{W}_k^0$ are either 1 or 0. Using $\mathbf{W}_k^0$, we note that $\mathbf{H}_k^0 = \mathbf{X}_k^0(\mathbf{X}_k^{0\prime}\mathbf{X}_k^0)^{-1}\mathbf{X}_k^{0\prime}$ is the projection matrix, where $\mathbf{X}_k^0 = (\mathbf{W}_k^0)^{1/2}\mathbf{X}$. Then we observe that $(\mathbf{W}_k^0)^{1/2}\mathbf{H}_k^0(\mathbf{W}_k^0)^{1/2}$ and $(\mathbf{W}_k^0)^{1/2}(\mathbf{I}-\mathbf{H}_k^0)(\mathbf{W}_k^0)^{1/2}$ are idempotent matrices and that $\{(\mathbf{W}_k^0)^{1/2}\mathbf{H}_k^0(\mathbf{W}_k^0)^{1/2}\}\{(\mathbf{W}_k^0)^{1/2}(\mathbf{I}-\mathbf{H}_k^0)(\mathbf{W}_k^0)^{1/2}\} = 0$. Next, applying results of Cochran (1934), we find that $\boldsymbol{\varepsilon}_k'(\mathbf{W}_k^0)^{1/2}\mathbf{H}_k^0(\mathbf{W}_k^0)^{1/2}\boldsymbol{\varepsilon}_k/(\sigma_k^0)^2$ and $\boldsymbol{\varepsilon}_k'(\mathbf{W}_k^0)^{1/2}(\mathbf{I}-\mathbf{H}_k^0)(\mathbf{W}_k^0)^{1/2}\boldsymbol{\varepsilon}_k/(\sigma_k^0)^2$ are independent chi-squared distributions with the degrees of freedom $\operatorname{tr}(\mathbf{H}_k^0\mathbf{W}_k^0) = \nu_{k1} = p^0$ and $\operatorname{tr}\{(\mathbf{I}-\mathbf{H}_k^0)\mathbf{W}_k^0\} = \delta_{k1} = \delta_{k2} = n_k^0 - p^0$, where $\boldsymbol{\varepsilon}_k$ is as defined in (A.2) in Appendix A. Consequently, $\nu_{k1} = p^0$, $\delta_{k1}/\delta_{k2} = 1$, and $\delta_{k1}^2/\delta_{k2} = n_k^0 - p^0$. When clusters are not separated, we approximate ν_{k1}, δ_{k1}/δ_{k2}, and $\delta_{k1}^2/\delta_{k2}$ by p^0, 1, and $n_k^0 - p^0$. Based on extensive simulation studies, we found these approximations to be reasonably accurate for the purpose of model selection. Using these approximations to simplify (4), we thus obtain the mixture regression criterion

$$\mathrm{MRC} = \sum_{k=1}^{K}\hat{n}_k\log(\hat{\sigma}_k^2) + \sum_{k=1}^{K}\frac{\hat{n}_k(\hat{n}_k+p_k)}{\hat{n}_k-p_k-2} - 2\sum_{k=1}^{K}\hat{n}_k\log(\hat{\alpha}_k), \qquad (5)$$

where $p_k = \operatorname{tr}(\hat{\mathbf{H}}_k)$ and $\hat{n}_k = \operatorname{tr}(\hat{\mathbf{W}}_k)$.

The first term of MRC measures the lack of fit, which can be reduced by including more variables in the candidate model so that $\hat{\sigma}_k^2$ becomes small. The second term balances this temptation to add variables by imposing a penalty for overfitting. The third term, which we call the clustering penalty function, provides a countervailing force to mitigate overclustering. To illustrate this point, we present two examples. First, let the mixing proportions be equal, that is, $\hat{\alpha}_k = 1/K$ for $k = 1, \ldots, K$. The resulting third term simplifies to $2n\log(K)$, which indicates that the clustering penalty increases with K. Second, let the mixing proportions of the first $K-1$ components be fixed. We investigate the effect of including one incremental component beyond the Kth one, ceteris paribus. As a result, the third term equals $-2\sum_{k=1}^{K}\hat{n}_k\log(\hat{\alpha}_k) = C_1 - 2\hat{n}_K\log(\hat{\alpha}_K) = C_1 - 2\{\rho\hat{n}_K\log(\hat{\alpha}_K)\} - 2\{(1-\rho)\hat{n}_K\log(\hat{\alpha}_K)\} \le C_1 - 2\{\rho\hat{n}_K\log(\rho\hat{\alpha}_K)\} - 2\{(1-\rho)\hat{n}_K\log((1-\rho)\hat{\alpha}_K)\}$, where $C_1 = -2\sum_{k=1}^{K-1}\hat{n}_k\log(\hat{\alpha}_k)$ and $\rho \in [0,1]$; that is, the third term tends to impose a larger penalty as an additional component is included incrementally. Next, we present the large-sample property of the MRC.

2.4 Asymptotic Efficiency of the Mixture Regression Criterion

Let $\mathcal{A}_n$ be the set of candidate models consisting of various combinations of components and variables. In other words, $\mathcal{A}_n = \{\xi : \xi = k\times\lambda, \lambda = \lambda_1\times\cdots\times\lambda_k, k\in\{1,2,\ldots,K\}, \lambda_k$ is a nonempty subset of $\{1,\ldots,q_n\}$, and $\mathbf{Y}_k = \mathbf{X}_k(\xi)\boldsymbol{\beta}_k(\xi) + \mathbf{e}_k\}$, where $\mathbf{Y}_k = \mathbf{Z}_k\mathbf{Y}$, $\mathbf{X}_k(\xi) = \mathbf{Z}_k\mathbf{X}(\xi)$, $\mathbf{Z}_k = \operatorname{diag}(z_{1k},\ldots,z_{nk})$, q_n is a positive integer that may depend on n, and $\mathbf{e}_k$ are random errors. In addition, let $\mathcal{A}_n^0 = \{\xi\in\mathcal{A}_n : E(\mathbf{Y}_k) = \mathbf{X}_k(\xi)\boldsymbol{\beta}_k(\xi) + o(1)$, and $k \ge K^0\}$ be a subset of $\mathcal{A}_n$ so that each ξ in $\mathcal{A}_n^0$ is associated with a model that is closest to the true model. Next, let $\hat{\xi}$ denote the model selected by the MRC based on its smallest value across all possible candidate models, and let $\xi^0 = K^0\times\lambda^0$ be the model in $\mathcal{A}_n^0$ with the smallest dimension. Hence the true model in the kth component, $\mathbf{Z}_k^0\mathbf{X}\boldsymbol{\beta}_k^*$, can be represented as $\mathbf{X}_k(\xi^0)\boldsymbol{\beta}_k(\xi^0) + o(1)$, where $\mathbf{Z}_k^0 = \operatorname{diag}(z_{1k}^0,\ldots,z_{nk}^0)$. To prove an asymptotic result, we make the following assumptions.

Assumption 1. When n is sufficiently large, $\hat{\xi}\in\mathcal{A}_n^0$.

Assumption 2. Let $L_{n,k}(\xi) = \|\hat{\mathbf{W}}_k^{1/2}\mathbf{X}(\xi)\hat{\boldsymbol{\beta}}_k(\xi) - (\mathbf{W}_k^0)^{1/2}\times\mathbf{X}_0\boldsymbol{\beta}_k^0\|^2/\hat{n}_k$. For all $\xi\in\mathcal{A}_n^0$ and $k = 1,\ldots,K^0$, $L_{n,k}(\xi) = \frac{(\sigma_k^0)^2p_k}{\hat{n}_k} + o_p(\frac{1}{\hat{n}_k})$, and there is a random variable ω_k independent of the current model such that

$$\hat{\sigma}_k^2 = \omega_k + L_{n,k}(\xi) - \frac{2p_k(\sigma_k^0)^2}{\hat{n}_k} + o_p(L_{n,k}(\xi)),$$

where $o_p(\cdot)$ denotes the convergence in probability for all $\xi\in\mathcal{A}_n$, the mean and variance of ω_k are finite, and $\omega_k > (\sigma_k^0)^2/2$ except for an event with probability tending to 0 with n.

Assumption 3. For all $\xi\in\mathcal{A}_n^0$, $\frac{q_n^2}{\min_{1\le k\le K}\hat{n}_k} = o_p(1)$, $\hat{n}_k/n = \alpha_k^0 + o_p(L_{n,k}(\xi))$, and $\hat{\alpha}_k = \alpha_k^0 + o_p(L_{n,k}(\xi))$, where $\alpha_k^0 > 0$ for $k = 1,\ldots,K^0$, $\alpha_k^0 = 0$ for $k = K^0+1,\ldots,K$, and $\hat{\sigma}_k/\hat{\alpha}_k$ is uniformly bounded away from 0 and ∞ for all $\xi\in\mathcal{A}_n$, except for an event with probability tending to 0 with n. To present the theorem, we introduce the notation $L_n(\xi) = W_n\sum_{k=1}^{K^0}B_kL_{n,k}(\xi)\prod_{j=1,j\ne k}^{K^0}A_j$, where

$$A_k = \left(\frac{\omega_k}{(\alpha_k^0)^2}\right)^{\alpha_k^0}\left[1+\frac{2}{n}\right],$$

$$B_k = \frac{2}{(\sigma_k^0)^2}\left(\frac{\omega_k}{(\alpha_k^0)^2}\right)^{\alpha_k^0} - \frac{1}{\alpha_k^0}\left(\frac{\omega_k}{(\alpha_k^0)^2}\right)^{\alpha_k^0-1}\left[1+\frac{2}{n}\right],$$

and

$$W_n = \left\{1 + \frac{2(K-K^0)}{n}\right\} \exp\left(\sum_{k=1}^{K^0} \alpha_k^0\right).$$

Theorem. Let $\bar{\xi}$ be the model such that $L_n(\bar{\xi}) = \min_{\xi \in \mathcal{A}_n} L_n(\xi)$. If $\bar{\xi} \in \mathcal{A}_n^0$ and Assumptions 1–3 are satisfied, then

$$\frac{L_n(\hat{\xi})}{L_n(\bar{\xi})} - 1 = o_p(1).$$

Appendix B provides the proof. Thus the model selected by the MRC is asymptotically efficient (see Remark 2 for further discussions). We close this section with six remarks that discuss the assumptions and relate the MRC to other criteria.

