KATHOLIEKE UNIVERSITEIT
LEUVEN

Faculty of Business and Economics

Order selection in finite mixtures of linear regressions

Nicolas Depraetere and Martina Vandebroek

DEPARTMENT OF DECISION SCIENCES AND INFORMATION MANAGEMENT (KBI)

KBI 1129

Order Selection in Finite Mixtures of Linear Regressions

Nicolas Depraetere[a,*], Martina Vandebroek[b]

[a] *Katholieke Universiteit Leuven, Faculty of Business and Economics, Naamsestraat 69, 3000 Leuven, Belgium*
[b] *Katholieke Universiteit Leuven, Faculty of Business and Economics, Naamsestraat 69, 3000 Leuven, Belgium*
Leuven Statistics Research Center, W. de Croylaan 54, 3001 Leuven-Heverlee, Belgium

Abstract

Finite mixture models can adequately model population heterogeneity when this heterogeneity arises from a finite number of relatively homogeneous clusters. A good example of such a situation is modeling market segmentation. Order selection in mixture models, i.e. selecting the correct number of components in the mixture model, however, is a problem which has not been satisfactorily resolved. Existing simulation results in the literature do not completely agree with each other. Moreover, it appears that the performance of different proposed selection methods is affected by the type of model and the parameter values. Furthermore, most existing results are based on simulations where the true generating model is identical to one of the models in the candidate set. In order to partly fill this gap we carried out a simulation study for finite mixture models of normal linear regressions. We included several types of model misspecification to study the robustness of 18 order selection methods. Furthermore, we compared the performance of these selection methods based on unpenalized and penalized estimates of the model parameters. The results indicate that order selection based on penalized estimates greatly improves the success rates of all order selection methods. The most successful methods were MRC, MRC_k, $MDL2$, ICL and $ICL\text{-}BIC$ but not one method was consistently good or best for all types of model misspecification.

Keywords: Finite mixture, Regression, Penalized likelihood, Order selection

1. Introduction

Finite mixtures present a very attractive modeling framework to increase model flexibility without the high-dimensional parameter spaces used in non-parametric or mixed modeling (Mclachlan and Peel 2000). Often, a regular statistical model is too rigid to adequately represent possible heterogeneity in the population. This heterogeneity can often be captured by a mixture of parametric models. Such mixtures have been successfully applied in a wide variety of fields. Wedel and Kamakura (1999) for instance, have spent two chapters of their book on market segmentation on this topic whereas Schlattmann (2009) has written an entire book about medical applications of finite mixture models. However, despite its popularity and frequent usage, there are still some complications with this type of model. The most important of these complications is that of selecting the correct number of components (Mclachlan and Peel 2000) which we will refer to as order selection. Not surprisingly, this has generated a lot of theoretical and applied research and

*Corresponding author
Email addresses: nicolas.depraetere@econ.kuleuven.be (Nicolas Depraetere), martina.vandebroek@econ.kuleuven.be (Martina Vandebroek)

many order selection methods have been suggested in the literature by now. However, most of the simulation results which have been presented either disagree with each other or were obtained in very idealized settings where model assumptions matched the simulation settings. Therefore, in this paper, we have investigated violations of standard model assumptions in finite mixtures of linear regression models, in the hope of partly filling this gap. We have compared several old and new order selection methods using two different types of estimation, unpenalized and penalized estimation. The rest of this paper is structured as follows. In section 2 some technical and practical background will be given about (fitting) a mixture model of linear regressions. In section 3 we present a non-exhaustive overview of various popular and some lesser known but rather effective methods to select the number of components in a mixture model. In this section we also give an overview of some published results. In section 4 the design and results of our simulation study will be presented and discussed.

2. Technical Background

2.1. Finite Mixtures of Linear Regressions

Suppose a population consists of K subpopulations S_k indexed by $k = 1, \cdots, K$. Within each of these subpopulations, suppose it makes sense to model a univariate[1] random variable Y as a linear combination of p explanatory variables denoted by the vector $\boldsymbol{x}$. Then, for a random sample of size n across the subpopulations, we have

$$\begin{cases} y_i = \beta_{01} + \beta_{11}x_{i1} + \cdots + \beta_{p1}x_{ip} + \epsilon_{i1} & \text{if} \quad y_i \in S_1 \\ \quad \vdots & \\ y_i = \beta_{0k} + \beta_{1k}x_{i1} + \cdots + \beta_{pk}x_{ip} + \epsilon_{ik} & \text{if} \quad y_i \in S_k \\ \quad \vdots & \\ y_i = \beta_{0K} + \beta_{1K}x_{i1} + \cdots + \beta_{pK}x_{ip} + \epsilon_{iK} & \text{if} \quad y_i \in S_K \end{cases} \tag{1}$$

where $i = 1, \cdots, n$. Note that the subpopulations are assumed to be mutually exclusive and exhaustive. The error terms within each component are assumed to be i.i.d. normal with mean 0 and variance σ_k^2 and independent across the subpopulations. The vector of regression coefficients will be denoted by $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_K)^T$ where $\boldsymbol{\beta}_k = (\beta_{0k}, \beta_{1k}, \cdots, \beta_{pk})^T$. Let $\boldsymbol{z}$ be a single trial realization of a random multinomial variable with parameter vector $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_K)^T$ which indicates from which subpopulation Y originates. Therefore, if an observation i belongs to component k, $\boldsymbol{z}_i$ is a vector of 0s with a 1 at the kth position. The parameters $\pi_k, k = 1, \cdots, K$, indicate the relative size of the subpopulations in the entire population under consideration. From its definition it follows that $\sum_{k=1}^{K} \pi_k = 1$ and that $\pi_k \geq 0, \forall k = 1, \cdots, K$. The joint distribution of y and $\boldsymbol{z}$, conditional on $\boldsymbol{x}$, can now be written as

$$f\left(y, \boldsymbol{z} | \boldsymbol{x}, \boldsymbol{\Psi}\right) = \prod_{k=1}^{K} \left[\pi_k \mathcal{N}\left(y | \boldsymbol{\beta}_k^T \boldsymbol{x}, \sigma_k^2\right)\right]^{z_{ik}} \tag{2}$$

[1] This can be readily extended to the multivariate case.

where $\mathcal{N}\left(y|\boldsymbol{\beta}_k^T\boldsymbol{x},\sigma_k^2\right)$ represents the normal distribution function of a variable y with mean $\boldsymbol{\beta}_k^T\boldsymbol{x}$ and variance σ_k^2, $\boldsymbol{x}$ includes an intercept term and $\boldsymbol{\Psi}=\left(\boldsymbol{\pi},\boldsymbol{\beta},\sigma_1^2,\cdots,\sigma_K^2\right)^T$ denotes the complete parameter vector. Note that one of the elements of $\boldsymbol{\pi}$ is redundant due to the summation restriction given above. The complete data log likelihood or joint log likelihood of y and $\boldsymbol{z}$ of the sample can then be expressed as

$$LL_c\left(\boldsymbol{\Psi}\right)=\sum_{i=1}^{n}\sum_{k=1}^{K}z_{ik}\{\log\left(\pi_k\right)+\log\left[\mathcal{N}\left(y_i|\boldsymbol{\beta}_k^T\boldsymbol{x}_i,\sigma_k^2\right)\right]\}. \tag{3}$$

A finite mixture model of linear regressions now arises when the subpopulation indicator variable $\boldsymbol{z}$ is not observed (or inherently unobservable). In this case, one has to resort to working with the marginal distribution of Y (marginalized over $\boldsymbol{Z}$) and the marginal distribution of y, conditional on x, becomes

$$f\left(y|\boldsymbol{x},\boldsymbol{\Psi}\right)=\sum_{k=1}^{K}\pi_k\mathcal{N}\left(y|\boldsymbol{\beta}_k^T\boldsymbol{x},\sigma_k^2\right) \tag{4}$$

and the corresponding observed log likelihood of the sample is

$$LL\left(\boldsymbol{\Psi}\right)=\sum_{i=1}^{n}\log\left[\sum_{k=1}^{K}\pi_k\mathcal{N}\left(y_i|\boldsymbol{\beta}_k^T\boldsymbol{x}_i,\sigma_k^2\right)\right]. \tag{5}$$

This model is the finite mixture model of normal linear regressions that was introduced by Desarbo and Cron (1988). The relative sizes of the subpopulations are called mixture proportions or mixture weights[2] and the densities in the subpopulations are called the component densities, which are conditional on component membership and the explanatory variables. Note that, in the absence of any other information, the mixture proportions are the a priori probabilities of belonging to a specific component for a randomly sampled subject.

Maximizing the observed log likelihood (5) can be done in a variety of ways (all iteratively as there is no closed-form solution) and is usually done by using the expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin 1977) which uses (3) rather than (5). Every iteration in the EM algorithm consists of two steps, an expectation step and a maximization step. In the expectation step the expected value of the complete data log likelihood (3), conditional on the vector of current parameter values and the observed data, is calculated. This expression is then subsequently maximized with respect to the model parameters in the maximization step, yielding a new set of parameter values. Dempster et al. (1977) showed that iterating these two steps is equivalent to maximizing the observed log likelihood, which is the goal. Calculating the conditional expected value of (3) is straightforward as the only random terms are the z_{ik} which are binary indicator variables and are linear in (3). So, for a general iteration $(t+1)$, the expectation step consists of calculating

[2]It is possible to generalize (4) by including explanatory variables to model the mixture proportions using a logistic regression model for instance. If these explanatory variables are different from the variables which model the component means they can be ignored for order selection as the marginal model is a mixture model with the same number of components (Bandeen-Roche, Miglioretti, Zeger, and Rathouz 1997).

$$E\left[Z_{ik}|y_i, \boldsymbol{x}_i, \boldsymbol{\Psi}^{(t)}\right] = \frac{\pi_k^{(t)} \mathcal{N}\left(y_i|\boldsymbol{\beta}_k^{T(t)}\boldsymbol{x}_i, \sigma_k^{2(t)}\right)}{\sum_{k=1}^{K} \pi_k^{(t)} \mathcal{N}\left(y_i|\boldsymbol{\beta}_k^{T(t)}\boldsymbol{x}_i, \sigma_k^{2(t)}\right)} \equiv \tau_{ik}^{(t+1)}. \tag{6}$$

The $\tau_{ik}^{(t+1)}$ can be viewed as the posterior probability of an observation with observed values y_i and $\boldsymbol{x}_i$ to belong to component k. Maximizing (3), with the z_{ik} replaced by the estimated $\tau_{ik}^{(t+1)}$, now yields the following closed-form solutions

$$\begin{aligned} \pi_k^{(t+1)} &= \frac{\sum_{i=1}^{n} \tau_{ik}^{(t+1)}}{n} \\ \boldsymbol{\beta}_k^{(t+1)} &= \left(\boldsymbol{X}^T \boldsymbol{W}_k^{(t+1)} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{W}_k^{(t+1)} \boldsymbol{y} \\ \sigma_k^{2(t+1)} &= \frac{\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_k^{(t+1)}\right)^T \boldsymbol{W}_k^{(t+1)} \left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}_k^{(t+1)}\right)}{\sum_{i=1}^{n} \tau_{ik}^{(t+1)}} \end{aligned} \tag{7}$$

where $\boldsymbol{X}$ is the $n \times (p+1)$ design matrix including an intercept column, $\boldsymbol{y}$ is the vector with the outcome variable and $\boldsymbol{W}_k^{(t+1)}$ is a diagonal matrix with diagonal elements $\tau_{1k}^{(t+1)}, \cdots, \tau_{nk}^{(t+1)}$. The updated parameter estimates can now be used for a new iteration by plugging them into (6). This algorithm is carried out until some convergence criterion is satisfied. A nice property of the EM algorithm is that the observed log likelihood cannot decrease (Dempster et al. 1977).

2.2. Mixture Regression in Practice

There are some considerations to be made for a practical implementation of finite mixture models. First of all, the log likelihood of all finite mixture models frequently has multiple local optima (Mclachlan and Peel 2000). Therefore, for a particular set of starting values, application of the EM algorithm can only guarantee you to find a local maximum (or a saddle point in pathological cases (Mclachlan and Krishnan 2008)) and not the global maximum (if this exists). In order to increase the probability of locating the desired optimum it is recommended to apply the EM algorithm from a variety of starting points (Mclachlan and Peel 2000) and select the solution with the highest log likelihood value. This strategy is, however, not a guarantee to success. Then there is still the matter of selecting appropriate starting values. While there has been some research on obtaining good start values (see for instance Karlis and Xekalaki (2003) for univariate normal and Poisson data and Biernacki, Celeux, and Govaert (2003) for multivariate normal data), as far as we know there are no results for mixtures of linear regressions. Viele and Tong (2002) proposed the following strategy for obtaining a random set of starting parameters:

- Generate the mixture proportions $\boldsymbol{\pi}$ as a random draw from a Dirichlet distribution with parameter vector $(1, \cdots, 1)$.
- For every component $k = 1, \cdots, K$, select a random sample of $p+1$ observations $(\boldsymbol{X}_r, \boldsymbol{y}_r)$ without replacement from the data. Obtain $\boldsymbol{\beta}_k$ as the solution of $\boldsymbol{\beta}_k = \boldsymbol{X}_r^{-1} \boldsymbol{y}_r$.
- Generate the component variances as random draws from a uniform distribution with support $[0, s_{(1)}^2]$. Here, $s_{(1)}^2$ denotes the estimated mean squared error obtained from a regular one-component regression analysis.

We have compared this procedure in some small simulation studies with two other procedures. The first alternative procedure only differs in how $\boldsymbol{\beta}$ is generated. For each component $k = 1, \cdots, K$, an intercept is randomly drawn from a uniform distribution with support $[min(y_i), max(y_i)], i = 1, \cdots, n$. All other coefficients are initialized as 0. The second alternative procedure consists of randomly assigning each sample point to one of the K components. We have done this by hard assignment (assign each observation to exactly one component) and by soft assignment (assign each observation to every component with random weights). The EM algorithm is then started from the M-step by considering the assignment as the initial E-step. In our results we found that the strategy of Viele and Tong (2002) performed favorably compared to the alternatives.

Second, the EM algorithm is generally known to converge slowly, linearly or even sublinearly (Mclachlan and Krishnan 2008). Usually, the algorithm is stopped when the log likelihood and/or the parameter estimates do not change much during the last iterations (Mclachlan and Peel 2000). However, due to its slow convergence rate, one can erroneously stop the algorithm too early, i.e. before convergence. Lindstrom and Bates (1988) call this a 'measure of lack of progress but not of actual convergence'. Böhning et al. (1994) used Aitken's acceleration to derive a suitable stopping criterion for a linear convergent sequence. At each iteration (starting from the third), one estimates the stationary value of the log likelihood by using the three last log likelihood values as $l_\infty^{(t+1)} = l^{(t)} + \frac{1}{1-a^{(t)}}(l^{(t+1)} - l^{(t)})$ where for simplicity of notation $l^{(t)}$ denotes the log likelihood value at iteration t and $a^{(t)} = \frac{l^{(t+1)} - l^{(t)}}{l^{(t)} - l^{(t-1)}}$ denotes the estimated rate of convergence of the sequence of log likelihood values. This method is also used to decrease the computation time caused by the multiple random starts as it predicts the stationary log likelihood without requiring the parameters to converge. Each set of starting parameters is iterated until the difference in the estimates of the stationary log likelihood value is smaller than 10^{-9}. The solution with the highest estimated stationary log likelihood is then taken as the optimal solution. However, for some selection criteria (see infra) accurate estimates of the parameter values are also necessary. Therefore, the best solution is then iterated further until the difference between the actual log likelihood and the estimated stationary log likelihood is smaller than 10^{-12} and the maximum absolute change in the estimated component variances is smaller than 10^{-9}. The latter criterion is added because Abbi et al. (2008) found that the variance parameters have the slowest convergence rate.

