Covariate Selection in Mixture Models with the Censored Response Variable

Zeng-Hua Lu

School of Mathematics and Statistics, University of South Australia, City West, GPO Box 2471, SA 5001, Australia (zen.lu@unisa.edu.au)

Abstract

This paper formulates a mixture model for modeling unobserved heterogeneity of explanatory mechanism. Our model allows for different sets of regressors, and/or different interactions among the same regressors in different regression regimes. The model is demonstrated with particular interest to the censored dependent variable. A two-step procedure is proposed for model identification. The first step is to identify the number of regression regimes with each regime including all regressors. The second step is to select regressors in the regression regimes. The results of our simulation studies suggest that the procedure works well. Two microeconometric applications are provided.

KEYWORDS: Mixture regression models; Switching regression models; Model selection criterion; EM algorithm
1. INTRODUCTION

There has been increasing interest in the econometric literature in modeling unobserved population heterogeneity through mixture regression models, particularly since the influential work by Heckman and Singer (1984). Here are some examples. In the framework of dynamic discrete choices modeling (see e.g. Keane and Wolpin 1997; Cameron and Heckman 1998, 2001; Gilleskie 1998; Eckstein and Wolpin 1999), mixture regression models are used to account for unobserved person-specific differences. In the stochastic frontier literature, (see e.g. Orea and Kumbhakar 2004; Tsionas and Kumbhakar 2004; Greene 2005), the mixture regression models are used for modeling the unobserved production heterogeneity across plants in manufacturing, or farms in agriculture, or industries in a country, or different countries. In the literature of marketing and transportation, the mixture regression mechanism has been incorporated into discrete choice models to account for unobserved taste heterogeneity underlying observed choice behaviour, (see e.g. Gupta and Chintagunta 1994; Bucklin, Gupta and Han 1995; Swait 2003).

However, econometricians have so far assumed that there is a same set of explanatory variables, which interact differently in explaining the outcome of dependent variable in each unobserved subpopulation. There are only two exceptions to this in the broad statistical literature, namely Thompson, Smith and Boyle (1998) and Roeder, Lynch and Nagin (1999).

While economic theories may well provide justifications for allowing different regressors in different regression regimes, statistical inference of the resultant model can be a challenge. If population heterogeneity is observable, then the problem would fit into the literature of variable selection in the non-mixture regression setting (see e.g. Miller 2002 for review of the literature). Challenge arises when population heterogeneity is unobservable. There are mainly two difficulties involved. The first is that the resultant model - mixture regression model inherits estimation complications associated with mixture models due to
failure of some regularity conditions, particularly when the number of regression regimes is unknown (see Titterington, Smith and Makov 1985; McLachlan and Peel 2000 for excellent review of the literature). The second is that the number of candidate models increases rapidly with increase in the number of possible explanatory variables, and/or the number of regression regimes. Given the estimation complications of mixture model, the best subset selection technique; selecting the best fit model among all candidate models is difficult to apply here.

In a mixture regression model where the number of regression regimes was known as simple as two, Thompson et al. (1998) identified the best fit model based on the deletion/addition approach. Roeder et al. (1999) studied a mixture of zero-inflated Poisson regression model with an unknown number of regression regimes. But they had different interest in that different sets of regressors are responsible for the probabilities associated with different regimes. Based on an approximation technique they proposed for estimation of their model, they suggested estimating the number of regression regimes first, then estimate all remaining candidate models and select the best fit model according to the BIC model selection criterion.

This paper studies a two-step procedure for model selection in a mixture model with an unknown number of mixture regression regimes and the censored dependent variable. Our proposed procedure combines the ideas presented in Thompson et al. (1998) and Roeder et al. (1999). We start with estimating the number of regimes in the model, in which each regime includes all possible regressors. The basis of the argument is that when a regressor is incorrectly included in a regression regime, the true value of its regression coefficient is zero. Thus, the incorrect inclusion of a regressor in a regression regime has little effect on the regression line when the sample size is large. We then adopt the stepwise deletion technique for variable selection. We suggest that, starting at a small threshold $z$ score of...
coefficient estimate, one gradually excludes regressors by increasing the threshold value until the resulting model is not better than the previous one according to a model selection criterion. To make sure that the process converges to the optimum model, the second step may be repeated at different starting threshold values and different incremental values.

A mixture model with the censored response variable was previously considered by Nagin and Tremblay (1999). However, they used a direct maximization procedure, which is dependent on a careful choice of starting values (Roeder et al. 1999). We have found that the direct evaluation of the Hessian matrix in our model often results in numerical underflow or overflow problems. This increases the difficulty of implementing a direct maximization procedure. We study an algorithm for estimation of our model, in which we approach both mixing parameter and unobserved part of the dependent variable based on the missing information principle in the EM algorithm. We also estimate the covariance matrix of our estimate based on Louis’ (1982) method.

The remainder of this paper is organised as follows. The model is described in Section 2. The asymptotic properties of our proposed procedure are discussed in Section 3. Section 4 provides an application of the EM algorithm for estimating our model. Section 5 reports the results of simulation studies and two empirical applications. Concluding remarks are made in Section 6.

2. THE MODEL

Consider a finite mixture regression model (FMRM), in which the $n$th ($n = 1, \ldots, N$) observation of the dependent random variable $y^*$ is attributed to one of $K$ finite mutually exclusive explanatory mechanisms,

$$y^*_n = \beta_k' x_{kn} + e_{kn}, \text{ with probability } \alpha_k,$$

(1)
where \( k = 1, \ldots, K \), \( x_{in} \) is a \( d_k \times 1 \) column vector of regressors, \( \beta_k \) a \( d_k \times 1 \) column vector of regression coefficients and \( e_{in} \) the error term. The mixing parameter \( \alpha_k \) is subject to the constraints \( 0 \leq \alpha_k \leq 1 \) and \( \sum_{k=1}^{K} \alpha_k = 1 \).

