Economics Bulletin

Volume 40, Issue 2

Determining the number of components in mixture regression models: an experimental design

Ana Brochado
Instituto Universitário de Lisboa (ISCTE-IUL), DINÂMIA CET – IUL

Vitorino Martins
Faculdade de Economia da Universidade do Porto (FEP-UP)

Abstract

Despite the popularity of mixture regression models, the decision of how many components to retain remains an open issue. This study thus sought to compare the performance of 26 information and classification criteria. Each criterion was evaluated in terms of that component's success rate. The research's full experimental design included manipulating 9 factors and 22 levels. The best results were obtained for 5 criteria: Akaike information criteria 3 (AIC3), AIC4, Hannan-Quinn information criteria, integrated completed likelihood (ICL) Bayesian information criteria (BIC) and ICL with BIC approximation. Each criterion's performance varied according to the experimental conditions.

Citation: Ana Brochado and Vitorino Martins, (2020) "Determining the number of components in mixture regression models: an experimental design", *Economics Bulletin*, Volume 40, Issue 2, pages 1465-1474
Contact: Ana Brochado - ana.brochado@iscte-iul.pt, Vitorino Martins - vmartins@fep.up.pt.
Submitted: February 11, 2020. **Published:** June 02, 2020.

1. Introduction

Models for mixture regression of normal data facilitate model-based clustering within econometric analysis. These models have attracted considerable interest in recent years, resulting in a vast number of both methodological developments and applications in different areas. Mixture models have been employed in economics and finance (e.g. Anh *et al.* 2018, Wu 2019). For instance, Yan and Han (2019) modelled stock returns, and Wu (2019) employed mixtures of Gaussian processes to model electricity demand and crop insurance. Despite the growing interest in mixture models, the best way to decide how many mixture components are required to represent the data adequately still needs more research (e.g. Kasahara and Shimotsu 2015, Li *et al.* 2016, Nasserinejad *et al.* 2017).

Determining the number of components is a necessary step in addressing unobserved heterogeneity across objects (Gu *et al*. 2018, Pan *et al*. 2020). A misspecification of model results in terms of under- or overestimation of the number of mixture components can mislead experts during decision-making processes.

Previous studies have developed a large number of criteria to help identify the mixture components needed. Information criteria (e.g. Akaike 1973) are used to balance the increase in model fit to the data against a larger number of parameters, while classification criteria can ensure that the solution produces well-separated clusters. A substantial majority of previous studies, with a few notable exceptions (e.g. Hawkins *et al.* 2001), have been designed to test new proposed measures' performance instead of comparing the existing components' performance. Therefore, the present study sought to compare the performance of 26 information and classification criteria in terms of identifying the correct number of mixture components and avoiding an expensive experimental design. This research adds to the existing literature (e.g. Hawkins *et al.* 2001) by comparing a larger number of criteria under a larger number of experimental conditions.

The paper below is organised as follows. The next section starts by briefly describing the theoretical background of mixture regression of normal data. This section also provides a general review of the information and classification criteria considered in the current study. In section three, the experimental design used to generate the simulated data is delineated. Section four discusses the results and presents some guidelines that guarantee, by applying the criteria in question, a more accurate decision about the number of components to include in mixture regression of normal data.

2. Theory/Calculation

2.1 Mixture Regression Model Based on Multivariate Normal Distribution with Notations

Finite mixture regressions facilitate the performance of model-based clustering for large heterogeneous populations. According to this approach, a sample of observations can be obtained with a specific set of unknown proportions. The finite mixture technique is used to decompose the sample into its mixture components.

A classical mixture regression model (Wedel and DeSarbo 1995) simultaneously allows the probabilistic classification of observations in terms of S mixture components and the estimation of separate regression models relating covariates to the observers' expectations of dependent variables within latent classes. The underlying idea is that a single set of regression coefficients across all observations may be inadequate and potentially misleading if the observations arise from a number of unknown groups with differing coefficients. Brief

mathematical derivations of and notations on the multivariate normal mixture regression used in the present study are presented below.