Third, for finite mixture models with normal components with component specific variance parameters (or covariance matrices) there exists no global maximum for the log likelihood (Mclachlan and Peel 2000). Recall that for a normal distribution the log likelihood is divided by the standard deviation. Therefore, in a mixture of normal linear regressions with $K > 1$ components, one can make the log likelihood infinite by taking any $p + 1$ sample points and put them in a separate component. The resulting fitted hyperplane in this component will then have a perfect fit and its estimated component variance will be 0. Such a solution is obviously neither desired nor useful. A simple solution to this problem would be to put an equality constraint on the variance terms across the components, but this might be too restrictive. Components for which the variance tends to 0 are however not really a problem in practice as the computer will detect these and one can just discard these 'solutions'. A far more serious problem is the potential existence of 'spurious' solutions. Mclachlan and Peel (2000, p. 99) describe these as 'solutions with relatively large local maxima that occur as a consequence of a fitted component having a very small (but

nonzero) variance'. Hence, these solutions converge to parameter values which are very close to, but not on, the edge of the parameter space (σ_k^2 and π_k close to 0). Usually, these solutions are not interesting as they accommodate some random local pattern but will most likely not generalize outside the sample. Despite that only a relatively small number of observations belong to these components, their contribution to the log likelihood may be so high that this solution has a larger sample log likelihood than the desired local maximum (containing meaningful components) and hence masks the desired solution. Dealing with such solutions (i.e. eliminating spurious solutions) will probably require some judgement from the researcher. However, in a simulation study this cannot be done. In our implementation, there are two ways for a local solution to be discarded. The first way is when the estimated component standard deviation becomes smaller than 10^{-10} to avoid singularities. A second way out is when an estimated mixture proportion becomes smaller than $\frac{p+1}{n}$ as this is the boundary value of the effective sample size with which a regression plane with $p+1$ coefficients can be estimated. Other solutions for the singular/spurious component problem include restricting the parameter space or penalizing the likelihood which is the subject of the next section.

2.3. Penalizing the Likelihood

Hathaway (1985) proposed to solve the unboundedness of the likelihood by constraining the parameter space such that $min_{k,k'}(\frac{\sigma_k}{\sigma_{k'}}) \geq c > 0$ for all combinations of $k, k^{'} = 1, \cdots, K$. This formulation ensures that there is a global maximum to the log likelihood which is not singular. Furthermore, by choosing the right c one can also get rid of the spurious solutions. On the other hand, implementing this constraint restricts the solution space and might exclude the desired solution if c is too large. A simpler approach seems to be to penalize the likelihood which has been proposed by Ciuperca, Ridolfi, and Idier (2003) and Chen, Tan, and Zhang (2008). For finite mixture models of univariate normal distributions Ciuperca et al. (2003) proved that in case K is known a priori, their penalized likelihood estimator is consistent and Chen et al. (2008) proved that their version of the penalized likelihood estimator is consistent even when K is unknown. The latter result was generalized to (unconditional) multivariate normal distributions by Chen and Tan (2009). In all three papers the conjugate prior distribution for the component variances is used as the penalty function which makes this method a variant of maximum a posteriori estimation. Ciuperca et al. (2003) showed in a small example how the penalized likelihood method can outperform the method from Hathaway (1985) in case c is too large. Chen et al. (2008) and Chen and Tan (2009) showed with simulation how their penalized estimator gives similar and sometimes better parameter estimates in terms of bias and variance compared to the unpenalized approach. In this paper, the approach of Chen et al. (2008) is adopted which results in a penalized log likelihood of the following form

$$LL\left(\mathbf{\Psi}\right) = \sum_{i=1}^{n} \log \left[\sum_{k=1}^{K} \pi_k \mathcal{N}\left(y_i | \boldsymbol{\beta}_k^T \boldsymbol{x}_i, \sigma_k^2\right) \right] - a_n \sum_{k=1}^{K} \left(\frac{s_{(1)}^2}{\sigma_k^2} + \log \sigma_k^2 \right) \tag{8}$$

where a_n is a constant which depends on the sample size and moderates the influence of the penalty function. The penalty function in (8) is equivalent to putting an inverse-gamma distribution on the component variances with mode at $s_{(1)}^2$. This mode is based on the sample data and is taken to be the estimated variance of the error term in a one-component regression. Maximizing (8) now results in a well-posed maximization problem with a global maximum in the interior of

the parameter space. This, however, does not make (8) concave (there can still be numerous local optima) and therefore does not rid us of the necessity of starting the EM algorithm from different points. The effect of penalizing the likelihood this way only modifies the estimation of the component variances in the M-step. All other equations in (6) and (7) remain the same. The new closed-form solution in an EM-iteration $(t+1)$ is

$$\sigma_k^{2(t+1)} = \frac{\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_k^{(t+1)}\right)^T \boldsymbol{W}_k^{(t+1)} \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_k^{(t+1)}\right) + 2a_n s_{(1)}^2}{\sum_{i=1}^n \tau_{ik}^{(t+1)} + 2a_n}. \tag{9}$$

From (9) it can be seen that when a_n is a function which goes to 0 for n going to ∞, the resulting penalized estimator is equivalent to the unpenalized estimator for large sample sizes. However, for a non-zero a_n in a finite sample, the component variances can never become 0. The resulting estimator in (9) looks similar to the James-Stein estimator which is known to decrease the mean squared error of an original estimator by introducing a relatively small bias (James and Stein 1961; Chen and Tan 2009).

In order to validate these results for mixtures of linear regressions and to select an appropriate a_n we carried out some simulations. We simulated 200 sets of true parameters for a mixture regression model with true number of components $K = 2$ and $K = 3$. The mixture proportions were uniformly drawn from the sets $\pi_1 \in \{0.2, 0.3, 0.4, 0.5\}$ for $K = 2$ and $(\pi_1, \pi_2) \in \{(0.2, 0.2), (0.2, 0.3), (0.2, 0.4), (0.3, 0.3), (0.3333, 0.3333)\}$ for $K = 3$. The regression coefficients were drawn from $U[-2, 2]$ and the component variances were drawn from $U[0.5, 2]$ where $U[a, b]$ denotes a continuous uniform distribution with support $[a, b]$. For each of these 2×200 sets of true parameters, a thousand samples were generated with sample sizes $n = 300$ and $n = 600$. Every sample had $p = 3$ explanatory variables which were drawn from $U[0, 10]$. A sample of size n is generated by drawing a single trial multinomial variable with the mixture proportions as parameter vector. Hence, each observation is labeled to belong to one specific component. Then, for each observation, the dependent variable y_i is drawn from a normal distribution with mean $\boldsymbol{\beta}_k^T \boldsymbol{x}_i$ and variance σ_k^2. Estimation was done using 9 random sets of start parameters and the true parameter vector using unpenalized estimation and penalized estimation with five specifications for $a_n = n^{-\frac{1}{j}}$ with $j = 1, \cdots, 5$ and where each estimator used the same start values. It is expected that the solution obtained by starting from the true parameter values will converge most of the times to the desired local optimum. The solutions of the random starts however, can converge to spurious solutions which may result in a larger sample log likelihood. The quality of the estimation procedures is therefore judged by their ability to recover the parameters which we measure by the mean squared error (MSE) of the estimates compared with the true parameters. Table 1 and table 2 present the average mean squared error over the 200 sets of random parameters for $K = 2$ and $K = 3$ respectively. The standard deviations of the MSE are shown in brackets. From table 1 one can see that larger penalties decrease the average MSE (except for the variance parameters) in the case of two components. Hence, including a penalty term decreases the risk of landing in a spurious solution. For the component variances, the average MSE decreases initially but then rapidly increases beyond the unpenalized average MSE. Hence, by including a larger penalty, the induced bias in the variance estimates offsets the decreased variance of the estimates. From table 2 it seems that the optimal penalty term with respect to average MSE is somewhere in between the two extremes for most parameters. Both tables demonstrate that the differences between the estimation methods become smaller for larger samples which is expected as larger

$K=2, n=300$						
Pen	-	n^{-1}	$n^{\frac{1}{2}}$	$n^{\frac{1}{3}}$	$n^{\frac{1}{4}}$	$n^{\frac{1}{5}}$
π_1	0.0017 (0.0089)	0.0011 (0.0031)	0.0011 (0.0025)	0.0010 (0.0022)	0.0010 (0.0021)	0.0010 (0.0020)
β_{01}	0.2127 (0.3260)	0.1910 (0.1402)	0.1888 (0.1251)	0.1882 (0.1222)	0.1879 (0.1200)	0.1881 (0.1195)
β_{02}	0.0993 (0.2450)	0.0798 (0.0609)	0.0784 (0.0478)	0.0778 (0.0427)	0.0777 (0.0423)	0.0770 (0.0369)
β_{11}	0.0026 (0.0051)	0.0022 (0.0018)	0.0022 (0.0015)	0.0021 (0.0013)	0.0021 (0.0013)	0.0021 (0.0013)
β_{12}	0.0012 (0.0036)	0.0009 (0.0006)	0.0009 (0.0005)	0.0009 (0.0004)	0.0009 (0.0004)	0.0009 (0.0004)
β_{21}	0.0029 (0.0084)	0.0024 (0.0036)	0.0023 (0.0029)	0.0023 (0.0027)	0.0023 (0.0027)	0.0023 (0.0026)
β_{22}	0.0012 (0.0037)	0.0010 (0.0014)	0.0010 (0.0010)	0.0010 (0.0010)	0.0010 (0.0009)	0.0009 (0.0008)
β_{31}	0.0026 (0.0058)	0.0022 (0.0019)	0.0022 (0.0016)	0.0022 (0.0015)	0.0022 (0.0014)	0.0021 (0.0014)
β_{32}	0.0012 (0.0037)	0.0009 (0.0007)	0.0009 (0.0006)	0.0009 (0.0004)	0.0009 (0.0004)	0.0009 (0.0004)
σ_1	0.0117 (0.0205)	0.0106 (0.0122)	0.0099 (0.0089)	0.0156 (0.0197)	0.0274 (0.0455)	0.0416 (0.0741)
σ_2	0.0062 (0.0244)	0.0044 (0.0062)	0.0042 (0.0037)	0.0056 (0.0056)	0.0089 (0.0128)	0.0128 (0.0217)

$K=2, n=600$						
Pen	-	n^{-1}	$n^{\frac{1}{2}}$	$n^{\frac{1}{3}}$	$n^{\frac{1}{4}}$	$n^{\frac{1}{5}}$
π_1	0.0005 (0.0012)	0.0005 (0.0006)	0.0005 (0.0005)	0.0005 (0.0005)	0.0005 (0.0004)	0.0005 (0.0004)
β_{01}	0.0986 (0.1315)	0.0899 (0.0550)	0.0898 (0.0546)	0.0897 (0.0543)	0.0897 (0.0539)	0.0897 (0.0536)
β_{02}	0.0401 (0.0389)	0.0374 (0.0165)	0.0373 (0.0164)	0.0373 (0.0164)	0.0373 (0.0163)	0.0373 (0.0163)
β_{11}	0.0011 (0.0010)	0.0010 (0.0006)	0.0010 (0.0006)	0.0010 (0.0006)	0.0010 (0.0006)	0.0010 (0.0006)
β_{12}	0.0005 (0.0006)	0.0004 (0.0002)	0.0004 (0.0002)	0.0004 (0.0002)	0.0004 (0.0002)	0.0004 (0.0002)
β_{21}	0.0012 (0.0025)	0.0011 (0.0009)	0.0011 (0.0009)	0.0010 (0.0008)	0.0010 (0.0008)	0.0010 (0.0008)
β_{22}	0.0005 (0.0005)	0.0004 (0.0003)	0.0004 (0.0003)	0.0004 (0.0002)	0.0004 (0.0002)	0.0004 (0.0002)
β_{31}	0.0011 (0.0013)	0.0010 (0.0006)	0.0010 (0.0006)	0.0010 (0.0006)	0.0010 (0.0006)	0.0010 (0.0006)
β_{32}	0.0005 (0.0006)	0.0004 (0.0002)	0.0004 (0.0002)	0.0004 (0.0002)	0.0004 (0.0002)	0.0004 (0.0002)
σ_1	0.0049 (0.0050)	0.0047 (0.0039)	0.0045 (0.0035)	0.0054 (0.0044)	0.0080 (0.0102)	0.0116 (0.0183)
σ_2	0.0021 (0.0028)	0.0020 (0.0011)	0.0019 (0.0009)	0.0021 (0.0011)	0.0028 (0.0026)	0.0037 (0.0048)

Table 1: Average MSE (standard deviation MSE) for 2 component models.

