UNIVERSITY OF MESSINA

DOCTORAL THESIS

Advances in finite mixture models: new methodologies and applications

Author:
Salvatore Daniele Tomarchio

Supervisor:
Prof. Antonio Punzo

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in

Economics, Management and Statistics
Curriculum: Statistics
XXXIII Cycle

Abstract

The extent of finite mixture models has widened considerably over the last century, from both a theoretical and a practical point of view. Their usefulness and flexibility is discussed in this thesis, which consists of a collection of four manuscripts that have as common background new methodologies and applications of finite mixture models. The first two manuscripts focus on specific economics and financial topics, and the finite mixture models are mainly used as a mathematical device for obtaining a flexible and tractable density. This has important consequences for the estimation of some commonly used risk measures. The other two manuscripts aim to use finite mixture models for clustering in a matrix-variate framework. In all the manuscripts, parameter estimation is carried by using the maximum-likelihood approach, implemented directly or via variants of the expectation-maximization algorithm. Both simulated and real datasets are used for illustrative purposes in each manuscript.

Acknowledgements

First and foremost, I would like to express my deepest appreciation to my supervisor Prof. Antonio Punzo. His presence and support during all my Ph.D. Program has been crucial for the completion of this thesis. I would also express my gratitude to him for all the suggestions and opportunities that made me grow professionally and personally. I will thank him forever.

I would like to show my gratitude also to Prof. Salvatore Ingrassia, for the opportunities and the support he gave me during my Ph.D. Program, Prof. Paul D. McNicholas, who hosted me at McMaster University during my visiting studies and for the supervision on one chapter of the thesis, and Prof. Luca Bagnato for the assistance on one chapter of the thesis. I would like to thank also Prof. Edoardo Otranto for the organizational support and the assistance during the Ph.D. Program.

I would like to acknowledge my family for the constant sustenance and understanding throughout the course of my education, especially in the most difficult moments. Lastly, but not least, I would like to thank my girlfriend and future wife Sara. She has always believed strongly in me and every day she gives me her support. Let this be just a new step to make our dreams come true.

Contents

List of Figures

List of Tables

List of Abbreviations

AECM	**A**lternating **E**xpectation-**C**onditional **M**aximization
ARI	**A**djusted **R**and **I**ndex
BFGS	**B**royden–**F**letcher–**G**oldfarb–**S**hanno
BIC	**B**ayesian **I**nformation **C**riterion
C-BK	**C**hen-**B**eta **K**ernel
CZ-BK	**C**alabrese **Z**enga-**B**eta **K**ernel
CWM	**C**luster **W**eighted **M**odel
DGP	**D**ata **G**enerating **P**rocess
EAD	**E**xposure **A**t **D**efault
ECM	**E**xpectation-**C**onditional **M**aximization
ECME	**E**xpectation-**C**onditional **M**aximization **E**ither
EGB2	**E**xponential **G**eneralized **B**eta of type **2**
EM	**E**xpectation-**M**aximization
FMR	**F**inite **M**ixtures of **R**egression models with fixed covariates
FMRC	**F**inite **M**ixtures of **R**egression models with **C**oncomitant variables
GB1	**G**eneralized **B**eta of type **1**
GK	**G**aussian **K**ernel
H-BK	**H**agmann-**B**eta **K**ernel
LGD	**L**oss **G**iven **D**efault
LN	**L**og **N**ormal
LN-DUC	**L**og **N**ormal-**D**ichotomous **U**nimodal **C**ompound
LR	**L**ikelihood-**R**atio
ML	**M**aximum **L**ikelihood
MSE	**M**ean **S**quared **E**rror
MV-FMR	**M**atrix-**V**ariate-**F**inite **M**ixtures of **R**egression models with fixed covariates
MVN	**M**atrix-**V**ariate **N**ormal
MVN-CWM	**M**atrix-**V**ariate **N**ormal-**C**luster **W**eighted **M**odel
MVN-Ms	**M**atrix-**V**ariate **N**ormal **M**ixtures
MVNSM	**M**atrix-**V**ariate **N**ormal **S**cale **M**ixture
MVSEN	**M**atrix-**v**ariate **S**hifted **E**xponential **N**ormal
MVSEN-Ms	**M**atrix-**v**ariate **S**hifted **E**xponential **N**ormal **M**ixtures
MVTIN	**M**atrix-**V**ariate **T**ail-**I**nflated **N**ormal
MVTIN-Ms	**M**atrix-**V**ariate **T**ail-**I**nflated **N**ormal **M**ixtures
MV*t*-Ms	**M**atrix-**V**ariate *t* **M**ixtures
NSM	**N**ormal **S**cale **M**ixture
PD	**P**robability of **D**efault
PDF	**P**robability **D**ensity **F**unction
PMF	**P**robability **M**ass **F**unction
PORT-MO$_p$	**P**eaks **O**ver a **R**andom **T**hreshold-**M**ean of **O**rder p
TVaR	**T**ail-**V**alue **A**t **R**isk
UG	**U**nimodal hump-shaped **G**amma
UG-DUC	**U**nimodal hump-shaped **G**amma-**D**ichotomous **U**nimodal **C**ompound
VaR	**V**alue **A**t **R**isk