In this methodology, $\boldsymbol{y}_n = (y_{nr})$ is a metric-dependent vector distributed as a finite mixture of *S* conditional multivariate normal densities as shown in Equation (1):

$$\boldsymbol{y}_n \sim \sum_{s=1}^{S} \lambda_s f_s(\mathbf{y}_n|\mathbf{x}_n, \boldsymbol{\beta}_s, \boldsymbol{\Sigma}_s) \quad (1)$$

in which f_s is defined by Equation (2):

$$f_s(\mathbf{y}_n|\mathbf{x}_n, \boldsymbol{\beta}_s, \boldsymbol{\Sigma}_s) = (2\pi)^{-R/2}|\boldsymbol{\Sigma}_s|^{-1/2}\exp[-1/2(\mathbf{y}_n - \mathbf{x}_n\boldsymbol{\beta}_s')\boldsymbol{\Sigma}_s^{-1}(\mathbf{y}_n - \mathbf{x}_n\boldsymbol{\beta}_s')'] \quad (2)$$

and λ_s, $s = 1,...,S$ define independent mixing proportions that satisfy two restrictions:

$$0 \le \lambda_s \le 1 \text{ and } \sum_{s=1}^{S} \lambda_s = 1 \quad (3)$$

and in which:

s = 1,..., *S* indicate the mixture components obtained

n = 1,..., *N* indicate cross-sectional units *c* (e.g. consumers, companies, countries and regions)

r = 1,..., *R* denote the repeated observations of object *n*

j = 1,..., *J* represent the independent variables

β_{js} = value of j^{th} regression coefficient for the s^{th} mixture component

$\boldsymbol{\beta}_s = (\beta_j)$ (Jx1) is the vector of regression coefficients for the s^{th} mixture component

$\boldsymbol{\Sigma}_s$ = (RxR) is the covariance matrix of mixture component *s*

y_{nr} = the value of dependent variable *y* for repeated measure *r* on cross-sectional unit *n*

$\mathbf{y}_n = (y_{nr})$ (Rx1) vector for object *n*

x_{nrj} = the value of the j^{th} independent variable for repeated measure *r* on cross-sectional unit *n*

$\boldsymbol{x}_n = (x_{rj})$ (RxJ) is the matrix of independent variables for cross-sectional unit *n*

Given a sample of N independent objects, the log likelihood is defined by Equation (4):

$$\ln L = \sum_{n=1}^{N} \ln \sum_{s=1}^{S} \lambda_s f_s\left(\mathbf{y}_n \middle| \mathbf{x}_n, \boldsymbol{\beta}_s, \boldsymbol{\Sigma}_s\right) \quad (4)$$

λ_s, $\boldsymbol{\Sigma}_s$ and $\boldsymbol{\beta}_s$ are estimated by maximising the likelihood that Equation (4) is subject to the restrictions in Equation (3).

The model is estimated with the expectation-maximisation (EM) algorithm (Dempster *et al.* 1997). This algorithm considers non-observed data to be defined as an indicator function z_{ns} that is assumed to be i.i.d multinomial. $z_{ns} = 1$ if cross-sectional unit *n* comes from latent class *s*, and $z_{ns} = 0$ otherwise.

The joint likelihood for the observed and non-observed data that are defined by $\mathbf{y}_n$ and $\mathbf{z}_n = (z_{ns})$ for all objects is represented as Equation (5):

$$\ln L_c = \sum_{n=1}^{N}\sum_{s=1}^{S} z_{ns} \ln\left[f_s\left(\mathbf{y}_n \middle| \mathbf{x}_n, \boldsymbol{\beta}_s, \boldsymbol{\Sigma}_s\right)\right] + \sum_{n=1}^{N}\sum_{s=1}^{S} z_{ns} \ln \lambda_s \quad (5)$$

Regarding the parameters' seed values, z_{ns} represents the missing data, and the EM algorithm's two steps – expectation step (E-step) and maximisation step (M-step) – are alternated until a convergent sequence of log-likelihood values is achieved. In the E-step, the expectation of Equation (5) is evaluated over the conditional distribution of the non-observed data z_{ns}, given $\boldsymbol{y}_n$, $\boldsymbol{x}_n$ and the estimates for λ_s, $\boldsymbol{\Sigma}_s$ and $\boldsymbol{\beta}_s$ obtained in the M-step. Each cross-sectional unit *n* is assigned to each latent class *S* via the estimated posterior probability (i.e. applying Bayes' rule), thereby incorporating fuzzy clustering into the E-step. The conditional expectation of z_{ns} is identical to that of p_{ns}.This can be written as Equation (6):

$$p_{ns} = \frac{\lambda_s f_s(\mathbf{y}_n|\mathbf{x}_n, \boldsymbol{\beta}_s, \boldsymbol{\Sigma}_s)}{\sum_{s=1}^{S} \lambda_s f_s(\mathbf{y}_n|\mathbf{x}_n, \boldsymbol{\beta}_s, \boldsymbol{\Sigma}_s)} \quad (6)$$

in which two conditions need to be satisfied:

$$\sum_{s=1}^{S} p_{ns} = 1 \text{ and } 0 \le p_{ns} \le 1 \tag{7}$$

In the E-step, the nonobserved discrete data z_{ns} are replaced by $p_{ns.}$ In the M-step, Equation (5) is maximised with respect to λ_s, $\boldsymbol{\Sigma}_s$ and $\boldsymbol{\beta}_{s.}$

