

On the Number of Components for Matrix-Variate Mixtures: A Comparison Among Information Criteria

Salvatore D. Tomarchio  and Antonio Punzo 

Department of Economics and Business, University of Catania, Catania, Italy

Corresponding to: Salvatore D. Tomarchio, Department of Economics and Business, University of Catania, Catania, Italy. Email: daniele.tomarchio@unict.it

Summary

This study explores the crucial task of determining the optimal number of components in mixture models, known as mixture order, when considering matrix-variate data. Despite the growing interest in this data type among practitioners and researchers, the effectiveness of information criteria in selecting the mixture order remains largely unexplored in this branch of the literature. Although the Bayesian information criterion (BIC) is commonly utilised, its effectiveness is only marginally tested in this context, and several other potentially valuable criteria exist. An extensive simulation study evaluates the performance of 10 information criteria across various data structures, specifically focusing on matrix-variate normal mixtures.

Key words: Information criteria; matrix-variate; mixture models; order selection.

1 Introduction

Finite mixture models hold a central position in statistical data analysis, as evidenced by the extensive literature dedicated to their application across various scientific disciplines. Their versatile nature serves as a convenient instrument for clustering and classification, capturing unobserved heterogeneity inherent in real-world datasets. For an in-depth survey about finite mixture models, refer to seminal texts such as McLachlan & Peel (2000) and McNicholas (2016).

A fundamental aspect of mixture modelling is accurately determining the number of components needed to describe the data, often termed mixture order. As McLachlan & Rathnayake (2014) discussed, discerning a mixture density with k components from those with fewer or more than k components might be empirically challenging. Moreover, the precise selection of a suitable number of components holds profound implications on the practical interpretations of the obtained results (Walaa & Ismail, 2022).

In both univariate and multivariate literature, the estimation of the mixture order has been largely discussed, with no single method receiving unanimous support (for some examples, see Akogul & Erisoglu, 2016, Chen & Khalili, 2009, Depraetere & Vandebroek, 2014, Hui *et al.*, 2015, McLachlan & Rathnayake, 2014, Walaa & Ismail, 2022). Nevertheless, the most common approach relies on the computation of penalised log-likelihood information criteria. In a nutshell, as each component is added to a mixture model, the log-likelihood increases.

However, to prevent overfitting due to an excessive number of parameters, a penalty term is generally introduced to adjust for the model's complexity.

In this study, we aim to evaluate the performance of information criteria in selecting the optimal number of components for analysing matrix-variate data. Interest in matrix-variate data analysis has continuously grown among practitioners and researchers, particularly within the context of mixture modelling. This type of data arises when each statistical unit is represented by a $p \times r$ matrix. As a result, when considering n observations, the data can be structured as a three-way array with dimensions $p \times r \times n$.

From a practical perspective, matrix-variate data emerge in many different contexts. For example, Viroli (2011a) analyses a dataset where $n = 58$ soybean lines are evaluated using $p = 2$ attributes across $r = 8$ environments, each representing a specific spatio-temporal configuration. The matrix-variate structure is particularly useful here because it enables the simultaneous modelling of interactions between environments and plant traits. This helps researchers identify how environmental conditions influence specific traits in a structured way, accounting for potential correlations across both dimensions.

Viroli (2011a) also examines two further datasets: one concerning Landsat satellite images and another on handwritten digits. In both cases, the data are structured as square matrices of pixels, with each matrix representing a single image. The matrix-variate framework is particularly well suited to image data because it preserves the spatial relationships between pixels, which is essential for recognising patterns, shapes and textures within images. This allows for more accurate image classification and segmentation, as the model captures pixel-level dependencies crucial for identifying features in satellite imagery and handwritten digits.

Melnykov & Zhu (2018) also analyse the Landsat satellite dataset alongside customer movie ratings, further demonstrating the flexibility of matrix-variate data in different domains. Specifically, the movie dataset includes scores for $n = 105$ comedies, with ratings categorised by gender ($p = 2$ levels: males and females) and age group ($r = 4$ levels: under 18, 18 to 29, 30 to 44 and over 44). The matrix-variate framework enables the simultaneous modelling of customer preferences across both demographic groups and movie titles, capturing complex interactions. For example, it may reveal that younger viewers of one gender prefer a specific type of comedy compared to older viewers of another gender, helping to uncover subtle trends in consumer behaviour. These datasets were also analysed by Sarkar *et al.* (2020), further validating the usefulness of matrix-variate methods in consumer behaviour analysis.

In the context of crime analysis, Melnykov & Zhu (2019) examine matrix-variate data related to crime trends in the United States. In this study, $p = 7$ crime-related variables were tracked across $r = 13$ years for $n = 236$ cities. The matrix-variate framework is particularly important here because it allows for the simultaneous modelling of multiple crime variables across time and cities. This helps capture the complex interplay between crime types and their temporal evolution across geographic locations, allowing for a deeper understanding of how crime patterns evolve and interact in different urban environments. Similarly, Viroli (2011a, 2011b) applies matrix-variate methods to analyse crime data in Italy, focusing on $n = 103$ provinces where $p = 4$ crime-related variables were measured over $r = 5$ years.

Other examples of matrix-variate data applications include economics and education frameworks, as explored by Tomarchio *et al.* (2021), Gallagher *et al.* (2022) and Tomarchio (2024). In these studies, matrix-variate data are used to model economic indicators or educational performance metrics, where p variables (e.g., economic indicators, test scores) are recorded over r time periods for n units (e.g., provinces, universities). The matrix-variate approach is particularly advantageous in these contexts because it enables researchers to model dependencies between variables across time, providing a more comprehensive understanding of how these variables co-evolve and influence each other over time.

While matrix-variate mixture models are gaining traction and recognition, the effectiveness of information criteria in determining mixture order remains underexplored in the literature. Indeed, existing studies often incorporate simulated analyses on this topic within broader papers introducing new mixture models and only for specific simulation factors. In other terms, these findings represent more of a tangential aspect rather than the primary focus of analysis.

For example, Viroli (2011a) presents results for the Bayesian information criterion (BIC; Schwarz, 1978), Akaike information criterion (AIC; Akaike, 1974) and integrated classification likelihood criterion (ICL; Biernacki *et al.*, 2000) when considering matrix-variate normal (MVN) mixtures. The findings suggest that the BIC marginally outperforms the ICL in recovering the mixture order, while both outperform the AIC.

Tomarchio (2024) compares AIC and BIC in the context of MVN mean-variance Birnbaum-Saunders mixtures, with the latter yielding superior results. Gallagher & McNicholas (2018), Sarkar *et al.* (2020), Tomarchio *et al.* (2022) and Amato *et al.* (2024) focus their investigations solely on BIC within specific contexts: matrix-variate skewed mixtures, parsimonious MVN mixtures, matrix-variate contaminated mixtures and matrix-variate mixtures for ordinal matrices, respectively. Likewise, Tomarchio *et al.* (2021), Gallagher *et al.* (2022) and Tomarchio & Gallagher (2024) exclusively investigate and assess the performance of the BIC within the framework of matrix-variate mixtures of regressions. Globally, these findings indicate that the BIC effectively recovers the correct mixture order.

As observed, the aforementioned contributions predominantly focus on the BIC for mixture order selection within this research domain. However, it is worth noting that several other information criteria exist in the literature, although they have not yet been applied in the matrix-variate framework. Furthermore, their behaviour should be evaluated by jointly considering and varying different experimental factors, such as sample size, overlap, true number of mixture components, matrix dimensionality and the presence of equal or not equal mixture weights. Notably, none of the above contributions examines all these factors simultaneously, and more in general, there are no studies aimed at comparing jointly the behaviour of the different information criteria. Thus, this research gap presents a significant opportunity for further exploration and analysis, particularly considering the unique characteristics and complexities inherent in matrix variate data.

For the sake of simplicity and practicality, our study exclusively considers MVN mixtures, using them as a representative example. Analysing the capacity of information criteria for every mixture model mentioned above would be overly complex and impractical to discuss in a single manuscript comprehensively. Thus, we begin by illustrating the characteristics of MVN mixtures and the maximum likelihood estimation of the model parameters in Section 2. Then, we present the information criteria under evaluation in Section 3. The simulation study design and the obtained results are reported in Section 4. Finally, Section 5 presents some concluding remarks and ideas for future research.

2 The Matrix-Variate Normal Mixture Model

A $p \times r$ random matrix $\mathcal{X}$, with realisation $\mathbf{X}$, arises from a finite mixture of k distributions if it has the following probability density function (pdf):

$$h(\mathbf{X}; \boldsymbol{\theta}) = \sum_{j=1}^k \pi_j f(\mathbf{X}; \boldsymbol{\Omega}_j), \quad (1)$$

where $\pi_j > 0$ is the mixing proportion, with $\sum_{j=1}^k \pi_j = 1$, $f(\mathbf{X}; \boldsymbol{\Omega}_j)$ is the pdf of the j th mixture component with parameters $\boldsymbol{\Omega}_j$ and $\boldsymbol{\theta}$ is the set containing all the parameters of the model.