Remark 1. We note that Assumptions 2 and 3 hold for single component normal linear regression models ($K = 1$). Specifically, when $K = 1$, $\mathbf{Y}_1 = \mathbf{Y}$, $\mathbf{X}_1(\xi) = \mathbf{X}(\xi)$, and $\hat{n}_k = n$, $\hat{\beta}_1(\xi)$ is the least squares estimate of $\beta_1(\xi)$. Consequently, the following results hold:

- $L_{n,1}(\xi) = \boldsymbol{\varepsilon}'\mathbf{H}\boldsymbol{\varepsilon}/n$ with $\mathbf{H} = \mathbf{X}(\xi)(\mathbf{X}(\xi)'\mathbf{X}(\xi))^{-1}\mathbf{X}(\xi)'$.
- $\hat{\sigma}_1^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}/n + L_{n,1}(\xi) - \frac{2}{n}\boldsymbol{\varepsilon}'\mathbf{H}\boldsymbol{\varepsilon} = \omega_1 + L_{n,1}(\xi) - \frac{2}{n}\boldsymbol{\varepsilon}'\mathbf{H}\boldsymbol{\varepsilon}$ with $\omega_1 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}/n$.

Under some regularity conditions, using an argument similar to that used by Shi and Tsai (1999, pp. 134–135), $\max_{\xi \in \mathcal{A}_n} |\frac{1}{n}\boldsymbol{\varepsilon}' \times \mathbf{H}\boldsymbol{\varepsilon} - \frac{p_1(\sigma_1^0)^2}{n}|/L_{n,1}(\xi) = o_p(1)$. If random errors are distributed normally, then $n\omega_1/(\sigma_1^0)^2$ is the chi-squared distribution with degrees of freedom n. Therefore, Assumption 2 is satisfied. Furthermore, when $K = 1$, Assumption 3 holds because $\alpha_1^0 = 1$, $\hat{n}_1/n = 1$, and $\hat{\alpha}_1 = 1$.

Remark 2. To prove the theorem, we adopted the approach of Shibata (1981, 1984) to show that the MRC is an efficient criterion, which means that an information criterion selects the candidate model such that the resulting average prediction error converges to the minimum of the average prediction error as the sample size increases. It is important to recognize that accurate predictive results can be obtained even when parameters are not estimated consistently or when the true model is not included in the list of candidate models. In this sense, the goals of prediction, as advocated by Akaike (1985), differ from the classical goals of consistent parameter estimation. Furthermore, consistency in the model selection context means that an information criterion (such as the BIC) selects the correct model with probability approaching 1 in large samples when the true model is in the family of candidate models. In data mining contexts, however, the class of models used need not necessarily contain the true model, and hence consistency becomes an irrelevant goal. Moreover, the conditions under which criteria such as the BIC or AIC work well depend on other factors, for example, the assumptions that researchers make about reality and their intentions for model-based inference in a given application (see Burnham and Anderson 2004). Further discussions on efficiency and consistency in model selection have been given by Burnham and Anderson (2002, sec. 6.4), Shao (1997), and McQuarrie and Tsai (1998).

Remark 3. The MRC extends the applicability of single-component information criteria to multiple-component regression models. Specifically, in single-component normal regression models ($K = 1$), we have that $\hat{n}_k = n$, $p_k = p$, and $\hat{\alpha}_k = 1$. Consequently, (5) can be expressed as $\mathrm{MRC}_{K=1} = n\log(\hat{\sigma}^2) + n(n+p)/(n-p-2)$, which equals the $\mathrm{AIC_c}$ of Hurvich and Tsai (1989, p. 300). Furthermore, on subtracting the constant n in $\mathrm{MRC}_{K=1}$, the penalty term can be written as $2\{n/(n-p-2)\}(p+1)$, which approaches $2(p+1)$ as $n \to \infty$. Hence MRC $\to$ AIC when $K = 1$ and n is large (except for the constant 2, which does not alter model selection).

Remark 4. To measure model complexity, Spiegelhalter, Best, Carlin, and van der Linde (2002) proposed the deviance information criterion (DIC), which approximates the expected predictive discrepancy so that users can select models with the best out-of-sample predictive power (Gelman, Carlin, Stern, and Rubin 2004, p. 183). When prior information is available, the DIC is a more appropriate criterion than the Bayes factor, because users may not know ex ante all possible candidate models. Spiegelhalter et al. (2002, p. 604) noted that when prior information is negligible, the DIC is equivalent to the AIC. In the discussion of their work, Burnham (2002, p. 629) stated that the DIC might need small-sample correction, and Richardson (2002, p. 627) showed that her computation of the DIC imposed an insufficient penalty for K-component mixture model selection. Note that the proposed MRC, also based on predictive likelihood (Akaike 1985), incorporates both small-sample correction (through its second term) and the clustering penalty (through the third term).

Remark 5. When explanatory variables contain noise, users may apply the denoised least squares (DLS) estimator to extract signals from noisy variables. Specifically, using equation (2.3) of Cai, Naik, and Tsai (2000, p. 1234), the observed $(\mathbf{X}, \mathbf{Y})$ can be denoised through wavelet transform to obtain $\mathbf{X}_{\mathrm{DLS}} = \mathbf{H}_{\mathbf{x}} \circ \mathbf{X}$ and $\mathbf{Y}_{\mathrm{DLS}} = \mathbf{H}_{\mathbf{y}} \circ \mathbf{Y}$, where $\mathbf{H}_{\mathbf{x}}$ and $\mathbf{H}_{\mathbf{y}}$ are smoothing matrices. Using the denoised $(\mathbf{X}_{\mathrm{DLS}}, \mathbf{Y}_{\mathrm{DLS}})$ as inputs, users can follow Section 2.2 for parameter estimation and then apply the MRC for model selection. When the sample size is smaller than the number of variables (e.g., in chemometrics or bioinformatics), users can apply the partial least squares (PLS) estimator (e.g., Helland and Almøy 1994; Naik and Tsai 2000; Hastie, Tibshirani, and Friedman 2001) within each component of the mixture regression model. Specifically, construct the $p_k \times q_k$ matrix $\mathbf{R}_k = (\mathbf{S}_{k1}, \mathbf{S}_{k2}\mathbf{S}_{k1}, (\mathbf{S}_{k2})^2\mathbf{S}_{k1}, \ldots, (\mathbf{S}_{k2})^{q_k-1}\mathbf{S}_{k1})$, where $p_k \times 1$ vector $\mathbf{S}_{k1}$ is the sample covariance of $(\tilde{\mathbf{X}}_k, \tilde{\mathbf{Y}}_k)$, the $p_k \times p_k$ matrix $\mathbf{S}_{k2}$ is the sample covariance of $\tilde{\mathbf{X}}_k$, and q_k is the dimension of subspace of $\tilde{\mathbf{X}}_k$. Then replace $\boldsymbol{\beta}_k$ with $\mathbf{b}_k = \mathbf{R}_k(\mathbf{R}_k'\mathbf{S}_{k2}\mathbf{R}_k)^{-1}\mathbf{R}_k'\mathbf{S}_{k1}$ in the $(m+1)$th iteration of Section 2.2 to obtain parameter estimators. Next, substituting p_k by q_k in the MRC, users can jointly determine the number of components and dimensions of subspace to retain.

Remark 6. We observe that, as in standard regression theory, there is no constraint on the maximum number of variables in the mixture model selection theory. That is, users in various fields can consider as many variables as they can estimate. The resulting high dimensionality may increase the ratio of the number of variables to the sample size. As we illustrate in the simulation studies, the MRC's performance improves as

this ratio increases. Hence, the MRC is well suited for high-dimensional models in which explanatory variables represent a moderate to large fraction of the sample size. Next, we study the finite-sample performance of the MRC.

3. SIMULATIONS AND APPLICATION

We begin this section by describing the simulation settings, model estimation, and selection procedure. We then illustrate the properties and performance of the MRC through simulation results. Finally, we present an empirical example to elucidate the MRC's applications and efficacy.

3.1 Simulation Settings

We consider the following scenarios: same covariates across components, different covariates across components, and a single-component regression.