2.2 Criteria

Information criteria are designed to account for over-parameterisation because likelihood increases with the addition of components to a mixture model when a larger number of mixture components are considered. These criteria balance the increase in the model fit to the data against the larger number of parameters estimated for models with more mixture components. The criteria take the general form of $\mathrm{IC}_{(s)} = -2\ln L + dk_{(s)}$, in which $k_{(s)}$ is the number of parameters associated with a solution with S mixture components and d is some constant or the 'marginal cost' per parameter (Bozdogan 1987). Information criteria encompass those that are estimates of relative Kullback-Leibler distance, including Akaike information criteria (AIC) (Akaike 1973), modified AIC 3 (Bozdogan 1994), modified AIC 4 (Bozdogan 1994) and $\mathrm{AIC_c}$ – small sample AIC (Hurvich and Tsai 1989, 1995), shown in Equations (8) through (11), respectively:

$$\mathrm{AIC} = -2\,ln\,L + 2k \tag{8}$$

$$\mathrm{AIC_3} = -2\,ln\,L + 3k \tag{9}$$

$$\mathrm{AIC_4} = -2\,ln\,L + 4k \tag{10}$$

$$\mathrm{AIC_c} = \mathrm{AIC} + \frac{[2k(k+1)]}{(N-k-1)} \tag{11}$$

Bayesian information criteria (BIC) (Schwartz 1978) and adjusted BIC (ABIC) (Sclove 1987) are developed within a Bayesian framework for model selection by using Equations (12) and (13):

$$\mathrm{BIC} = -2\,ln\,L + k\,ln\,N \tag{12}$$

$$\mathrm{ABIC} = -2\,ln\,L + k\,ln[(N+2)/24] \tag{13}$$

Information criteria also cover consistent criteria including consistent AIC (CAIC) (Bozdogan 1987), CAIC with Fisher information (CAICF) (Bozdogan 1987), information complexity criterion (ICOMP) (Bozdogan 1994), Hannan-Quinn (HQ) information criterion (Hannan and Quinn 1979), minimum description length 2 ($\mathrm{MDL_2}$)(Liang *et al.* 1992) and minimum description length 5 ($\mathrm{MDL_5}$) (Liang *et al.* 1992). These criteria are defined by Equations (14) through (19), respectively:

$$\mathrm{CAIC} = -2\ln L + k\left[\left(\ln N\right) + 1\right] \tag{14}$$

$$\mathrm{CAICF} = \mathrm{AIC} + k\log N + \log|\mathbf{F}| \tag{15}$$

$$\mathrm{ICOMP} = -2\,ln\,L + k\,ln\left[\frac{tr(F^{-1})}{k}\right] - ln|\boldsymbol{F}^{-1}| \tag{16}$$

$$\mathrm{HQ} = -2\,ln\,L + 2k\,ln(ln\,N) \tag{17}$$

$$\mathrm{MDL_2} = -2\,ln\,L + 2k\,ln\,N \tag{18}$$

$$\mathrm{MDL_5} = -2\,ln\,L + 5k\,ln\,N \tag{19}$$

in which F denotes the inverse Fisher information matrix.

Classification criteria evaluate a mixture model's ability to provide well-separated clusters of components based on the degree of separation in estimated posterior probabilities. Probabilistic indices include the measure of entropy (Es) (DeSarbo *et al.* 1992), logarithm of the partition probability (LP) (Biernacki 1997), entropy (E) (Liang *et al.* 1992), normalised entropy criterion (NEC) (Celeux and Soromenho 1996), classification criterion (C) (Biernacki and Govaert 1997), classification likelihood criterion (CLC) (Biernacki and Govaert 1997) and

approximate weight of evidence (AWE) (Banfield and Raftery 1993). Two other relevant indices are integrated completed likelihood (ICL)-BIC (Biernacki *et al.* 2000) and ICL with BIC approximation (ICOMPLBIC) (Dias 2004). Equations (20) through (28) represent each of these indices:

$$\text{Es} = 1 - \left[\sum_{n=1}^{N} \sum_{s=1}^{S} -p_{ns} \ln p_{ns} \right] \Big/ N \ln S \tag{20}$$