$K = 3, n = 300$						
Pen	-	n^{-1}	$n^{\frac{1}{2}}$	$n^{\frac{1}{3}}$	$n^{\frac{1}{4}}$	$n^{\frac{1}{5}}$
π_1	0.0015 (0.0046)	0.0011 (0.0019)	0.0010 (0.0014)	0.0010 (0.0014)	0.0010 (0.0015)	0.0011 (0.0016)
π_2	0.0013 (0.0034)	0.0010 (0.0017)	0.0010 (0.0014)	0.0010 (0.0013)	0.0010 (0.0012)	0.0010 (0.0011)
π_3	0.0018 (0.0044)	0.0013 (0.0023)	0.0013 (0.0019)	0.0012 (0.0017)	0.0012 (0.0017)	0.0012 (0.0017)
β_{01}	0.9442 (6.8421)	0.3238 (0.2501)	0.3111 (0.1814)	0.3094 (0.1663)	0.3126 (0.1640)	0.3177 (0.1646)
β_{02}	0.3444 (1.1443)	0.2255 (0.1844)	0.2164 (0.1387)	0.2188 (0.1490)	0.2223 (0.1566)	0.2263 (0.1607)
β_{03}	0.2209 (0.5593)	0.1389 (0.0761)	0.1375 (0.0726)	0.1372 (0.0720)	0.1377 (0.0715)	0.1388 (0.0716)
β_{11}	0.0298 (0.3191)	0.0045 (0.0069)	0.0044 (0.0068)	0.0047 (0.0104)	0.0048 (0.0108)	0.0049 (0.0109)
β_{12}	0.0062 (0.0270)	0.0034 (0.0068)	0.0032 (0.0065)	0.0035 (0.0103)	0.0035 (0.0106)	0.0036 (0.0106)
β_{13}	0.0046 (0.0174)	0.0021 (0.0037)	0.0019 (0.0028)	0.0019 (0.0030)	0.0019 (0.0031)	0.0019 (0.0032)
β_{21}	0.0387 (0.4527)	0.0041 (0.0047)	0.0039 (0.0038)	0.0039 (0.0035)	0.0040 (0.0036)	0.0044 (0.0057)
β_{22}	0.0047 (0.0152)	0.0028 (0.0039)	0.0027 (0.0035)	0.0026 (0.0032)	0.0027 (0.0033)	0.0030 (0.0053)
β_{23}	0.0038 (0.0115)	0.0019 (0.0020)	0.0018 (0.0016)	0.0018 (0.0017)	0.0018 (0.0019)	0.0019 (0.0022)
β_{31}	0.0116 (0.0532)	0.0049 (0.0095)	0.0044 (0.0063)	0.0045 (0.0087)	0.0046 (0.0088)	0.0046 (0.0089)
β_{32}	0.0060 (0.0287)	0.0034 (0.0080)	0.0032 (0.0069)	0.0034 (0.0090)	0.0034 (0.0091)	0.0035 (0.0090)
β_{33}	0.0059 (0.0295)	0.0024 (0.0074)	0.0020 (0.0028)	0.0019 (0.0024)	0.0019 (0.0024)	0.0019 (0.0023)
σ_1	0.0231 (0.0414)	0.0186 (0.0190)	0.0176 (0.0135)	0.0447 (0.0580)	0.0937 (0.1263)	0.1499 (0.1999)
σ_2	0.0148 (0.0279)	0.0121 (0.0122)	0.0119 (0.0098)	0.0314 (0.0398)	0.0673 (0.0938)	0.1088 (0.1510)
σ_3	0.0114 (0.0256)	0.0084 (0.0110)	0.0078 (0.0060)	0.0156 (0.0198)	0.0312 (0.0481)	0.0500 (0.0802)

$K = 3, n = 600$						
Pen	-	n^{-1}	$n^{\frac{1}{2}}$	$n^{\frac{1}{3}}$	$n^{\frac{1}{4}}$	$n^{\frac{1}{5}}$
π_1	0.0006 (0.0017)	0.0005 (0.0006)	0.0005 (0.0005)	0.0005 (0.0005)	0.0005 (0.0007)	0.0005 (0.0008)
π_2	0.0005 (0.0011)	0.0005 (0.0006)	0.0005 (0.0006)	0.0005 (0.0007)	0.0005 (0.0008)	0.0005 (0.0008)
π_3	0.0007 (0.0012)	0.0006 (0.0008)	0.0006 (0.0007)	0.0006 (0.0007)	0.0006 (0.0007)	0.0006 (0.0006)
β_{01}	0.1625 (0.1948)	0.1458 (0.0824)	0.1447 (0.0789)	0.1444 (0.0772)	0.1451 (0.0768)	0.1465 (0.0783)
β_{02}	0.1401 (0.5011)	0.1013 (0.0580)	0.1008 (0.0579)	0.1014 (0.0585)	0.1030 (0.0643)	0.1050 (0.0759)
β_{03}	0.0689 (0.0520)	0.0657 (0.0331)	0.0655 (0.0326)	0.0654 (0.0323)	0.0654 (0.0322)	0.0656 (0.0322)
β_{11}	0.0022 (0.0056)	0.0017 (0.0014)	0.0017 (0.0016)	0.0019 (0.0043)	0.0022 (0.0083)	0.0023 (0.0096)
β_{12}	0.0017 (0.0069)	0.0012 (0.0012)	0.0012 (0.0015)	0.0014 (0.0044)	0.0017 (0.0083)	0.0018 (0.0096)
β_{13}	0.0011 (0.0039)	0.0008 (0.0008)	0.0008 (0.0007)	0.0008 (0.0007)	0.0008 (0.0007)	0.0008 (0.0007)
β_{21}	0.0020 (0.0039)	0.0016 (0.0009)	0.0016 (0.0008)	0.0016 (0.0008)	0.0016 (0.0008)	0.0016 (0.0008)
β_{22}	0.0015 (0.0043)	0.0011 (0.0007)	0.0011 (0.0006)	0.0011 (0.0006)	0.0011 (0.0006)	0.0011 (0.0006)
β_{23}	0.0009 (0.0019)	0.0008 (0.0005)	0.0008 (0.0005)	0.0008 (0.0005)	0.0008 (0.0005)	0.0008 (0.0005)
β_{31}	0.0027 (0.0105)	0.0017 (0.0013)	0.0017 (0.0014)	0.0019 (0.0035)	0.0021 (0.0066)	0.0022 (0.0077)
β_{32}	0.0022 (0.0127)	0.0013 (0.0017)	0.0013 (0.0017)	0.0015 (0.0038)	0.0017 (0.0068)	0.0017 (0.0077)
β_{33}	0.0017 (0.0107)	0.0009 (0.0017)	0.0008 (0.0009)	0.0008 (0.0009)	0.0008 (0.0009)	0.0008 (0.0008)
σ_1	0.0089 (0.0148)	0.0079 (0.0064)	0.0074 (0.0050)	0.0122 (0.0129)	0.0238 (0.0330)	0.0390 (0.0555)
σ_2	0.0057 (0.0060)	0.0054 (0.0042)	0.0051 (0.0038)	0.0082 (0.0069)	0.0163 (0.0196)	0.0272 (0.0375)
σ_3	0.0042 (0.0087)	0.0036 (0.0032)	0.0035 (0.0023)	0.0047 (0.0037)	0.0080 (0.0100)	0.0125 (0.0184)

Table 2: Average MSE (standard deviation MSE) for 3 component models.

samples decrease the number of spurious solutions and the value of the penalty term. Furthermore, we can see that the average MSE is larger for 3 component models than for 2 component models. This seems logical as complexer models will likely introduce more local optima and hence probably more spurious optima. Larger sample sizes also decrease the size of the average MSE which reflects the consistency of both estimators. It is also apparent that the intercept parameters are estimated relatively poorly. This is most likely caused by the fact that these parameters are estimated at the boundary of the design space of the explanatory variables. If one is interested in estimating this parameter precisely, better experimental designs are warranted. Note also the very large average and standard deviation of the intercept terms for the unpenalized estimator in the upper part of table 2. This is due to one very large outlier (estimated MSE almost 96) for which the unpenalized method deviated extremely from the true solution despite the inclusion of the true parameters in the start values.

From tables 1 and 2 it appears that including a penalty term pays off with respect to the MSE. However, it is hard to determine the optimal value of the penalty constant from these tables. We have summarized the results even more by summing the parameter-wise average MSE. As the different types of parameters have different ranges, the MSEs were first divided by the square of their range to make the errors comparable. Figure 1 shows the relative total average MSE with

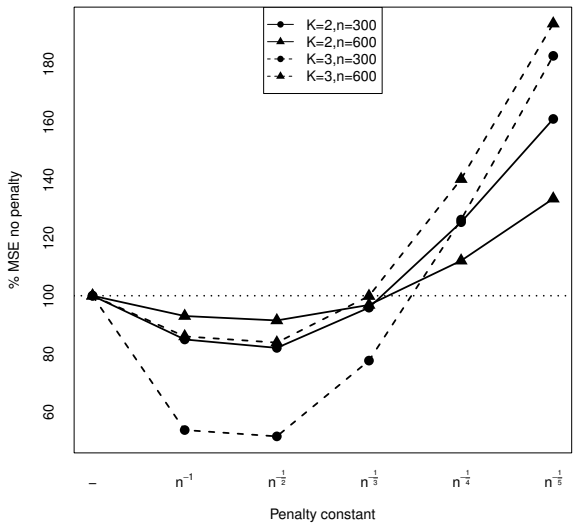

Figure 1: Total relative average MSE relative to the unpenalized estimator.

respect to the unpenalized estimator. From this plot it appears that a penalty constant of $n^{-\frac{1}{2}}$ performs best for our very limited grid search although the difference with n^{-1} is very small. Chen and Tan (2009) also found in their simulations that a penalty constant of $n^{-\frac{1}{2}}$ performed best relative to no penalty and n^{-1}. It might pay dividends to find the optimal penalty constant over a much finer grid (and an optimal penalty function) but this is beyond the scope of this paper. As was shown empirically, the penalized maximum likelihood estimator has on average a smaller MSE and has the ability to steer the EM algorithm away from spurious optima. Therefore, we hypothesize that this estimator can improve model selection for finite mixture models as most

of the (non-Bayesian) selection criteria are based on the maximum log likelihood and/or the maximum likelihood estimates.

3. Order Selection

3.1. Order Selection Criteria

Mixture models in general can be used for two main purposes, namely density estimation or approximation and model-based clustering (Mclachlan and Peel 2000). A mixture model can be used to 'semi-parametrically' estimate densities as any distributional form can be mimicked by adding enough components (see for instance Marron and Wand (1992)). Mixture models can also be used to perform model-based clustering. In model-based clustering, the components represent real but unobserved (or perhaps inherently unobservable) groups in a population and thus have a meaningful interpretation. In both cases the number of components is often unknown a priori. Order selection in finite mixture models consists of finding the appropriate number of components based on the observed data. Order selection for density estimation has mostly been resolved as criteria such as AIC and BIC appear to select a suitable number of components (Mclachlan and Peel 2000). In a model-based clustering context however, order selection is a hard problem for which still no general solution exists (Mclachlan and Peel 2000; Nylund et al. 2007).

An obvious method to determine the number of components would seem to use the likelihood ratio test because a model with K components is nested in a model with $K+1$ components. Unfortunately, the limiting distribution of the test statistic is not the usual χ^2 distribution with degrees of freedom equal to the difference in numbers of parameters. The reason for this is that the regularity conditions which are used in the derivation of the limiting distribution, are violated in the case of mixture models (Ghosh and Sen 1985)[3]. Moreover, Seidel et al. (2000a), Seidel et al. (2000b) and Seidel and Sevcikova (2004) have demonstrated that the distribution of the likelihood ratio test statistic depends on the particular implementation of the EM algorithm. They showed how different start strategies, different stopping rules and different ways of handling spurious components affect this distribution in mixtures of exponential distributions. As a way out, McLachlan (1987) suggested a parametric bootstrap approach. In such a procedure, one generates B datasets under the null hypothesis ($H_0 : K = K_0$) and subsequently calculates the likelihood ratio test statistic for each bootstrap sample. Unfortunately, the number of bootstrap samples B will likely have to be high in order to achieve sufficient power. Furthermore, for every bootstrap sample one has to implement the same estimation procedure used on the original sample which generally will require multiple starts. This results in a computationally burdensome procedure, especially in a simulation setting[4], and therefore this selection method will not be used here. Burnham and Anderson (2002) give another justification for this decision, as they vehemently argue throughout their book that hypothesis testing procedures are not designed for model selection. Therefore, these tests lack theoretical justification in model selection whereas information criteria such as AIC are specifically designed for model selection and should be more suited for order selection in mixture models. Furthermore, Sarstedt (2008) searched applications of mixture regression models in marketing journals between 2000 and 2006 and found that none of the 32 articles he found used

[3]For more on this topic, see for instance Mclachlan and Peel (2000, section 6.4) or Garel (2007).

[4]Bootstrapping the likelihood ratio test may however be very useful if one has enough time and/or computing power. Nylund et al. (2007) presented very favorable results from their simulation study.

a likelihood ratio test or a bootstrapped version for model selection. In most articles BIC was used to select the number of components, followed by AIC and some variants of that suggesting that in practice the bootstrap test is not really used. Another type of model selection methods which will not be considered here are methods based on the Fisher information matrix because approximations to this matrix are only valid for very large samples, especially for mixture models (Mclachlan and Peel 2000) and inaccurate estimates will only introduce extra variability in the order selection[5]. In what follows, the selection methods which were used in our simulation study will be discussed.

Burnham and Anderson (2002) classify model selection criteria into three broad classes, namely optimization of a selection criterion, hypothesis testing and ad hoc methods. As mentioned previously, hypothesis testing will not be used here. We will start with reviewing some criteria which belong to the first class, the information criteria. Most of these criteria were derived for general statistical models and not for order selection in finite mixture models specifically. It should also be noted that for all subsequent criteria, the model in the candidate model set for which the respective criterion is minimized is the selected model. The best known information criterion is most likely AIC which stands for Akaike's information criterion[6]. AIC is defined as

$$-2LL\left(\hat{\mathbf{\Psi}}\right) + 2n_p \tag{10}$$

where $LL\left(\hat{\mathbf{\Psi}}\right)$ is the log likelihood of the data evaluated at the maximum likelihood estimates and n_p denotes the number of parameters in the model which is equal to $(p+3)K-1$ for mixture regressions with p explanatory variables in each component. Akaike (1974) derived AIC as an estimate of the (directed) Kullback-Leibler divergence[7] between the true model and the fitted model. The term n_p is a bias-correction term as the maximized log likelihood is a positively biased estimator of the expected Kullback-Leibler information. Despite popular belief, AIC does not require that the true model is in the set of candidate models (Konishi and Kitagawa 1996; Burnham and Anderson 2002) but the approximations in the derivation do require the same regularity conditions as are needed for the likelihood ratio test (Titterington, Smith, and Makov 1985; Mclachlan and Peel 2000). Several authors have noticed that it tends to overfit, i.e. select too many components, in a finite mixture context (Mclachlan and Peel 2000) but it is still used as shown by Sarstedt (2008). AIC is only asymptotically correct and Burnham and Anderson (2002) warn against using AIC when the ratio $\frac{n}{n_p}$ is smaller than 40. To remedy this, Hurvich and Tsai (1989) developed a small-sample version of AIC for regular linear models with normal errors. Burnham and Anderson (2002) however, also advocate its use in other contexts unless the underlying probability distribution deviates strongly from a normal one. Finite mixtures of normal distributions however, are not normal and can be multimodal, skewed, ... Hence, it would seem that this small sample improvement will not work well in the mixture context. The small-sample

[5]The most widely known criterion of this type is probably $ICOMP$ (Bozdogan 1993) which is defined as $-2LL\left(\hat{\mathbf{\Psi}}\right) + n_p log\left[n_p^{-1} trace\left(\mathcal{I}^{-1}\right)\right] - log\left(|\mathcal{I}^{-1}|\right)$ where $\mathcal{I}$ denotes the expected information matrix, n_p is the number of parameters and $|.|$ is the determinant.

[6]Akaike himself actually called it 'An information criterion'(Burnham and Anderson 2002).

[7]The Kullback-Leibler divergence between distributions f and g is defined as $I(f,g) = \int f(x)\log f(x)dx - \int f(x)\log g(x|\theta)dx$ and represents the lost information when approximating f by g (Kullback and Leibler 1951; Burnham and Anderson 2002).

AIC, denoted by AIC_c is equal to $AIC + \frac{2n_p(n_p+1)}{n-n_p-1}$. It is straightforward to see that the penalty will be larger than that of AIC for finite sample sizes and tends to 0 as the sample size increases.