Same Covariates. In this scenario we consider equal and unequal sample sizes across various components of the mixture distribution. Specifically, in the equal sample case, the true model consists of three components ($K^0 = 3$) with 100 observations per component ($n_k^0 = 100$). Each component has a regression model with four explanatory variables ($p^0 = 4$). The true regression parameters are $\boldsymbol{\beta}_1^0 = (1, 1, 1, 1)'$, $\boldsymbol{\beta}_2^0 = (1, 2, 3, 4)'$, and $\boldsymbol{\beta}_3^0 = (5, 6, 7, 8)'$. The true explanatory variables are stored in $n_k \times 4$ matrices, $\mathbf{X}_1^0$, $\mathbf{X}_2^0$, and $\mathbf{X}_3^0$, with elements generated from U(0, 5), U(5, 10), and U(10, 15), where U(a, b) denotes the uniform distribution on interval $[a, b]$. The dependent variable for each component is generated from $\mathbf{Y}_k = \mathbf{X}_k^0\boldsymbol{\beta}_k^0 + \boldsymbol{\varepsilon}_k^0$, for $k = 1$, 2, and 3, where $\boldsymbol{\varepsilon}_k^0 \sim \mathrm{N}(\mathbf{0}, (\sigma_k^0)^2\mathbf{I}_{n_k})$, $\sigma_k^0 = 1$, and $\mathbf{I}_{n_k}$ is an identity matrix of dimension $n_k^0 \times n_k^0$. The errors, $\boldsymbol{\varepsilon}_k^0$, are independent of the explanatory variables, $\mathbf{X}_k^0$, and the total number of observations is $n = \sum_{k=1}^3 n_k^0 = 300$. In the unequal case, we change the foregoing sample sizes to $n_1^0 = 50$, $n_2^0 = 75$, and $n_3^0 = 100$.

In each of the three components, seven variables are stored in an $n_k^0 \times 7$ matrix, $\mathbf{X}_k$. The first four columns of $\mathbf{X}_k$ are the same as $\mathbf{X}_k^0$, whereas the last three columns are generated from the same distributions, namely U(0, 5) for component 1, U(5, 10) for component 2, and U(10, 15) for component 3. The observed matrix, $\mathbf{X}$, is constructed by stacking $\mathbf{X}_1$, $\mathbf{X}_2$, and $\mathbf{X}_3$ on top of each other. Analogously, the observed dependent variable, $\mathbf{Y}$, is obtained by stacking $\mathbf{Y}_1$, $\mathbf{Y}_2$, and $\mathbf{Y}_3$ on top of each other. We consider five sets of candidate components, $K = 1, \ldots, 5$. A family of candidate regression models includes up to seven variables from $\mathbf{X}$ in a sequentially nested fashion. Hence we have 35 possibilities—7 nested regression models in each of the 5 sets of candidate components—from which to select the true model. (Note that we slightly abuse notation $\mathbf{Y}_k$, $\mathbf{X}_k^0$, and $\mathbf{X}_k$ with respect to their dimensions for the sake of simplicity.)

Different Covariates. Although we derived the MRC by assuming that $p_k = p$, it naturally generalizes to permit selection of different variables across different components (i.e., $p_k \neq p$). To study this more general case, the true model consists of two components with two variables in the first one and four variables in the second component. The simulation setting from the same covariates case is modified by retaining the last two components (i.e., $k = 2$ and 3) and deleting the last two variables in the component $k = 2$. Thus the resulting $\boldsymbol{\beta}_1^0 = (1, 2)'$, $\boldsymbol{\beta}_2^0 = (5, 6, 7, 8)'$, $K^0 = 2$, and $n = n_1 + n_2 = 200$. The set of candidate models is similar to that in the same covariates case except that $\mathbf{X}_k$ now includes the first five columns of $\mathbf{X}_k$ given in that case, and only three sets of candidate components ($K = 1, 2, 3$) are considered. For each K, the five candidate regression models include variables of $\mathbf{X}$ in a sequentially nested manner.

Single-Component Regression. As Zhu and Zhang (2004, p. 5) noted, one key issue is to determine whether two or more mixture regressions are warranted. Toward this end, we generate data from a standard regression model using the same covariates setting with $K^0 = 1$ and fit candidate models with either one or two components. We next describe the estimation and selection procedure for calibrating these mixture regression models.

3.2 Estimation and Selection Procedure

We use the following procedure to simultaneously determine the number of components and variables in mixture regression models. First, for the given $\{(K, p_k) : K = 1, \ldots, 5, p_k = 1, \ldots, 7\}$, we use the K-means algorithm to classify observations from a candidate matrix $\mathbf{X}$ into K groups so that the initial probabilities can be estimated to start the EM algorithm. Then we apply the EM algorithm (with $c = .1$ as in Hathaway 1985) to estimate the mixture regression model (see Sec. 2.2). Next, we compute the MRC by substituting the parameter estimates into (5). Finally, we generate 1,000 realizations from the true models described in Section 3.1 and compute the selection criteria for each realization. For some random realizations, $(\hat{n}_k - p_k - 2)$ can become negative, in which case we replace it with 10^{-2} to ensure a positive penalty in (5).

3.3 Simulation Results

Table 1 presents the joint frequency of component and variable selection for the same covariates across components. In the top part of the table, we see that MRC correctly selects the true components and variables on 924 occasions out of 1,000 realizations. To provide insight into the source of this improvement, the bottom part of the table displays the MRC's selection

Table 1. Frequency of Components and Variables Jointly Selected by the MRC in 1,000 Realizations for the Same Covariates Setting With Equal Sample Sizes

	p							Column
K	1	2	3	$p^0 = 4$	5	6	7	sum
MRC								
1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0
$K_0 = 3$	0	0	0	924	52	16	8	1,000
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0
MRC without the clustering penalty								
1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0
$K_0 = 3$	0	0	0	0	0	0	0	0
4	0	0	0	1	0	1	1	3
5	0	0	0	167	258	336	236	997

frequency *without* the clustering penalty. When it is ignored, the MRC tends to not only retain the largest number of components, but also select irrelevant variables (833 times out of 1,000 realizations). Thus the clustering penalty function plays an important role in the proper selection of mixture regression models, yielding a marked improvement. A qualitatively similar improvement due to the clustering penalty is observed for the unequal sample size, and hence we do not elaborate on the results here.

We next discuss the results for the case with different covariates in the two components ($K^0 = 2, p_1^0 = 2$, and $p_2^0 = 4$). This scenario allows $p_k \neq p$ for all k and considers 155 possibilities (5 nested variable selections in $K = 1$, 5×5 variable selections when $K = 2$, and $5 \times 5 \times 5$ variable selections for $K = 3$) from which to identify the true model. Consequently, determining components and variables becomes complicated. To reduce this complexity, we implement a two-stage procedure: first, determine the best number of components by retaining all variables in the regression models, and then select the best set of covariates for the chosen components. Simulation results show that this two-stage procedure compares well with the exhaustive search of 155 cluster–variable combinations. Furthermore, the MRC selects the number of components correctly 999 times out of 1,000 realizations, whereas it chooses the covariates correctly in the first and second components on 783 and 844 realizations.

As for the results of the single-component regression, we find an average MRC of 105.10 for the single-component model and 207.57 for the two-component regression model, implying that it correctly selects the true model on average. More specifically, the MRC identifies the single-component model almost always and jointly selects the correct number of components and variables 797 times, compared with 65 times for the AIC (out of 1,000 realizations). This result supports the findings of the empirical example given in Section 3.4.

Finally, we compare the MRC with other prominent information criteria. Specifically, in single-component regression models, the AIC and BIC are $\text{AIC} = -2\{\log f(\mathbf{Y}; \mathbf{X}, \hat{\phi})\} + 2p$ and $\text{BIC} = -2\{\log f(\mathbf{Y}; \mathbf{X}, \hat{\phi})\} + p\log(n)$, where $\log f(\mathbf{Y}; \mathbf{X}, \hat{\phi})$ is the maximized log-likelihood. To account for the $(K - 1)$ estimated mixing proportions as well as the estimated regression coefficients and error variances in each of the K components, applied users replace p with the penalty $d = (K-1) + K(p+1)$ and consider the resulting criterion a "heuristic figure of merit" (DeSarbo and Cron 1988, p. 259). We illustrate the relative performance of these criteria by considering three components with small sample size ($n = 30$), high-dimensional models (large $p = 10$ and moderate $n = 75$), and large sample size ($n = 300$). In both the small- and large-sample cases, the true and candidate models are identical to those in the same-covariates setting detailed in Section 3.1. In the high-dimensional case, we let the true $\boldsymbol{\beta}_1^0 = (1, \ldots, 1)'$, $\boldsymbol{\beta}_2^0 = (1, 2, \ldots, 10)'$, and $\boldsymbol{\beta}_3^0 = \boldsymbol{\beta}_1^0 + \text{U}(0, 1)$ and consider up to 15 variables in candidate models.

Table 2 indicates that the AIC performs poorly in all three cases because it lacks the appropriate penalty to prevent overclustering. Further examining the joint frequency of variable and component selections (not presented here), shows that the AIC tends to choose incorrectly the maximum candidate components ($K = 5$) as much as 75% of the time, thereby segmenting the data too finely. Similarly, it selects more variables than necessary (instead of too few). Thus an important insight here is not just that the AIC overclusters, but that the AIC chooses the variables incorrectly because of spurious clustering. Analogously, the BIC also performs poorly when the sample size is small or the model dimensionality is high (see Table 2). However, it performs better than the AIC in large samples, as it should because of its consistency property. This finding extends the single-component simulation results given by McQuarrie and Tsai (1998, sec. 9.2). Moreover, Table 2 reveals that the MRC outperforms both the AIC and BIC for small samples and high-dimensional models and is competitive with the BIC in large samples even though it is not a consistent criterion. The MRC's performance improves as the sample size decreases from $n = 300$ to $n = 30$. This "small n bias correction" is due to the parameter penalty (i.e., the second term of the MRC), which increases as the $p_k/\hat{n}_k$ ratio increases. This finding corroborates the results of Hurvich and Tsai (1989) for the single-component case. Hence we recommend using the MRC when datasets are small or explanatory variables represent a moderate to large fraction of the sample size.