We assume that \( \{e_{in}\} \) is independent normally distributed, \( e_{in} \sim i.i.d.N(0, \sigma^2_k) \), and \( y^* \) is only observed \( y = y^* \) when \( y^* > 0 \), and \( y = 0 \) when \( y^* \leq 0 \). Let \( \phi_k \) and \( \Phi_k \) be the probability density function (pdf), and cumulative distribution function (cdf) of \( e_k \), respectively. Then \( y_n \) follows the \( K \)-component mixture distribution,

\[
f_K(y_n | x_n; \theta) = \begin{cases} 
\sum_{k=1}^{K} \alpha_k \{1 - \Phi_k (\beta_k' x_{in})\} & \text{if } y_n = 0, \\
\sum_{k=1}^{K} \alpha_k \phi_k (y - \beta_k' x_{in}) & \text{if } y_n > 0,
\end{cases}
\]

where \( \theta = (\alpha_1, \ldots, \alpha_{K-1}, \beta_1', \ldots, \beta_K', \sigma_1^2, \ldots, \sigma_K^2)' \) and \( \alpha_K = 1 - \sum_{k=1}^{K-1} \alpha_k \). Note that when \( K \) is unknown one needs to assume parametric component distributions, i.e. parametric distribution forms for all \( e_k \), because of the issue of model identification (see e.g. Heckman and Singer 1984 for discussions).

3. STRONG CONSISTENCY

There is a parameter identifiability problem when \( K > K_0 \). For example, suppose the true model is \( y_n = \beta_{01} + \beta_{11} x_{in} + e_n \). If one attempts to fit a two-component mixture with an extra component \( y_n = \beta_{02} + \beta_{12} x_{in} + \beta_{22} x_{2n} + e_n \), then the true parameter value is not unique; the true model can be recovered with \( \alpha_2 = 0 \), or \( \beta_{02} = \beta_{01} \), \( \beta_{12} = \beta_{11} \) and \( \beta_{22} = 0 \). Because of such problem, statistical inference of mixture models often becomes very challenging and attracts ongoing research (see e.g. Zhu and Zhang 2006; Chambaz 2006 for the recent literature). In relation to estimation of \( K \), there have been various techniques suggested in the literature (see e.g. Roeder 1994; Chen, Chen and Kalbfleisch 2004; Charnigo and Sun 2004...

5
and references therein). Given the complexity of the problem considered in this paper, we adopt the model selection criterion based technique for the job. A number of authors (e.g. Dasgupta and Raftery 1998; Keribin 2000; Wong and Li 2001) found that the BIC model selection technique worked well in the context of mixture models. Based on the work of Feng and McCulloch (1996), Leroux (1992) and Keribin (2000), we show that our proposed two-step model selection procedure based on the BIC model selection criterion possesses the strong consistency property.

Define $\rightarrow_{a.s.}$ as “converge almost surely (a.s.)”, “hat” for estimator, the symbol with the subscript “0” for the true parameter value or space, and “$E$” proceeding a symbol as the expected value.

Let the parameter space be $\Theta = \{\theta : 0 \leq \alpha_k \leq 1, \beta_k \in R^{d_k}, \sigma_k^2 > 0\}$, and the sub parameter space $\Theta_0 = \{\theta : f_K(y, x; \theta) = f_{K_0}(y, x; \theta_0) \subset \Theta\}$, where $\theta_0$ is the parameter in the true model with $K_0$.

**Theorem 1.** If Assumptions A.1-A.6 are satisfied, then the MLE $\hat{\theta}$ possesses the strong consistency property such that $\hat{\theta} - \hat{\theta}_0(\hat{\theta}) \rightarrow_{a.s.} 0$, where $\hat{\theta}_0(\hat{\theta}) \in \Theta_0$ is the point in $\Theta_0$ that is closest to $\hat{\theta}$ in the Euclidean distance.

Feng and McCulloch (1996) derived such results in the non-regression setting. The proof of Theorem 1 can be easily done by extending their work and is outlined in Appendix A.

As each point in $\Theta_0$ represents the true model, the result in Theorem 1 is equivalent to establish the strong consistency property of the estimator of distribution, $\hat{f}_K \rightarrow_{a.s.} f_{K_0}$.

Suppose one wishes to identify the true model by maximizing an information criterion, $l_N \{f_K(\theta)\} - b_{KN}$, where $l_N(\cdot) = \sum_{n=1}^N \ln \{f_K(y_n | x_n; \theta)\}$ and $b_{KN}$ is a penalty term. The next theorem show the strong consistency result of the MPLE.
Theorem 2. If Assumptions A.1-A.6 are satisfied, and if \( b_{KN} \) satisfies the following conditions, (i) \( b_{(k+1)N} \geq b_{KN} > 0 \), for \( \forall N \), (ii) \( \lim_{N \to \infty} \ln b_{KN} = \infty \), \( b_{KN} = o(N) \) and 
\[
\lim_{N \to \infty} \ln N / b_{KN} = 0, \text{ for } \forall K \text{ and } (iii) \lim_{k \to \infty} b_{KN} / b_{K_{0}N} > 1 \text{ for } K_{0} < K < K, \text{ where } K \text{ is a known upper bound, then the MPLE } K \xrightarrow{a.s.} K_{0} \text{ and } \hat{\theta}_{K} \xrightarrow{a.s.} \theta_{0}.
\]

The results of Theorem 2 were essentially provided in Leroux (1992) and Keribin (2000), for completeness of presentation we outline the proof in Appendix A.

Biernacki and Govaert (1997) and Biernacki, Celeux and Govaert (2000) suggested two model selection criteria in the context of mixture models, named as the classification likelihood information criterion (CLC) and the integrated classification likelihood (ICL) criterion respectively. Their aim is to (i) to detect separated mixture components and (ii) to be more robust than BIC (or AIC) in case of misspecified model for detecting the number of components. This paper defines the BIC, Akaike’s Information Criterion (AIC), CLC and ICL as 
\[
\text{BIC} = l_{N}(\theta) - (\hat{q} / 2) \ln N, \quad \text{AIC} = l_{N}(\theta) - \hat{q}, \quad \text{CLC} = l_{N}(\theta) + \sum_{k=1}^{K} \sum_{n=1}^{N} \tau_{kn} \ln(\tau_{kn}) \quad \text{and} \quad \text{ICL} = \text{CLC} - (\hat{q} / 2) \ln N,
\]
where \( \hat{q} \) is the number of free parameters to be estimated and \( \tau_{kn} = E(z_{kn}) \) evaluated at the estimate, where \( z_{kn} \) is defined in Section 4.1 below. We compare the performance of these four criteria for correctly identifying \( K \). because only the penalty term in the BIC satisfies the conditions on the rate of growth of \( b_{KN} \) in Theorem 2. Therefore, we expect the BIC to perform best in the first step of our procedure. After finding \( K \) we then compare the performance of the BIC and AIC for correctly identifying regressors.