Whereas AIC is derived by looking at the directed Kullback-Leibler divergence between the truth and the approximating model, Cavanaugh (1999) used the symmetric Kullback-Leibler divergence[8] between truth and approximation. He showed that optimizing this criterion leads to $KIC = -2LL\left(\hat{\mathbf{\Psi}}\right) + 3n_p$ which is short for Kullback information criterion and has a larger penalty than AIC. Cavanaugh (2004) derived also a small sample version $KIC_c = -2LL\left(\hat{\mathbf{\Psi}}\right) + nlog\left(\frac{n}{n-n_p+1}\right) + \frac{n\{(n-n_p+1)(2n_p+1)-2\}}{(n-n_p-1)(n-n_p+1)}$ and showed that it can also hold as an approximation for non-linear models. Cavanaugh (1999) argued that KIC might be a more sensitive measure of departure from the truth than AIC. Interestingly enough, Bozdogan (1993) conjectured that the asymptotic log likelihood ratio for nested mixture models is distributed as a non-central χ^2 distribution. From this he derived that the penalty in (10) should be $3n_p$, which is the same formula as Cavanaugh's KIC. Another modification of AIC was suggested by Bhansali and Downham (1977) who suggested to increase the penalty term to $4n_p$, based on simulations of autoregressive models, which we will denote by $AIC4$.

One 'drawback' of AIC is that it is not a consistent criterion[9]. A consistent model selection criterion is a criterion which, as the sample size grows, asymptotically selects the true model with probability 1 provided that the true model is in the candidate set of models (Burnham and Anderson 2002). Several of such consistent criteria have been derived in the literature. It should also be noted that by requiring a criterion to be consistent, it no longer is an estimator of the relative Kullback-Leibler divergence and is hence no longer efficient (Burnham and Anderson 2002; Yang 2005). Efficiency here means that as the sample size tends to infinity, an efficient information criterion will select the model in the candidate model set which has the smallest expected squared prediction error. Hannan and Quinn (1979) derived the consistent HQ criterion which replaces $2n_p$ by $2n_p log\left(log(n)\right)$ in (10) and has a larger penalty than AIC for sample sizes larger than 15. Bozdogan (1987) proposed another consistent modification of (10), namely $CAIC = -2LL\left(\hat{\mathbf{\Psi}}\right) + n_p\left[log(n)+1\right]$ which increases the penalty function for any sample size. Perhaps the most famous among the consistent criteria is BIC (Schwarz 1978), known as Bayesian information criterion or Schwarz criterion, which is defined as

$$BIC = -2LL\left(\hat{\mathbf{\Psi}}\right) + n_p log(n) \tag{11}$$

and can be derived as a large sample approximation of the logarithm of the integrated likelihood (integrated over the parameter space). Using BIC implies selecting the model with the largest posterior probability without specifying priors. Mclachlan and Peel (2000) note that the derivation of (11) requires regularity conditions which break down for finite mixture models. However, as AIC, BIC is still used in practice as indicated by Sarstedt (2008). It has been reported that BIC underfits finite mixtures (i.e. selects a model with too few components) for small sample sizes (Mclachlan and Peel 2000). BIC was independently derived by Rissanen (1986) based on coding theory and is known as minimum description length in this field. There also exists an adjusted

[8]The symmetric Kullback-Leibler divergence $J(f,g)$ between f and g is defined as $J(f,g) = I(f,g) + I(g,f)$.

[9]Burnham and Anderson (2002) argue that in most realistic situations, it is impossible that the true model is in the set of candidate models and show furthermore by simulation that in case it is, AIC also selects the true model with high probability.

version of BIC, denoted by $aBIC$, which mitigates underfitting in small samples where sample size n in (11) is replaced by $\frac{n+2}{24}$ (Sclove 1987). Liang, Jaszczak, and Coleman (1992) mention two other modifications of (11) where the penalty term is $2n_p log(n)$ and $5n_p log(n)$. These criteria will be denoted by $MDL2$ and $MDL5$ respectively. For non-trivial sample sizes (larger than 55) we can order most of these criteria from the smallest penalty function to the largest penalty function as AIC, KIC, $AIC4$, BIC,$CAIC$,$MDL2$,$MDL5$. AIC_c,KIC_c,$aBIC$ and HQ are somewhere in between depending on the sample size and the dimension of the parameter vector. In general we can say that AIC would select larger models as it has the lowest penalty term which may cause problems with overfitting, as is reported in the literature. $MDL5$ on the other hand will select small models as its penalty term is by far the largest and will therefore be most prone to underfitting. Burnham and Anderson (2002) state that a lot of simulation results report overfitting of AIC because it is improperly used. They argue that the small sample version AIC_c should have been used in many cases. However, as mentioned earlier, the derivation of AIC_c was done for regular linear models and is theoretically incorrect for other types of models. Burnham and Anderson (2002) therefore recommend using a corrected AIC specifically developed for mixture models. Naik, Shi, and Tsai (2007) derived such a mixture regression criterion for simultaneous selection of the number of components and the number of explanatory variables per component, MRC, which has the following formula

$$MRC = \sum_{k=1}^{K} n\hat{\pi}_k \log(\hat{\sigma}_k^2) + \sum_{k=1}^{K} \frac{n\hat{\pi}_k(n\hat{\pi}_k + p_k)}{n\hat{\pi}_k - p_k - 2} - 2\sum_{k=1}^{K} n\hat{\pi}_k \log(\hat{\pi}_k) \tag{12}$$

where $p_k = trace\left(\hat{\boldsymbol{X}}_k \left(\hat{\boldsymbol{X}}_k^T \hat{\boldsymbol{X}}_k\right)^{-1} \hat{\boldsymbol{X}}_k^T\right), \hat{\boldsymbol{X}}_k = \hat{\boldsymbol{W}}_k^{1/2}\boldsymbol{X}$ and $\hat{\boldsymbol{W}}_k$ is a diagonal matrix with elements $\hat{\tau}_{1k}, \cdots, \hat{\tau}_{nk}$. The first term in (12) measures the lack of fit and hence minimizing it will lead to larger models. This tendency is countered by the second term which penalizes retaining many explanatory variables and by the third term which penalizes the number of components. When $K = 1$, (12) is equal to AIC_c and for large samples it is equivalent to AIC. Similar to Cavanaugh (1999), Hafidi and Mkhadri (2010) derived an information criterion based on the symmetric Kullback-Leibler divergence which we will call MRC_k and which is defined as $MRC + \sum_{k=1}^{K}(p_k + 1)$.

Next to the information criteria we will also consider some classification based methods which were also specifically developed for finite mixture models but not for mixtures of (linear) regressions. These methods take classification into account and tend to select models which are able to convincingly classify the observations. It can be shown that the estimated complete data log likelihood is equal to the sample log likelihood minus the entropy of the posterior classification matrix of the estimated posterior probabilities (Hathaway 1986):

$$LL_c\left(\hat{\boldsymbol{\Psi}}, \hat{\boldsymbol{\tau}}\right) = LL\left(\hat{\boldsymbol{\Psi}}\right) + \sum_{k=1}^{K}\sum_{i=1}^{n} \hat{\tau}_{ik} \log \hat{\tau}_{ik} \tag{13}$$

where $\hat{\boldsymbol{\tau}}$ denotes the matrix of posterior probabilities and the second term on the right hand side is the negative of the estimated entropy $EN(\hat{\boldsymbol{\tau}})$. Biernacki and Govaert (1997) suggested using this for order selection. By multiplying (13) by -2 one obtains the classification likelihood criterion (CLC). Biernacki and Govaert (1997) found that this criterion works well for well

separated components and equal mixture proportions. Banfield and Raftery (1993) also used the classification likelihood to derive an approximate Bayesian criterion called the approximate weight of evidence $AWE = -2LL_c\left(\hat{\boldsymbol{\Psi}}, \hat{\boldsymbol{\tau}}\right) + 2n_p(\frac{3}{2} + \log n)$. Note that the penalty term in AWE is very large. Celeux and Soromenho (1996) propose to use the entropy directly to select the correct number of components by using the normalized entropy criterion $NEC = \frac{EN(\hat{\boldsymbol{\tau}})}{LL(\hat{\boldsymbol{\Psi}}) - LL_{(1)}}$ where $LL_{(1)}$ denotes the maximized log likelihood for a one-component model. As this criterion is undefined for $K = 1$, Biernacki, Celeux, and Govaert (1999) modified it by setting NEC at 1 in this case. As CLC and NEC don't penalize for model complexity these methods might tend to overfit which can be overcome by including a penalty for model complexity. Furthermore, BIC does not take the mixture context into account. Biernacki, Celeux, and Govaert (1998) proposed to solve these problems with the integrated classification likelihood criterion which is defined as

$$ICL = CLC + 2n\sum_{k=1}^{K}\hat{\pi}_k \log(\hat{\pi}_k) + (n_p - K + 1)\log(n) - 2K(n\hat{\pi}_1, \cdots, n\hat{\pi}_K) \tag{14}$$

$$K(n_1, \cdots, n_K) = \sum_{k=1}^{K}\log\left(\Gamma(n_k + \alpha)\right) - \log\left(\Gamma(n + K\alpha)\right) - g\log\left(\Gamma(\alpha)\right) + \log\left(\Gamma(K\alpha)\right) \tag{15}$$

where $\Gamma(.)$ is the gamma function and α represents the parameter of a prior Dirichlet distribution on $\boldsymbol{\pi}$. Jeffrey's non-informative prior takes α as $1/2$ which is also what Biernacki et al. (1998) use and what will be used here. Biernacki et al. (1998) also provide a large sample BIC approximation to ICL which is $ICL - BIC = CLC + n_p\log(n)$ and they have found that this approximation doesn't differ much from using (14). An overview of all order selection criteria considered can be found in table 3.

3.2. Previous Results

Selecting the correct number of components has been extensively studied in the literature. These simulation studies vary in the type of models considered, the selection methods used and the settings of the simulation design (experimental factors). In this section, some of these studies will be reviewed.

In the context of mixtures of multinomial distributions (also known as latent class analysis) several extensive simulation studies have been performed. Yang (2006) found that $aBIC$ was generally the best criterion. For large samples BIC and $CAIC$ also performed well. Dias (2007) concluded that BIC outperforms several complete information based criteria. Yang (2007) also found that $aBIC$ was the best performing information criterion and also mention KIC as a good alternative. Cutler and Windham (1994) simulated mixtures of multivariate normal components. They found that $ICOMP$ was superior to both AIC and BIC. In a small scale simulation McLachlan and Ng (2000) found that ICL and ICL-BIC outperformed BIC and AIC and showed that AIC tends to overfit. Celeux and Soromenho (1996) also performed some simulations for both univariate and multivariate mixtures of normal distributions. They found that AIC has a slight tendency to select too many components, that BIC tends to select too few and that NEC and $ICOMP$ generally perform best. Nylund et al. (2007) concluded that BIC is the best information criterion for both mixtures of contingency tables and mixtures of normal distributions. They also showed that a parametric bootstrap of the likelihood ratio test outperforms BIC. An interesting study is that of Fonseca and Cardoso (2007) where they compared the performance of

Criterion	Formula	Source
AIC	$-2LL\left(\hat{\boldsymbol{\Psi}}\right)+2n_p$	Akaike (1974)
AIC_c	$AIC+\frac{2n_p(n_p+1)}{n-n_p-1}$	Hurvich and Tsai (1989)
KIC	$-2LL\left(\hat{\boldsymbol{\Psi}}\right)+3n_p$	Cavanaugh (1999)
KIC_c	$-2LL\left(\hat{\boldsymbol{\Psi}}\right)+nlog\left(\frac{n}{n-n_p+1}\right)+\frac{n\{(n-n_p+1)(2n_p+1)-2\}}{(n-n_p-1)(n-n_p+1)}$	Cavanaugh (2004)
$AIC4$	$-2LL\left(\hat{\boldsymbol{\Psi}}\right)+4n_p$	Bhansali and Downham (1977)
HQ	$-2LL\left(\hat{\boldsymbol{\Psi}}\right)+2n_plog\left(log(n)\right)$	Hannan and Quinn (1979)
$CAIC$	$-2LL\left(\hat{\boldsymbol{\Psi}}\right)+n_p\left[log(n)+1\right]$	Bozdogan (1987)
BIC	$-2LL\left(\hat{\boldsymbol{\Psi}}\right)+n_plog(n)$	Schwarz (1978), Rissanen (1986)
$aBIC$	$-2LL\left(\hat{\boldsymbol{\Psi}}\right)+n_plog(\frac{n+2}{24})$	Sclove (1987)
$MDL2$	$-2LL\left(\hat{\boldsymbol{\Psi}}\right)+2n_plog(n)$	Liang et al. (1992)
$MDL5$	$-2LL\left(\hat{\boldsymbol{\Psi}}\right)+5n_plog(n)$	Liang et al. (1992)
MRC	$\sum_{k=1}^{K}n\hat{\pi}_k\log(\hat{\sigma}_k^2)+\sum_{k=1}^{K}\frac{n\hat{\pi}_k(n\hat{\pi}_k+p_k)}{n\hat{\pi}_k-p_k-2}-2\sum_{k=1}^{K}n\hat{\pi}_k\log(\hat{\pi}_k)$	Naik et al. (2007)
MRC_k	$MRC+\sum_{k=1}^{K}\left(p_k+1\right)$	Hafidi and Mkhadri (2010)
CLC	$-2LL\left(\hat{\boldsymbol{\Psi}}\right)-2\sum_{k=1}^{K}\sum_{i=1}^{n}\hat{\tau}_{ik}\log\hat{\tau}_{ik}$	Biernacki and Govaert (1997)
AWE	$CLC+2n_p(\frac{3}{2}+\log n)$	Banfield and Raftery (1993)
NEC	$\frac{-\sum_{k=1}^{K}\sum_{i=1}^{n}\hat{\tau}_{ik}\log\hat{\tau}_{ik}}{LL(\hat{\boldsymbol{\Psi}})-LL_{(1)}}$	Celeux and Soromenho (1996)
ICL	$CLC+2n\sum_{k=1}^{K}\hat{\pi}_k\log(\hat{\pi}_k)+(n_p-K+1)\log(n)-2K(n\hat{\pi}_1,\cdots,n\hat{\pi}_K)$	Biernacki et al. (1998)
$ICL\text{-}BIC$	$CLC+n_plog(n)$	Biernacki et al. (1998)

Table 3: Overview of order selection criteria. $K(n_1,\cdots,n_K)=\sum_{k=1}^{K}\log\left(\Gamma(n_k+\alpha)\right)-\log\left(\Gamma(n+K\alpha)\right)-g\log\left(\Gamma(\alpha)\right)+\log\left(\Gamma(K\alpha)\right)$.

several selection measures on 42 real datasets where the true number of components is known. For the categorical datasets, they found that KIC worked best as it selected the correct number of components in 95% of the cases. For continuous data, they used multivariate normal models and found that BIC works best with a success rate of 77%. In the datasets with mixed types of data (both continuous and categorical) they found that $ICL\text{-}BIC$ performed best (80% success rate). They also noted that the performance of the AIC family of information criteria and $ICL\text{-}BIC$ varied a lot across the different types of data. From their results, it can be seen that BIC has the highest average success rate followed by $CAIC$. CLC on the other hand performs worst on average, followed by AWE. Jedidi, Jagpal, and DeSarbo (1997) found that BIC and to a lesser extent $CAIC$ work well in mixtures of structural equation models. Andrews and Currim (2003a) showed that KIC outperforms $ICOMP$, BIC and a validation sample method in mixtures of logistic regressions. In the context of mixtures of growth models, Lubke and Neale (2006) found that AIC and $aBIC$ outperform BIC and $CAIC$. Tofighi and Enders (2008) also found that $aBIC$ works well and BIC performs poorly for this type of models.