Table 2. Correct Selections in 1,000 Realizations by the AIC, BIC, and MRC

	Small-sample	*High-dimensional models*	*Large-sample*
AIC	1	0	20
BIC	70	114	995
MRC	990	999	924

3.4 Empirical Application

Here we illustrate that the AIC tends to overcluster and overfit the real data. We investigate the problem of evaluating sales territory performance (see, e.g., Cravens, Woodruff, and Stamper 1972) and analyze the dataset given by Dielman (2001, p. 493), which contains information on 25 sales territories. In each territory, we observe the unit sales ($\mathbf{Y}$), the company's advertising effort ($\mathbf{X}_1$), the salesperson's effort ($\mathbf{X}_2$) (which equals a salesperson's workload per account multiplied by the number of accounts in that territory), the salesperson's experience ($\mathbf{X}_3$), and the salesperson's ability ($\mathbf{X}_4$) as rated by his or her supervisor. First, we center the data by subtracting from each variable its corresponding mean, to eliminate the need for estimating the intercept coefficients. Given the small sample of 25 observations, we estimate mixture regression models with two components and consider the retention of four variables. We apply the MRC, AIC, and BIC to jointly determine the components and variables to be retained.

Based on the results in Table 3, the MRC retains the single-component model with two variables. In contrast, the AIC retains the two-component mixture regression model with all four variables, which corroborates the phenomena observed in simulation studies, namely that the AIC overclusters and overfits the data. Although the BIC suggests the single-component model (as selected by the MRC), it lends substantial support to the two-component model with four variables (as selected by the AIC), because the difference between the smallest and the next-smallest BIC values equals $332.93 - 332.54 = .39 < 2$ (see

Table 3. AIC, BIC, and MRC Values for the Sales Territory Example

	p			
K	1	2	3	4
AIC				
1	351.01	328.88	329.84	330.02
2	353.06	334.88	327.85	319.52
BIC				
1	353.44	332.54	334.72	336.11
2	359.16	343.42	338.82	332.93
MRC				
1	376.55	355.02	356.84	358.17
2	400.55	389.29	374.29	379.77

Burnham and Anderson 2002, p. 70; Spiegelhalter et al. 2002, p. 613). Hence the BIC is ambivalent. Next, we highlight the adverse decision making consequences of overclustering and overfitting.

In Table 4, the estimated single-component model indicates that sales are driven by both advertising and the salesperson's effort in a sales territory. Although a salesperson's experience and ability increase sales, these effects are not statistically significant. Substantively, these results suggest that the company's management should allocate its scarce resources to sales compensation (to motivate sales effort) and trade advertising (to support selling activities) rather than investing those resources in sales training programs (to improve the salespersons' abilities). When we consider the two-component model retained by the AIC, we see that the parameter estimate for the salesperson effort in the second component ($k = 2$) has a wrong sign (i.e., suggesting that increased sales effort leads to lower sales). Consequently, the management would be misled into reducing the salesperson's workload, the number of accounts handled, or both. In addition, the estimated effects of salesperson experience are large and significant for both of the components. This finding would lead to incorrect hiring decisions, because management would target older, more experienced salespersons instead of younger ones. Finally, in the second component, both the effect and the significance of salesperson ability are overstated, potentially leading to overinvestment in sales training programs. Thus, in the absence of an appropriate clustering penalty function as in the MRC, the users of mixture regression models incur the risk of detecting spurious components and fitting regressions in nonexistent components.

4. CONCLUDING REMARKS

We have examined the joint determination of the number of components and variables for mixture regression models and have derived the MRC model selection criterion, which has a desirable asymptotic property and performs well in finite samples. Here we discuss extension of the MRC to non-Gaussian data and time series models.

Table 4. Estimated Effects for the Sales Territory Models for K = 1 and K = 2

	K = 1	K = 2	
Variables	*One segment*	*Segment k = 1*	*Segment k = 2*
Advertising effort	.22	.17	.25
	(4.31)	(2.92)	(23.46)
Salesperson effort	.81	1.14	−.17
	(4.22)	(6.22)	(−3.24)
Salesperson experience	2.26	7.47	7.70
	(1.16)	(2.98)	(16.81)
Salesperson ability	194.44	147.12	325.79
	(1.38)	(.93)	(11.94)

NOTE: *t*-values are reported in parentheses.

4.1 Mixture Quasi-Likelihood Model

The quasi-likelihood approach (Nelder and Pregibon 1987; McCullagh and Nelder 1989) augments the scope of linear models by including various non-Gaussian models (e.g., logistic regression models, Poisson regression models). Here we further extend the approach of McCullagh and Nelder (1989, pp. 336, 350) and consider the mixture of quasi-likelihood models in the exponential family,

$$f(y_j;\phi) = \sum_{k=1}^{K}\left[\alpha_k\left\{-\frac{1}{2}\log(\sigma_k)^2 - \frac{1}{2}\frac{D(y_j;\mu_{jk})}{{\sigma_k}^2}\right\}\right], \quad (6)$$

where $\alpha_k > 0$, $\sum_k \alpha_k = 1$; $D(y_j;\mu_{jk}) = 2\{Q(y_j;y_j) - Q(y_j;\mu_{jk})\}$; $Q(y_j;\mu_{jk}) = \{y_j\theta_{jk} - b(\theta_{jk}) + c(y_j)\}$; $b(\cdot)$ and $c(\cdot)$ are suitably chosen functions; $\mu_{jk} = E_k(y_j) = \partial b(\theta_{jk})/\partial\theta_{jk}$; $g(\mu_{jk}) = \eta_{jk} = \mathbf{x}_j'\boldsymbol{\beta}_k$; $\mathbf{x}_j$ is a $p \times 1$ explanatory vector for $j = 1,\ldots,n$, and $\boldsymbol{\beta}_k$ is a conformable parameter vector. Relegating the details to Appendix C, we show that the MRC in (5) serves as the selection criterion for model (6). Note that in the single-component case ($K = 1$), this extended MRC is identical to Hurvich and Tsai's (1994) AIC_c criterion. Thus this extended MRC can be used to select non-Gaussian mixture regression models (e.g., Wedel and Kamakura 2000, chap. 7; McLachlan and Peel 2000, chap. 5).

4.2 Mixture Autoregressive Time Series Model

We consider two broad classes of time series models. The Gaussian mixture transition distribution model (GMTD) captures various nonlinear phenomena, including flat stretches, bursts, and outliers in time series data (Le et al. 1996). The mixture autoregressive models (MARs), which include the GMTD model as a special case, incorporate additional features, such as cycles and conditional heteroscedasticity (Wong and Li 2000). Following Wong and Li (2000, eq. 2.1), we specify the MAR model,

$$F(y_t|\mathcal{F}_{t-1}) = \sum_{k=1}^{K}\alpha_k\Phi\left\{\frac{y_t - \gamma_{k1}y_{t-1} - \cdots - \gamma_{kp_k}y_{t-p_k}}{\sigma_k}\right\},$$
$$t = \tilde{p}+1,\ldots,n, \quad (7)$$

where $\tilde{p} = \max(p_1,\ldots,p_K)$, $\mathcal{F}_{t-1}$ is the information set up to time $t-1$ and $\Phi(\cdot)$ is the distribution function of a standard normal variate. The parameters of the MAR model can be estimated using an EM algorithm (see Wong and Li 2000). To determine the components K and select the order p_k, we apply the derivation in Section 2.3 to the model structure in (7). Relegating the details to Appendix D, we show that the resulting criterion has the same form as (5). Following Chen, Chen, and Kalbfleisch (2004), we test the quintessential hypothesis of homogeneity versus K-component mixture autoregressive models, which would otherwise require a computationally intensive bootstrap approach (McLachlan 1987). Specifically, we consider Hurvich and Tsai's (1989) simulation setting to generate 30 observations from an AR(2) process, $y_t = .99y_{t-1} - .8y_{t-2} +$

ε_t, $\varepsilon_t \sim \mathrm{N}(0,1)$, and select across 35 different MAR models with $p = 1, \ldots, 7$ lagged variables and $K = 1, \ldots, 5$ components. The Monte Carlo results show that the AIC retains the correct model 0 times, the BIC retains the correct model 142 times, and the MRC yields marked improvement with 862 correct selections out of 1,000 realizations.

We close by identifying three avenues for further research. The first of these is to derive the MRC for mixtures of single-index models in which $f_k(y_j; \mathbf{x}_j, \boldsymbol{\beta}_k, \sigma_k)$ in (1) is the normal density with mean $h_k(\mathbf{x}_j'\boldsymbol{\beta}_k)$, where $h_k(\cdot)$ is an unknown differentiable function. (For the single-component case, see Naik and Tsai 2001; for the partial linear regression model, see Chen and Jin 2006.) The second avenue is to adopt Shi and Tsai's (2002) residual likelihood approach or Hjort and Claeskens' (2003) frequentist model average method to further extend the applicability of the MRC. The third avenue is to extend Green's (1995) reversible-jump Markov chain Monte Carlo strategies and Brooks et al.'s (2003) simulated annealing algorithm for joint determination of mixture models and regression variables. We believe that such efforts would enhance the usefulness of mixture regression models.