Chapter 1

Introduction

The extent of finite mixture models has widened considerably over the last century, from both a theoretical and a practical point of view. Indeed, finite mixtures models have provided a mathematical-based tool for the statistical modeling of a broad variety of random phenomena. Fields in which they have been successfully studied and applied include for instance medicine, biology, marketing, finance, engineering and social sciences (for examples see, Schlattmann, 2009; Wedel and Kamakura, 2012; Bouguila and Fan, 2020). Their usefulness and flexibility is also discussed in this thesis, which consists of a collection of four manuscripts that have as common background new methodologies and applications of finite mixture models.

After a brief presentation of some preliminary concepts in Chapter 2, a family of zero-and-one inflated mixture models is considered in Chapter 3 for the modelization of the loss given default (LGD) distribution. This chapter is based on the publication Tomarchio and Punzo (2019). The LGD is an important parameter that banks and other financial institutions have to properly estimate. However, its distribution has posed substantial challenges, since it is generally defined between 0 and 1 (both included), often exhibits bimodality and, more in general, multimodality and it has a high amount of observations at the boundary values 0 and 1. With the zero-and-one inflated mixture models proposed, it is possible to take into account all these peculiar characteristics. To allow for more flexible shapes of the mixture components on $(0,1)$, other then considering distributions already defined on $(0,1)$, distributions defined on $(-\infty,\infty)$ and mapped to $(0,1)$ via the inverse-logit transformation are investigated. This yields to a family of thirteen zero-and-one inflated mixture models, which are applied to two real data sets: one from an European Bank and the other from the Bank of Italy. The best models, selected via a classical information criterion, are then compared with several well-established semi-parameteric/nonparametric approaches via a convenient simulation-based procedure.

Chapter 4 discusses the modelization of the insurance losses distribution, that is crucial in the insurance industry, and it based on the publication Tomarchio and Punzo (2020). This distribution is generally highly positively skewed, unimodal hump-shaped, and with a heavy right tail. A profitable way to accommodate these characteristics is by using a dichotomous unimodal compound model. It consists of 2-component mixture model, in which the first component (defined on a positive support, reparameterized with respect to the mode and to another parameter related to the distribution variability) is mixed with an inflating component (with the same support and mode, scaled variability parameter and small prior probability). The proposed model can also allow for automatic detection of typical and atypical losses via a simple procedure based on maximum *a posteriori* probabilities. The unimodal gamma and log-normal distributions are considered as examples for the mixture components. The resulting models are firstly evaluated in a sensitivity study and then fitted to two real insurance loss data sets, along with several competitors. The

comparisons are made via a classical information criterion and by the computation of some well-known risk measures.