Hawkins, Allen, and Stromberg (2001) were the first to systematically investigate model selection in finite mixtures of univariate linear regressions using an extensive simulation study. The factors in the experiment were the true number of mixing components (1 to 4), the mixture proportions and the parameters in the component regression models which were condensed in one measure of separation between the components. They compared order selection based on 22 selection criteria which were based on the log likelihood, an approximation to the Fisher information matrix and several approximations to the complete data log likelihood and the complete data Fisher information matrix. They also included two classification-based measures. In general they concluded that model selection performance of all criteria decreased as the true number of components increased and in the presence of small mixture proportions. The performance increased on the other hand when the components were better separated. For a small number of components (1 or 2) they found that $ICOMP$ was the second worst criterion (only better than the log likelihood itself). BIC and to a lesser extent AWE performed the best in that situation. For larger numbers of components no criterion outperformed the others in all circumstances. They could however conclude that AIC, KIC, $ICOMP$, BIC and AWE as a group performed better than the other measures which were based on approximations of the complete data Fisher information matrix or on the posterior probabilities. Finally, they also noted that KIC did not systematically outperform AIC or the other way around. Andrews and Currim (2003b) investigated the performances of AIC, KIC, BIC, $CAIC$, $ICOMP$, a validation sample log likelihood and NEC in a simulation of linear regression with repeated observations per subject. They varied eight factors: the true number of components, the mean separation between component coefficients, the number of subjects, the number of observations per subject, the number of explanatory variables, R^2 within the components, the minimum mixture proportion and the measurement level of the explanatory variables. They found that KIC was the best criterion in all experimental conditions followed by BIC and the validation log likelihood. $ICOMP$, NEC and AIC on the other hand did not perform well. Oliveira-brochado and Martins (2008) performed a similar simulation study as Andrews and Currim (2003b). They added another experimental factor differentiating between normal errors and uniform errors. Furthermore they compared 26 selection criteria. They found that overall, KIC, $ICL\text{-}BIC$, HQ and $AIC4$ (in that order) performed best and that AIC, AIC_c and $ICOMP$ had the largest tendency to overfit. Most of the classification-based criteria on the

other hand showed high rates of underfitting. Both studies also showed that generally the performance of the criteria increases when the true model is less complex, i.e. fewer components and explanatory variables, the separation between the components increases, the sample size grows and the absence of very small components. Surprisingly, Oliveira-brochado and Martins (2008) found that the effect of error misspecification only had a small negative effect. Finally, Sarstedt (2008) investigated the performance of AIC, KIC,BIC and $CAIC$ in mixtures of univariate regressions while varying the sample size systematically between 100 and 500. In this study it was found that $CAIC$ and to a lesser degree BIC performed well across all sample sizes. KIC only performed well for sample sizes larger than 250 and AIC performed poorly in all experimental conditions. All these results suggest that it might be impossible to find one selection criterion to work best in all situations, let alone for all types of models[10].

4. Simulation Study

4.1. Experimental Design

The design of our simulation study largely follows Hawkins et al. (2001). The number of explanatory variables p is set to 3 in all true models. All explanatory variables are drawn from uniform distributions with support $[0, 10]$. The regression coefficients (including the intercept) and the component variances are drawn from uniform distributions with support $[-2, 2]$ and $[0.5, 2]$ respectively to increase generalizability. As a measure of separation for the components we calculated the average distance between the component regression hyperplanes as in Hawkins et al. (2001). The distance between 2 components k and l at some specific point $\boldsymbol{x}$ is equal to

$$M = \sqrt{\frac{\left(\boldsymbol{\beta}_k^T \boldsymbol{x} - \boldsymbol{\beta}_l^T \boldsymbol{x}\right)^2}{\sigma_k^2 + \sigma_l^2}}. \tag{16}$$

We evaluated this at 50 evenly spaced grid points between 0 and 10 in each of the 3 dimensions and took the average as the separation between component k and l.

The experimental factors and levels are:

- K^*, true number of components: 1, 2 or 3;
- n, sample size: 300 or 600;
- $\boldsymbol{\pi}$, the mixture proportions: equal $(1/K^*)$ or unequal with $\boldsymbol{\pi} = (0.34, 0.66)$ for $K^* = 2$ and $\boldsymbol{\pi} = (0.25, 0.25, 0.5)$ for $K^* = 3$;
- t, type of model (mis)specificaton: 1-8.

A level of 1 for t indicates no misspecification and is the only specification (together with $t = 5$) where the true model is in the set of candidate models. Level 2 means that after the true data generation 3 independent explanatory variables were added to the sample. This is a situation which frequently arises when researchers are unsure which variables are relevant. The data used for estimation thus contain superfluous, uninformative variables. A misspecification level 3 indicates that after data generation one of the explanatory variables was dropped from the sample (we have

[10]Bootstrapping might be one but its computational burden makes it impractical.

arbitrarily taken the last one). This mimics a situation were an important variable is unknown to be related to the dependent variable. In both cases, due to the independency of the explanatory variables, it would be expected that the regression coefficients could still be estimated without bias when the model is estimated with K^* components. It is however expected that with type 2 misspecification the order selection procedures can capitalize on the higher dimensionality of the parameter space and hence prefer models with more components which would lead to overfitting. In situation 3, the parameter space has a smaller dimension and therefore it might be harder to pick up the true number of components. As the importance of the dropped explanatory variable is not uniform across the components (the regression coefficient varies across the components) it might also be the case that specific components become much harder to find for a large $|\beta_{3k}|$ whereas detection of others might hardly be influenced for small $|\beta_{3k}|$. It is therefore expected that this would increase the rate of underfitting for the selection procedures. Misspecification type 4 means that the true data generation mechanism includes an interaction (arbitrarily taken between explanatory variables 2 and 3) whereas it is estimated without this effect. The estimated regression coefficients will no longer be unbiased as the explanatory variables are correlated with the unincluded but real interaction effect. It is unclear how this will affect model selection. Type 5 is not a real model misspecification as it indicates that the explanatory variables are correlated. The design matrix for this factor level was generated according to Falk (1999) with all correlations put to 0.5. For all types 1 to 5 the errors are normally distributed as specified earlier. Misspecification of type 6 indicates that the normal error terms are transformed to have a higher kurtosis and type 7 that they are transformed to have skewed errors. The transformations were done according to Fleishman's method (Fleishman 1978). The type 6 errors were transformed to have excess kurtosis of 4 whereas the type 7 errors were transformed to have excess kurtosis of 4 and skewness of 1.5[11].

The effect of these transformations is illustrated in figure 2 for standard normal variables. It can readily be seen that type 6 makes the tails of the error distribution heavier with respect to a normal distribution. On the one hand this makes it easier to find the real components but on the other hand this may lead to extra components which accommodate the outlying observations. It is therefore expected that this type of misspecification will lead to overfitting. For type 7 of model misspecification, the error terms are asymmetric which will most likely also lead to overfitting. Titterington et al. (1985) for instance, showed how it is practically impossible to differentiate a lognormal distribution (which is skewed) from a mixture of 2 normal distributions. The final type of model misspecification (8) is a case where the errors within a component are heteroskedastic. This was achieved by multiplying the error of observation i belonging to component k with $\exp(\frac{\sum_{j=1}^{p} x_{ij}}{5p} - 0.3)$. Afterwards the errors were multiplied by the appropriate scaling factor to make them have the required average variance within each component. It is expected that this will also lead to increased overfitting as the regions with higher error variability might accommodate multiple components. An overview of the different model specifications can be found in table 4. The design is full factorial and was executed with 1000 replications. For each replication and combination of factor settings, a set of parameters and a design matrix was generated as specified above. Component membership was generated by drawing a sample of size n from a multinomial distribution with parameter vector $\boldsymbol{\pi}$. The dependent variables y_i where then generated as a

[11] It is not possible to set skewness independently from kurtosis (Headrick 2002).

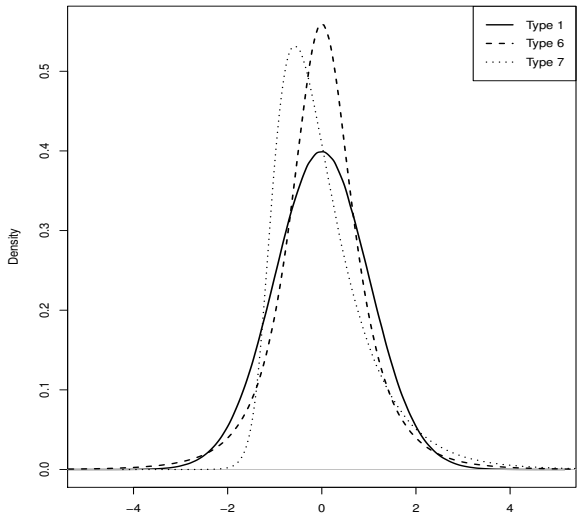

Figure 2: Kernel density plot of error distributions.

Code	Type of misspecification
1	-
2	3 superfluous explanatory variables
3	1 missing explanatory variable
4	missing interaction
5	multicollinearity
6	heavy tailed errors
7	skewed errors
8	heteroskedastic errors within each component

Table 4: Overview of model (mis)specifications.

draw from a normal distribution with mean $\boldsymbol{\beta}_k^T \boldsymbol{x}_i$ for the relevant component k and a (potentially transformed) variance. Models where $K^* = 1$ were fitted with $K = 1 - 3$, models with $K^* = 2$ were fitted with $K = 1 - 5$ and models with $K^* = 3$ were fitted with $K = 1 - 6$ where K^* denotes the true number of components. Estimation was done with an unpenalized and a penalized EM algorithm with 200 random starts for $K > 1$. The penalty constant was taken to be $n^{-\frac{1}{2}}$. As measures of performance we will look at the relative root mean squared error of estimation and the success rate of the order selection criteria compared to the known true number of components K^*. However, as the correct model is not always in the set of candidate models it might be that a model with $K \neq K^*$ is a more appropriate model. Therefore we will also look at a validation sample of size 1000 generated from the true data generating model. The estimated model with the highest log likelihood in the validation sample is then taken as a success with respect to out of sample prediction as this is an estimate of the Kullback-Leibler divergence up to a constant (Burnham and Anderson 2002).

4.2. Results and Discussion

Before we analyze the model selection results, we take a look at the convergence of the unpenalized and the penalized estimators. From table 5 it can be seen that there is no difference between both estimation procedures (in terms of convergence) when the true number of components is 1 as for every dataset a non-spurious solution could be found for $k = 1 - 3$. For a higher number of true components, however there is a large difference between both procedures. Both estimators converged to non-spurious solutions for models where $K \leq K^*$ with some rare exceptions for the penalized estimator. This is not necessarily a big problem. The instances when this happened occurred when one component was very close to another component which made them virtually indistinguishable (separation < 1) and happened 8 out of 9 times for a sample size of 300. Nevertheless this serves as an indication that penalizing the likelihood can be problematic if the penalty term is too large. From table 5 it can also be observed that a penalized likelihood es-

		Type															
		1		2		3		4		5		6		7		8	
K^*	K	Unp	Pen	Unp	Pen	Unp	Pen	Unp	Pen	Unp	Pen	Unp	Pen	Unp	Pen	Unp	Pen
1	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	3	1.00	0.86	1.00	0.92	1.00	1.00	1.00	0.86	1.00	0.85	1.00	0.98	1.00	0.99	1.00	1.00
	4	0.97	0.67	0.99	0.82	1.00	0.98	0.98	0.69	1.00	0.72	0.99	0.90	0.99	0.93	1.00	0.97
	5	0.82	0.44	0.77	0.58	0.96	0.82	0.83	0.46	0.91	0.51	0.88	0.69	0.87	0.65	0.93	0.83
3	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	3	1.00	1.00^{5}	1.00	1.00^{1}	1.00	1.00	1.00	1.00	1.00	1.00^{3}	1.00	1.00	1.00	1.00	1.00	1.00
	4	0.99	0.73	1.00	0.87	1.00	1.00	0.99	0.72	1.00	0.79	0.99	0.94	1.00	0.96	1.00	0.99
	5	0.84	0.40	0.79	0.55	0.96	0.90	0.83	0.40	0.93	0.50	0.89	0.72	0.89	0.76	0.96	0.91
	6	0.17	0.06	0.32	0.19	0.46	0.36	0.17	0.06	0.37	0.13	0.26	0.17	0.25	0.18	0.37	0.32

Table 5: Rates of properly converged estimations. Cells where estimation with the correct number of components did not converge are indicated with a superscript which denotes the number of failures out of a total of 4000.

timator can partly serve as an order selection tool by not converging to any non-spurious solution with $K > K^*$. This also happens for the non-penalized estimator but with much lower frequency. Obviously, solutions which did not converge are not considered for order selection and hence for these datasets, that particular number of components cannot be selected.

Next we take a look at the root mean squared error of both estimators when $K^* > 1$ and $K = K^*$ because for one-component models, both estimators are identical (closed-form solution) and for $K \neq K^*$ the model structure is too different from the truth to be easily compared. As before, the results are adjusted for the different ranges of the different types of parameters. From table 6 it is clear that on most occasions the penalized estimates have smaller RMSE than the

			n			
			300		600	
type	K^*	$\boldsymbol{\pi}$	Unp	Pen	Unp	Pen
1	2	Equal	0.135	0.132	0.094	0.094
		Unequal	0.142	0.141	0.099	0.098
	3	Equal	0.228	0.221	0.159	0.147
		Unequal	0.276	0.408	0.168	0.158
2	2	Equal	0.173	0.167	0.117	0.114
		Unequal	0.184	0.179	0.130	0.126
	3	Equal	0.319	0.278	0.216	0.197
		Unequal	0.357	0.316	0.230	0.200
3	2	Equal	2.911	2.909	2.926	2.922
		Unequal	2.924	2.923	2.912	2.912
	3	Equal	3.711	3.664	3.657	3.642
		Unequal	3.595	3.526	3.612	3.608
4	2	Equal	0.134	0.133	0.094	0.092
		Unequal	0.146	0.142	0.095	0.095
	3	Equal	0.243	0.218	0.157	0.148
		Unequal	0.246	0.229	0.168	0.157
5	2	Equal	0.122	0.120	0.083	0.083
		Unequal	0.128	0.127	0.091	0.087
	3	Equal	0.234	0.211	0.151	0.152
		Unequal	0.243	0.214	0.155	0.143
6	2	Equal	0.169	0.165	0.122	0.121
		Unequal	0.171	0.170	0.126	0.125
	3	Equal	0.297	0.264	0.201	0.194
		Unequal	0.340	0.300	0.231	0.226
7	2	Equal	0.176	0.170	0.122	0.120
		Unequal	0.192	0.189	0.137	0.135
	3	Equal	0.296	0.266	0.232	0.221
		Unequal	0.325	0.294	0.250	0.230
8	2	Equal	0.470	0.474	0.448	0.454
		Unequal	0.597	0.607	0.589	0.594
	3	Equal	1.099	1.038	1.003	0.995
		Unequal	1.263	1.202	1.232	1.162

Table 6: Average relative RMSE.

unpenalized estimates. Furthermore, when the unpenalized estimator is better, the difference is small expect for the cell with $n = 300$, type$= 1, K^* = 3$ and $\boldsymbol{\pi}$ is unequal. This large deviation is caused by 3 large outliers and the median difference is only 0.003. Hence, once again one can see that penalization generally leads to better estimation. Furthermore one can notice that the estimators perform better for equal mixture proportions, larger samples and fewer components. Finally, it can be seen that type 3 misspecification leads to the worst estimates (by far) of all types of misspecification followed by type 8. Other misspecifications have only minor detrimental

effects (if at all).