APPENDIX A: DERIVATION OF THE MIXTURE REGRESSION CRITERION

We note that $E_{\mathbf{Y}}[E_{\mathbf{Y}^*}\{L^0(\phi^0; \mathbf{Z}^*, \mathbf{Y}^*, \mathbf{X})\}]$ in (3) does not depend on the fitted candidate model and can be viewed as a constant (see McLachlan and Peel 2000, chap. 6.8; Burnham and Anderson 2002, chap. 2.1.2). Ignoring this constant, (3) can be expressed as

$$\begin{aligned} d_{\mathrm{CL}} &= -2E_{\mathbf{Y}}\big[E_{\mathbf{Y}^*}\{L(\hat{\phi}; \hat{\boldsymbol{\tau}}, \mathbf{Y}^*, \mathbf{X})\}\big] \\ &= -2E_{\mathbf{Y}}E_{\mathbf{Y}^*}\left\{\sum_{k=1}^{K}\sum_{j=1}^{n}\hat{\tau}_{jk}\big[\log\hat{\alpha}_k + \log f_k(y_j^*; \mathbf{x}_j, \hat{\boldsymbol{\beta}}_k, \hat{\sigma}_k)\big]\right\}, \end{aligned} \tag{A.1}$$

where $\hat{\tau}_{jk}$, $\hat{\alpha}_k$, $\hat{\boldsymbol{\beta}}_k$, and $\hat{\sigma}_k^2$ are EM estimators of τ_{jk}, α_k, $\boldsymbol{\beta}_k$, and σ_k^2 (see Sec. 2.2). After algebraic simplifications, we have

$$\begin{aligned} d_{\mathrm{CL}} &= E_{\mathbf{Y}}\left\{\sum_{k=1}^{K^0}\left[\mathrm{tr}(\hat{\mathbf{W}}_k)\left\{\log\left(\frac{\hat{\sigma}_k^2}{\hat{\alpha}_k^2}\right) + \frac{(\sigma_k^0)^2}{\hat{\sigma}_k^2}\right\}\right.\right. \\ &\qquad \left.\left. + \frac{(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^*)'\hat{\mathbf{X}}_k'\hat{\mathbf{X}}_k(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^*)}{\hat{\sigma}_k^2}\right]\right\} \\ &= E_{\mathbf{Y}}\left[\sum_{k=1}^{K^0}\left\{\mathrm{tr}(\hat{\mathbf{W}}_k)\log\left(\frac{\hat{\sigma}_k^2}{\hat{\alpha}_k^2}\right)\right\}\right] \\ &\quad + E_{\mathbf{Y}}\left\{\sum_{k=1}^{K^0}\left[\{\mathrm{tr}(\hat{\mathbf{W}}_k)\}^2\frac{(\sigma_k^0)^2}{\boldsymbol{\varepsilon}_k'\hat{\mathbf{W}}_k^{1/2}(\mathbf{I} - \hat{\mathbf{H}}_k)\hat{\mathbf{W}}_k^{1/2}\boldsymbol{\varepsilon}_k}\right]\right\} \\ &\quad + E_{\mathbf{Y}}\left\{\sum_{k=1}^{K^0}\left[\mathrm{tr}(\hat{\mathbf{W}}_k)\frac{\boldsymbol{\varepsilon}_k'\hat{\mathbf{W}}_k^{1/2}\hat{\mathbf{H}}_k\hat{\mathbf{W}}_k^{1/2}\boldsymbol{\varepsilon}_k}{\boldsymbol{\varepsilon}_k\hat{\mathbf{W}}_k^{1/2}(\mathbf{I} - \hat{\mathbf{H}}_k)\hat{\mathbf{W}}_k^{1/2}\boldsymbol{\varepsilon}_k}\right]\right\}, \end{aligned} \tag{A.2}$$

where $\boldsymbol{\varepsilon}_k = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_k^*$.

Under regularity conditions, $\hat{\boldsymbol{\tau}}_k$ is a consistent estimate of $\boldsymbol{\tau}_k^0 = (\tau_{1k}^0, \ldots, \tau_{nk}^0)'$ (see Redner and Walker 1984; Leroux 1992), where $\tau_{jk}^0 = E(z_{jk}^0|\mathbf{Y})$. Therefore, we replace $\hat{\mathbf{W}}_k$ and $\hat{\mathbf{H}}_k$ in the second and third terms of the right side of (A.2) by $\mathbf{W}_k^0 = \mathrm{diag}(\tau_k^0)$ and $\mathbf{H}_k^0$, where $\mathbf{H}_k^0 = \mathbf{X}_k^0(\mathbf{X}_k^{0\prime}\mathbf{X}_k^0)^{-1}\mathbf{X}_k^{0\prime}$ and $\mathbf{X}_k^0 = (\mathbf{W}_k^0)^{1/2}\mathbf{X}$. Then we apply Satterthwaite's (1946) results on the two-moment chi-squared approximation for quadratic forms, $\boldsymbol{\varepsilon}_k'(\mathbf{W}_k^0)^{1/2}\mathbf{H}_k^0(\mathbf{W}_k^0)^{1/2}\boldsymbol{\varepsilon}_k$ and $\boldsymbol{\varepsilon}_k'(\mathbf{W}_k^0)^{1/2}(\mathbf{I} - \mathbf{H}_k^0)(\mathbf{W}_k^0)^{1/2}\boldsymbol{\varepsilon}_k$ (also see Loader 1999, p. 161). Furthermore, we adopt Cleveland and Devlin's (1988) results to approximate the ratio of quadratic forms,

$$\frac{\boldsymbol{\varepsilon}_k'(\mathbf{W}_k^0)^{1/2}\mathbf{H}_k^0(\mathbf{W}_k^0)^{1/2}\boldsymbol{\varepsilon}_k}{\boldsymbol{\varepsilon}_k'(\mathbf{W}_k^0)^{1/2}(\mathbf{I} - \mathbf{H}_k^0)(\mathbf{W}_k^0)^{1/2}\boldsymbol{\varepsilon}_k},$$

by an F distribution. Consequently, the second term can be written as

$$\begin{aligned} &E_{\mathbf{Y}}\left\{\sum_{k=1}^{K^0}\left[\{\mathrm{tr}(\mathbf{W}_k^0)\}^2\frac{(\sigma_k^0)^2}{\boldsymbol{\varepsilon}_k'(\mathbf{W}_k^0)^{1/2}(\mathbf{I} - \mathbf{H}_k^0)(\mathbf{W}_k^0)^{1/2}\boldsymbol{\varepsilon}_k}\right]\right\} \\ &\qquad \approx \sum_{k=1}^{K^0}\{\mathrm{tr}(\mathbf{W}_k^0)\}^2\frac{\delta_{k1}/\delta_{k2}}{\delta_{k1}^2/\delta_{k2} - 2}, \end{aligned} \tag{A.3}$$

where $\delta_{k1} = \mathrm{tr}\{(\mathbf{I}_k - \mathbf{H}_k^0)\mathbf{W}_k^0\}$ and $\delta_{k2} = \mathrm{tr}[\{(\mathbf{I} - \mathbf{H}_k^0)\mathbf{W}_k^0\}^2]$. The third term can be written as

$$\begin{aligned} &E_{\mathbf{Y}}\left\{\sum_{k=1}^{K^0}\left[\mathrm{tr}(\mathbf{W}_k^0)\frac{\boldsymbol{\varepsilon}_k'(\mathbf{W}_k^0)^{1/2}\mathbf{H}_k^0(\mathbf{W}_k^0)^{1/2}\boldsymbol{\varepsilon}_k}{\boldsymbol{\varepsilon}_k'(\mathbf{W}_k^0)^{1/2}(\mathbf{I} - \mathbf{H}_k^0)(\mathbf{W}_k^0)^{1/2}\boldsymbol{\varepsilon}_k}\right]\right\} \\ &\qquad \approx \sum_{k=1}^{K^0}\mathrm{tr}(\mathbf{W}_k^0)\frac{\nu_{k1}(\delta_{k1}/\delta_{k2})}{\delta_{k1}^2/\delta_{k2} - 2}, \end{aligned} \tag{A.4}$$

where $\nu_{k1} = \mathrm{tr}(\mathbf{H}_k^0\mathbf{W}_k^0)$. Finally, substituting (A.3) and (A.4) into the second and third terms on the right side of (A.2), and replacing K^0 and $\mathbf{W}_k^0$ by the candidate component K and the EM estimator $\hat{\mathbf{W}}_k$, we obtain an estimate of d_{CL}, which is stated in (4).