Herein, the MVN distribution is adopted for $f(\mathbf{X}; \boldsymbol{\Omega}_j)$ in (1). In detail, the MVN distribution has the following pdf:

$$f(\mathbf{X}; \boldsymbol{\Omega}) = (2\pi)^{-\frac{pr}{2}} |\boldsymbol{\Sigma}|^{-\frac{r}{2}} |\boldsymbol{\Psi}|^{-\frac{p}{2}} \exp \left\{ -\frac{1}{2} \text{tr} [\boldsymbol{\Sigma}^{-1} (\mathbf{X} - \mathbf{M}) \boldsymbol{\Psi}^{-1} (\mathbf{X} - \mathbf{M})'] \right\},$$

where $\mathbf{M}$ is the $p \times r$ mean matrix, $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$ are the $p \times p$ and $r \times r$ covariance matrices related to the rows and columns, respectively, and $\text{tr}(\cdot)$ is the trace operator. In symbols, $\mathcal{MVN}_{p \times r} \sim (\mathbf{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi})$.

It's noteworthy that after data vectorisation, the MVN and multivariate normal (MN) distributions are related as follows:

$$\mathcal{MVN}_{p \times r} \sim (\mathbf{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi}) \Leftrightarrow \mathcal{MN}_{pr} \sim (\boldsymbol{\mu} = \text{vec}(\mathbf{M}), \boldsymbol{\Phi} = \boldsymbol{\Psi} \otimes \boldsymbol{\Sigma}),$$

where $\mathcal{MN}_{pr}$ refers to the MN distribution, $\text{vec}(\cdot)$ is the vectorisation operator and $\otimes$ is the Kronecker product. However, the MVN distribution has the following desirable characteristics:

- C.1 It separately models and identifies the variability between the variables and within the occasions via $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$, respectively. Such information would be lost if the unique covariance matrix $\boldsymbol{\Phi}$ would be considered.
- C.2 It reduces the number of free covariance parameters from $pr(pr+1)/2$ to $[p(p+1)/2] + [r(r+1)/2] - 1$. This parsimonious advantage is particularly relevant for the computation of information criteria, as they incorporate a penalty term dependent on the number of parameters.

We recall that the -1 term in C.2 arises from a constraint introduced to solve an identifiability issue. Indeed, $\boldsymbol{\Psi} \otimes \boldsymbol{\Sigma} = \boldsymbol{\Psi}^* \otimes \boldsymbol{\Sigma}^*$ if $\boldsymbol{\Sigma}^* = a\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}^* = a^{-1}\boldsymbol{\Psi}$. As a result, matrices $\boldsymbol{\Psi}$ and $\boldsymbol{\Sigma}$ are identifiable up to a positive multiplier a . Several ways have been used in the literature to handle this problem. For example, Viroli (2011b) imposes the trace of $\boldsymbol{\Sigma}$ equal to p , Sarkar *et al.* (2020) and Tomarchio *et al.* (2022) adopt the constraint $|\boldsymbol{\Psi}| = 1$, while Gallagher & McNicholas (2018) set the first diagonal element of $\boldsymbol{\Sigma}$ to 1.

2.1 Maximum Likelihood Estimation

Maximum likelihood parameter estimation for model (1) can be efficiently performed through the expectation conditional-maximisation (ECM) algorithm (Meng & Rubin, 1993). Let $\mathcal{S} = \{\mathbf{X}_i\}_{i=1}^n$ be a sample of n independent observations from model (1). Within an ECM framework, $\mathcal{S}$ is viewed as incomplete because, for each observation, we do not know its component membership. To address this aspect, we use an indicator vector $\mathbf{z}_i = (z_{i1}, \dots, z_{ik})$, where $z_{ij} = 1$ if observation i is in group j , and $z_{ij} = 0$ otherwise. Therefore, the complete data are $\mathcal{S}_c = \{\mathbf{X}_i, \mathbf{z}_i\}_{i=1}^n$, and the complete-data log-likelihood is

$$l_c(\boldsymbol{\Theta}) = \sum_{i=1}^n \sum_{j=1}^k z_{ij} \ln(\pi_j) + \sum_{i=1}^n \sum_{j=1}^k z_{ij} \left\{ -\frac{pr}{2} \ln(2\pi) - \frac{r}{2} \ln |\boldsymbol{\Sigma}_j| - \frac{p}{2} \ln |\boldsymbol{\Psi}_j| - \frac{1}{2} \text{tr} [\boldsymbol{\Sigma}_j^{-1} (\mathbf{X}_i - \mathbf{M}_j) \boldsymbol{\Psi}_j^{-1} (\mathbf{X}_i - \mathbf{M}_j)'] \right\}.$$

In the forthcoming algorithm, parameters labelled with one dot denote updates from the previous iteration, whereas those marked with two dots signify updates from the current iteration.

2.1.1 E-Step

At the E-step, we have to update the posterior probability that a point $\mathbf{X}_i$ belongs to the j th component of the considered model, that is,

$$\ddot{z}_{ij} = E_{\boldsymbol{\theta}}(Z_{ij}|\mathbf{X}_i) = \frac{\dot{\pi}_j f(\mathbf{X}_i; \dot{\boldsymbol{\Omega}}_j)}{\sum_{h=1}^k \dot{\pi}_h f(\mathbf{X}_i; \dot{\boldsymbol{\Omega}}_h)}.$$

2.1.2 CM-Step 1

At the first CM-step, we maximise the expectation of the complete-data log-likelihood with respect to $\{\pi_j, \mathbf{M}_j, \boldsymbol{\Sigma}_j\}_{j=1}^k$, keeping fixed $\{\boldsymbol{\Psi}_j\}_{j=1}^k$. Thus, we have the following updates:

$$\ddot{\pi}_j = \frac{\sum_{i=1}^n \ddot{z}_{ij}}{n}, \quad \ddot{\mathbf{M}}_j = \frac{\sum_{i=1}^n \ddot{z}_{ij} \mathbf{X}_i}{\sum_{i=1}^n \ddot{z}_{ij}}, \quad \text{and} \quad \ddot{\boldsymbol{\Sigma}}_j = \frac{\sum_{i=1}^n \ddot{z}_{ij} (\mathbf{X}_i - \ddot{\mathbf{M}}_j) \dot{\boldsymbol{\Psi}}_j^{-1} (\mathbf{X}_i - \ddot{\mathbf{M}}_j)'}{r \sum_{i=1}^n \ddot{z}_{ij}}.$$

2.1.3 CM-Step 2

Now, with the updated parameters from the CM-Step 1, we compute

$$\ddot{\boldsymbol{\Psi}}_j = \frac{\sum_{i=1}^n \ddot{z}_{ij} (\mathbf{X}_i - \ddot{\mathbf{M}}_j)' \ddot{\boldsymbol{\Sigma}}_j^{-1} (\mathbf{X}_i - \ddot{\mathbf{M}}_j)}{p \sum_{i=1}^n \ddot{z}_{ij}}.$$

The E-step and CM-steps are alternated repeatedly until the difference of the observed-data log-likelihoods at two consecutive iterations changes by an arbitrarily small amount, which we set to 10^{-3} .

An important aspect of any EM-based algorithm is the selection of initial values. In the matrix-variate framework, a frequently used initialisation approach (see, e.g., Melnykov & Zhu, 2018, Tomarchio *et al.*, 2020,2023, Zhu *et al.*, 2022) is the EM-EM strategy introduced by Biernacki *et al.* (2003). This method involves several short runs of the algorithm from random initial points for a predefined number of iterations. The run with the highest log-likelihood is then selected as the starting point for the final algorithm. In our simulations, the short runs are restricted to 5 iterations, with 50 random starting points used.

3 Information Criteria

As outlined in Section 1, a pivotal aspect of employing mixture models involves selecting the appropriate number of components. When this number cannot be predetermined, one can resort to information criteria. Hereafter, we present the list of ten information criteria that we will consider in this manuscript. Notice that, in the following sections, for each fitted model, the log-likelihood is evaluated once and then used to compute all information criteria.

One of the most commonly adopted is the AIC, which is centred on estimating the Kullback-Leibler (K-L) divergence between the true and the estimated density. This estimation revolves around the expected (observed-data) log-likelihood and is defined as

$$\text{AIC} = -2l(\hat{\boldsymbol{\theta}}) + 2m,$$

where $l(\hat{\boldsymbol{\theta}})$ is the maximised (observed-data) log-likelihood value, which acts as a measure of model fit, while m denotes the number of model parameters, serving as a penalty term.