Table 7 presents the percentages of underfitting, correct fitting and overfitting with respect to the true number of components for each type of misspecification, estimation and order selection method. A striking difference can be noticed between penalized estimation and unpenalized estimation. In all but two settings, the model selection criteria have a higher or equal probability of selecting the correct number of components when using the penalized estimator and when they do worse, it is only by 1 percentage point at most. Therefore, the rest of the discussion will be about the results from the penalized estimation. Another very noticeable effect is that neglecting to include an important explanatory variable decreases the performance of all criteria by a sizable percentage. The criteria which are most robust to this type of misspecification are $MDL2$ with a 72% succes rate, $CAIC$ with a 68% success rate and BIC with a 63% success rate. Furthermore, in this setting, the largest rates of underfitting can be observed, especially for those selection criteria which perform well (80% overall success rate or higher). Including superfluous explanatory variables does not seem to affect the better selection methods on the other hand whereas it has a substantial detrimental effect on the estimators of the relative or symmetric Kullback-Leibler divergence (but not on those derived specifically for mixture models) and on the sample size adjusted BIC and HQ. This outcome would suggest that, all things considered, one is better of with too many explanatory variables than with too few, although this is a conclusion which would need to be investigated in more detail. Surprisingly enough neglecting an interaction does not seem to decrease the performance of the selection criteria. Moreover, many of them actually increase their success rate by a percentage point in this situation. On the other hand, multicollinearity of the explanatory variables appears to have a small negative effect. Many criteria do surprisingly well in the case of error misspecification although the highest success rates are somewhat lower for these situations. However, none of the criteria appear to be very robust to all three types of misspecification. $MDL5$ and AWE appear to be unaffected by the heavier tails and skewness but drop substantially in case of heteroskedasticity. The most stable criteria here are MRC_k, MRC and the integrated classification criteria ICL and $ICL\text{-}BIC$. Furthermore, it can be seen that AIC and AIC_c are by far the least successful criteria with high rates of overfitting and that AIC is dominated by AIC_c by a small margin. Their symmetric counterparts KIC and KIC_c perform better but still not well compared to other criteria and it can be seen that KIC_c dominates KIC. The two criteria derived especially for mixtures of linear regressions (MRC and MRC_k) outperform these criteria substantially but amongst themselves they don't differ much. The larger penalties in $AIC4$, $CAIC$ and BIC make these measures very performant for situations 1, 2, 4 and 5, among the better performers in situation 3 but decrease this excellent performance substantially in case of error misspecification. Similar behaviour can be observed for CLC and HQ. It should also be noted that $aBIC$ is dominated by BIC. NEC is an interesting case as it never performs really well with a maximum succes rate of 80% but, except for type 3 misspecification, seems to keep itself at a respectable level in all other situations. All things considered, it appears that MRC, MRC_k, $MDL2$, $MDL5$, ICL and $ICL\text{-}BIC$ are the criteria which consistently perform with a high success rate (except for misspecification level 3). Unfortunately, there is not a single criterion which outperforms the others in all scenarios and never drops below 80%. The performance of the best criteria is graphically presented in figure 3.

Table 8 presents the percentages of underfitting, correct fitting and overfitting of the penalized

		AIC			AIC_c			MRC			KIC			KIC_c			MRC_k			AIC4			CAIC			HQ		
Type	Pen	U	C	O	U	C	O	U	C	O	U	C	O	U	C	O	U	C	O	U	C	O	U	C	O	U	C	O
1	P	0.00	0.58	0.42	0.00	0.62	0.37	0.06	0.94	0.00	0.00	0.84	0.16	0.00	0.87	0.13	0.06	0.94	0.00	0.00	0.97	0.03	0.01	0.99	0.00	0.00	0.94	0.06
	U	0.00	0.01	0.99	0.00	0.02	0.98	0.05	0.75	0.20	0.00	0.04	0.96	0.00	0.06	0.94	0.05	0.82	0.12	0.00	0.14	0.86	0.01	0.77	0.23	0.00	0.09	0.91
2	P	0.00	0.39	0.61	0.00	0.49	0.51	0.06	0.94	0.00	0.00	0.67	0.33	0.00	0.73	0.27	0.06	0.94	0.00	0.01	0.86	0.13	0.02	0.98	0.00	0.00	0.78	0.22
	U	0.00	0.00	1.00	0.00	0.00	1.00	0.05	0.44	0.52	0.00	0.01	0.99	0.00	0.01	0.99	0.05	0.57	0.38	0.00	0.02	0.98	0.01	0.43	0.56	0.00	0.01	0.99
3	P	0.00	0.08	0.92	0.00	0.11	0.89	0.44	0.54	0.01	0.01	0.26	0.73	0.01	0.29	0.69	0.45	0.54	0.01	0.04	0.42	0.55	0.13	0.68	0.19	0.02	0.36	0.62
	U	0.00	0.01	0.99	0.00	0.01	0.99	0.42	0.49	0.09	0.00	0.02	0.98	0.00	0.04	0.96	0.43	0.51	0.06	0.01	0.10	0.89	0.11	0.62	0.26	0.00	0.06	0.94
4	P	0.00	0.58	0.42	0.00	0.63	0.37	0.05	0.95	0.00	0.00	0.83	0.17	0.00	0.87	0.13	0.06	0.95	0.00	0.00	0.97	0.03	0.01	0.99	0.00	0.00	0.93	0.06
	U	0.00	0.01	0.99	0.00	0.02	0.98	0.05	0.75	0.21	0.00	0.05	0.95	0.00	0.07	0.93	0.05	0.82	0.13	0.00	0.14	0.86	0.01	0.76	0.24	0.00	0.09	0.91
5	P	0.00	0.55	0.45	0.00	0.60	0.40	0.09	0.91	0.00	0.00	0.83	0.17	0.00	0.86	0.13	0.09	0.91	0.00	0.01	0.96	0.03	0.01	0.99	0.00	0.00	0.93	0.06
	U	0.00	0.00	1.00	0.00	0.00	1.00	0.08	0.70	0.22	0.00	0.02	0.98	0.00	0.03	0.97	0.08	0.78	0.14	0.00	0.09	0.91	0.01	0.73	0.26	0.00	0.05	0.95
6	P	0.00	0.07	0.93	0.00	0.09	0.91	0.05	0.89	0.06	0.00	0.14	0.86	0.00	0.17	0.83	0.05	0.90	0.05	0.00	0.24	0.76	0.01	0.58	0.42	0.00	0.19	0.81
	U	0.00	0.00	1.00	0.00	0.00	1.00	0.04	0.78	0.18	0.00	0.01	0.99	0.00	0.01	0.99	0.05	0.81	0.14	0.00	0.03	0.97	0.00	0.26	0.74	0.00	0.02	0.98
7	P	0.00	0.03	0.97	0.00	0.03	0.97	0.04	0.87	0.09	0.00	0.04	0.96	0.00	0.05	0.95	0.05	0.88	0.08	0.00	0.08	0.92	0.00	0.27	0.73	0.00	0.06	0.94
	U	0.00	0.00	1.00	0.00	0.00	1.00	0.04	0.76	0.21	0.00	0.00	1.00	0.00	0.00	1.00	0.04	0.79	0.18	0.00	0.01	0.99	0.00	0.12	0.88	0.00	0.00	1.00
8	P	0.00	0.02	0.98	0.00	0.02	0.98	0.18	0.82	0.00	0.00	0.07	0.93	0.00	0.11	0.89	0.18	0.81	0.00	0.00	0.19	0.80	0.02	0.62	0.36	0.00	0.13	0.87
	U	0.00	0.00	1.00	0.00	0.00	1.00	0.17	0.73	0.11	0.00	0.01	0.99	0.00	0.01	0.99	0.17	0.75	0.08	0.00	0.03	0.97	0.01	0.39	0.60	0.00	0.01	0.99

		BIC			aBIC			MDL2			MDL5			CLC			AWE			ICL			ICL-BIC			NEC		
Type	Pen	U	C	O	U	C	O	U	C	O	U	C	O	U	C	O	U	C	O	U	C	O	U	C	O	U	C	O
1	P	0.01	0.99	0.00	0.00	0.81	0.19	0.02	0.98	0.00	0.07	0.93	0.00	0.05	0.94	0.00	0.09	0.91	0.00	0.06	0.94	0.00	0.06	0.94	0.00	0.20	0.80	0.00
	U	0.00	0.58	0.42	0.00	0.04	0.96	0.02	0.98	0.01	0.07	0.93	0.00	0.03	0.14	0.83	0.09	0.91	0.00	0.06	0.62	0.32	0.06	0.63	0.31	0.18	0.36	0.45
2	P	0.01	0.99	0.00	0.00	0.64	0.35	0.03	0.97	0.00	0.15	0.85	0.00	0.05	0.93	0.01	0.12	0.88	0.00	0.07	0.93	0.00	0.07	0.93	0.00	0.20	0.80	0.01
	U	0.00	0.21	0.79	0.00	0.00	1.00	0.03	0.95	0.02	0.15	0.85	0.00	0.02	0.02	0.96	0.11	0.89	0.00	0.05	0.30	0.65	0.05	0.31	0.64	0.16	0.27	0.57
3	P	0.10	0.63	0.27	0.01	0.21	0.78	0.24	0.72	0.04	0.44	0.56	0.00	0.42	0.54	0.04	0.51	0.49	0.00	0.46	0.53	0.00	0.47	0.53	0.00	0.48	0.51	0.01
	U	0.07	0.50	0.42	0.00	0.02	0.98	0.24	0.72	0.04	0.44	0.56	0.00	0.36	0.30	0.34	0.51	0.49	0.00	0.45	0.48	0.07	0.45	0.48	0.07	0.43	0.37	0.20
4	P	0.01	0.99	0.00	0.00	0.81	0.19	0.02	0.98	0.00	0.07	0.93	0.00	0.05	0.95	0.00	0.09	0.92	0.00	0.06	0.94	0.00	0.06	0.94	0.00	0.19	0.80	0.00
	U	0.00	0.57	0.43	0.00	0.04	0.96	0.02	0.98	0.01	0.07	0.93	0.00	0.03	0.13	0.84	0.08	0.92	0.00	0.05	0.61	0.33	0.05	0.62	0.33	0.18	0.36	0.46
5	P	0.01	0.99	0.00	0.00	0.80	0.20	0.03	0.97	0.00	0.11	0.89	0.00	0.08	0.92	0.00	0.13	0.87	0.00	0.10	0.90	0.00	0.10	0.90	0.00	0.23	0.76	0.00
	U	0.00	0.52	0.47	0.00	0.01	0.99	0.03	0.96	0.01	0.10	0.90	0.00	0.05	0.09	0.86	0.13	0.87	0.00	0.09	0.55	0.37	0.09	0.55	0.36	0.21	0.30	0.49
6	P	0.00	0.47	0.52	0.00	0.11	0.89	0.02	0.88	0.10	0.07	0.93	0.00	0.04	0.64	0.32	0.08	0.92	0.00	0.06	0.86	0.09	0.06	0.86	0.08	0.18	0.73	0.09
	U	0.00	0.14	0.86	0.00	0.01	0.99	0.02	0.79	0.19	0.07	0.93	0.00	0.02	0.25	0.73	0.07	0.88	0.04	0.04	0.43	0.53	0.04	0.44	0.52	0.17	0.41	0.42
7	P	0.00	0.19	0.81	0.00	0.03	0.97	0.01	0.62	0.37	0.06	0.93	0.01	0.03	0.70	0.26	0.07	0.92	0.01	0.05	0.86	0.09	0.05	0.87	0.09	0.19	0.76	0.05
	U	0.00	0.06	0.94	0.00	0.00	1.00	0.01	0.55	0.44	0.06	0.93	0.01	0.02	0.50	0.48	0.07	0.90	0.03	0.04	0.67	0.29	0.04	0.67	0.29	0.18	0.63	0.19
8	P	0.01	0.49	0.49	0.00	0.04	0.96	0.07	0.86	0.07	0.22	0.78	0.00	0.16	0.80	0.03	0.26	0.74	0.00	0.20	0.79	0.00	0.20	0.79	0.00	0.29	0.70	0.01
	U	0.00	0.22	0.78	0.00	0.00	1.00	0.06	0.85	0.09	0.22	0.78	0.00	0.11	0.45	0.44	0.26	0.74	0.00	0.18	0.55	0.27	0.18	0.56	0.26	0.25	0.48	0.27

Table 7: Rates of underfitting (U), correct fitting (C) and overfitting (O) by misspecification type and method for the penalized (P) and the unpenalized (U) estimator with respect to the true number of components in the generating model.