APPENDIX B: PROOF OF THE THEOREM

Before proving the theorem, we introduce two lemmas, the proofs of which are straightforward and can be obtained from the third author. Let $r_k = \frac{\hat{n}_k}{n} - \alpha_k^0$ and $Q = \exp\{\sum_{k=K^0+1}^{K}\frac{\hat{n}_k}{n}\log(\hat{\sigma}_k^2/\hat{\alpha}_k^2)\}$. Also, define $R = \exp\{\sum_{k=1}^{K^0} r_k\log(\hat{\sigma}_k^2/\hat{\alpha}_k^2\}$, $S_1 = \exp\{\sum_{k=1}^{K^0}\frac{\hat{n}_k(\hat{n}_k+p_k)}{n(\hat{n}_k-p_k-2)}\}$, $S_2 = \exp\{\sum_{k=K^0+1}^{K}\frac{\hat{n}_k(\hat{n}_k+p_k)}{n(\hat{n}_k-p_k-2)}\}$, and $\delta = \prod_{k=1}^{K^0}(\hat{\sigma}_k^2/\hat{\alpha}_k^2)^{\alpha_k^0} - \prod_{k=1}^{K^0}(\hat{\sigma}_k^2/(\alpha_k^0)^2)^{\alpha_k^0}$.

Lemma B.1. Under Assumptions 1, 2, and 3,

$$\begin{aligned} &Q = 1 + o_p(L_n(\xi)), \\ &R = 1 + o_p(L_n(\xi)), \\ &\exp\left\{\frac{\hat{n}_k(\hat{n}_k + p_k)}{n(\hat{n}_k - p_k - 2)}\right\} \\ &\quad = \exp(\alpha_k^0)\left\{1 + \frac{2}{n} + \frac{2L_{n,k}(\xi)}{(\sigma_k^0)^2} + o_p(L_{n,k}(\xi))\right\} \\ &\qquad \text{for } k = 1, \ldots, K^0, \\ &S_1 = \exp\left(\sum_{k=1}^{K^0}\alpha_k^0\right)\left\{1 + \sum_{k=1}^{K^0}\left[\frac{2}{n} + \frac{2L_{n,k}(\xi)}{(\sigma_k^0)^2}\right] + o_p(L_n(\xi))\right\}, \\ &S_2 = \exp\left\{\frac{2(K - K^0)}{n} + o_p(L_n(\xi))\right\} \\ &\quad = 1 + \frac{2(K - K^0)}{n} + o_p(L_n(\xi)), \end{aligned}$$

and

$$\delta = o_p(L_n(\xi)).$$

Lemma B.2. Under Assumptions 1, 2, and 3,

$$\prod_{k=1}^{K^0}\left\{\left(\frac{\omega_k}{(\alpha_k^0)^2}-\frac{L_{n,k}(\xi)}{(\alpha_k^0)^2}+o_p(L_{n,k}(\xi))\right)^{\alpha_k^0}\times\left[1+\frac{2}{n}+\frac{2}{(\sigma_k^0)^2}L_{n,k}(\xi)+o_p(L_{n,k}(\xi))\right]\right\}=\prod_{k=1}^{K^0}\left[A_k+B_kL_{n,k}(\xi)+o_p(L_{n,k}(\xi))\right] \tag{B.1}$$

and

$$\prod_{k=1}^{K^0}\left[A_k+B_kL_{n,k}(\xi)+o_p(L_{n,k}(\xi))\right]=\prod_{k=1}^{K^0}A_k+\prod_{k=1}^{K^0}B_kL_{n,k}(\xi)\prod_{j\neq k}^{K^0}A_j+o_p(L_n(\xi)). \tag{B.2}$$

To prove the theorem, we re-express the *MRC* in (5) as

$$MRCE=\exp\left(\frac{1}{n}MRC\right). \tag{B.3}$$

For $\xi\in\mathcal{A}_n^0$, it follows from Lemma B.1 that

$$\begin{aligned}&MRCE\\&=\exp\left\{\sum_{k=1}^{K}\frac{\hat n_k}{n}\log\left(\frac{\hat\sigma_k^2}{\hat\alpha_k^2}\right)+\sum_{k=1}^{K}\frac{\hat n_k(\hat n_k+p_k)}{n(\hat n_k-p_k-2)}\right\}\\&=\prod_{k=1}^{K^0}\left[\left(\frac{\hat\sigma_k^2}{(\alpha_k^0)^2}\right)^{\alpha_k^0}\exp\left\{\frac{\hat n_k(\hat n_k+p_k)}{n(\hat n_k-p_k-2)}\right\}\right]S_2QR+\delta S_1S_2QR\\&=\prod_{k=1}^{K^0}\left\{\left(\frac{\omega_k}{(\alpha_k^0)^2}-\frac{L_{n,k}(\xi)}{(\alpha_k^0)^2}+o_p(L_{n,k}(\xi))\right)^{\alpha_k^0}\times\exp(\alpha_k^0)\left(1+\frac{2}{n}+\frac{2L_{n,k}(\xi)}{(\sigma_k^0)^2}+o_p(L_{n,k}(\xi))\right)\right\}\\&\quad\times\left(1+\frac{2(K-K^0)}{n}+o_p(L_n(\xi))\right)+o_p(L_n(\xi)).\end{aligned} \tag{B.4}$$

Equations (B.3) and (B.4) together with Lemma B.2 yield

$$\begin{aligned}MRCE&=W_n\prod_{k=1}^{K^0}A_k+W_n\sum_{k=1}^{K^0}B_kL_{n,k}(\xi)\prod_{j\neq k}^{K^0}A_j+o_p(L_n(\xi))\\&=W_n\prod_{k=1}^{K^0}A_k+L_n(\xi)+o_p(L_n(\xi)).\end{aligned} \tag{B.5}$$

Next, using Assumption 1, we obtain

$$MRCE(\hat\xi)=\min_{\xi\in\mathcal{A}_n}MRCE(\xi)=\min_{\xi\in\mathcal{A}_n^0}MRCE(\xi), \tag{B.6}$$

for large n. Therefore, we omit the constant term $W_n\prod_{k=1}^{K^0}A_k$ from $MRCE(\xi)$ in (B.5). Using (B.6), we can show that

$$L_n(\hat\xi)+R_n(\hat\xi)=\min_{\xi\in\mathcal{A}_n^0}\bigl(L_n(\xi)+R_n(\xi)\bigr), \tag{B.7}$$

where $R_n(\xi)=o_p(L_n(\xi))$.

Finally, it follows from (B.6) and (B.7) that

$$\begin{aligned}1&=\frac{L_n(\bar\xi)}{L_n(\bar\xi)}\\&\geq\frac{\inf_{\xi\mathcal{A}_n^0}(L_n(\xi)+R_n(\xi))}{L_n(\bar\xi)}-\frac{R_n(\bar\xi)}{L_n(\bar\xi)}\\&=\frac{L_n(\hat\xi)+R_n(\hat\xi)}{L_n(\bar\xi)}-\frac{R_n(\bar\xi)}{L_n(\bar\xi)}\\&\geq\frac{L_n(\hat\xi)}{L_n(\bar\xi)}-\frac{|R_n(\hat\xi)|}{L_n(\bar\xi)}-\frac{|R_n(\bar\xi)|}{L_n(\bar\xi)}\\&\geq\frac{L_n(\hat\xi)}{L_n(\bar\xi)}-\sup_{\xi\in\mathcal{A}_n^0}\frac{|R_n(\xi)|}{L_n(\xi)}\cdot\frac{L_n(\hat\xi)}{L_n(\bar\xi)}-\frac{|R_n(\bar\xi)|}{L_n(\bar\xi)}\\&=\frac{L_n(\hat\xi)}{L_n(\bar\xi)}\left(1-\sup_{\xi\in\mathcal{A}_n^0}\frac{|R_n(\xi)|}{L_n(\xi)}\right)-\frac{|R_n(\bar\xi)|}{L_n(\bar\xi)}.\end{aligned}$$

This inequality further implies that

$$\begin{aligned}1&=\frac{L_n(\bar\xi)}{L_n(\bar\xi)}\\&\geq\frac{L_n(\hat\xi)}{L_n(\bar\xi)}\left(1-\sup_{\xi\in\mathcal{A}_n^0}\frac{|R_n(\xi)|}{L_n(\xi)}\right)-\sup_{\xi\in\mathcal{A}_n^0}\frac{|R_n(\xi)|}{L_n(\xi)}\longrightarrow\frac{L_n(\hat\xi)}{L_n(\bar\xi)}\geq1\end{aligned}$$

in probability as n tends to infinity. This completes the proof of the theorem.