Chapter 5 introduces two new matrix-variate distributions, which are subsequently used as components of the corresponding mixture models. This chapter is based on the publication Tomarchio *et al.* (2020). Both distributions are heavy-tailed generalization of the matrix-variate normal distribution, with respect to which are characterized by only one additional parameter that governs the tail-weight. The resulting mixtures, being able to handle data with atypical observations in a better way than the matrix-variate normal mixture, can avoid the disruption of the true underlying group structure. Different variants of the well-known expectation-maximization (EM) algorithm are implemented for parameter estimation and tested in terms of computational times and parameter recovery. These mixture models are fitted to simulated and real data sets, and their fitting and clustering performances are analyzed and compared to those obtained by other well-established competitors.

By continuing within the matrix-variate framework, Chapter 6 introduces the first matrix-variate cluster weighted model (CWM). This chapter is based on a paper currently under review at the *Journal of Classification*. Specifically, it consider finite mixture of regression models. The traditional way of regressing data in presence of an underlying grouping structure is via the finite mixtures of regressions with fixed covariates. However, they assume assignment independence, i.e. the allocation of data points to the clusters is made independently of the distribution of the covariates. In order to take into account this last aspect, finite mixtures of regressions with random covariates, also known as CWMs, have been proposed in the univariate and multivariate literature. Here, the CWM approach is extended to matrix data by using the matrix normal distribution both for the cluster-specific conditional distribution of the responses given the covariates and the cluster-specific marginal distribution of the covariates. Maximum likelihood parameter estimates are derived by using an ECM algorithm. The parameter recovery and the classification assessment of the algorithm are analyzed on simulated data. Finally, two real data applications concerning educational indicators and the Italian non-life insurance market are presented.

Lastly, Chapter 7 drawn some conclusions as well as possible extensions for each the four manuscripts discussed.

Chapter 2

Background on finite mixture models

This chapter contains background aspects that are useful for a better understanding of the next chapters. In detail, it includes concepts that will mentioned or used in more than one chapter throughout the thesis. This should avoid their repetition in different chapters and make easier the reading.

2.1 A general overview of finite mixture models

A random variable X arises from a finite mixture model if its probability density function (PDF) can be written

$$g(x;\boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_k f(x;\boldsymbol{\phi}_k), \tag{2.1}$$

where π_k is the mixing proportion of the k-th component, with $\pi_k > 0$ and such that $\sum_{k=1}^{K} \pi_k = 1$, $f(x;\boldsymbol{\phi}_k)$ is the PDF of the k-th component with parameters $\boldsymbol{\phi}_k$, and $\boldsymbol{\Theta}$ contains all of the parameters of the mixture. Notice that, as commonly done in the mixture modeling literature, the component densities are taken to be of the same type. Extensive details on finite mixture models can be found in the well-known texts by Titterington *et al.*, 1985; McLachlan and Peel, 2000; Frühwirth-Schnatter, 2006; McNicholas, 2016.

According to Titterington *et al.* (1985), finite mixture models are generally used in two different ways. In the *indirect applications*, they are used as a mathematical device in order to obtain a flexible and tractable density. Typical examples are mixture models with smooth curve-fitting design, which resemble a nonparametric density technique, or 2-component mixture models in which one of the components has an inflated variance, so that the overall model can approximate an intractable heavy-tailed distribution. Both examples are extremely close to the models discussed in Chapters 3 and 4. Therefore, they represent an indirect application of finite mixture models.

In the *direct applications*, finite mixture models are considered a powerful tool for clustering, and each mixture component is assumed to represent a group (or cluster) in the original data. The aim is to recover the underlying grouping structure and to evaluate the classification produced by the model. Therefore, the manuscripts discussed in Chapters 5 and 6 can be considered mainly devoted to a *direct application*. However, it is important to notice that both *applications* can be considered two sides of the same coin, since the dividing line between them is not always clear and, in any case, relies on which of the two is more applicable by the researcher.