4. ESTIMATION

The EM algorithm has been a dominant technique for the estimation of mixture models in the literature (e.g. Dempster, Laird and Rubin 1977; McLachlan and Krishnan 1997; Arcidiacono and Jones 2003), where the missing information is assumed for a categorical
variable representing observations on mixture components. In our model, we further treat the unobserved information on \( y^* \) when \( y^* \leq 0 \) as missing information. This allows us to maximise the exploitation of the property of the monotonic increase in the EM iterations and the simplicity of the quadratic expression of the log-likelihood function of the Gaussian components. In relation to the estimation of the covariance matrix of our estimator, we found that Louis’ method is preferred to the direct evaluation of the Fisher’s information matrix, and the method suggested by McLachlan and Peel (2000, pp. 64-66). The latter two are more likely to result in numerical underflows and overflows.

4.1. EM Algorithm

Let \( \{z_{kn}\} \) be a \( K \times N \) dimension random variables such that \( z_{kn} = 1 \) indicates \( y_n \) is explained by the \( k \)th explanatory mechanism; 0 otherwise. Let \( y_{n_1}^* \) be the pseudo-observed \( y \) value when \( y = 0 \). Following Meng and Rubin (1991), we reparameterize the variance parameter by \( \gamma_k = \ln(\sigma_k^2) \) to give a better normal approximation to the log-likelihood function. The complete data log-likelihood function is

\[
l_c(\theta) = -\frac{N \ln(2\pie)}{2} + \sum_{k=1}^{K} \sum_{n}^{N_k} \left\{ \ln \alpha_k - \frac{\gamma_k}{2} - \frac{(y_{n_k} - \beta_k'x_{kn_k})^2 \exp(-\gamma_k)}{2} \right\} \\
+ \sum_{n_2}^{N} \left\{ \ln \alpha_k - \frac{\gamma_k}{2} - \frac{(y_{n_2}^* - \beta_k'x_{kn_2})^2 \exp(-\gamma_k)}{2} \right\}
\]  

(3)

where \( n_1 \) indicates when \( y > 0 \) is observed; \( n_2 \) indicates when \( y = 0 \), and \( N_1 + N_2 = N \).

**E-step:**

The E-step at the \( (h+1) \)th iteration requires taking the expectation of \( l_c(\theta) \) conditional on the observed \( y \), using the current fit value \( \theta^{(h)} \) for \( \theta \). The expectation related to \( z_{kn} \) and \( y_{n_1}^* \) can be calculated as follows.
\[ E(z_{mk}^{(h)} | y = 0) = \frac{\alpha_k^{(h)} \{1 - \Phi_k^{(h)}(\beta_k^{(h)} x_{kn2})\}}{\sum_{k=1}^{K} \alpha_k^{(h)} \{1 - \Phi_k^{(h)}(\beta_k^{(h)} x_{kn2})\}}, \]

\[ E(z_{kn}^{(h)} | y > 0) = \frac{\alpha_k^{(h)} \phi_k^{(h)}(y_n - \beta_k^{(h)} x_{kn})}{\sum_{k=1}^{K} \alpha_k^{(h)} \phi_k^{(h)}(y_n - \beta_k^{(h)} x_{kn})}, \]

\[ E(y_{n2}^{(h)} | k) = \beta_k^{(h)} x_{kn2} - \exp\left(\frac{r_k^{(h)}}{2}\right) \frac{\overline{\phi}(\lambda_{kn2}^{(h)})}{\Phi(\lambda_{kn2}^{(h)})}, \]

\[ E\{y_{n2}^{(h)} | k\} = \exp(r_k^{(h)}) + (\beta_k^{(h)} x_{kn2})^2 - \exp\left(\frac{r_k^{(h)}}{2}\right)(\beta_k^{(h)} x_{kn2}) \frac{\overline{\phi}(\lambda_{kn2}^{(h)})}{\Phi(\lambda_{kn2}^{(h)})}, \]

where \( \lambda_{kn2} = -\beta_k^{(h)} x_{kn2} \exp(r_k / 2) \), and \( \overline{\phi} \) and \( \Phi \) are the pdf and cdf of the standard normal distribution.

**M-step:**

The results of the maximisation are as follows.

\[ \alpha_k^{(h+1)} = \frac{\sum_{n1} E(z_{kn1}^{(h)}) + \sum_{n2} E(z_{kn2}^{(h)})}{N}, \]

\[ \beta_k^{(h+1)} = \left\{ \sum_{n1} \{E(z_{kn1}^{(h)}) x_{kn1} x_{kn1}'\} + \sum_{n2} \{E(z_{kn2}^{(h)}) x_{kn2} x_{kn2}'\} \right\}^{-1} \]

\[ \{ \sum_{n1} \{E(z_{kn1}^{(h)}) y_n1 x_{kn1}\} + \sum_{n2} \{E(z_{kn2}^{(h)}) y_{n2} x_{kn2}\} \}, \]

\[ r_k^{(h+1)} = \ln \frac{\sum_{n1} \{E(z_{kn1}^{(h)}) (y_n1 - \beta_k^{(h+1)} x_{kn1})^2 \} + \sum_{n2} \{E(z_{kn2}^{(h)}) E(y_{n2} - \beta_k^{(h+1)} x_{kn2})^2 \}}{\sum_{n1} E(z_{kn1}^{(h)}) + \sum_{n2} E(z_{kn2}^{(h)})}. \]

The iterations continue until the difference between the estimates obtained in two consecutive runs is less than an allowable error. The process then repeats at different starting values of the parameter, and the largest local maximum found is the global maximum.