$$\text{LP} = -\sum_{n=1}^{N} \sum_{s=1}^{S} z_{ns} \ln p_{ns} \tag{21}$$

$$\text{E} = -\sum_{n=1}^{N} \sum_{s=1}^{S} p_{ns} \ln p_{ns} \tag{22}$$

$$\text{NEC}(s) = \text{E}(s) \big/ \ln L(s) - \ln L(1) \tag{23}$$

$$\text{C} = -2 \ln L + 2\text{E} \tag{24}$$

$$\text{CLC} = -2 \ln L + 2\text{LP} \tag{25}$$

$$\text{AWE} = -2 \ln L_c + 2k \left(3/2 + \ln N \right) \tag{26}$$

$$\text{ICL-BIC} = -2 \ln L + 2\text{LP} + k \ln N \tag{27}$$

$$\text{ICOMPLBIC} = -2 \ln L + 2\text{E} + k \ln N \tag{28}$$

A second group of indexes have been developed in the literature on fuzzy logic (Bezdek *et al*. 1997), including partition coefficient (PC) (Bezdek 1981), partition entropy (PE) (Bezdek 1981), normalised partition entropy (NPE) (Bezdek 1981), non-fuzzy index (NFI) (Roubens 1978), minimum hard tendency (Min_{ht}) (Rivera *et al.* 1990) and mean hard tendency (Mean_{ht}) (Rivera *et al.* 1990). These indexes are calculated using Equations (29) through (34), respectively:

$$\text{PC} = \sum_{n=1}^{N} \sum_{s=1}^{S} p_{ns}^2 \Big/ N \tag{29}$$

$$\text{PE} = \left[\sum_{n=1}^{N} \sum_{s=1}^{S} p_{ns} \ln p_{ns} \right] \Big/ N \tag{30}$$

$$\text{NPE} = \text{PE} \big/ \left[1 - S/N \right] \tag{31}$$

$$\text{NFI} = \left[S \left(\sum_{n=1}^{N} \sum_{s=1}^{S} p_{ns}^2 \right) - N \right] \Big/ \left[N(S-1) \right] \tag{32}$$

$$\text{Min}_{ht} = \max_{1 \le s \le S} \left\{ -\log_{10}(T_s) \right\} \tag{33}$$

$$\text{Mean}_{ht} = \sum_{s=1}^{S} -\log_{10}(T_s) \big/ S \tag{34}$$

in which T_s is the mean of all the first and second p_{ns} maxima for the cross-sectional units included in the mixture component s.

3. Methods

When dealing with real world data, the true number of mixture components is unknown, so the selected criteria's effectiveness as a guide during model selection cannot be evaluated without an experimental design. This research's experimental design included manipulating 9 factors and 22 levels. Factor 1 was the true number of mixture components (i.e. 2 or 3). Factor 2 was the mean separation between latent classes (i.e. low = 0.5; medium = 1.0; or high = 1.5). Factor 3 was the number of individuals (i.e. 100 or 300). Factor 4 was the number of observations per individual (i.e. 5 or 10). Factor 5 was the number of independent variables (i.e. 2, 6 or 10). Factor 6 was the levels of measurement for independent variables (i.e. binary, metric or half-binary and half-metric). Factor 7 was error variance (i.e. 20% or 60%). Factor 8 was minimum segment size (i.e. 5–10%, 10–20% or 20–30%). Factor 9 was error distribution (i.e. normal vs uniform).

As the design was factorial with three replications (i.e. datasets) per cell, the overall total generated was 7,776 (i.e. $2^5 3^5$). The likelihood was maximised using the EM algorithm, which

was run repeatedly with three replications in order to avoid its convergence to local maxima. Then, the best solution was retained for each possible number of mixture components.

For each object *n* and all replications, the programme generated U = Xβ. Next, an error was added to these true values (i.e. U, Y = U+ε). The error term's variance was obtained from Equation (35), in which *PEV* is the percent of error variance, σ_u^2 is the variance of *U* and σ_ε^2 is the variance of the error term *ε*:

$$PEV = \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \sigma_u^2} \Rightarrow \sigma_\varepsilon^2 = \left(\frac{PEV}{1 - PEV} \right) \sigma_u^2 \tag{35}$$

To improve the level of general inference, parameter values were randomly generated for each segment and each data set.