2.2 The scale mixture model

Historically, the most popular finite mixture model has made use of the normal distribution for $f(x;\boldsymbol{\phi}_k)$ in (2.1) (McNicholas, 2016). However, for many real phenomena, the tails of the normal distribution are lighter than required, with a direct effect on the corresponding finite mixture model. A common way to generalize the normal distribution, in order to obtain distributions with heavier tails, is by means of the normal scale mixture model (McLachlan and Peel, 2000). Specifically, a random variable X arises from a normal scale mixture (NSM) model if its PDF is

$$f_{\text{NSM}}(x;\mu,\sigma^2,\boldsymbol{\nu}) = \int_0^\infty f_{\text{N}}(x;\mu,\sigma^2/w)h(w;\boldsymbol{\nu})dw, \tag{2.2}$$

where $f_{\text{N}}(x;\cdot)$ is the PDF of the normal distribution and $h(w;\cdot)$ is the PDF or PMF (probability mass function) of a mixing random variable W with positive support. Therefore, the PDF in (2.2) is a finite/continuous mixture of normal distributions on σ^2 obtained via a convenient discrete/continuous mixing distribution, whose parameter(s) $\boldsymbol{\nu}$ control the tailedness of the model. Several well-known distributions can be obtained from (2.2), such as the t, Pearson type VII, variance gamma, logistic, symmetric generalized hyperbolic and power exponential distributions (Boris Choy and Chan, 2008; Dang *et al.*, 2015; Lee and McLachlan, 2019).

It is clear that model (2.2) can be also used to generalize distributions different from the normal one. For example, Punzo *et al.* (2018) consider unimodal hump-shaped positively skewed distributions, defined on a positive support, instead of $f_{\text{N}}(x;\cdot)$. Therefore, a scale mixture model can be intended as a general tool for re-weighting the tails of a distribution.

This model is the basis on which the manuscripts in Chapters 4 and 5 are based.

2.3 Parameter estimation of finite mixture models

2.3.1 The EM algorithm and its variants

Since the advent of the expectation-maximization (EM) algorithm by Dempster *et al.* (1977), the maximum likelihood (ML) approach has been by far the most commonly used to estimate the parameters of a finite mixture model. Let $\mathcal{S} = \{X_i\}_{i=1}^N$ be a sample of N independent observations from model (2.1). Then, the incomplete-data likelihood function is

$$L(\boldsymbol{\Theta}) = \prod_{i=1}^N g(x_i;\boldsymbol{\Theta}) = \prod_{i=1}^N \left[\sum_{k=1}^K \pi_k f(x;\boldsymbol{\phi}_k)\right]. \tag{2.3}$$

In the context of the EM algorithm, $\mathcal{S}$ is considered incomplete because, for each observation, we do not know its component membership. Let $\mathbf{z}_i = (z_{i1},\ldots,z_{iK})$ be the component membership vector such that $z_{ik} = 1$ if X_i comes from group k and $z_{ik} = 0$ otherwise. Now, the complete-data are $\mathcal{S}_c = \{X_i,\mathbf{z}_i\}_{i=1}^N$, and the complete-data likelihood is

$$L_c(\boldsymbol{\Theta}) = \prod_{i=1}^N \prod_{k=1}^K [\pi_k f(x;\boldsymbol{\phi}_k)]^{z_{ik}}. \tag{2.4}$$

The EM algorithm approaches the problem of solving the incomplete-data likelihood function in (2.3) indirectly, by proceeding iteratively in terms of the logarithm of the complete-data likelihood in (2.4), say $l_c(\boldsymbol{\Theta})$. Since the latter is unobservable,

it is replaced by its conditional expectation given the observed data and using the current fit for $\boldsymbol{\Theta}$. More specifically, let $\boldsymbol{\Theta}^0$ be an initial value for $\boldsymbol{\Theta}$. Then on the first iteration, the E-step requires the calculation of

$$Q\left(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(0)}\right) = E_{\boldsymbol{\Theta}^{(0)}}\left\{l_c\left(\boldsymbol{\Theta}\right)|\mathcal{S}\right\}. \tag{2.5}$$

The M-step requires the maximization of $Q\left(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(0)}\right)$ with respect to $\boldsymbol{\Theta}$, i.e.

$$\boldsymbol{\Theta}^{(1)} = \arg\max_{\boldsymbol{\Theta}} Q\left(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(0)}\right). \tag{2.6}$$