A potential drawback of the AIC is its tendency to overestimate the number of mixture components (McLachlan & Rathnayake, 2014). Furthermore, a perceived limitation of AIC is its inconsistency, a factor that may be deemed significant when determining the optimal k (Buckby *et al.*, 2023). In pursuit of consistency, meaning the ability to increasingly favour the best model from the suggested set as the sample size grows, Bozdogan (1987) introduced the consistent Akaike's information criterion (CAIC), which is defined as

$$\text{CAIC} = -2l(\hat{\boldsymbol{\theta}}) + m[\ln(n) + 1].$$

A small sample correction of AIC is the corrected AIC (AICc; Sugiura, 1978; Hurvich & Tsai, 1989), defined as

$$\text{AICc} = -2l(\hat{\boldsymbol{\theta}}) + 2m\left(\frac{n}{n - m - 1}\right).$$

When the ratio n/m is sufficiently large, AICc will tend to the AIC (Buckby *et al.*, 2023). Note that AICc can be calculated only if $m \leq n - 1$, as exceeding this threshold alters the intended behaviour of the penalty term.

Conversely, McQuarrie *et al.* (1997) proposed an approximate unbiased estimator of the K-L divergence for large samples when the true model is finite-dimensional. It is calculated as

$$\text{AICu} = -2l(\hat{\boldsymbol{\theta}}) + 2m\left(\frac{n}{n - m - 1}\right) + n \ln\left(\frac{n}{n - m - 1}\right).$$

Note that AICu can be calculated only if $m < n - 1$. Lastly, Bozdogan (1994) introduced another variation, called AIC3, which incorporates a penalty term of $3m$ instead of $2m$ as seen in AIC. Therefore, we have

$$\text{AIC3} = -2l(\hat{\boldsymbol{\theta}}) + 3m.$$

From a Bayesian perspective, another popular information criterion is the BIC, which is formulated as

$$\text{BIC} = -2l(\hat{\boldsymbol{\theta}}) + m \ln(n).$$

Typically, BIC tends to choose fewer mixture components compared to AIC, due to its more stringent penalisation term. This disparity becomes particularly relevant in complex models, and as commented on in Section 1, it is extensively used in empirical analyses.

In the context of the classification likelihood approach, Banfield & Raftery (1993) derived the approximate weight of evidence (AWE) criterion, defined as

$$\text{AWE} = -2l(\hat{\boldsymbol{\theta}}) + 2m\left[\frac{3}{2} + \ln(n)\right].$$

The complexity of the penalty term within the AWE criterion results in the selection of more parsimonious models compared to BIC (Walaa & Ismail, 2022).

Using the estimated entropy function as a basis, Celeux & Soromenho (1996) proposed the normalised entropy criterion (NEC), which is equal to

$$\text{NEC} = \frac{\text{EN}}{l(\hat{\boldsymbol{\theta}}) - l_1(\hat{\boldsymbol{\theta}})},$$

where $l_1(\hat{\theta})$ is the maximised (observed-data) log-likelihood in a one-component mixture and $EN = -\sum_{j=1}^k \sum_{i=1}^n \hat{z}_{ij} \ln(\hat{z}_{ij})$, being $\hat{z}_{ij}$ the value of z_{ij} at convergence, is an entropy measure based on the estimated posterior probabilities. This entropy term can be viewed as an indicator of the model's capacity to generate a meaningful partition of the data. It is noteworthy that NEC is undefined when $k = 1$, and it is convention to set $NEC = 1$ in such a case. Biernacki *et al.* (1999) proposed a practical guideline for addressing this issue: Let k^* be the value minimising NEC_j , for $j = 2, \dots, k^{\sup}$, where $k^{\sup}$ serves as an upper bound for the number of mixture components. Then, k^* clusters are chosen if the corresponding NEC is equal to or less than 1; otherwise, no clustering structures in the data are declared.

Another common information criterion is the ICL, which is defined as

$$ICL = BIC - 2 \sum_{j=1}^k \sum_{i=1}^n \text{MAP}(\hat{z}_{ij}) \ln(\hat{z}_{ij}),$$

where $\text{MAP}(\cdot)$ is the maximum a posteriori probabilities operator

$$\text{MAP}(\hat{z}_{ij}) = \begin{cases} 1 & \text{if } \max_h \{\hat{z}_{ih}\} \text{ occurs in group } h = j, \\ 0 & \text{if otherwise.} \end{cases}$$

Thus, this criterion is essentially the BIC adjusted with an entropy term that measures the overlap of the clusters. It is interesting to report that a slightly modified version of the ICL, called ICL-BIC criterion, is often considered in the literature (McLachlan & Peel, 2000; Bacci *et al.*, 2014; Walaa & Ismail, 2022). Specifically, this criterion replaces the entropy term of the ICL with that presented in the NEC criterion. Put differently, the ICL-BIC can be expressed as

$$ICL\text{-}BIC = BIC + 2EN.$$

To enhance the clarity of the results discussed in the following sections, we will label the ICL as ICL1 and the ICL-BIC as ICL2.

4 Simulated Data Analysis

In what follows, we explain the process (Section 4.1) and findings (Section 4.2) of a Monte Carlo simulation study designed to assess the effectiveness of the 10 information criteria discussed in Section 3 in determining the optimal number of components for MVN mixtures.

4.1 Simulation Procedure and Design

To explore the impact of various data structures commonly encountered in real-data analyses on each information criterion, we manipulated several factors in the data-generation process. Specifically, we generated data from MVN mixtures by varying the following five factors:

- F.1 sample size: $n \in \{250, 500, 1000\}$;
- F.2 matrix dimensionality: 2×3 and 7×6 ;
- F.3 number of latent groups: $k \in \{2, 4, 6\}$;
- F.4 mixture weights: equal and unequal;
- F.5 overlap: close and separated groups.

As a result, we have a design with $3^2 \times 2^3 = 72$ unique data configurations. For each configuration, we draw 500 samples, with varying sets of parameters, resulting in a total of 36 000

simulated datasets. Details on all the other parameters used to simulate the datasets are available in the supplementary material.

Regarding factor F.5, to the best of our knowledge, no overlap measures are currently defined for use within the matrix-variate framework. Consequently, controlling this factor in such a setting is a challenge. Here, we follow an approach similar to that of Amato *et al.* (2024), where, keeping other parameters fixed, the values of the mean matrices are fine-tuned to achieve a certain degree of overlap among the clusters, measured in terms of the Adjusted Rand Index (ARI; Hubert & Arabie, 1985). To this aim, two ARI intervals are defined: [0.80–0.85] for the ‘close’ group setting and [0.95–1.00] for the ‘separated’ setting. The mean matrix values are adjusted such that, for a given data configuration, when datasets are simulated and the true model is fitted using the true parameters as initial values, the resulting ARI values generally fall within the defined intervals. To ensure that all 500 datasets for each data configuration have an ARI within the specified intervals, we further adopt the following scheme during the simulations:

- generate a dataset and fit the true model using the true parameters as initial values;
- compute the ARI. If it falls within the target ARI interval, retain the dataset; otherwise, discard it and simulate a new one.

In each data configuration, MVN mixtures are then fitted for $k \in \{1, \dots, k^* + 2\}$, where k^* is the true number of latent groups in the data. The decision to limit the maximum number of components to $k^* + 2$ is motivated by the need to balance model flexibility with the risk of spurious cluster solutions. It is well documented that EM-based algorithms in mixture models can converge to spurious local maxima, especially when the number of components is unnecessarily large. These spurious solutions often manifest as clusters with very few observations, which reflect random patterns in the data rather than meaningful group structure. Such clusters are usually of limited practical value and can be difficult to interpret (McLachlan & Peel, 2000). Moreover, the presence of spurious clusters can lead to computational challenges during the estimation process, such as issues with matrix inversion when the covariance matrix of a group is nearly singular. Extending the range beyond $k^* + 2$ is unlikely to improve model selection or uncover additional meaningful structure, but it would increase the risk of incurring such issues.

The functions used for initialising, fitting and simulating from the considered models have been bundled into the **MatVarMix** package for the R statistical software (R Core Team, 2023). This package is available at <https://github.com/danieletomarchio/Matrix-and-multivariate-normal-mixtures.git>.

4.2 Results

To simplify our discussion and results, we organise them into six blocks based on the combination of matrix dimensionality and number of groups in the data. Within each block, we vary all other factors. In detail, we have

- B.1 2×3 matrices and $k = 2$ groups;
- B.2 2×3 matrices and $k = 4$ groups;
- B.3 2×3 matrices and $k = 6$ groups;
- B.4 7×6 matrices and $k = 2$ groups;
- B.5 7×6 matrices and $k = 4$ groups;
- B.6 7×6 matrices and $k = 6$ groups.

In the following subsections, we aim to evaluate the performance of information criteria from different points of view.

4.2.1 Fitting, Underfitting or Overfitting the Number of Groups

Following the approach outlined by Bacci *et al.* (2014), each criterion is categorised as either *underfitting*, *fitting* or *overfitting*, based on the relationship between k^* and the chosen k by that criterion. To clarify, *underfitting* indicates that the selected number of groups is less than the true value, *fitting* means that the correct number of groups is picked, whereas *overfitting* denotes that the chosen number of groups exceeds the true value.