Regarding the coefficients for explanatory variables in each group (i.e. $\boldsymbol{\beta}_s = (\beta_{sj})$, $s = 1, ..., S$), the parameters' vector for the first segment $\boldsymbol{\beta}_1$ was randomly generated within the intervals -1.5 to 1.5. Next, a separation vector $\boldsymbol{\delta}_i$ ($i = L, M, H$) was generated, with an average interval of 0.5, 1.0 or 1.5 and a standard deviation equal to 10% of the average (i.e. 0.05, 0.1 and 0.15), as well as a signs vector $\mathbf{S}_-^+$, with positives (1) and negatives (-1), for $\boldsymbol{\delta}_i$ ($i = L, M, H$). The vectors with low (0.5), medium (1.0) and large (1.5) separations were denoted, respectively, by $\boldsymbol{\delta}_S$, $\boldsymbol{\delta}_M$ and $\boldsymbol{\delta}_E$. Vector coefficients for segments 2 and 3 were obtained using $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_1 + \mathbf{S}_-^+ \boldsymbol{\delta}_i$, $i = S, M, L$ and $\boldsymbol{\beta}_3 = \boldsymbol{\beta}_1 - \mathbf{S}_-^+ \boldsymbol{\delta}_i$, $i = B, M, E$, respectively.

Cross-sectional units were assigned to mixture components on the basis of randomly determined segment sizes for each data set. The smallest segments consisted of 5–10%, 10–20% or 20–30% of the sample, depending on Factor 7's level. Normal and uniform distributed errors (i.e. Factor 9) were generated with an average of 0 and standard deviation of 1.

The experimental design applied allowed each model to be fitted with the correct and incorrect number of latent classes. Information and classification criteria were calculated for each solution based on the estimated log-likelihoods and memberships after the EM estimators converged. Then, using the decision rules for information and classification criteria (i.e. maximise or minimise), the identified number of mixture components for each model was retained.

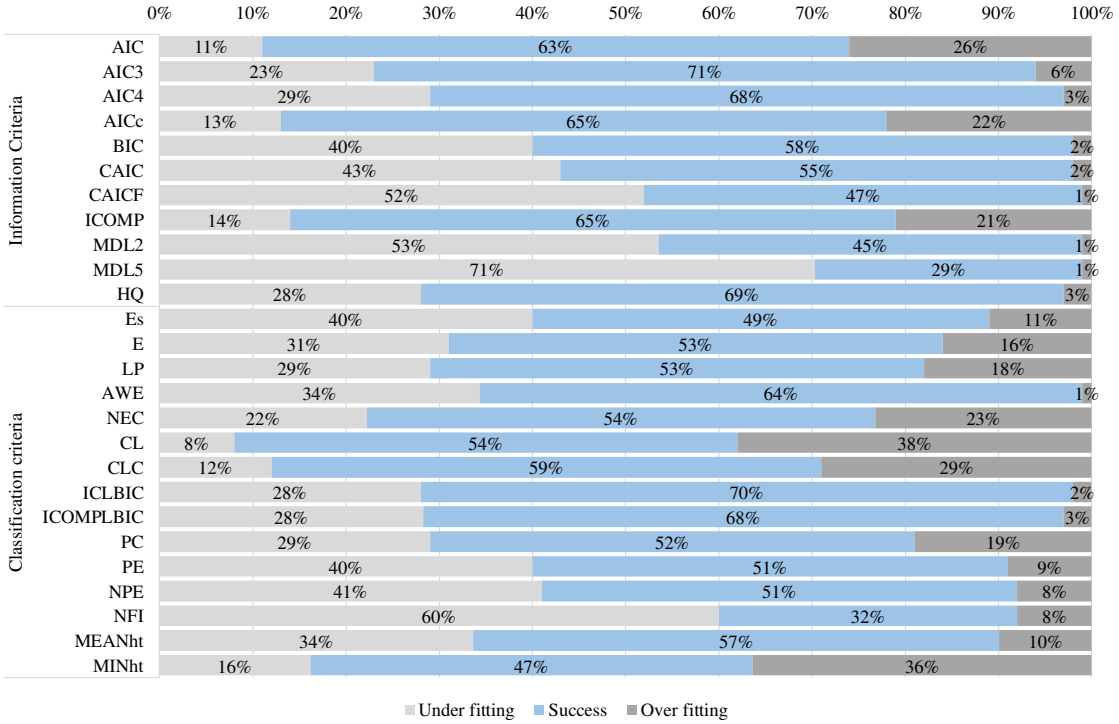
The information and classification criteria were compared in terms of their success rate, namely, the percentage of experimental datasets in which the criteria identified the correct number of latent classes out of a set of competing solutions. Given two criteria with the same success rate, underfitting was preferred to overfitting because previous studies have reported that overfitting produces more parameter bias than underfitting does and that overfitting results in extremely small latent classes with large or unstable parameter values (Cutler and Windham 1994).