APPENDIX C: THE MIXTURE REGRESSION CRITERION FOR THE MIXTURE QUASI–LIKELIHOOD MODEL

The complete-data log-likelihood discrepancy of the mixture quasi-likelihood model (6) is

$$\begin{aligned}d_{\mathrm{CL}}&=-2E_{\mathbf{Y}}\bigl[E_{\mathbf{Y}^*}\{L(\hat\phi;\hat{\boldsymbol\tau},\mathbf{Y}^*,\mathbf{X})\}\bigr]\\&=-2E_{\mathbf{Y}}E_{\mathbf{Y}^*}\left\{\sum_{k=1}^{K}\sum_{j=1}^{n}\hat\tau_{jk}\bigl[\log\hat\alpha_k+\log f_k(y_j^*;\mathbf{x}_j,\hat{\boldsymbol\beta}_k,\hat\sigma_k)\bigr]\right\}\\&=\sum_{k=1}^{K^0}E_{\mathbf{Y}}\left[\operatorname{tr}(\hat{\mathbf{W}}_k)\log\left(\frac{\hat\sigma_k^2}{\hat\alpha_k^2}\right)+\frac{E_{\mathbf{Y}^*}\{\mathbf{Y}^{*\prime}\hat{\mathbf{W}}_k\mathbf{Y}^*-\mathbf{1}'\hat{\mathbf{W}}_kb(\mathbf{Y}^*)\}}{\hat\sigma_k^2}\right.\\&\qquad\left.-\frac{\boldsymbol\mu_k^{0\prime}\hat{\mathbf{W}}_k\hat{\boldsymbol\theta}_k-\mathbf{1}'\hat{\mathbf{W}}_kb(\hat{\boldsymbol\theta}_k)}{\hat\sigma_k^2}\right],\end{aligned} \tag{C.1}$$

where $\mathbf{Y}^*$ is defined as in Section 2.3; $\boldsymbol\mu_k^0=(\mu_{1k}^0,\ldots,\mu_{nk}^0)'$; $\mu_{jk}^0=\partial b(\theta_{jk}^0)/\partial\theta_{jk}^0$; θ_{jk}^0 is the jth element of the canonical parameter in the kth component of the true model; $\mathbf{1}=(1,\ldots,1)'$; $\hat\alpha_k$, $\hat\tau_{jk}$, and $\hat{\boldsymbol\beta}_k$ can be obtained through the EM algorithm for the generalized linear mixture model (see Wedel and Kamakura 2000, chap. 7); $\hat\sigma_k^2=(\tilde{\mathbf{Y}}_k-\tilde{\boldsymbol\mu}_k)'(\tilde{\mathbf{Y}}_k-\tilde{\boldsymbol\mu}_k)/\operatorname{tr}(\hat{\mathbf{W}}_k)$; $\tilde{\mathbf{Y}}_k=\tilde{\mathbf{W}}_k^{1/2}\mathbf{Y}$; $\tilde{\boldsymbol\mu}_k=\tilde{\mathbf{W}}_k^{1/2}\hat{\boldsymbol\mu}_k$; $\boldsymbol\mu_k=(\mu_{1k},\ldots,\mu_{nk})'$; $\tilde{\mathbf{W}}_k=\hat{\mathbf{V}}_k\hat{\mathbf{W}}_k$; $\mathbf{V}_k=\operatorname{diag}\{(\partial\boldsymbol\mu_k/\partial\eta_k)^2(\partial^2b(\boldsymbol\theta_k)/\partial\boldsymbol\theta_k\,\partial\boldsymbol\theta_k')\}$ is the $n\times n$ matrix; the jth diagonal element of $\mathbf{V}_k$ is $(\partial\mu_{jk}/\partial\eta_{jk})^2(\partial^2b(\theta_{jk})/\partial\theta_{jk}\,\partial\theta_{jk})$; $\boldsymbol\theta_k=(\theta_{1k},\ldots,\theta_{nk})'$; $\hat{\boldsymbol\theta}_k$, $\hat{\boldsymbol\mu}_k$, and $\hat{\mathbf{V}}_k$ are $\boldsymbol\theta_k$, $\boldsymbol\mu_k$, and $\mathbf{V}_k$ evaluated at $\hat{\boldsymbol\tau}_k$, $\hat\alpha_k$, and $\hat{\boldsymbol\beta}_k$; and $\hat{\boldsymbol\tau}_k$ and $\hat{\mathbf{W}}_k$ are as defined as in Section 2.3.

Applying the quadratic approximations of $\hat{\boldsymbol\theta}_k$ and $b(\hat{\boldsymbol\theta}_k)$ at $\boldsymbol\beta_k=\boldsymbol\beta_k^*$ and then replacing the $\hat{\mathbf{W}}_k$ in the last two terms of (C.1) by $\mathbf{W}_k^0$, we

have

$$d_{\mathrm{CL}} \approx \sum_{k=1}^{K^0} E_{\mathbf{Y}}\left[\operatorname{tr}(\hat{\mathbf{W}}_k)\log\left(\frac{\hat{\sigma}_k^2}{\hat{\alpha}_k^2}\right) + \operatorname{tr}(\mathbf{W}_k^0)\frac{(\sigma_k^0)^2}{\hat{\sigma}_k^2} + \frac{(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^*)'\tilde{\mathbf{X}}_k'\tilde{\mathbf{X}}_k(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^*)}{\hat{\sigma}_k^2}\right],$$

where $(\sigma_k^0)^2 = E_{\mathbf{Y}^*}\{\mathbf{Y}^{*\prime}\mathbf{W}_k^0\mathbf{Y}^* - \mathbf{1}'\mathbf{W}_k^0 b(\mathbf{Y}^*) - \boldsymbol{\mu}_k^{0\prime}\mathbf{W}_k^0\boldsymbol{\theta}_k^0 + \mathbf{1}' \times \mathbf{W}_k^0 b(\boldsymbol{\theta}_k^0)\}/\operatorname{tr}(\mathbf{W}_k^0)$ and $\tilde{\mathbf{X}} = \tilde{\mathbf{W}}_k^{1/2}\mathbf{X}$. Subsequently, adapting Jorgensen's (1987, sec. 4) asymptotic results, we have

$$\begin{aligned} d_{\mathrm{CL}} \approx{}& E_{\mathbf{Y}}\left[\sum_{k=1}^{K^0}\left\{\operatorname{tr}(\hat{\mathbf{W}}_k)\log\left(\frac{\hat{\sigma}_k^2}{\hat{\alpha}_k^2}\right)\right\}\right] \\ &+ E_{\mathbf{Y}}\left\{\sum_{k=1}^{K^0}\left[\operatorname{tr}(\hat{\mathbf{W}}_k)\operatorname{tr}(\mathbf{W}_k^0)\frac{(\sigma_k^0)^2}{\boldsymbol{\varepsilon}_k'\tilde{\mathbf{W}}_k^{1/2}(\mathbf{I}-\tilde{\mathbf{H}}_k)\tilde{\mathbf{W}}_k^{1/2}\boldsymbol{\varepsilon}_k}\right]\right\} \\ &+ E_{\mathbf{Y}}\left\{\sum_{k=1}^{K^0}\left[\operatorname{tr}(\hat{\mathbf{W}}_k)\frac{\boldsymbol{\varepsilon}_k'\tilde{\mathbf{W}}_k^{1/2}\tilde{\mathbf{H}}_k\tilde{\mathbf{W}}_k^{1/2}\boldsymbol{\varepsilon}_k}{\boldsymbol{\varepsilon}_k'\tilde{\mathbf{W}}_k^{1/2}(\mathbf{I}-\tilde{\mathbf{H}}_k)\tilde{\mathbf{W}}_k^{1/2}\boldsymbol{\varepsilon}_k}\right]\right\}, \end{aligned} \tag{C.2}$$

where $\tilde{\mathbf{H}}_k = \tilde{\mathbf{X}}_k(\tilde{\mathbf{X}}_k'\tilde{\mathbf{X}}_k)^{-1}\tilde{\mathbf{X}}_k'$ and $\boldsymbol{\varepsilon}_k = \mathbf{Y} - \boldsymbol{\mu}_k^0$. Because (C.2) is similar to (A.2), we can apply techniques analogous to those used after (A.2) in Appendix A to obtain the MRC as given in (5), where $\hat{n}_k = \operatorname{tr}(\tilde{\mathbf{W}}_k)$ and $p_k = \operatorname{tr}(\tilde{\mathbf{H}}_k)$.

APPENDIX D: THE MIXTURE REGRESSION CRITERION FOR THE MIXTURE AUTOREGRESSIVE TIME SERIES MODEL

Adapting the approach of Brockwell and Davis' (1991, chaps. 8.8–8.10), we first define $\mathbf{Y} = (y_{\tilde{p}+1}, \ldots, y_n)'$, $\mathbf{X} = (\mathbf{x}_{\tilde{p}+1}, \ldots, \mathbf{x}_n)'$, $\mathbf{x}_j = (y_{j-1}, \ldots, y_{j-p_k})'$, and $\boldsymbol{\beta}_k = (\gamma_{k1}, \ldots, \gamma_{kp_k})'$, where $j = \tilde{p}+1, \ldots, n$ and $k = 1, \ldots, K$. Then, the resulting complete-data log-likelihood discrepancy of the mixture autoregressive time series model (7) is

$$\begin{aligned} d_{\mathrm{CL}} &= -2E_{\mathbf{Y}}\left[E_{\mathbf{Y}^*}\{L(\hat{\phi}; \hat{\boldsymbol{\tau}}, \mathbf{Y}^*, \mathbf{X})\}\right] \\ &= -2E_{\mathbf{Y}}E_{\mathbf{Y}^*}\left\{\sum_{k=1}^{K}\sum_{j=\tilde{p}+1}^{n}\hat{\tau}_{jk}\left[\log\hat{\alpha}_k + \log f_k(y_j^*; \mathbf{x}_j, \hat{\boldsymbol{\beta}}_k, \hat{\sigma}_k)\right]\right\}, \end{aligned}$$

where $\hat{\phi}$, $\hat{\boldsymbol{\tau}}$, and $\mathbf{Y}^*$ are defined as in Section 2.3. This expression is the same as (A.1) except for the summation, which ranges from $\tilde{p}+1$ to n. Next, applying the asymptotic results of Brockwell and Davis (1991, chaps. 8.9 and 8.10) together with algebraic simplifications, we obtain

$$(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^*)'\hat{\mathbf{X}}_k'\hat{\mathbf{X}}_k(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^*) \approx \boldsymbol{\varepsilon}_k'\hat{\mathbf{W}}_k^{1/2}\hat{\mathbf{H}}_k\hat{\mathbf{W}}_k^{1/2}\boldsymbol{\varepsilon}_k$$

and

$$\hat{\sigma}_k^2 \approx \frac{\boldsymbol{\varepsilon}_k'\hat{\mathbf{W}}_k^{1/2}(\mathbf{I}-\hat{\mathbf{H}}_k)\hat{\mathbf{W}}_k^{1/2}\boldsymbol{\varepsilon}_k}{\operatorname{tr}(\hat{\mathbf{W}}_k)},$$

where $\boldsymbol{\beta}_k^*$, $\hat{\mathbf{X}}_k$, $\hat{\mathbf{H}}_k$, and $\hat{\mathbf{W}}_k^{1/2}$ are defined as in Section 2.3 and $\boldsymbol{\varepsilon}_k$ is defined as in (A.2). Consequently, d_{CL} is approximately the same as the right side of (A.2). Applying the techniques used in Appendix A, we finally obtain the MRC as in (5).