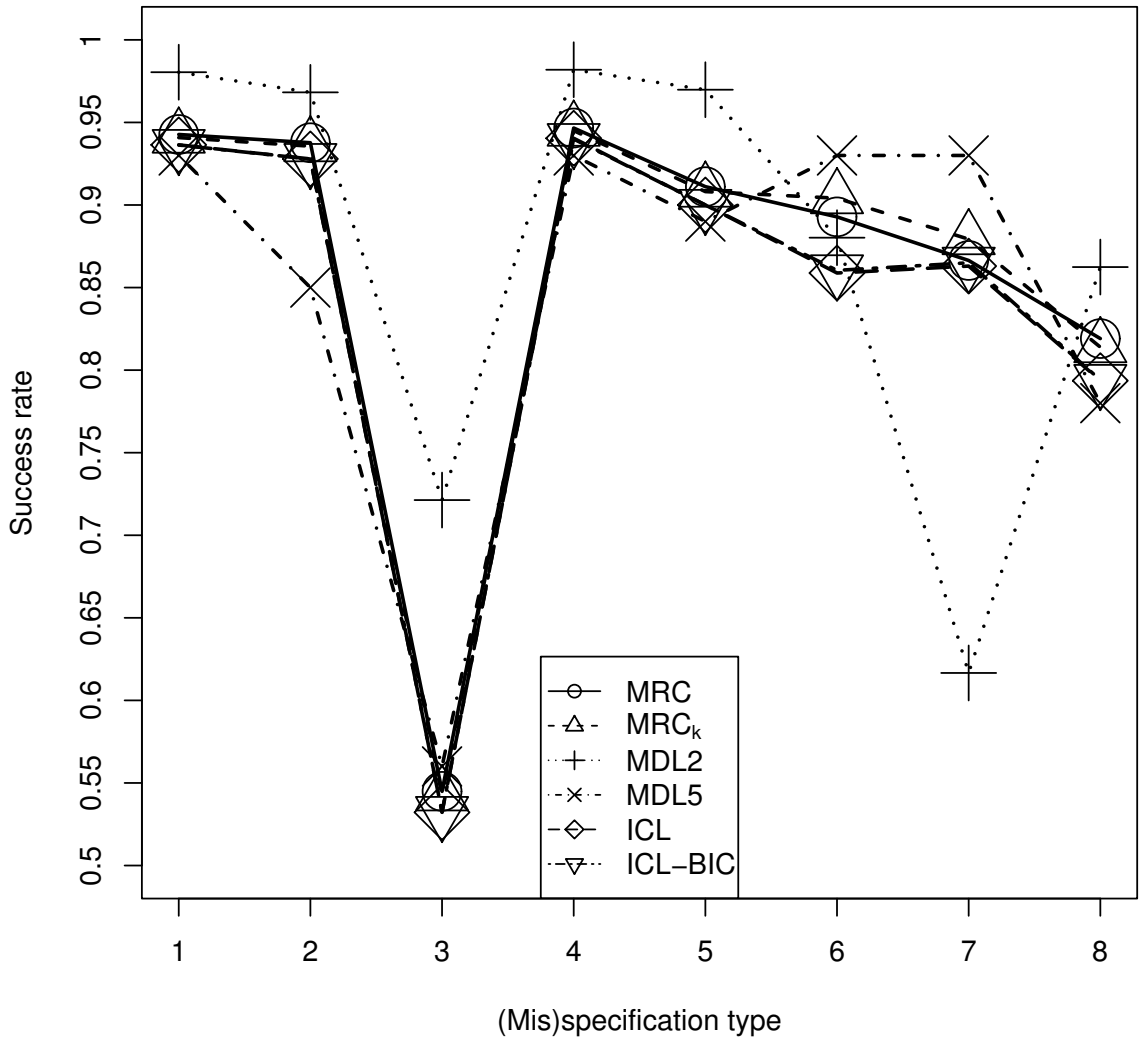

Figure 3: Success rates of best performing criteria using the penalized estimator for the different model specifications.

	AIC			AIC_c			MRC			KIC			KIC_c			MRC_k			AIC4			CAIC			HQ		
Type	U	C	O	U	C	O	U	C	O	U	C	O	U	C	O	U	C	O	U	C	O	U	C	O	U	C	O
1	0.03	0.56	0.42	0.03	0.60	0.37	0.08	0.92	0.00	0.03	0.81	0.16	0.03	0.84	0.13	0.08	0.92	0.00	0.03	0.94	0.03	0.04	0.96	0.00	0.03	0.91	0.06
2	0.01	0.38	0.61	0.01	0.48	0.51	0.07	0.93	0.00	0.01	0.66	0.33	0.01	0.72	0.27	0.07	0.93	0.00	0.01	0.86	0.13	0.02	0.98	0.00	0.01	0.77	0.22
3	0.08	0.36	0.56	0.10	0.39	0.51	0.82	0.18	0.01	0.21	0.49	0.30	0.24	0.50	0.26	0.82	0.17	0.01	0.35	0.51	0.14	0.62	0.37	0.01	0.30	0.51	0.20
4	0.03	0.55	0.42	0.03	0.60	0.37	0.08	0.92	0.00	0.03	0.81	0.17	0.03	0.84	0.13	0.08	0.92	0.00	0.03	0.94	0.03	0.04	0.96	0.00	0.03	0.91	0.06
5	0.02	0.53	0.45	0.03	0.58	0.40	0.11	0.89	0.00	0.03	0.80	0.17	0.03	0.84	0.13	0.11	0.89	0.00	0.03	0.94	0.03	0.04	0.96	0.00	0.03	0.90	0.07
6	0.16	0.30	0.55	0.18	0.30	0.52	0.77	0.21	0.02	0.26	0.33	0.41	0.29	0.34	0.38	0.77	0.21	0.02	0.35	0.39	0.26	0.54	0.39	0.06	0.32	0.36	0.32
7	0.10	0.57	0.34	0.12	0.55	0.32	0.93	0.06	0.01	0.19	0.53	0.28	0.22	0.53	0.25	0.93	0.06	0.01	0.30	0.51	0.19	0.60	0.37	0.04	0.25	0.52	0.23
8	0.09	0.49	0.42	0.12	0.48	0.40	0.91	0.09	0.00	0.22	0.45	0.33	0.27	0.44	0.30	0.91	0.09	0.00	0.38	0.40	0.22	0.73	0.24	0.03	0.32	0.42	0.27

	BIC			aBIC			MDL2			MDL5			CLC			AWE			ICL			ICL-BIC			NEC		
Type	U	C	O	U	C	O	U	C	O	U	C	O	U	C	O	U	C	O	U	C	O	U	C	O	U	C	O
1	0.03	0.97	0.00	0.03	0.78	0.19	0.05	0.95	0.00	0.10	0.90	0.00	0.08	0.92	0.00	0.11	0.89	0.00	0.09	0.91	0.00	0.09	0.91	0.00	0.22	0.78	0.00
2	0.02	0.98	0.00	0.01	0.63	0.36	0.03	0.97	0.00	0.15	0.85	0.00	0.06	0.93	0.01	0.12	0.88	0.00	0.08	0.92	0.00	0.08	0.92	0.00	0.20	0.80	0.01
3	0.56	0.41	0.03	0.19	0.46	0.35	0.74	0.26	0.00	0.83	0.17	0.00	0.80	0.18	0.02	0.85	0.15	0.00	0.83	0.17	0.00	0.83	0.17	0.00	0.83	0.16	0.01
4	0.03	0.97	0.00	0.03	0.78	0.19	0.04	0.96	0.00	0.10	0.90	0.00	0.07	0.92	0.00	0.11	0.89	0.00	0.08	0.92	0.00	0.08	0.92	0.00	0.21	0.79	0.00
5	0.04	0.96	0.00	0.03	0.77	0.20	0.05	0.95	0.00	0.13	0.87	0.00	0.10	0.89	0.00	0.15	0.85	0.00	0.12	0.88	0.00	0.12	0.88	0.00	0.25	0.75	0.00
6	0.49	0.41	0.09	0.25	0.32	0.43	0.71	0.28	0.01	0.79	0.21	0.00	0.70	0.15	0.15	0.79	0.21	0.00	0.77	0.20	0.03	0.77	0.20	0.03	0.80	0.14	0.06
7	0.51	0.42	0.07	0.17	0.54	0.28	0.80	0.19	0.00	0.95	0.05	0.00	0.87	0.07	0.06	0.95	0.05	0.00	0.93	0.05	0.01	0.93	0.05	0.01	0.94	0.04	0.02
8	0.66	0.28	0.06	0.20	0.44	0.36	0.86	0.14	0.00	0.93	0.07	0.00	0.90	0.08	0.02	0.94	0.06	0.00	0.92	0.08	0.00	0.92	0.08	0.00	0.93	0.06	0.00

Table 8: Rates of underfitting (U), correct fitting (C) and overfitting (O) by misspecification type with respect to the maximum log likelihood in validation sample.

estimator with respect to the maximized log likelihood in the validation sample. An interesting pattern can be found here. The criteria which were very bad in selecting the true number of components, AIC, AIC_c, KIC, KIC_c, $AIC4$, HQ and $aBIC$, have relatively high success rates here for types 3, 6, 7 and 8 of model misspecification whereas MRC, MRC_k, $MDL2$, $MDL5$, CLC, AWE, ICL and $ICL\text{-}BIC$ do very bad here. In conclusion, a trade-off appears to be noticeable between selecting the number of components and selecting a model which predicts future samples best. Hence, AIC and its relatives in fact do what they are designed to do. Unfortunately, the success rates are not overwhelming ranging between 30% and 57%. Furthermore, there is no clear best criterion here too. Perhaps, with larger sample sizes, this performance would increase and if Burnham and Anderson (2002) are right in the sense that there don't exist any simple models (i.e. truth has nearly an infinite number of parameters), the AIC family of efficient selection criteria would be preferred. However, selecting the correct number of components can also be very important and we feel it would be preferrable to remedy misspecification by data transformations or different model specifications rather than by adding components which are not represented in the population.

In table 9 the results are presented at a lower level of detail for the case where the true number

Type	1		2		3		4		5		6		7		8	
n	300	600	300	600	300	600	300	600	300	600	300	600	300	600	300	600
AIC	0.02	0.01	0.00	0.00	0.06	0.04	0.02	0.01	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00
AIC_c	0.04	0.01	0.00	0.00	0.07	0.04	0.04	0.02	0.05	0.01	0.00	0.00	0.00	0.00	0.00	0.00
MRC	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.94	0.96	0.88	0.99	0.99	1.00
KIC	0.53	0.40	0.01	0.00	0.24	0.18	0.47	0.40	0.51	0.41	0.01	0.00	0.00	0.00	0.00	0.00
KIC_c	0.62	0.45	0.05	0.01	0.26	0.19	0.58	0.45	0.60	0.46	0.01	0.00	0.00	0.00	0.00	0.00
MRC_k	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96	0.97	0.89	0.99	1.00	1.00
$AIC4$	0.92	0.87	0.52	0.36	0.36	0.27	0.92	0.87	0.91	0.86	0.07	0.00	0.00	0.00	0.01	0.00
$CAIC$	1.00	1.00	1.00	1.00	0.61	0.41	1.00	1.00	1.00	1.00	0.42	0.08	0.01	0.00	0.34	0.00
HQ	0.77	0.79	0.13	0.15	0.31	0.26	0.75	0.79	0.74	0.78	0.03	0.00	0.00	0.00	0.00	0.00
BIC	1.00	1.00	1.00	1.00	0.51	0.38	1.00	1.00	1.00	1.00	0.28	0.04	0.00	0.00	0.14	0.00
$aBIC$	0.22	0.54	0.00	0.01	0.15	0.21	0.19	0.56	0.21	0.56	0.00	0.00	0.00	0.00	0.00	0.00
$MDL2$	1.00	1.00	1.00	1.00	0.96	0.66	1.00	1.00	1.00	1.00	0.91	0.53	0.29	0.00	0.96	0.35
$MDL5$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.89	1.00	1.00
CLC	0.99	0.99	0.97	0.99	0.99	1.00	0.98	0.99	0.98	0.99	0.56	0.90	0.78	0.99	0.97	1.00
AWE	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.99	1.00	1.00	1.00
ICL	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.88	0.96	0.91	0.99	1.00	1.00
$ICL\text{-}BIC$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.87	0.96	0.91	0.99	1.00	1.00
NEC	0.99	0.99	0.97	0.99	0.99	1.00	0.98	0.99	0.98	0.99	0.56	0.90	0.78	0.99	0.97	1.00

Table 9: Success rates with respect to the true number of components when $K^* = 1$.

of components is 1 ($K^* = 1$). Order selection in this case entails the important decision whether there is actually heterogeneity in the population in the form of multiple groups or not. It can be seen that the best performing criteria here are AWE, $MDL5$, MRC_k, MRC, ICL and $ICL\text{-}BIC$. It should however be noted that the performance of these criteria in this case is not a completely reliable quality measure as a success rate of 100% can be achieved by making the penalty term on the number of parameters large enough. On the other side of the spectrum one can see that AIC and AIC_c perform dreadfully as they overfit in nearly every case. A curious result is that the performance of several criteria decreases or does not increase when the sample size is larger among all types of true model specification. These criteria are AIC, AIC_c, KIC, KIC_c, $AIC4$, $CAIC$, BIC and $MDL2$. This is not a desirable result as more information should lead to better inference. With respect to the type of model misspecification, one can observe that even without any misspecification, AIC, AIC_c, KIC, KIC_c and $aBIC$ have poor performances.

Among the selection criteria that consistently perform well, there is not much of a drop comparing no misspecification to the various types of misspecification. Furthermore, it can be seen that dropping an explanatory variable has a smaller negative impact than error misspecification. This is a logical result as the criteria can only make a mistake in one direction, i.e. overfitting, and dropping an explanatory variable generally increases the rate of underfitting.

In order to study the results in more detail, table 10 presents the odds ratios for the experimen-

Factor	n	K^*	$\boldsymbol{\pi}$	type							min s	max est
Odds ratio	300 vs 600	2 vs 3	equal vs unequal	2 vs 1	3 vs 1	4 vs 1	5 vs 1	6 vs 1	7 vs 1	8 vs 1	-	-
AIC	1.13*	0.25*	1.03	0.40*	0.05*	0.82*	1.00	0.03*	0.01*	0.01*	1.11*	0.27*
AIC_c	1.62*	0.22*	1.04	0.53*	0.05*	0.86*	1.00	0.02*	0.01*	0.02*	1.10*	0.29*
MRC	0.92*	1.27*	0.91*	0.98	0.02*	0.56*	0.57*	0.52*	0.37*	0.20*	2.06*	0.81*
KIC	2.06*	0.26*	1.08*	0.42*	0.03*	0.85*	0.94	0.01*	0.00*	0.01*	1.05*	0.36*
KIC_c	3.67*	0.23*	1.08*	0.56*	0.02*	0.94	1.01	0.01*	0.00*	0.01*	1.04*	0.40*
MRC_k	0.91*	1.33*	0.92*	0.98	0.01*	0.54*	0.56*	0.61*	0.44*	0.19*	2.18*	0.82*
AIC4	4.85*	0.32*	1.15*	0.54*	0.01*	0.93	0.96	0.00*	0.00*	0.00*	0.99	0.49*
CAIC	3.18*	0.47*	1.31*	0.69*	0.04*	0.95	0.69*	0.02*	0.01*	0.04*	1.01	0.68*
HQ	2.83*	0.29*	1.11*	0.46*	0.02*	0.92	0.95	0.01*	0.00*	0.00*	1.01	0.43*
BIC	3.48*	0.39*	1.31*	0.71*	0.02*	0.98	0.71*	0.01*	0.00*	0.02*	0.99*	0.63*
aBIC	0.60*	0.25*	1.06*	0.45*	0.03*	0.91	0.98	0.01*	0.00*	0.01*	1.05*	0.30*
MDL2	1.11*	1.09*	1.31*	0.66*	0.06*	0.88	0.67*	0.32*	0.07*	0.31*	1.19*	0.78*
MDL5	0.15*	7.05*	1.34*	0.30*	0.01*	0.40*	0.54*	1.33*	1.47*	0.11*	2.95*	0.69*
CLC	0.65*	1.37*	0.92*	0.95	0.04*	0.81*	0.66*	0.11*	0.13*	0.25*	1.30*	0.73*
AWE	0.50*	2.40*	0.98	0.69*	0.00*	0.37*	0.47*	1.26*	1.35*	0.10*	3.18*	0.75*
ICL	0.79*	1.39*	0.93*	0.93	0.02*	0.60*	0.57*	0.43*	0.41*	0.20*	1.84*	0.81*
ICL-BIC	0.79*	1.39*	0.93*	0.93	0.02*	0.59*	0.57*	0.44*	0.42*	0.20*	1.85*	0.81*
NEC	1.24*	11.20*	0.93*	0.79*	0.02*	0.51*	0.59*	0.55*	0.52*	0.20*	2.04*	2.01*