The number of times each information criterion underfits, fits or overfits is reported using bar plots from Figures 1 to 6, for blocks from B.1 to B.6, respectively. Starting with the analysis of Figure 1, it is evident that AIC, AICc and NEC (though less pronounced) tend to overfit the number of groups, regardless of the overlap and the sample size. In constraint, AWE and, to a lesser extent, ICL1 and ICL2 tend to underfit when $n = 250$, and the groups are close. This

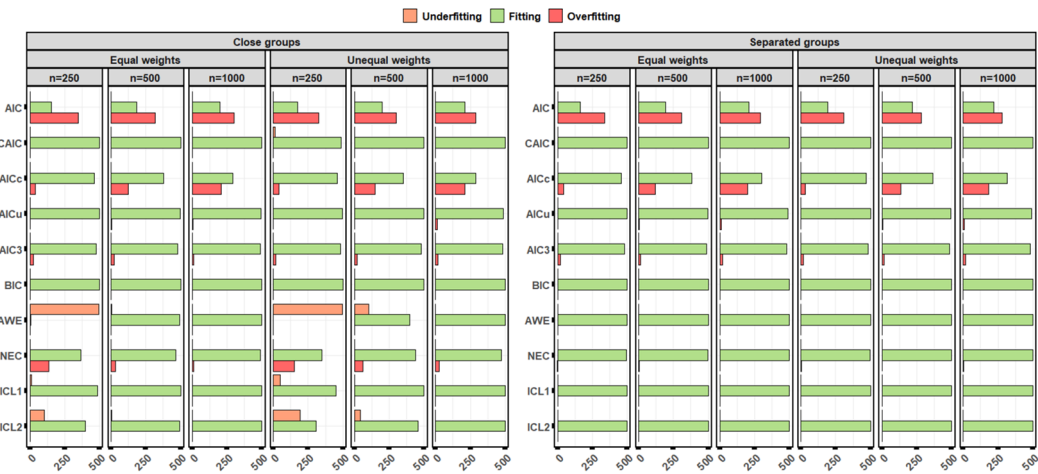


FIGURE 1. Bar plots reporting the number of times each information criterion underfits, fits or overfits the true number of groups in the data when the dimensionality is 2×3 and $k = 2$.

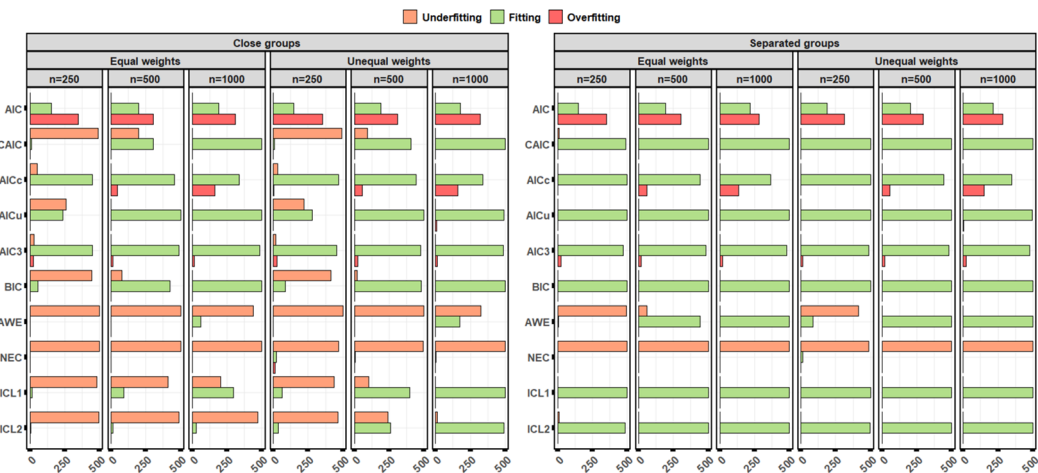


FIGURE 2. Bar plots reporting the number of times each information criterion underfits, fits or overfits the true number of groups in the data when the dimensionality is 2×3 and $k = 4$.



FIGURE 3. Bar plots reporting the number of times each information criterion underfits, fits or overfits the true number of groups in the data when the dimensionality is 2×3 and $k = 6$.



FIGURE 4. Bar plots reporting the number of times each information criterion underfits, fits or overfits the true number of groups in the data when the dimensionality is 7×6 and $k = 2$.

leads to the selection of $k = 1$, suggesting a failure to identify distinct groups within the data. However, this outcome does not imply that there are no groups present; instead, it may indicate that the criteria are not adequately capturing the underlying complexity of the data. Such a scenario can have serious consequences, as it could lead to overlooking the true grouping structure and result in misleading interpretations of the results.

Nevertheless, excluding AICc, their performance improves as n increases. This improvement is particularly relevant in scenarios with poorly separated groups. For instance, AWE shows significant improvement when transitioning from $n = 250$ to $n = 500$, and similarly, ICL1 and ICL2 also perform better under these conditions. Generally, as group separation increases, AWE, NEC, ICL1 and ICL2 improve, AIC and AICc persist in overfitting the number of groups, whereas BIC and the other AIC-based criteria maintain the same performance.



FIGURE 5. Bar plots reporting the number of times each information criterion underfits, fits or overfits the true number of groups in the data when the dimensionality is 7×6 and $k = 4$.



FIGURE 6. Bar plots reporting the number of times each information criterion underfits, fits or overfits the true number of groups in the data when the dimensionality is 7×6 and $k = 6$.

An interesting observation is the superior effectiveness of ICL1 compared to ICL2. Among the AIC-based criteria, CAIC regularly identifies the correct number of groups, followed closely by AICu and AIC3. Overall, within these data configurations, CAIC, AICu and BIC emerge as the most reliable information criteria.

Moving to Figure 2, similar tendencies are observed for AIC and AICc (in a lesser way), which overfit the true number of groups, and AWE, which underfits instead. Conversely, NEC generally underfits now. CAIC and BIC, which were among the best information criteria in Figure 1, now face an underfitting tendency when $n = 250$ in the close group configurations. Among the AIC-based criteria, AIC3 maintains good performance, followed by the AICu and CAIC. Furthermore, ICL1 consistently outperforms ICL2.

In Figure 3, we observe a general degradation in the performance of all information criteria, particularly in scenarios with close groups. On a related note, only the AIC-based criteria seem to correctly identify the true number of groups most of the time, especially as n increases. Among these, AIC3 and AICc perform the best. AWE, NEC, ICL1, ICL2 and CAIC tend to regularly underfit, regardless of the sample size, while BIC improves its performance when $n = 1000$.

As in previous sections, an increase in group separation generally leads to better results. However, AWE and NEC continue to underfit, while AIC remains the only criterion that overfits.

Moving now to configurations characterised by higher matrix dimensionality, in Figure 4, several information criteria struggle to identify the correct number of groups, particularly when $n = 250$. In these cases, NEC repeatedly overestimates the number of groups, while CAIC, AICc, AICu, BIC, AWE, ICL1 and ICL2 tend to underfit. In this context, underfitting results in the selection of $k = 1$, which indicates that no mixture components are detected. As discussed before, this can have significant consequences if there are indeed groups present in the data, as such misidentification may lead to an oversimplified interpretation that overlooks the true structure of the data and potentially results in misleading conclusions.

Among the AIC-based criteria, AIC3 consistently identifies the correct number of groups. Notably, AIC shows improved performance compared to previous analyses, though it weakens as n increases—unlike other criteria that typically perform better with larger sample sizes. As in earlier results, ICL1 continues to outperform ICL2, and NEC overestimates the number of groups even when different sample sizes are considered.

As group separation increases, all information criteria show better performance. However, AWE still tends to underfit, while AICc and AICu behave similarly to how they did in close-group scenarios.

In Figure 5, most information criteria encounter significant issues, particularly in close-group configurations. When $n = 250$ and $n = 500$, they frequently underfit the true number of groups, except for AIC and AIC3. This seems to suggest that criteria with weaker penalties tend to perform better. Additionally, as n increases, AIC begins to overfit, while AIC3 consistently performs well. When $n = 1000$, both AICc and AICu correctly identify the number of groups.

As the separation between groups increases, some criteria, such as AICc and AICu, maintain consistent performance, while others improve, with AIC3 emerging as the best-performing criterion.

Lastly, in Figure 6, we observe configurations characterised by the highest number of parameters. Similar to the discussion of Figure 5, only AIC performs well when n exceeds 500, and AIC3 only when $n = 1000$. All other criteria consistently underfit. This appears to indicate that the penalty terms in most information criteria are overly stringent, significantly outweighing the effectiveness in accurately determining the correct number of groups. Once again, an increase in group separation leads to improved results, though these improvements are generally observed only when $n = 1000$.