However, there are a number of issues to consider. First, as the algorithm is computationally
intensive, we suggest that one begins searching for local maxima with a large allowable error (say 0.01). When the global maximum is found, the allowable error is then restored to a required value for computation. Second, when judging for a global maximum, one would need to be very cautious if taking a global maximum from a point with very small values of the estimate of $\sigma_k^2$ or very large negative values of $r_k$. Such global maximum may not be genuine because of the unboundness property of the ML function of mixture models (Day 1969; Quandt and Ramsey 1978; Kiefer 1978), and spurious correlation between the dependent variable and regressors in a very small proportion of the sample (Hathaway 1985, Sec. 1).

In our applications reported in Section 5, we take the starting value $\alpha_k = 1/K$, and generate the starting values for $\vartheta_k = (\beta_k', \sigma_k^2)' \forall k$ through $\vartheta_k = \vartheta_k^* + u_k$, where $\vartheta_k^*$ is the ordinary least squares estimate of the regression $y_n = \beta_k x_{kn} + e_{kn}$ and $u_k$ is generated from a centred normal distribution. As the convergence can be extremely slow near the boundary of the parameter space, one could speed up the searching by bounding the solution for $\alpha_k$, $\forall k$ away from 0 (say 0.0001).

4.2. Covariance Matrix of the Estimator

We have studied three methods for estimating the covariance matrix of the estimator: inverting the observed Hessian matrix, the method suggested by McLachlan and Peel (2000), which expresses the observed information matrix in terms of the conditional expectation of the gradient vector of the complete data log-likelihood function and Louis’ (1982) method. We found that Louis’ method is the preferred, as the other two often result in numerical underflows and overflows.

In Louis’ method the observed information matrix is expressed as $I = I_c - I_m$, where $I_c = E(-\partial^2 l_c / \partial \theta^2 | \theta, y, y^*, z)$ is the complete information matrix and
\[ I_M = Var(\frac{\partial l_c}{\partial \theta}|_\theta, y, y^*, z) \] is the missing information matrix. The formulae for computing \( I_c \) and \( I_M \) are provided in Appendix B.

5. APPLICATIONS

5.1. Simulation Studies

We have conducted simulation studies to assess the suggested procedure for model identification. The results appear to suggest that the proposed procedure works well, particularly when the component distributions are reasonably separated. For clarity, we only report the results of the two-component FMRMs in Figure 1. The results for FMRMs with three or more components, which are not reported here, suggest similar findings.

Consider the model,

\[
y^*_n = 0.5 + x_{1n} + e_{1n}, \text{ with probability } \alpha,
\]
\[
y^*_n = \beta_0 + \beta_1 x_{2n} + e_{2n}, \text{ with probability } 1 - \alpha,
\]

where \( x_{1n} = 5D_z \) and \( x_{2n} = 5D_u \), \( D_z \) and \( D_u \) are sampled from a standard normal distribution and a standard uniform distribution, respectively, and \( e_{kn} \sim N(0, \sigma^2), k = 1, 2 \). Let
\[
y_n = \max(y^*_n, 0).
\]

Choose \( (\beta_0, \beta_1) \) to be \((-2.5, 1)\) and \((-7.5, 3)\), so that the distance between the two regression lines increases as \( \beta_1 \) increases. Note that \( \beta_0 \) is chosen to ensure a proper balance between the zero and non-zero observations of \( y \). Figure 1 reports the simulation results under \( \alpha = 0.1, 0.5 \) and \( \sigma = 2, 1 \) for sample sizes \( N = 100, 300, 500, 1000 \).

With each regression line including both regressors \( x_{1n} \) and \( x_{2n} \), we first estimate the percentage of correctly identifying \( K \) in 1000 replications. The appended letter “c” in the graphs represents the performance of each model selection criterion in identifying \( K \). The results suggest that the BIC shows far superior performance than the others. Occasionally the
AIC performs slightly better than the BIC when $N$ is small. But the AIC seems to do a poor job when $N$ is large.

We then estimate the percentage of correctly identifying the final model. Note that we consider only those replications in which $K$ has been correctly identified using the BIC, for model identification in the second step. This is because the BIC has been found to outperform the others in the first step and in fact the models selected by the BIC almost include the models selected by the other criteria during the iteration process. We report the results for each of the four model selection criteria with the appended letter “f”. Again the BIC dominates the others as suggested by the proportion of its correct identification approaching 1 faster than the others as $N$ increases. We also observe that the performance of the BIC in the first step for identifying $K$ is better than that in the second step for identifying regressors after $K$ has been identified.

When $\alpha$ is small, the finite sample performance may be poor. A possible explanation for this is the lack of observations for the associated mixture component. Increases in $N$ and/or in the separation of component distributions through the separation in the component regression lines or a decrease in the variance of component distributions, results in the percentage of correctly identifying the true model with the use of the BIC increasing toward 1.

5.2. Empirical Application: Extramarital Affairs

The first empirical example we are interested in is a textbook example of a censoring regression model (Greene 2003, sec. 22.3.6), which originally appeared in Fair (1977, 1978). The data is a sample of 601 observations on men and women, who were then currently married for the first time, in response to a survey question about extramarital affairs. The dependent variable is the number of affairs an individual had in the past year. The regressors are ‘gender dummy’ ($\bar{z}_1 = 1$ for male; 0 for female), ‘individual age’ ($\bar{z}_2$), ‘number of years
married’ ($\tilde{z}_1$), ‘children dummy’ ($\tilde{z}_4 = 1$ for individual having children; 0 otherwise),
‘religiousness’ ($\tilde{z}_5 = 1$, anti, …, = 5 very), ‘years of education’ ($\tilde{z}_6$), ‘occupation’ ($\tilde{z}_7$,
measured in Hollingshead scale), and ‘self-rating of marriage’ ($\tilde{z}_8 = 1$, very unhappy, …, =
5, very happy).