4. Results and Discussion

4.1 Overall Results

The best overall success rates were obtained for 3 information criteria – AIC_3 (71%), HQ (69%) and AIC_4 (68%) – and 2 classification criteria – ICLBIC (70%) and ICOMPLBIC (68%). Some information criteria, such as AIC, AIC_c and ICOMP, showed a tendency to overestimate the number of mixture components, while others, such as MDL_5, MDL_2, CAICF, CAIC and BIC, tended more strongly towards underfitting than overfitting. In general, almost all the classification criteria (i.e. E_s, E, LP, AWE, NEC, ICL, ICLBIC, ICOMPLBIC, PC, PE, NPE, NFI and $MEAN_{ht}$) presented higher rates of underfitting than of overfitting (see Figure 1).

Figure 1. Overall success rates – underfitting and overfitting by information and classification criterion

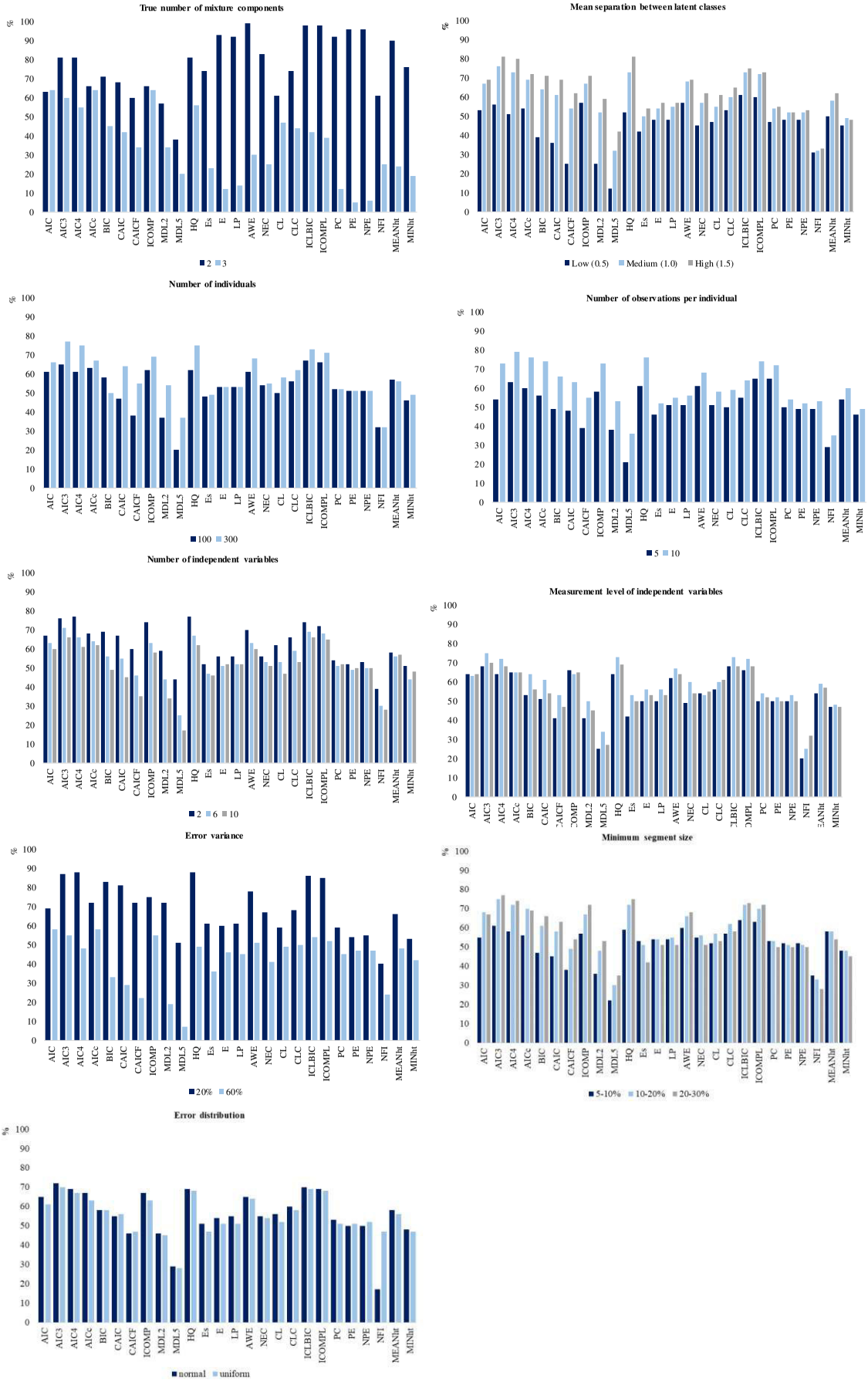


4.2 Results by Experimental Condition

An analysis of success rates by experimental condition revealed that no single criterion is best for all experimental conditions. For Factor 1, AWE (99%) registers the best success rate for a 2-mixture component solution, while AIC, AIC_c and ICOMP (64%) produces the best results for a 3-segment solution. With regard to Factor 2, ICLBIC (61%) registers the highest success rate for low separation between classes, but AIC_3 (76%) is best for medium – and both AIC_3 and HQ (81%) for high – separation between classes. In terms of Factor 3, ICLBIC (67%) produces the highest rate for 100 individuals, but AIC_3 registers the best performance (77%) for 300 individuals. The results for Factor 4 show that the criteria with the highest success rate for 5 observations per individual are ICLBIC and ICOMPLBIC (65%), and, for 10 observations, AIC_3 (79%) is the best.

Regarding Factor 5, the criteria that register the highest response rates for 2 independent variables are AIC_4 and HQ, but, for 6, AIC_3 (71%) and, for 10, AIC_3 and ICLBIC (66%) show the best performance. The results for Factor 6 reveal that, when the independent variables are all metric, the criteria AIC_3 and ICLBIC (68%) register the highest success rates, while AIC_3 (75%) is the best criteria for binary variables, as well as when the variables are half-metric and half-binary (70%). When Factor 7 is taken into account, AIC_4 and HQ (88%) are the best criteria for an error variance of 20% (R-squared [R^2] = 80%), and AIC and AIC_3 offer the highest success rates for an error variance of 40% (R^2 = 60%). With regard to Factor 8, the criteria best suited for a minimum segment size of 5–10% is ICLBIC (64%), but, at 10–20%, and 20–30%, AIC_3 (75% and 77%, respectively) has the highest success rates. Finally, the best criteria for Factor 9 is AIC3 (72.5% and 70%) for normal and uniform error, respectively (see Figure 2).

Figure 2. Success rates by criterion and experimental condition



5. Conclusion

The correct number of mixture components is unknown in real-world applications, so researchers rely on heuristic devices such as information- and classification-based criteria to help them select the best model. Therefore, a thorough understanding of these measures' performance across different data characteristics is of utmost importance. Researchers who take the wrong measures into consideration will conduct incorrect data analyses, which can subsequently contribute to imprudent managerial decisions.