If we were to summarise the results presented up to now into a single value, we could calculate the average number of times, across all examined data configurations, each criterion selects the correct number of mixture components, that is, the average of the green bars from Figures 1 to 6. Such information is detailed using a bar plot in Figure 7. Upon examination, it is clear that the commonly used BIC does not emerge as the optimal choice, but it ranks as the second best. On the contrary, AIC3 exhibits the highest average fitting performance, while NEC and AWE demonstrate relatively poor and similar results. ICL1 outperforms ICL2 and is closely followed by AICu, AIC, CAIC and AICc, all of which show substantially similar performance.

Despite the insights provided by the average results, none of the information criteria uniformly demonstrate superior performance across all the considered data configurations. In light

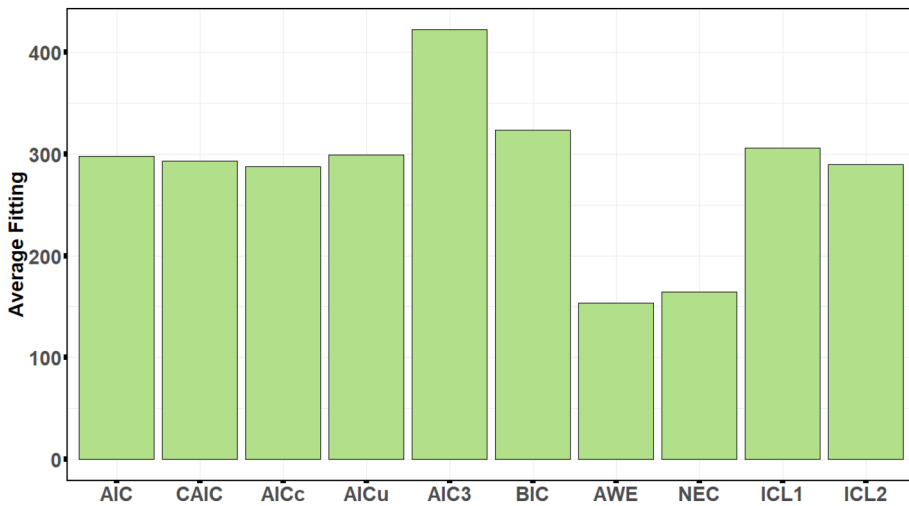


FIGURE 7. Bar plot reporting the number of times, on average across all data configurations, each information criterion fits the true number of groups in the data.

of this, it is advisable to take into account the data and model characteristics, in terms of the number of components to be fitted, when selecting the most appropriate set of information criteria to use. That said, we remand to Section 5 for a schematic overview of the results discussed here.

4.2.2 Pairwise Comparisons

Additionally, we performed an analysis of the agreement ratios between pairs of information criteria. Contingency tables were constructed to summarise the selections of the number of components made by each pair of criteria across all datasets within each data configuration. For each pair of criteria, the agreement ratio was calculated as the proportion of agreements (i.e., cases where both criteria selected the same number of components) relative to the total number of datasets. This analysis provides insights into the comparative performance of the information criteria in selecting the same number of components in mixture models.

Results are presented via heatmaps from Figures 8 to 13, for blocks from B.1 to B.6, respectively. Each figure contains three heatmaps corresponding to three different sample sizes. The upper triangular matrix of each heatmap displays results for the equal mixture weights configuration, while the lower triangular matrix shows results for the unequal mixture weights configuration. It is important to note that these results pertain to scenarios with close groups; results for separated groups are provided in the supplementary material.

Figure 8, which represents the configurations with 2×3 matrices and $k = 2$, reveals that the agreement ratios are generally high across all sample sizes. Notably, as the sample size increases, the ratios mostly improve. AIC, AICc and AWE generally select a different number of mixture components than the other criteria, reflecting the results already presented in Figure 1.

In Figure 9, the complexity increases with the number of groups rising to $k = 4$. Here, we observe a decline in agreement ratios compared to Figure 8, particularly in the lower triangular matrix when $n = 250$. Larger sample sizes generally show an upward trend in agreement ratios, and the differences between the equal and unequal weight scenarios become less pronounced.



FIGURE 8. Heatmaps of agreement ratios between pairs of information criteria when the dimensionality is 2×3 , $k = 2$, and the groups are close. The upper triangular matrices represent equal mixture weights configurations, while the lower triangular matrices represent unequal mixture weights configurations. The results are shown for the three sample sizes.



FIGURE 9. Heatmaps of agreement ratios between pairs of information criteria when the dimensionality is 2×3 , $k = 4$ and the groups are close. The upper triangular matrices represent equal mixture weights configurations, while the lower triangular matrices represent unequal mixture weights configurations. The results are shown for the three sample sizes.

As we move to Figure 10, where $k = 6$, the impact of increased complexity becomes evident. The agreement ratios drop significantly, highlighting the divergences in the selections made by the information criteria. Although larger sample sizes improve the ratios, good agreement is observed in only a few cases. Examples include the pairs AICu-AIC3, ICL1-ICL2, CAIC-ICL1 and CAIC-ICL2.

In Figure 11, with 7×6 matrices and $k = 2$ groups, the agreement ratios are more favourable, especially with larger sample sizes. This suggests that even with larger matrices, a simpler model with fewer groups can still produce strong consensus among the criteria. Notably, for $n = 250$, most information criteria are in agreement, except for AIC and AIC3, which themselves show a high degree of concordance (as seen in Figure 4). Conversely, for $n = 500$, an increase in agreement is observed only in equal mixture component scenarios, while in unequal cases,



FIGURE 10. Heatmaps of agreement ratios between pairs of information criteria when the dimensionality is 2×3 , $k = 6$ and the groups are close. The upper triangular matrices represent equal mixture weights configurations, while the lower triangular matrices represent unequal mixture weights configurations. The results are shown for the three sample sizes.



FIGURE 11. Heatmaps of agreement ratios between pairs of information criteria when the dimensionality is 7×6 , $k = 2$ and the groups are close. The upper triangular matrices represent equal mixture weights configurations, while the lower triangular matrices represent unequal mixture weights configurations. The results are shown for the three sample sizes.

the agreement declines. Finally, at $n = 1000$, the differences between the two mixture weight scenarios largely disappear, and a significant improvement in agreement values is observed.

In Figure 12, excluding the $n = 250$ configurations, a decline in agreement ratios is observed. Interestingly, the agreement among pairs of information criteria decreases rapidly as the sample size increases, except for a few notable cases. This decreasing trend is further illustrated in Figure 13, where we generally see the lowest levels of agreement across all considered configurations. The complexities introduced by the higher dimensions and increased number of groups significantly hinder the uniformity of component selection in the majority of cases.

To summarise, for a fixed dimensionality, the agreement generally declines as the number of groups increases. Additionally, when dimensionality is low, larger sample sizes tend to result in higher agreement. However, at higher dimensionalities, this pattern is only observed when $k = 2$.



FIGURE 12. Heatmaps of agreement ratios between pairs of information criteria when the dimensionality is 7×6 , $k = 4$ and the groups are close. The upper triangular matrices represent equal mixture weights configurations, while the lower triangular matrices represent unequal mixture weights configurations. The results are shown for the three sample sizes.



FIGURE 13. Heatmaps of agreement ratios between pairs of information criteria when the dimensionality is 7×6 , $k = 6$ and the groups are close. The upper triangular matrices represent equal mixture weights configurations, while the lower triangular matrices represent unequal mixture weights configurations. The results are shown for the three sample sizes.

4.2.3 Assessing the Impact of Starting Values

To assess the influence of starting values on the selection of the number of components, MVN mixtures are now fitted to each dataset by using the true values as the starting points. When the model being fitted has more components than the true model, the additional components are initialised by setting $\pi_j = 0.01$, with the other parameters initialised using the same random values obtained in the random approach. To ensure the mixture weights sum to one, the weights of the true components are slightly rescaled accordingly. Conversely, when the model being fitted has fewer components than the true model, only the true parameters of the components with the largest mixture weights are used for initialisation.

Results are shown in bar plots from Figures 14 to 19, for blocks from B.1 to B.6, respectively. More specifically, for each information criterion, a bar in these plots represents the difference in

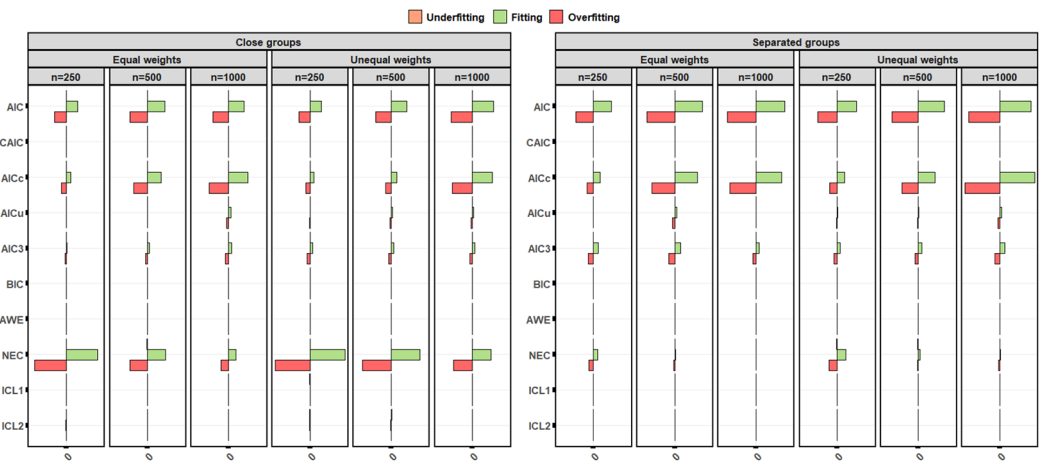


FIGURE 14. Bar plots reporting the differences (random minus true initialisation) in underfitting, fitting and overfitting counts for each information criterion when the dimensionality is 2×3 and $k = 2$. Positive values indicate higher counts for random initialisation, while negative values indicate higher counts for true initialisation.