We start with identifying the number of mixture components, each of which includes all
the regressors above, by using the BIC. The results reported in Table 1 suggest that the two-
component FMRM provides a slightly better fit than others. Note that FMRM reduces to the
Tobit model (Tobin 1958) when $K = 1$. The next step is to identify regressors by using the
BIC again from the two-component FMRM identified above. The estimate and its standard
error of the final best fit model are reported in Table 2. The best fit Tobit model includes the
regressors $\tilde{z}_2, \tilde{z}_3, \tilde{z}_5, \tilde{z}_7$ and $\tilde{z}_8$ with the BIC value being -705.576 (Greene 2003, p. 776).
Our best fit FMRM improves the BIC value by about 19 and suggests that the factors
‘religiousness’ ($\tilde{z}_5$) and ‘self-rating of marriage’ ($\tilde{z}_8$) alone accounted for about 84% of
individuals’ outcome of extramarital affairs, while all covariates significantly contribute to
the outcome of extramarital affairs for the remaining 16% of individuals.

Table 1. The estimated log-likelihood and BIC value for the extramarital affairs data

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log-likelihood</td>
<td>-704.731</td>
<td>-669.285</td>
<td>-635.246</td>
<td>-614.102</td>
</tr>
<tr>
<td>BIC</td>
<td>-736.724</td>
<td>-736.470</td>
<td>-737.624</td>
<td>-751.672</td>
</tr>
</tbody>
</table>

NOTE: the mixture models presented in this table possess a same set of regressors in each component.
Table 2. The MLE and standard error (in bracket) of the best fit model for the extramarital affairs data

<table>
<thead>
<tr>
<th></th>
<th>Order of the components</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>0.158 (0.019)</td>
</tr>
<tr>
<td>$\ln(\sigma^2)$</td>
<td>-2.321 (0.234)</td>
</tr>
<tr>
<td>Const dummy</td>
<td>-34.212 (0.550)</td>
</tr>
<tr>
<td>$z_1$</td>
<td>-8.414 (0.138)</td>
</tr>
<tr>
<td>$z_2$</td>
<td>-0.237 (0.010)</td>
</tr>
<tr>
<td>$z_3$</td>
<td>1.770 (0.018)</td>
</tr>
<tr>
<td>$z_4$</td>
<td>14.023 (0.162)</td>
</tr>
<tr>
<td>$z_5$</td>
<td>-2.858 (0.048)</td>
</tr>
<tr>
<td>$z_6$</td>
<td>2.793 (0.030)</td>
</tr>
<tr>
<td>$z_7$</td>
<td>-1.244 (0.039)</td>
</tr>
<tr>
<td>$z_8$</td>
<td>-5.404 (0.042)</td>
</tr>
<tr>
<td>Log-likelihood</td>
<td>-676.492</td>
</tr>
<tr>
<td>BIC</td>
<td>-724.482</td>
</tr>
</tbody>
</table>

5.3. Empirical Application: Individual Labour Supply

The Tobit model is commonly used to study individual labour supply (see e.g. Amemiya 1981 or Greene 2003, chap. 22). The example we are interested in here is the study of the female labour supply in the UK. More detailed descriptions of the data set can be found in Blundell and Smith (1989). The data are a sample of 2539 observations drawn from the UK Family Expenditure Survey for 1981.

The dependent variable is the weekly hours an individual female participated in the labour force. There were 1460 working women ($y > 0$) and 1079 non-working women ($y = 0$). The regressors are ‘other household income’ ($Income$), ‘female age’ ($Age$), ‘female age squared’ ($Age-sq$), ‘female education’ ($Ed$), ‘female education squared’ ($Ed-sq$) and child dummy variables ($D1$, $D2$, and $D3$) representing the presence of pre-school children, children of ages 5-10, and children of age 11 and above, respectively. $Income$ is measured
weekly in English pounds and contains the husband’s income, unearned income, and dissaving. *Age* is the actual individual female’s age minus 40, then divided by 10. *Ed* is the number of years of full-time education, minus 8. The dummy variables *Da* = 1, for *a* = 1, 2, 3, if the number of children in the particular category of children’s ages is more than zero, otherwise *Da* = 0.

The results reported in Table 3 suggest that the 5-component FMRM is identified as the best fit model in the first step of our procedure. The final best fit model is reported in Table 4. It is worth noting that the final model further improves the BIC value by about 179 compared to the 5-component FMRM reported in Table 3, which already has an approximate increase of 580 in the BIC value from the Tobit model, the estimation results of which are not reported here.

### Table 3. The estimated log-likelihood and BIC value for the labour supply data

<table>
<thead>
<tr>
<th>K</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log-likelihood</td>
<td>-6372.149</td>
<td>-6295.925</td>
<td>-6250.450</td>
<td>-6228.026</td>
</tr>
<tr>
<td>BIC</td>
<td>-6497.582</td>
<td>-6464.475</td>
<td>-6462.118</td>
<td>-6482.811</td>
</tr>
</tbody>
</table>

*NOTE:* the mixture models presented in this table possess a same set of regressors in each component.

There are 5 distinct groups identified. ‘*Income*’, ‘*Ed*’ and ‘*D3*’ were the three driving factors for individual female labour supply for the largest group of about 40% of women. Interestingly, there was a group of individuals, about 17% of women, for whom the status of having a pre-school child (*D1*) was the sole factor driving their labour participating decision, while another group, about 12% of women, for whom ‘*Age-sq*’ was the sole contributing factor explaining their decision. In the latter group, the older and younger women were more likely to take part in the labour force and work longer than women around 40. In the remaining two groups, ‘*D1*’ and ‘*D2*’ played roles in the individual labour participation.
decision making process, but contributed differently. In a group accounting for about 24% of women, they affected the dependent variable with ‘Income’ and ‘Age’. In a group accounting for 7% of women, they affected the dependent variable along the regressors $Ed$ and $Ed$-sq, the level of individual education attainment, which exhibited a nonlinear effect.