The present study addressed this problem by comparing 26 classification and information criteria's performance, using an experimental design that manipulated 9 factors. The best overall success rates for all data sets were obtained for 3 information criteria (i.e. AIC_3, HQ and AIC_4) and 2 classification criteria (i.e. ICLBIC and ICOMPLBIC). In the context of multilevel mixture models, Lukočieė and Verut's (2009) study found that AIC_3 performs best.

The overall results reveal that data characteristics affect information and classification criteria's performance in terms of identifying the number of classes to include in mixture regression of normal data (Hawkins *et al.* 2001). More specifically, the results demonstrate that most information and classification criteria exhibit higher success rates when certain conditions are met. These include two mixture components, fewer explanatory variables, metric and mixed binary variables, a larger separation between latent classes, larger sample sizes, smaller error variance and normal distributed errors.

This study's findings provide researchers and practitioners with a better understanding of 26 criteria's effectiveness in terms of identifying the best number of classes to include in mixture regression of normal data. However, as no single criterion is able to identify the exact number of mixture components, additional research is needed to find the best criteria. In addition, significant room is left for improvement in current practices, and continued research is required on model selection criteria's performance to ensure more practical procedural guidelines.

Model selection decisions should be based on the data at hand and the results provided by information and classification criteria, as well as practical and theoretical considerations, in order to provide meaningful research conclusions. Various other research topics thus need to be considered, such as assessing the performance of criteria that address the problem of jointly selecting the number of components and variables in mixture regression models. The present study could also be extended further by considering different scenarios for distribution misspecifications to assess their influence on criteria's performance.

References

Akaike, H. (1973) "Information Theory as an Extension of the Maximum Likelihood Principle" in *Second International Symposium on Information Theory, Budapest, 1973*, by B.N. Petrov, F. Csaki and A. Kiado, Eds., 267-281

Anh, L.H., L.S. Dong, V. Kreinovich and N.N. Thach (2018) *Econometrics for Financial Applications*, Springer: Berlin

Banfield, J. and A.E. Raftery (1993) "Model-based Gaussian and non-Gaussian clustering" *Biometrika* **49**, 803-821

Bezdek, J.C. (1981) *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press: New York

Bezdek, J.C., W.Q. Li, Y. Attikiouzel and M. Windham (1997) "A geometric approach to cluster validity for normal mixtures" *Soft Computing* **1**, 166-179