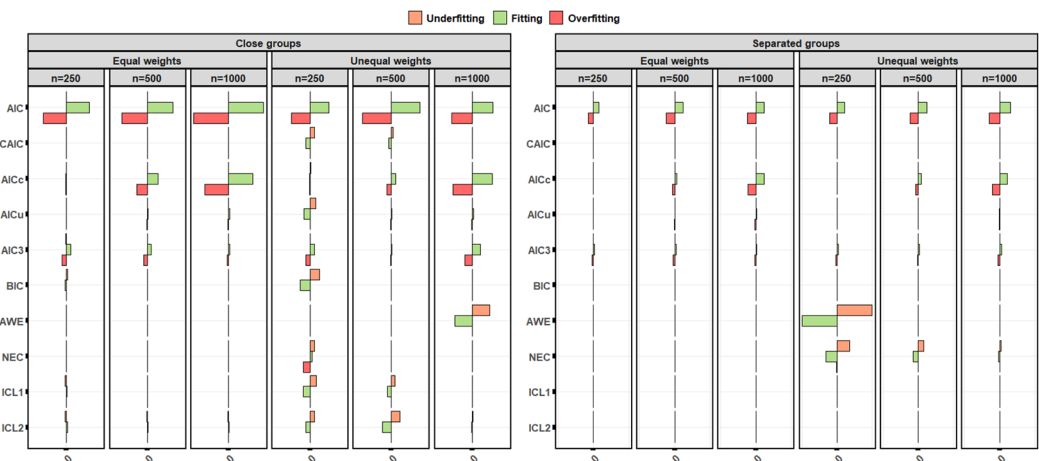


FIGURE 15. Bar plots reporting the differences (random minus true initialisation) in underfitting, fitting and overfitting counts for each information criterion when the dimensionality is 2×3 and $k = 4$. Positive values indicate higher counts for random initialisation, while negative values indicate higher counts for true initialisation.

counts between the two initialisation methods for a specific outcome: underfitting, fitting or overfitting. Thus, the first bar indicates the difference in underfitting occurrences, the second shows the difference in fitting occurrences and the third represents the difference in overfitting occurrences. A positive bar means that the random initialisation resulted in more occurrences of that outcome compared to the true initialisation, while a negative bar indicates that the true initialisation produced more occurrences. If a bar is zero, it signifies that both initialisation methods yielded the same number of occurrences for that particular outcome.

Starting with Figure 14, we observe that AIC, AICc and NEC generally produce better fitting results with random initialisation compared to the true initialisation. The latter tends to overfit more frequently, regardless of group separation, mixture weights or sample size.

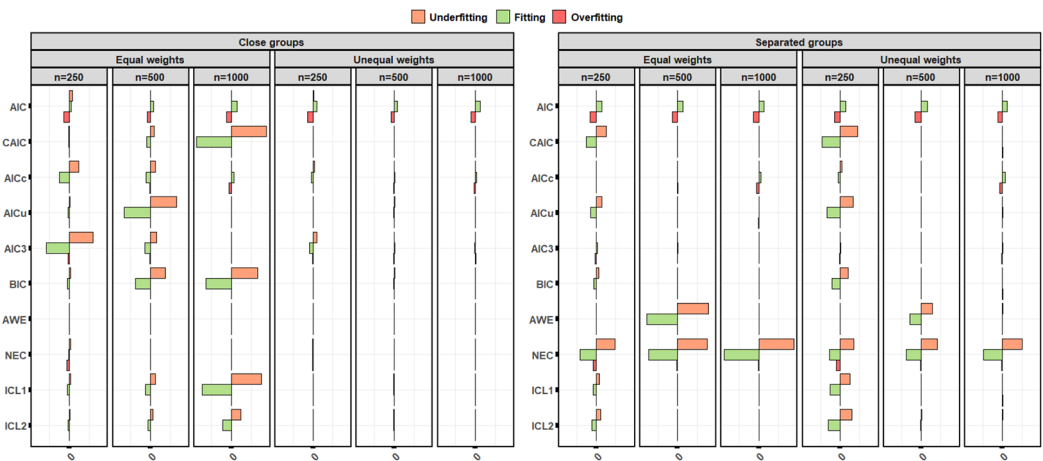


FIGURE 16. Bar plots reporting the differences (random minus true initialisation) in underfitting, fitting and overfitting counts for each information criterion when the dimensionality is 2×3 and $k = 6$. Positive values indicate higher counts for random initialisation, while negative values indicate higher counts for true initialisation.



FIGURE 17. Bar plots reporting the differences (random minus true initialisation) in underfitting, fitting and overfitting counts for each information criterion when the dimensionality is 7×6 and $k = 2$. Positive values indicate higher counts for random initialisation, while negative values indicate higher counts for true initialisation.

Figure 15 shows more targeted differences. AIC and AICc still show better fitting with random initialisation, especially in close groups configurations. In the unequal weights settings, the differences between the initialisations are minor, with the true initialisation slightly outperforming random initialisation in most cases.

In Figure 16, the results are more variable. Here, when AIC is considered, the random strategy still presents slightly better results. In the close groups scenarios, no relevant differences emerge in the unequal mixture weights cases, while some differences are present in the equal weights cases. Indeed, depending on the sample size, some specific information criteria provide better fitting results when the true initialisation is used. The same applies in the separated groups configurations, particularly when $n = 250$.

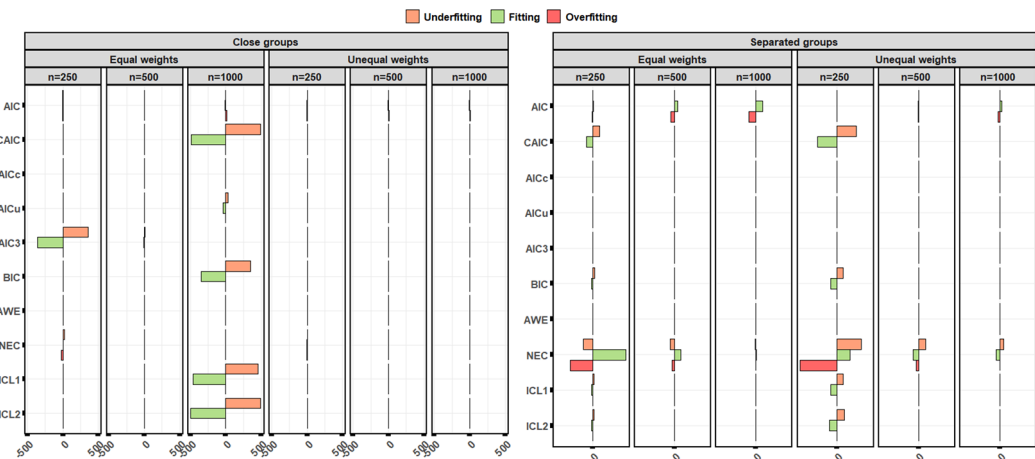


FIGURE 18. Bar plots reporting the differences (random minus true initialisation) in underfitting, fitting and overfitting counts for each information criterion when the dimensionality is 7×6 and $k = 4$. Positive values indicate higher counts for random initialisation, while negative values indicate higher counts for true initialisation.