Table 4. The MLE and standard error (in bracket) of the best fit model for the labour supply data

<table>
<thead>
<tr>
<th>Order of the components</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>0.075</td>
<td>0.122</td>
<td>0.166</td>
<td>0.239</td>
<td>0.398</td>
</tr>
<tr>
<td>(0.006)</td>
<td>(0.009)</td>
<td>(0.009)</td>
<td>(0.010)</td>
<td>(0.011)</td>
<td></td>
</tr>
<tr>
<td>$\ln(\sigma^2)$</td>
<td>-4.413</td>
<td>4.801</td>
<td>0.836</td>
<td>3.563</td>
<td>4.928</td>
</tr>
<tr>
<td>(0.127)</td>
<td>(0.133)</td>
<td>(0.091)</td>
<td>(0.085)</td>
<td>(0.063)</td>
<td></td>
</tr>
<tr>
<td>$Const_{dummy}$</td>
<td>39.387</td>
<td>17.081</td>
<td>37.430</td>
<td>31.524</td>
<td>29.028</td>
</tr>
<tr>
<td>(0.019)</td>
<td>(1.614)</td>
<td>(0.104)</td>
<td>(0.656)</td>
<td>(0.927)</td>
<td></td>
</tr>
<tr>
<td>$Income$</td>
<td>-0.120</td>
<td>-0.895</td>
<td></td>
<td>(0.010)</td>
<td>(0.012)</td>
</tr>
<tr>
<td>(0.010)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Age$</td>
<td></td>
<td></td>
<td></td>
<td>-2.810</td>
<td></td>
</tr>
<tr>
<td>(0.428)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Age$-sq</td>
<td></td>
<td></td>
<td>3.502</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.946)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Ed$</td>
<td>0.739</td>
<td>1.291</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.007)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$Ed$-sq</td>
<td>-0.198</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.001)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$D1$</td>
<td>-53.827</td>
<td>-55.307</td>
<td>-62.764</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.051)</td>
<td>(0.323)</td>
<td>(1.153)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$D2$</td>
<td>-20.005</td>
<td></td>
<td>-11.700</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(0.022)</td>
<td></td>
<td>(0.906)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$D3$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>12.376</td>
</tr>
<tr>
<td>(1.222)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log-likelihood</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-6177.301</td>
</tr>
<tr>
<td>BIC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-6283.134</td>
</tr>
</tbody>
</table>
6. CONCLUDING REMARKS

Although researchers may find it easy to justify the assumption of non-same-set of regressors across regression regimes, the statistical inference issues that arise from this can be difficult. The number of candidate models with \( q \) regressors and \( K \) components is

\[
\binom{2^q}{K} + K,
\]

which expands rapidly as \( q \) or \( K \) increases. This is a difficult enough situation when \( K \) is known, it is much more difficult if \( K \) is unknown. This paper studies a two-step model selection procedure in a mixture regression model with the censored dependent variable. Although our investigation is illustrated with a particular reference to a fixed censoring point at zero, our results should be readily extended to the situations of the non-zero fixed censoring or non-censoring dependent variable. Our model can also be extended to the situation where we have random censoring such as in duration models. In such applications, care should be taken with respect to the conditional expected value in relation to the unobservable part of \( y^* \) in the E-step of the EM algorithm and the estimation of the covariance matrix. The findings from our studies of two microeconometric examples suggest that failure to account for population heterogeneity can result in severe consequence of model misspecification, hence lead to very different interpretations of social and economic phenomena.

ACKNOWLEDGMENTS

This research was supported in part by the Australian Research Council's Discovery project grant DP0666677. The author thanks Azhong Ye and Brenton Dansie for helpful discussions. Comments from participants in the seminars in the school of economics, the University of Adelaide, the department of econometrics and business statistics, Monash University, the Australasian Meeting of the Econometric Society, 2006 and the Far Eastern Meeting of Econometric Society, 2006, are gratefully acknowledged.
We begin with making the following assumptions.

**Assumption A.1.** The parameter space $\Theta$ is compact.

**Assumption A.2.** $f_k(y,x;\theta)$, $(x,y) \in \Omega$, is integrable with respect to a $\sigma$-finite measure $\mu$ over $(x,y)$ for each $\theta \in \Theta$.

**Assumption A.3.** $f_k(y,x;\theta_0) = f_k(y,x;\hat{\theta}_0)$ for all $\theta_0 \neq \hat{\theta}_0$ and $\theta_0, \hat{\theta}_0 \in \Theta$, where $\Theta$ is the complement of $\Theta_0$.

**Assumption A.4.** $f_k(y,x;\theta) \neq f_k(y,x;\hat{\theta})$ for all $\theta \neq \hat{\theta}$ and $\theta, \hat{\theta} \in \Theta_0$, where $\Theta_0$ is the complement of $\Theta$.

**Assumption A.5.** $f_k(y,x;\theta)$ is a.s. continuous in $\theta$. When $\theta \in \Theta_0$, the continuity holds at least from an appropriate direction.

**Assumption A.6.** Let the covariate space be $X_K \subset R^{\sum_{k=0}^K d_k}$, then $X_{K_0} \subset X_K$, for $K \geq K_0$.

**Remark:** Assumptions A.1-A.5 are the modifications of the assumptions stated in Feng and McCulloch (1996) to accommodate the regression setting. Assumption A.6 is to ensure that the true regressors are included in the mixture components.

Proof of Theorem 1.

Let $p(x)$, $x \in X_K$, is the pdf over $x$, then $f_k(y,x;\theta) = f_k(y|x;\theta)p(x)$. Assumptions A1-A5 are clearly satisfied for the model (1). The sample log-likelihood function of our model is

$$l_N(\theta) = \sum_{n=0}^N \ln \left[ \sum_{k=1}^K \alpha_k \{1 - \Phi(\beta'_k x_{n0})\} \right] + \sum_{n>0} \ln \left[ \sum_{k=1}^K \alpha_k \phi(y_n - \beta'_k x_{n0}) \right].$$

Define the random variable $v$ such that $v = 1$ if $y = 0$; $v = 0$ otherwise, and

$L_N = N^{-1}l_N(\theta)$. Then we have

$$L_N = N^{-1} \sum_{n=1}^N v_n \ln \left[ \sum_{k=1}^K \alpha_k \{1 - \Phi(\beta'_k x_{n0})\} \right] + N^{-1} \sum_{n=1}^N (1-v_n) \ln \left[ \sum_{k=1}^K \alpha_k \phi(y_n - \beta'_k x_{n0}) \right].$