[Received July 2004. Revised May 2006.]

REFERENCES

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," in *The 2nd International Symposium on Information Theory*, eds. B. N. Petrov and F. Csaki, Budapest: Akademia Kiado, pp. 267–281.

——— (1985), "Prediction and Entropy," in *A Celebration of Statistics*, eds. A. C. Atkinson and S. E. Fienberg, New York: Springer-Verlag, pp. 1–24.

Altham, F. M. E. (1984), "Improving the Precision of Estimation by Fitting a Model," *Journal of the Royal Statistical Society*, Ser. B, 46, 118–119.

Biernacki, C., Celeux, G., and Govaert, G. (2000), "Assessing a Mixture Model for Clustering With the Integrated Completed Likelihood," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 719–725.

Brockwell, P. J., and Davis, R. A. (1991), *Time Series: Theory and Methods* (2nd ed.), New York: Springer-Verlag.

Brooks, S. P., Friel, N., and King, R. (2003), "Classical Model Selection via Simulated Annealing," *Journal of the Royal Statistical Society*, Ser. B, 65, 503–520.

Bryant, P. G., and Williamson, J. A. (1978), "Asymptotic Behaviour of Classification Maximum Likelihood Estimates," *Biometrika*, 65, 273–281.

Burnham, K. P. (2002), Discussion of "Bayesian Measures of Model Complexity and Fit," by D. J. Spieglehalter, N. G. Best, B. P. Carlin, and A. van der Linde, *Journal of the Royal Statistical Society*, Ser. B, 64, 583–639.

Burnham, K. P., and Anderson, D. R. (2002), *Model Selection and Inference: A Practical Information-Theoretic Approach* (2nd ed.), New York: Springer-Verlag.

——— (2004), "Multimodel Inference: Understanding AIC and BIC in Model Selection," *Sociological Methods & Research*, 33, 261–304.

Cai, Z., Naik, P. A., and Tsai, C. L. (2000), "Denoised Least Squares Estimators: An Application to Estimating Advertising Effectiveness," *Statistica Sinica*, 10, 1231–1241.

Chen, H., Chen, J., and Kalbfleisch, J. D. (2004), "Testing for a Finite Mixture Model With Two Components," *Journal of the Royal Statistical Society*, Ser. B, 66, 95–115.

Chen, K., and Jin, Z. (2006), "Partial Linear Regression Models for Clustered Data," *Journal of the American Statistical Association*, 101, 195–204.

Cleveland, W. S., and Devlin, S. J. (1988), "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting," *Journal of the American Statistical Association*, 83, 596–610.

Cochran, W. G. (1934), "The Distribution of Quadratic Forms in a Normal System, With Applications to the Analysis of Covariance," *Proceedings of the Cambridge Philosophical Society*, 30, 178–191.

Cravens, D. W., Woodruff, R. B., and Stamper, J. C. (1972), "An Analytical Approach for Evaluating Sales Territory Performance," *Journal of Marketing*, 36, 31–37.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

DeSarbo, W. S., and Corn, W. L. (1988), "A Maximum Likelihood Methodology for Clusterwise Linear Regression," *Journal of Classification*, 5, 249–282.

Dielman, T. E. (2001), *Applied Regression Analysis for Business and Economics* (3rd ed.), New York: Duxbury.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004), *Bayesian Data Analysis* (2nd ed.), New York: Chapman & Hall/CRC.

Green, P. J. (1995), "Reversible-Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 711–732.

Hartigan, J. A. (1975), *Clustering Algorithms*, New York: Wiley.

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, New York: Springer-Verlag.

Hathaway, R. J. (1985), "A Constraint Formulation of Maximum-Likelihood Estimation for Normal Mixture Distributions," *The Annals of Statistics*, 13, 795–800.

Helland, I. S., and Almøy, T. (1994), "Comparison of Prediction Methods When Only a Few Components Are Relevant," *Journal of the American Statistical Association*, 89, 583–591.

Hjort, N., and Claskens, G. (2003), "Frequentist Model Average Estimators" (with discussion), *Journal of the American Statistical Association*, 98, 879–899.

Hurvich, C. M., and Tsai, C. L. (1989), "Regression and Time Series Model Selection in Small Samples," *Biometrika*, 76, 297–307.

——— (1994), "Model Selection for Extended Quasi-Likelihood Models in Small Samples," *Biometrics*, 51, 1077–1084.

Jorgensen, B. (1987), "Exponential Dispersion Models" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 49, 127–162.

Le, N. D., Martin, R. D., and Raftery, A. E. (1996), "Modeling Flat Stretches, Bursts, and Outliers in Time Series Using Mixture Transition Distribution Models," *Journal of the American Statistical Association*, 91, 1504–1514.

Leroux, B. (1992), "Consistent Estimation of a Mixing Distribution," *The Annals of Statistics*, 20, 1350–1360.

Loader, C. (1999), *Local Regression and Likelihood*, New York: Springer-Verlag.

MacQueen, J. (1967), "Some Methods for Classification and Analysis of Multivariate Observations," in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1, Berkeley, CA: University of California Press, pp. 281–297.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman & Hall.

McLachlan, J. G. (1987), "On Bootstrapping the Likelihood Ratio Test Statistics for the Number of Components in a Normal Mixture," *Applied Statistics*, 36, 318–324.

McLachlan, J. G., and Peel, D. (2000), *Finite Mixture Models*, New York: Wiley.

McQuarrie, A. D. R., and Tsai, C. L. (1998), *Regression and Time Series Model Selection*, Singapore: World Scientific Publishing.

Naik, P. A., and Tsai, C. L. (2000), "Partial Least Squares Estimator for Single-Index Models," *Journal of the Royal Statistical Society*, Ser. B, 62, 763–771.

——— (2001), "Single-Index Model Selections," *Biometrika*, 88, 821–832.

Nelder, J. A., and Pregibon, D. (1987), "An Extended Quasi-Likelihood Function," *Biometrika*, 74, 221–232.

Raftery, A. E., and Dean, N. (2006), "Variable Selection for Model-Based Clustering," *Journal of the American Statistical Association*, 101, 168–179.

Redner, R. A., and Walker, H. F. (1984), "Mixture Densities, Maximum Likelihood and the EM Algorithm," *SIAM Review*, 26, 195–239.

Richardson, S. (2002), Discussion of "Bayesian Measures of Model Complexity and Fit," by D. J. Spieglehalter, N. G. Best, B. P. Carlin, and A. van der Linde, *Journal of the Royal Statistical Society*, Ser. B, 64, 583–639.

Satterthwaite, F. E. (1946), "An Approximate Distribution of Estimates of Variance Components," *Biometrics Bulletin*, 2, 110–114.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Shao, J. (1997), "An Asymptotic Theory for Linear Model Selection" (with discussion), *Statistica Sinica*, 7, 221–264.

Shi, P., and Tsai, C. L. (1999), "Semiparametric Regression Model Selections," *Journal of Statistical Planning and Inference*, 77, 119–139.

——— (2002), "Regression Model Selection: A Residual Likelihood Approach," *Journal of the Royal Statistical Society*, Ser. B, 64, 237–252.

Shibata, G. (1981), "An Optimal Selection of Regression Variables," *Biometrika*, 68, 45–54.

——— (1984), "Approximate Efficiency of a Selection Procedure for the Number of Regression Variables," *Biometrika*, 71, 43–49.

Spieglehalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002), "Bayesian Measures of Model Complexity and Fit" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 64, 583–639.

Tadesse, M. G., Sha, N., and Vannucci, M. (2005), "Bayesian Variable Selection in Clustering High-Dimensional Data," *Journal of the American Statistical Association*, 100, 602–617.

Tibshirani, R., Walther, G., and Hastie, T. (2001), "Estimating the Number of Clusters in a Data Set via the Gap Statistic," *Journal of the Royal Statistical Society*, Ser. B, 63, 411–423.

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

Wedel, M., and Kamakura, W. (2000), *Market Segmentation: Conceptual and Methodological Foundations* (2nd ed.), Boston: Kluwer Academic.

Wong, C. S., and Li, W. K. (2000), "On a Mixture Autoregressive Model," *Journal of the Royal Statistical Society*, Ser. B, 62, 95–115.

Wu, C. F. J. (1983), "On the Convergence Properties of the EM Algorithm," *The Annals of Statistics*, 11, 95–103.

Zhu, H. T., and Zhang, H. (2004), "Hypothesis Testing in Mixture Regression Models," *Journal of the Royal Statistical Society*, Ser. B, 66, 3–16.