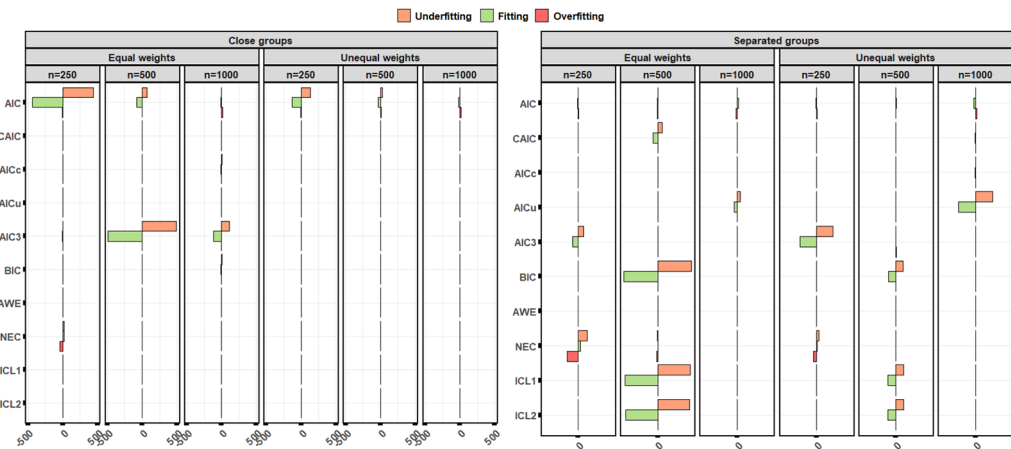


FIGURE 19. Bar plots reporting the differences (random minus true initialisation) in underfitting, fitting and overfitting counts for each information criterion when the dimensionality is 7×6 and $k = 6$. Positive values indicate higher counts for random initialisation, while negative values indicate higher counts for true initialisation.

In contrast to the previous blocks, in Figure 17, we observe the fewer differences between the two initialisation strategies. The outcomes are largely similar, except for NEC and, to a lesser extent, AIC, where random initialisation shows a slight advantage.

Figure 18 presents similar results. In most cases, there is little difference between the two initialisation methods. In close groups configurations, true initialisation leads to better fitting for AIC3 when $n = 250$ and for CAIC, BIC, ICL1 and ICL2 when $n = 1000$. In separated configurations, AIC and NEC favour random initialisation, while CAIC, BIC, ICL1 and ICL2 slightly favour true initialisation when $n = 250$ in the unequal mixture weights setting.

Finally, in Figure 19, notable differences are only observed for AIC (when $n = 250$) and AIC3 (when $n = 500$) in close groups settings, favouring true initialisation. Similarly, BIC, ICL1 and ICL2 favour the true strategy in separated settings when $n = 500$.

To summarise, in most cases, both initialisation methods produce similar results. Where differences arise, they sometimes favour random initialisation and sometimes the true initialisation, depending on the configuration. No single information criterion consistently shows a preference for one initialisation strategy over the other.

4.2.4 Assessing the Impact of Data Vectorisation

In this subsection, we investigate the impact of vectorising matrix-variate data on the performance of information criteria when fitting MN mixtures. While the focus of this paper is on MVN mixtures, this comparison highlights the potential consequences of treating matrix-variate data as vectorized. As discussed in Section 2, when matrix-variate data that follow an MVN distribution are vectorized, the resulting vector follows an MN distribution, with the covariance matrix adopting a Kronecker product structure that reflects the original row and column covariances. By comparing this vectorized approach to the matrix-variate case, we aim to assess its influence on model selection criteria.

For the sake of simplicity, we limit our discussion to the blocks B.1 and B.3, that is, where $k = 2$. Specifically, the 2×3 matrices in B.1 are transformed into 6-dimensional vectors, while the 7×6 matrices in B.3 become 42-dimensional vectors. The results for the vectorized block B.1 are reported in Figure 20, while those about the vectorisation of B.3 are depicted in Figure 21.

Beginning our discussion with the analysis of Figure 20, we immediately observe several differences compared to the results of Figure 1. When focusing on the close-group configurations, AWE, ICL1 and ICL2 almost regularly underestimate the number of groups, closely mimicked by CAIC and BIC which properly perform only when $n = 1000$. Excluding CAIC, and to a lesser extent AIC, we observe that AIC-based information criteria generally select the correct number of groups as n increases. There are no notable differences between the balanced and unbalanced structures, although the overall performance tends to improve as the separation between groups increases.

When Figure 21 is considered, it becomes evident that (almost) none of the information criteria yield reliable results. The explosion in the number of parameters heavily affects their penalty term, resulting in biased outcomes. The computation of AICu and AICc is substantially

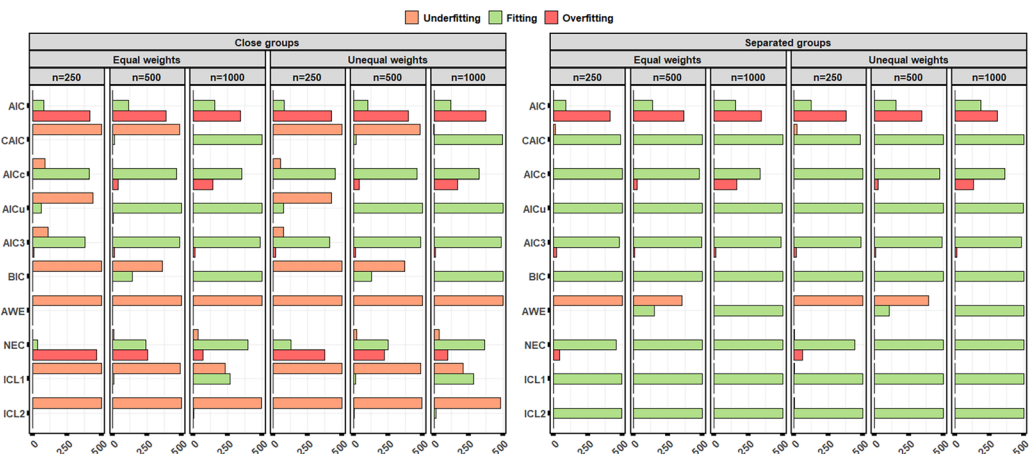


FIGURE 20. Bar plots reporting the number of times each information criterion underfits, fits or overfits the true number of groups in the data upon vectorisation of the 2×3 matrices and $k = 2$.

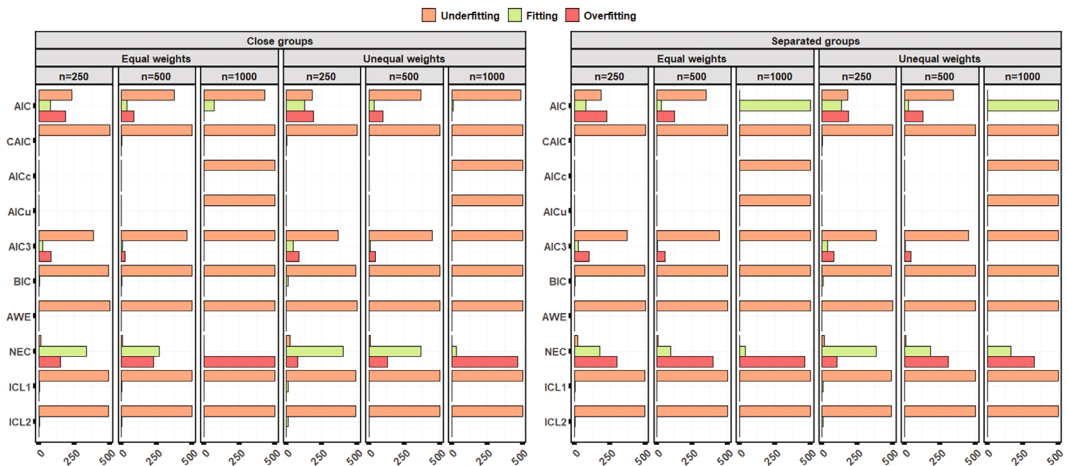


FIGURE 21. Bar plots reporting the number of times each information criterion underfits, fits or overfits the true number of groups in the data upon vectorisation of the 7×6 matrices and $k = 2$.

impossible when $n = 250$ and $n = 500$ (refer to Section 3). Only NEC, which does not include the number of parameters in its definition, appears to estimate the correct number of groups better than the others. However, its performance declines as the sample size increases. Relatedly, when $n = 1000$, only AIC properly works in the separated configurations.

To summarise, the vectorisation process, along with the loss of interpretability of the two sources of variability and the increase in the number of parameters (see Section 2), leads to the improper functioning of the information criteria. Hence, we stress the importance of avoiding such an approach when dealing with data presenting a matrix-variate structure.

5 Conclusions

In this paper, we aimed to investigate a critical issue involving finite mixture models, that is, the selection of the number of components. We conducted our analysis in the context of matrix-variate data, using mixtures of matrix-variate normal distributions, given the lack of results within this branch of the literature. We considered 10 information criteria and set up an extensive simulation study to assess their effectiveness in selecting the correct number of components. Based on the obtained results, the following considerations can be drawn:

- 1 In terms of fitting, underfitting and overfitting, no single information criterion emerges as the optimal choice across all the investigated data setups.
 - The BIC, which has been the most commonly adopted information criterion in this framework (refer to Section 1) should not be indiscriminately used. In particular, as the matrix dimensionality and the number of groups increases, it tends to select a lower number of mixture components.
 - The AWE should be avoided in the matrix-variate framework since it tends to underfit even in lower dimensional settings.
 - The ICL1 and ICL2, which are BIC-adjusted criteria, generally perform worse than the BIC. Nevertheless, ICL1 provides better results than ICL2, despite the large use of ICL2 in the model-based clustering literature.
 - NEC, the only criteria that does not incorporate the number of parameters in the penalty term, tends to overfit in most of the data configurations.