Further we define the pdf
\[ q(y, x, v; \theta) = g(y|x, v; \theta)b(v|x; \theta)p(x), \]

where

\[
g(y|x, v=1; \theta) = \begin{cases} 1 & \text{if } y = 0, \\ 0 & \text{if } y > 0, \end{cases}
\]

\[
g(y|x, v = 0; \theta) = \begin{cases} 0 & \text{if } y = 0, \\ \sum_{k=1}^{K} \alpha_k \Phi_k (y - \beta_k' x_k) & \text{if } y > 0, \end{cases}
\]

\[
b(v|x; \theta) = \begin{cases} \sum_{k=1}^{K} \alpha_k \{1 - \Phi_k (\beta_k' x_k)\}, & \text{if } v = 1, \\ \sum_{k=1}^{K} \alpha_k \Phi_k (\beta_k' x_k), & \text{if } v = 0. \end{cases}
\]

After applying the results of Theorem 1 of Feng and McCulloch (1996), we have

\[
L_N \xrightarrow{a.s.} Q(\theta) = \int \int p(x)dx\int \int g(y|x, v; \theta)b(v|x; \theta)\ln[g(y|x, v; \theta)b(v|x; \theta)]dvdy \\
= \int \int g(y|x, v; \theta)b(v|x; \theta)\ln[g(y|x, v; \theta)b(v|x; \theta)]dvdy,
\]

for all \( \theta \in \Theta \), where \( v \in V \) and \( V \) is the set containing only 0 and 1, hence

\[
\hat{\theta} - \hat{\theta}_0(\hat{\theta}) \xrightarrow{a.s.} 0.
\]

In the situation that a regressor is incorrectly included in a component, the above result still holds by noting the true regression coefficient associated with the regressor being zero.

Proof of Theorem 2.

Based on the results in Theorem 1, the proof of Theorem 2 follows easily Leroux (1992) and Keribin (2000). The proof is outlined as follows.

If \( K < K_0 \), the true model is not nested in the model and

\[
L_N(f_{K_0}) - L_N(f_K) \xrightarrow{a.s.} D(f_{K_0}, f_K),
\]
where $D(f_{\theta}, f_K)$ is the Kullback-Leibler measure of the divergence of $f_K$ from $f_{\theta}$ and is positive. Leroux (1992) and Keribin (2000) showed that $K$ is not underestimated, i.e.

$\hat{K} \geq K_0$ a.s..

If $K \geq K_0$, we have shown the strong consistency of our estimator in Theorem 1. Then by Lemma 3.3 and Proposition 4.2 of Keribin (2000), $\{l_N(\hat{f}_K) - l_N(f_{K_0})\} / \ln N$ is bounded a.s.. Together with the conditions on $a_{KN}$, this gives the term

$$\frac{l_N(\hat{f}_K) - l_N(f_{K_0}) \ln \ln N}{\ln N} - \frac{b_{KN}}{b_{K_0N}} + 1,$$

tends to a strictly negative term a.s.. Because the probability for the estimate $\hat{K}$ being greater than its true value $K_0$ can be expressed as

$$\Pr(\hat{K} > K_0) \leq \sum_{K=K_0+1}^{\infty} \Pr(l_N(\hat{f}_K) - a_{KN} > l_N(f_{K_0}) - b_{K_0N})$$

$$= \sum_{K=K_0+1}^{\infty} \Pr(\frac{l_N(\hat{f}_K) - l_N(f_{K_0})}{b_{KN} b_{K_0N}} + 1 > 0)$$

$$= \sum_{K=K_0+1}^{\infty} \Pr(\frac{l_N(\hat{f}_K) - l_N(f_{K_0}) \ln \ln N}{b_{KN} b_{K_0N}} + 1 > 0)$$

Where $K$ is the known upper bound of $K$, we have $\lim_{N \to \infty} \Pr(\hat{K} > K_0) = 0$, which implies

$\hat{K} \xrightarrow{a.s.} K_0$.

Now consider MPLE. Because

$$\Pr(\hat{\theta}_K - \theta_{0, K_0} | (\hat{K}) > \delta) \leq \Pr(\hat{K} \neq K_0) + \Pr(\hat{\theta}_K - \theta_{0, K_0} | (\hat{K}) > \delta),$$

$\Pr(\hat{K} \neq K_0)$ tends to 0 as $\hat{K} \xrightarrow{a.s.} K_0$. If the penalty term $a_{KN}$ does not involve parameter $\theta$, then Theorem 1 also suggests MPLE $\hat{\theta}_{K_0} \xrightarrow{a.s.} \theta_{0, K_0}$. Therefore $\hat{\theta}_K - \theta_{0, K_0} (\hat{K}) \xrightarrow{a.s.} 0$.

**APPENDIX B: EVALUATION OF THE COVARIANCE MATRIX**

20
We present the formulae for evaluating the non-zero elements in $I_c$ and $I_m$, which will be evaluated at the estimate $\hat{\theta}$. Define $\phi_{kn} = \phi_{kn}/\Phi_{kn}$. Note that the third and fourth moments of $y_n^*$ about $\beta_k'x_{kn}$ have also been used in the process of derivation, i.e.,

$$E(y_n^* - \beta_k'x_{kn})^3 = \sigma_k^2 o_{kn}\{(\beta_k'x_{kn})^2 - 2\sigma_k^2\},$$

and

$$E(y_n^* - \beta_k'x_{kn})^4 = \sigma_k^2 \{3\sigma_k^2 + 3\sigma_k^2 o_{kn}(\beta_k'x_{kn}) + o_{kn}(\beta_k'x_{kn})^3\}.$$

For evaluating $I_c$, we have,

$$E - \partial^2 l_c / \partial \alpha_k^2 = \sum_{n=1}^{N} \tau_{kn} / \alpha_k^2 + \tau_{kn} / \alpha_k^2,$$

$$E - \partial^2 l_c / \partial \alpha_k \partial \gamma_k = (1/2) \{ \sum_{n=1}^{N} \tau_{kn} / \alpha_k^2 \},$$

$$E - \partial^2 l_c / \partial \beta_k \partial \gamma_k = -\sum_{n=1}^{N} \tau_{kn} o_{kn} x_{kn}.$$ 