Biernacki, C. (1997) *Choix de Modèles en Classification* (doctoral dissertation), Université de Technologie de Compiègne: Compiègne, France http://www-math.univ-fcomte.fr/pp_Annu/CBIERNACKI/

Biernacki, C., G. Celeux and G. Govaert (2000) "Assessing a mixture model for clustering with the integrated completed likelihood" *IEEE Transactions Pattern Analysis and Machine Intelligence* **22**, 719-725

Biernacki, C. and G. Govaert (1997) "Using the classification likelihood to choose the number of clusters" *Journal on Scientific Computing* **29**, 451-457

Bozdogan, H. (1987) "Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions" *Psychometrika* **52**, 345-370

Bozdogan, H. (1994) "Mixture-model Cluster Analysis Using Model Selection Criteria and a New Information Measure of Complexity" in *Proceedings of the Fist US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach, Vol. 2*, by H. Bozdogan, Ed., Kluwer Academic Publishers: Boston, 69-113

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

Cutler, A. and M.P. Windham (1994) "Information-based Validity Functionals for Mixture Analysis" in *Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modelling: An Informational Approach* by H. Bozdogan, Ed., Kluwer Academic Publishers: Boston, 1994, 149-170

Dempster, A.P., N.M. Laird and D.B. Rubin (1997) "Maximum likelihood from incomplete data via EM-algorithm" *Journal of the Royal Statistical Society: Series B* **39**, 1-38

DeSarbo, W.S., M. Wedel, M. Vriens and V. Ramaswamy (1992) "Latent class metric conjoint analsysis" *Marketing Letters* **3**, 273-288

Dias, J.G. (2004) *Finite Mixture Models. Review, Applications, and Computer-intensive Methods. Research School Systems, Organization and Management* (doctoral dissertation), Groningen University: Groningen, the Netherlands

Gu, J., R. Koenker and S. Volgushev (2018) "Testing for homogeneity in mixture models" *Econometric Theory*, **34**, 850-895

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

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

Hurvich, C.M. and C.L. Tsai (1989) "Regression and time series model selection in small samples" *Biometrika* ***76***, 297-307

Hurvich, C.M. and C.L. Tsai (1995) "Model selection for extended quasi-likelihood models in small samples" *Biometrics* **51**, 1077-1084

Kasahara, H. and K. Shimotsu (2015) "Testing the number of components in normal mixture regression models" *Journal of the American Statistical Association* **110**, 1632-1645

Li, M., S. Xiang and W. Yao (2016) "Robust estimation of the number of components for mixtures of linear regression models" *Computational Statistics* **31**, 1539-1555

Liang, Z., Jaszczak, R.J. and R.E. Coleman (1992) "Parameter estimation of finite mixture models using the EM algorithm and information criteria with applications to medical image processing" *IEEE Transactions on Nuclear Science* **39**, 1126-1133

Lukočieė, O. and J.K. Verut (2009) "Determining the Number of Components in Mixture Models for Hierarchical Data" in *Advances in Data Analysis, Data Handling and Business Intelligence* by A. Fink, B. Lausen, W. Seidel and A. Ultsch, Eds., Springer: Berlin, 241-249

Nasserinejad, K., J. van Rosmalen, W. de Kort and E. Lesaffre (2017) "Comparison of criteria for choosing the number of classes in Bayesian finite mixture models" *PLoS ONE* **12**, e0168838. https://doi.org/10.1371/journal.pone.0168838

Pan, L., Y. Li, K. He, Y. Li and Y. Li (2020) "Generalized linear mixed models with Gaussian mixture random effects: inference and application" *Journal of Multivariate Analysis* **175**, 104555

Rivera, F.F., E.L.E. Zapata and J.M. Carazo (1990) "Cluster validity based on the hard tendency of the fuzzy classification" *Pattern Recognition Letters* **11**, 7-12

Roubens, M. (1978) "Pattern classification problems and fuzzy sets" *Fuzzy Sets Systems* **1** 239-253

Schwartz, G. (1978) "Estimating the dimension of a model" *The Annals of Statistics* **6**, 461-464

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

Wedel, M. and W.S. DeSarbo (1995) "A mixture likelihood approach for generalized linear models" *Journal of Classification* **12,** 21-55

Wu, W. (2019) *Three Essays on Mixture Model and Gaussian Processes* (doctoral dissertation), Texas A & M University: College Station, TX. http : / /hdl .handle .net /1969 .1 /184913

Yan, H. and L. Han (2019) "Empirical distributions of stock returns: mixed normal or kernel density?" *Physica A: Atatistical Mechanics and Its Applications* **514**, 473-486