- The AIC, which is known to overfit, exhibits this behaviour in the matrix-variate context as well. However, its less stringent penalty term renders it the most promising criterion when matrix dimensionality is high.
 - Among the AIC-based criteria, AICc generally shows the poorest performance. In contrast, AIC3 emerges as the most viable option across most data configurations. Although CAIC, AICc and AICu exhibit some differences in certain cases, their overall performance is largely similar. In light of these results, the suggestion is to consider the data and model characteristics, in terms of the number of components to be fitted, for choosing which set of information criteria to use. Blinding selecting an information criterion could result in an underfitting or overfitting.
- 2 The above results were obtained using the random initialisation strategy for the ECM algorithm. When a different approach, using the true data-generating parameters as starting values, was applied, we obtained practically similar results. Where differences did occur, they sometimes favoured random initialisation and other times the true initialisation, depending on the configuration. No single information criterion consistently preferred one initialisation strategy over the other.
 - 3 From the analysis of agreement ratios between pairs of information criteria, we observe that, for a given dimensionality, the agreement between criteria generally decreases as the number of groups increases. Furthermore, when dimensionality is low, an increase in sample size often leads to higher agreement. However, when dimensionality is higher, this trend appears only when $k = 2$. Indeed, the complexities introduced by the higher dimensions and increased number of groups significantly hinder the uniformity of component selection in the majority of cases.
 - 4 Vectorising the data and then using multivariate models should be avoided. Indeed, alongside the interpretational drawbacks highlighted in Section 2, this approach impacts the correct selection by the information criteria. This issue becomes increasingly significant as the matrix dimensionality increases. Notably, when the dimensionality is high, only NEC seems to estimate the correct number of groups more effectively than the others; however, its performance deteriorates as the sample size grows. Additionally, when $n = 1000$, only AIC functions appropriately in the separated configurations.

Concerning further developments of the present work, several avenues could be explored. First, incorporating covariates into the mixture models could be pursued, allowing for an analysis of information criterion performance within a matrix-variate mixture of regressions framework. Additionally, investigating scenarios and models that depart from the normality assumption would be also valuable.

Acknowledgements

Salvatore D. Tomarchio acknowledges the ‘PIAno di inCentivi per la Ricerca di Ateneo 2020/2022’, Linea di intervento 3 – STARTING GRANT 2020, University of Catania. A. Punzo has been partially supported by MUR, grant no. 2022XRHT8R (CUP: E53D23005950006)–The SMILE project: Statistical Modelling and Inference to Live the Environment, funded by the European Union–Next Generation EU. Open access publishing facilitated by Università degli Studi di Catania, as part of the Wiley - CRUI-CARE agreement.

Conflict of Interest

None of the authors have a conflict of interest to disclose.

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**(6), 716–723.
- Akogul, S. & Erisoglu, M. (2016). A comparison of information criteria in clustering based on mixture of multivariate normal distributions. *Mathematical and Computational Applications*, **21**(3), 34.
- Amato, F., Jacques, J. & Prim-Allaz, I. (2024). Clustering longitudinal ordinal data via finite mixture of matrix-variate distributions. *Statistics and Computing*, **34**(2), 81.
- Bacci, S., Pandolfi, S. & Pennoni, F. (2014). A comparison of some criteria for states selection in the latent Markov model for longitudinal data. *Advances in Data Analysis and Classification*, **8**, 125–145.
- Banfield, J.D. & Raftery, A.E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 803–821.
- Biernacki, C., Celeux, G. & Govaert, G. (1999). An improvement of the NEC criterion for assessing the number of clusters in a mixture model. *Pattern Recognition Letters*, **20**(3), 267–272.
- Biernacki, C., Celeux, G. & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(7), 719–725.
- Biernacki, C., Celeux, G. & Govaert, G. (2003). 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.
- Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, **52**(3), 345–370.
- Bozdogan, H. (1994). Mixture-model cluster analysis using model selection criteria and a new informational measure of complexity. In *Proceedings of the first US/Japan conference on the frontiers of statistical modeling: An informational approach: Volume 2 multivariate statistical modeling*, pp. 69–113, Springer.
- Buckby, J., Wang, T., Fletcher, D., Zhuang, J., Takeo, A. & Obara, K. (2023). Finding the number of latent states in hidden Markov models using information criteria. *Environmental and Ecological Statistics*, **30**(4), 797–825.
- Celeux, G. & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, **13**, 195–212.
- Chen, J. & Khalili, A. (2009). Order selection in finite mixture models with a nonsmooth penalty. *Journal of the American Statistical Association*, **104**(485), 187–196.
- Depraetere, N. & Vandebroek, M. (2014). Order selection in finite mixtures of linear regressions: Literature review and a simulation study. *Statistical Papers*, **55**(3), 871–911.
- Gallaughier, M.P.B. & McNicholas, P.D. (2018). Finite mixtures of skewed matrix variate distributions. *Pattern Recognition*, **80**, 83–93.
- Gallaughier, M.P.B., Tomarchio, S.D., McNicholas, P.D. & Punzo, A. (2022). Model-based clustering via skewed matrix-variate cluster-weighted models. *Journal of Statistical Computation and Simulation*, **92**(13), 2645–2666.
- Hubert, L. & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, **2**, 193–218.
- Hui, F.K.C., Warton, D.I. & Foster, S.D. (2015). Order selection in finite mixture models: Complete or observed likelihood information criteria? *Biometrika*, **102**(3), 724–730.
- Hurvich, C.M. & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**(2), 297–307.
- McLachlan, G.J. & Peel, D. (2000). *Finite Mixture Models*. John Wiley & Sons.
- McLachlan, G.J. & Rathnayake, S. (2014). On the number of components in a Gaussian mixture model. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **4**(5), 341–355.
- McNicholas, P.D. (2016). *Mixture Model-Based Classification*. Chapman and Hall/CRC.
- McQuarrie, A., Shumway, R. & Tsai, C.-L. (1997). The model selection criterion AICu. *Statistics & Probability Letters*, **34**(3), 285–292.
- Melnykov, V. & Zhu, X. (2018). On model-based clustering of skewed matrix data. *Journal of Multivariate Analysis*, **167**, 181–194.
- Melnykov, V. & Zhu, X. (2019). Studying crime trends in the USA over the years 2000–2012. *Advances in Data Analysis and Classification*, **13**, 325–341.
- Meng, X.-L. & Rubin, D.B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, **80**(2), 267–278.
- R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria. <https://www.R-project.org/>
- Sarkar, S., Zhu, X., Melnykov, V. & Ingrassia, S. (2020). On parsimonious models for modeling matrix data. *Computational Statistics & Data Analysis*, **142**, 106822.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**(2), 461–464.
- Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections: Further analysis of the data by Akaike's. *Communications in Statistics-Theory and Methods*, **7**(1), 13–26.

- Tomarchio, S.D. (2024). Matrix-variate normal mean-variance Birnbaum–Saunders distributions and related mixture models. *Computational Statistics*, **39**, 405–432.
- Tomarchio, S.D. & Gallagher, M.P.B. (2024). Mixtures of regressions using matrix-variate heavy-tailed distributions. *Advances in Data Analysis and Classification*, 1–28.
- Tomarchio, S.D., Gallagher, M.P.B., Punzo, A. & McNicholas, P.D. (2022). Mixtures of matrix-variate contaminated normal distributions. *Journal of Computational and Graphical Statistics*, **31**(2), 413–421.
- Tomarchio, S.D., McNicholas, P.D. & Punzo, A. (2021). Matrix normal cluster-weighted models. *Journal of Classification*, **38**(3), 556–575.
- Tomarchio, S.D., Punzo, A. & Bagnato, L. (2020). Two new matrix-variate distributions with application in model-based clustering. *Computational Statistics & Data Analysis*, **152**, 107050.
- Tomarchio, S.D., Punzo, A. & Maruotti, A. (2022). Parsimonious hidden Markov models for matrix-variate longitudinal data. *Statistics and Computing*, **32**(3), 53.
- Tomarchio, S.D., Punzo, A. & Maruotti, A. (2023). Matrix-variate hidden Markov regression models: Fixed and random covariates. *Journal of Classification*, **41**, 429–454.
- Viroli, C. (2011a). Finite mixtures of matrix normal distributions for classifying three-way data. *Statistics and Computing*, **21**, 511–522.
- Viroli, C. (2011b). Model based clustering for three-way data structures. *BAYESIAN ANALYSIS*, **6**(4), 573–602.
- Walaa, EL-S & Ismail, M. (2022). Order selection in a finite mixture of Birnbaum–Saunders distributions. *Austrian Journal of Statistics*, **51**(3), 98–120.
- Zhu, X., Sarkar, S. & Melnykov, V. (2022). Mattransmix: An R package for matrix model-based clustering and parsimonious mixture modeling. *Journal of Classification*, 1–24.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

[Received May 2024; accepted December 2024]