For evaluating $I_m$, we have,

$$E(\partial l_c / \partial \alpha_k) = \sum_{n=1}^{N} \tau_{kn} (1 - \tau_{kn}) / \alpha_k^2 + \tau_{kn} (1 - \tau_{kn}) / \alpha_k^2 - 2 \tau_{kn} \tau_{kn} / \alpha_k^2 \alpha_k,$$

$$E \partial l_c \partial l_c / \partial \alpha_k \partial \alpha_k = \sum_{n=1}^{N} \frac{\tau_{kn} \tau_{mn} - \tau_{kn} \tau_{mn} \tau_{kn} - \tau_{mn} \tau_{mn} + \tau_{kn} (1 - \tau_{kn})}{\alpha_k^2 \alpha_k^2},$$

$$E \partial l_c \partial l_c / \partial \beta_k \partial \beta_k' = \exp(-\gamma_k) \{ \sum_{n=1}^{N} \tau_{kn} (1 - \tau_{kn}) x_{kn} x_{kn} + \sum_{n=1}^{N} \tau_{kn} (1 - \tau_{kn}) (1 + \beta_k'x_{kn} o_{kn}) x_{kn} x_{kn} \},$$

$$E \partial l_c \partial l_c / \partial \beta_k \partial \beta_m' = \sum_{n=1}^{N} \tau_{kn} \tau_{mn} o_{kn} o_{mn} x_{kn} x_{mn},$$

where $m = 1,\ldots,K$ and $k \neq m$.
\[ E\left( \frac{\partial l}{\partial \gamma_k} \right)^2 = 2^{-1} \sum_{n_2=1}^{N_2} \tau_{kn_2} (1-\tau_{kn_2}) + 4^{-1} \sum_{n_2=1}^{N_2} \tau_{kn_2} (1-\tau_{kn_2}) \left\{2 + o_{kn_2} \left( \beta'_k x_{kn_2} \right) + o_{kn_2} \left( \beta'_k x_{mn_2} \right)^3 \exp(-\gamma_k') \right\}, \]

\[ E\left( \frac{\partial l}{\partial \gamma_k} \right) \frac{\partial l}{\partial \gamma_m} = 4^{-1} \sum_{n_2=1}^{N_2} \tau_{kn_2} \tau_{mn_2} o_{kn_2} o_{mn_2} \left( \beta'_k x_{kn_2} \right) \left( \beta'_m x_{mn_2} \right), \]

\[ E\left( \frac{\partial l}{\partial \alpha_k} \right) \frac{\partial l}{\partial \beta'_k} = \sum_{n_2=1}^{N_2} \left( \frac{\tau_{kn_2} \tau_{Km_2}}{\alpha_K} - \frac{\tau_{kn_2} (1-\tau_{kn_2})}{\alpha_k} \right) o_{mn_2} x_{kn_2}, \]

\[ E\left( \frac{\partial l}{\partial \alpha_k} \right) \frac{\partial l}{\partial \beta'_m} = \sum_{n_2=1}^{N_2} \left( \frac{\tau_{mn_2} \tau_{Kn_2}}{\alpha_K} - \frac{\tau_{mn_2} \tau_{Km_2}}{\alpha_k} \right) o_{mn_2} x_{mn_2}, \]

\[ E\left( \frac{\partial l}{\partial \gamma_k} \right) \frac{\partial l}{\partial \gamma_m} = 2^{-1} \sum_{n_2=1}^{N_2} \left( \frac{\tau_{kn_2} (1-\tau_{kn_2})}{\alpha_k} - \frac{\tau_{kn_2} \tau_{Kn_2}}{\alpha_K} \right) o_{mn_2} \left( \beta'_k x_{kn_2} \right), \]

\[ E\left( \frac{\partial l}{\partial \gamma_k} \right) \frac{\partial l}{\partial \gamma_m} = 2^{-1} \sum_{n_2=1}^{N_2} \left( \frac{\tau_{mn_2} \tau_{Kn_2}}{\alpha_K} - \frac{\tau_{mn_2} \tau_{Km_2}}{\alpha_k} \right) o_{mn_2} \left( \beta'_m x_{mn_2} \right), \]

\[ E\left( \frac{\partial l}{\partial \beta'_k} \right) \frac{\partial l}{\partial \gamma_k} = 2^{-1} \sum_{n_2=1}^{N_2} \tau_{kn_2} (1-\tau_{kn_2}) o_{kn_2} \left( (\beta'_k x_{kn_2})^2 \exp(-\gamma_k') - 1 \right) x_{kn_1}, \]

\[ E\left( \frac{\partial l}{\partial \beta'_m} \right) \frac{\partial l}{\partial \gamma_m} = -2^{-1} \sum_{n_2=1}^{N_2} \tau_{kn_2} \tau_{mn_2} o_{kn_2} x_{kn_2} \left( \beta'_m x_{mn_2} \right) o_{mn_2}. \]

References


Figure 1. Percentage of correctly identifying $K$ (with the appended letter “c”) and the final model (with the appended letter “f”) under different model selection criteria as a function of sample size.

\[
\beta_0 = -2.5, \beta_1 = 1, \sigma = 2, \alpha = 0.1
\]

\[
\beta_0 = -2.5, \beta_1 = 1, \sigma = 1, \alpha = 0.1
\]
$\beta_0 = -7.5, \beta_1 = 3, \sigma = 2, \alpha = 0.1$

$\beta_0 = -7.5, \beta_1 = 3, \sigma = 1, \alpha = 0.1$

$\beta_0 = -2.5, \beta_1 = 1, \sigma = 2, \alpha = 0.5$
$\beta_0 = -2.5, \beta_1 = 1, \sigma = 1, \alpha = 0.5$

$\beta_0 = -7.5, \beta_1 = 3, \sigma = 2, \alpha = 0.5$

$\beta_0 = -7.5, \beta_1 = 3, \sigma = 1, \alpha = 0.5$