Table 10: Odds ratios of selecting the true number of components by logistic regression for $K^* = 2, 3$. Entries marked with a * are significant at 5%. Min s denotes the minimum pairwise separation and max est denotes the highest component model for which the estimation converged to an acceptable solution.

tal factors for the cases where $K^* > 1$. These odds ratios were calculated from logistic regressions for each order selection method[12]. Correctly selecting the true number of components was taken as a success. Two of the factors in the model warrant some clarification. First, the minimum separation between the components is included in the models. In case $K^* = 2$ this is simply the separation between components 1 and 2. In case $K^* = 3$, the minimum of the three pairwise separations is taken because the components for which the separation is minimal will be harder to separate. Second, the factor 'max est' represents the maximum number of components for which a proper solution was found and is taken as a continuous effect. This factor was included in the models as it limits the possible amount of overfitting. To illustrate the interpretation of the table entries, consider the estimated odds ratio of AIC with respect to the sample size factor n. This odds ratio was estimated at 1.127 and indicates that the odds of a success, i.e. selecting the true number of components, when using AIC was approximately 1.13 times larger in a sample of size 300 than in a sample of size 600, controlling for the other experimental factors. Table entries marked by a $*$ are significantly different from 1 at a significance level of 5%. Similar to the case where $K^* = 1$, AIC, AIC_c, KIC, KIC_c, $AIC4$, $CAIC$, BIC and $MDL2$ perform worse in larger samples. This group is joined by HQ and NEC. The effect of the true number of components is also very dissimilar across the different selection criteria and several of the estimated odds ratios

[12]Presenting these results in a high dimensional contingency table would be unwieldy.

are very far from 1. Equal or unequal mixture proportions also have different effects across all methods but the size of these effects is much smaller than the effects of the sample size or the number of components. We can conclude that in most cases the criteria which performed better for smaller samples also perform better in case of equal mixture proportions and with 3 true components. Conversely, selection methods which perform better for larger samples tend to perform better in case of unequal mixture proportions and with 2 true components. This distinction largely coincides with a criterion's proneness to respectively overfit or underfit. It can be noted that the odds of successfully selecting the true number of components increase when the minimum separation increases as would be expected. The criteria which do not perform well across experimental conditions seem to be less affected by the separation however (BIC and $AIC4$ even performed better when the minimum separation was smaller). Furthermore, the performance of all criteria decreased as the range of models which could be fitted increased. Again, there is an exception here, namely NEC, the criterium which showed the highest rate of underfitting across all types of model specification which would seem to indicate that this selection method is highly conservative. Focussing on the group of order selection criteria which, on average, performed best (MRC, MRC_k, $MDL2$, $MDL5$, ICL and $ICL\text{-}BIC$), one can see that, controlling for all other factors, the effect of the various model misspecifications compared to no model misspecification is much larger than it appeared earlier. Including superfluous explanatory variables strongly affects $MDL2$ and $MDL5$. Omitting a relevant explanatory variable has a very large negative effect on all these criteria and $MDL2$ was least affected here. The effect of excluding a real interaction and multicollinearity seems to be largely similar across these methods and again, $MDL2$ appears to be most robust here. For most of these criteria, heteroskedasticity within the components seems to have the largest negative effect of all error misspecifications with the exception of $MDL2$, which seems more affected by skewed errors. Curiously enough, $MDL5$ actually performed better for heavier tailed or skewed error specifications relative to no misspecification. This would indicate that such misspecifications counter $MDL5$'s tendency to underfit due to its large penalty term. On the other hand, this criterion was affected most by the heteroskedastic errors.

5. Conclusion

Order selection in finite mixture models is not a simple problem which seems to be confirmed in our simulation. Different experimental settings influence the order selection criteria differently. Some results however are obtained on which criteria seek to select the number of components rather than minimizing the expected prediction error. For order selection it appears that the newly developed mixture criteria (MRC and MRC_k) perform rather well on most occasions. Similar things can be said about $MDL2$, $MDL5$, ICL and $ICL\text{-}BIC$. The traditional model selection criteria, AIC, AIC_c, BIC and $aBIC$ on the other hand performed very poorly. Therefore, based on our findings, we would recommend using selection methods which have been specifically derived for finite mixture models or the lesser known $MDL2$ and $MDL5$. Furthermore, there is some evidence that including irrelevant explanatory variables, excluding interaction effects or multicollinearity are not very detrimental to order selection if one chooses a correct criterion. Not including an important explanatory variable on the other hand does have a substantial negative effect on all criteria. We have also found that distributional misspecification of the error terms has a non-uniform effect on the selection criteria. In conclusion, we found that none of the selection criteria was robust to every sort of misspecification we tested. A limitation of our

simulation was that all misspecifications were present in all components. Furthermore, we only tested for one particular 'amount' of misspecification each time, rather than a range of mild to severe misspecifications. Both of these settings could be interesting avenues to explore further. There is one constant positive effect present in all our results: it pays to penalize. We have found that appropriately penalizing the likelihood resulted in fewer spurious solutions. This had a positive effect on the estimation error of the model parameters and on the performance of the order selection criteria. Obviously, in practical situations it would be recommended to study all local solutions which have been found in detail. Nevertheless, we think it would be useful to further investigate the choice of penalizing constant(s), the data dependent element(s) and the functional form of this penalty function.

6. References

R. Abbi, E. El-Darzi, C. Vasilakis, and P. Millard. Analysis of stopping criteria for the EM algorithm in the context of patient grouping according to length of stay. In *2008 4th International IEEE Conference Intelligent Systems*, pages 9–14, Varna, Bulgaria, 2008. IEEE.

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

R. L. Andrews and I. S. Currim. A comparison of segment retention criteria for finite mixture logit models. *Journal of Marketing Research*, 40(2):235–243, 2003a.

R. L. Andrews and I. S. Currim. Retention of latent segments in regression-based marketing models. *International Journal of Research in Marketing*, 20(4):315–321, 2003b.

K. Bandeen-Roche, D. L. Miglioretti, S. L. Zeger, and P. J. Rathouz. Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*, 92(440):1375–1386, 1997.

J. D. Banfield and A. E. Raftery. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49(3):803, 1993.

R. J. Bhansali and D. Y. Downham. Some properties of the order of an autoregressive model selected by a generalization of Akaike's EPF criterion. *Biometrika*, 64(3):547, 1977.

C. Biernacki and G. Govaert. Using the classification likelihood to choose the number of clusters. *Computing Science and Statistics*, 29(2):451–457, 1997.

C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated classification likelihood. Technical Report 3521, No. 3521. Rhône-Alpes:INRIA, 1998.

C. Biernacki, G. Celeux, and G. Govaert. An improvement of the NEC criterion for assessing the number of clusters in a mixture model. *Pattern Recognition Letters*, 20(3):267–272, 1999.

C. Biernacki, G. Celeux, and G. Govaert. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics & Data Analysis*, 41(3-4):561–575, 2003.

D. Böhning, E. Dietz, R. Schaub, P. Schlattmann, and B. G. Lindsay. The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics*, 46(2):373–388, 1994.

H. Bozdogan. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370, 1987.

H. Bozdogan. Choosing the number of component clusters in the mixture-model using a new informational complexity criterion of the Inverse-Fisher information matrix. In O. Opitz, B. Lausen, and R. Klar, editors, *Information and Classification*, pages 40–54. Springer-Verlag, Heidelberg, 1993.

K. P. Burnham and D. R. Anderson. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer, 2nd edition, 2002.

J. E. Cavanaugh. A large-sample model selection criterion based on Kullback's symmetric divergence. *Statistics & Probability Letters*, 42(4):333–343, 1999.

Joseph E. Cavanaugh. Criteria for linear model selection based on Kullback's symmetric divergence. *Australian New Zealand Journal of Statistics*, 46(2):257–274, 2004.

G. Celeux and G. Soromenho. An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, 13(2):195–212, 1996.

J. Chen and X. Tan. Inference for multivariate normal mixtures. *Journal of Multivariate Analysis*, 100(7):1367–1383, 2009.

J. Chen, X. Tan, and R. Zhang. Inference for normal mixtures in mean and variance. *Statistica Sinica*, 18(2):443–465, 2008.

G. Ciuperca, A. Ridolfi, and J. Idier. Penalized maximum likelihood estimator for normal mixtures. *Scandinavian Journal of Statistics*, 30(1):45–59, 2003.

A. Cutler and M. P. Windham. Information-based validity functionals for mixture analysis. In H Bozdogan, editor, *Proceedings of the First US/Japan Conference for Mixture Analysis*, pages 149–170. Amsterdam: Kluwer, 1994.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1): 1–38, 1977.

W. S. Desarbo and W. L. Cron. A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, 5(2):249–282, 1988.

J. G. Dias. Performance Evaluation Of Information Criteria For The Naive-Bayes Model In The Case Of Latent Class Analysis: A Monte Carlo Study. *Journal of the Korean Statistical Society*, 36(3):435–445, 2007.

M. Falk. A simple approach to the generation of uniformly distributed random variables with prescribed correlations. *Communications in Statistics - Simulation and Computation*, 28(3): 785–791, 1999.

Allen I. Fleishman. A method for simulating non-normal distributions. *Psychometrika*, 43(4): 521–532, 1978.

J. R. S. Fonseca and M. G. M. S. Cardoso. Mixture-model cluster analysis using information theoretical criteria. *Intelligent Data Analysis*, 11(2):155–173, 2007.

B. Garel. Recent asymptotic results in testing for mixtures. *Computational Statistics & Data Analysis*, 51(11):5295–5304, 2007.

J. K. Ghosh and P. K. Sen. On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Vol 2*, number 1467, pages 789–806, Monterey: Wadsworth, 1985. Citeseer.

B. Hafidi and A. Mkhadri. The Kullback information criterion for mixture regression models. *Statistics & Probability Letters*, 80(9-10):807–815, 2010.

E. J. Hannan and B. G. Quinn. The determination of the order of an autoregression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):190–195, 1979.

R. J. Hathaway. A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *The Annals of Statistics*, 13(2):795–800, 1985.

R. J. Hathaway. Another interpretation of the EM algorithm for mixture distributions. *Statistics & Probability Letters*, 4(2):53–56, 1986.

D. S. Hawkins, D. M. Allen, and A. J. Stromberg. Determining the number of components in mixtures of linear models. *Computational Statistics & Data Analysis*, 38(1):15–48, 2001.

T. C. Headrick. Fast fifth-order polynomial transforms for generating univariate and multivariate nonnormal distributions. *Computational Statistics & Data Analysis*, 40(4):685–711, 2002.

C. M. Hurvich and C.-L. Tsai. Regression and time series model selection in small samples. *Biometrika*, 76(2):297, 1989.

W. James and C. Stein. Estimation with quadratic loss. In J Neyman, editor, *Proceedings Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume I*, volume 1, pages 361–379. University of California Press, 1961.

K. Jedidi, H. S. Jagpal, and W. S. DeSarbo. Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity. *Marketing Science*, 16(1):39–59, 1997.

D. Karlis and E. Xekalaki. Choosing initial values for the EM algorithm for finite mixtures. *Computational Statistics & Data Analysis*, 41(3-4):577–590, 2003.

S. Konishi and G. Kitagawa. Generalized information criteria in model selection. *Biometrika*, 83 (4):875–890, 1996.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

Z. Liang, R. J. Jaszczak, and R. E. Coleman. Parameter estimation of finite mixtures using the EM algorithm and information criteria with application to medical image processing. *IEEE Transactions on Nuclear Science*, 39(4):1126–1133, 1992.

M. J. Lindstrom and D. M. Bates. Newton-Raphson and EM algorithms for linear mixed models for repeated-measures data. *Journal of the American Statistical Association*, 83(404):1014–1022, 1988.

G. H. Lubke and M. C. Neale. Distinguishing between latent classes and continuous factors: Resolution by maximum likelihood? *Multivariate Behavioral Research*, 41(4):499–532, 2006.

J. S. Marron and M. P. Wand. Exact mean integrated squared error. *The Annals of Statistics*, 20 (2):712–736, 1992.

G. J. McLachlan. On bootstrapping the likelihood ratio test stastistic for the number of components in a normal mixture. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 36(3):318–324, 1987.

G. J. Mclachlan and T. Krishnan. *The EM algorithm and extensions*. Wiley-Interscience, 2nd edition, 2008.

G. J. McLachlan and S. K. Ng. A comparison of some information criteria for the number of components in a mixture model. Technical report, University of Queensland, Brisbane, 2000.

G. J. Mclachlan and D. Peel. *Finite mixture models*. Wiley-Interscience, 2000.

P. A. Naik, P. Shi, and C.-L. Tsai. Extending the Akaike information criterion to mixture regression models. *Journal of the American Statistical Association*, 102(477):244–254, 2007.

K. L. Nylund, T. Asparouhov, and B. O. Muthén. Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, 14(4):535–569, 2007.

A. Oliveira-brochado and F. V. Martins. Determining the Number of Market Segments Using an Experimental Design. 2008.

J. Rissanen. Stochastic complexity and modeling. *The Annals of Statistics*, 14(3):1080–1100, 1986.

M. Sarstedt. Market segmentation with mixture regression models: Understanding measures that guide model selection. *Journal of Targeting, Measurement and Analysis for Marketing*, 16(3): 228–246, 2008.

P. Schlattmann. *Medical applications of finite mixture models*. Statistics for Biology and Health. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

S. L. Sclove. Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, 52(3):333–343, 1987.

W. Seidel and H. Sevcikova. Types of likelihood maxima in mixture models and their implication on the performance of tests. *Annals of the Institute of Statistical Mathematics*, 41(4):85–654, 2004.

W. Seidel, K. Mosler, and M. Alker. Likelihood ratio tests based on subglobal optimization: A power comparison in exponential mixture models. *Statistical Papers*, 41(1):85–98, 2000a.

W. Seidel, K. Mosler, and M. Alker. A cautionary note on likelihood ratio tests in mixture models. *Annals of the Institute of Statistical Mathematics*, 52(3):481–487, 2000b.

D. M. Titterington, a. F. M. Smith, and U. E. Makov. *Statistical analysis of finite mixture distributions.*, volume 42. Wiley, 1985.

D. Tofighi and C. K. Enders. Identifying the correct number of classes in growth mixture models. In G. R. Hancock and K. M. Samuelsen, editors, *Advances in Latent Variable Mixture Models*, pages 317–341. Information Age Publishing Inc., 2008.

K. Viele and B. Tong. Modeling with mixtures of linear regressions. *Statistics and Computing*, 12:315–330, 2002.

M. Wedel and W. A. Kamakura. *Market segmentation: Concepts and methodological foundations.* Springer Verlag, 1999.

C. Yang. Evaluating latent class analysis models in qualitative phenotype identification. *Computational Statistics & Data Analysis*, 50(4):1090–1104, 2006.

C.-C. Yang. Separating latent vlasses by information criteria. *Journal of Classification*, 24:183–203, 2007.

Y. Yang. Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950, 2005.