Then, both steps are carried out again, by replacing $\boldsymbol{\Theta}^{(0)}$ with $\boldsymbol{\Theta}^{(1)}$ in (2.5), and $\boldsymbol{\Theta}^{(1)}$ with $\boldsymbol{\Theta}^{(2)}$ in (2.6). This procedure is repeated until the difference between two consecutive likelihood values is lower than an arbitrary small threshold. In Chapters 5 and 6, and by following the notation used by Melnykov and Zhu (2019), the parameters marked with one dot correspond to the updates at the previous iteration and those marked with two dots represent the updates at the current iteration.

The EM algorithm, as previously described, is used for parameter estimation in Chapter 3. However, in some situations the EM algorithm cannot be directly implemented, and modified versions of the algorithm must be adopted. An example is the expectation-conditional maximization (ECM) algorithm proposed by Meng and Rubin (1993). The only difference with respect to the EM algorithm is that the M-step is replaced by a sequence of simpler and computationally convenient CM-steps. It will be used for parameter estimation in Chapters 5 and 6. Other examples, are the expectation-conditional maximization either (ECME) algorithm (Liu and Rubin, 1994) and the alternating expectation-conditional maximization (AECM) algorithm (Meng and Van Dyk, 1997). Both algorithms generalize the ECM and will be considered for parameter estimation in Chapter 5. More specifically, the ECME allows to maximize on some or all of the CM-steps the incomplete-data log-likelihood, while the AECM algorithm always maximize the complete-data log-likelihood in all the CM-steps, but the complete data are allowed to be different on each CM-step.

2.3.1.1 Initialization strategies

The choice of the starting values constitutes an important aspect for any EM-based algorithm, since the solution at convergence can highly depend on its starting position. This topic has been vastly investigated in the literature (for details, see e.g. McLachlan and Peel, 2000; Biernacki *et al.*, 2003; Melnykov and Melnykov, 2012) and, in a broad way, it is possible to group the initialization strategies in two main categories, depending on which of the following two paths is followed:

1. start from the M-step by providing an initial value to the quantities involved in (2.5);

2. start from the E-step by providing an initial value to the parameters of the model.

For each category, several methods have been proposed. If we start from the M-step, and by assuming that only the updates for the posterior probabilities z_i are required in (2.5), a possibility is to use the classification produced by some clustering algorithm such as k-means or, say, a hierarchical procedure. Alternatively, such "hard" values can be randomly generated via a multinomial distribution with probabilities

$(1/K, \ldots, 1/K)$ (Mazza *et al.*, 2018). Another way of specifying initial values for the z_i is by using "soft" values, which can be randomly generated by a uniform distribution, and that are subsequently normalized in order to sum to 1. It is clear that when (2.5) does not require only the updates for the z_i, applying such approach may become more complicated, and should be carefully evaluated according to the case under consideration.

When we start from the E-step, a common strategy consists in randomly draw values for some or all the parameters. For example, in the case of multivariate Gaussian mixtures, McLachlan and Peel (2000) suggest to randomly generate the component means from a Gaussian distribution having mean and covariance matrix equal to the corresponding mean and covariance sample estimates. The component covariance matrices are then all initialized by using the sample covariance matrix and the mixture proportions are all assumed to be $1/K$. Another strategy, introduced by Biernacki *et al.* (2003), is the so-called short-EM initialization. It consists in H short runs of the EM algorithm for different random set of parameters. The term "short" means that the algorithm is run for a very small number of iterations s, without waiting for convergence. Then, the parameter set producing the highest log-likelihood value is used to initialize the final EM algorithm.

It should be noted that, regardless of the chosen approach, the log-likelihood for finite mixture models usually has multiple roots, corresponding to local maxima. Therefore, in order to find the global maximum, the algorithm should be run multiple times from a wide choice of starting values and then the solution corresponding to the largest log-likelihood must be selected.

2.3.2 A glimpse on alternative estimation approaches

The EM algorithm and its variants are the most commonly used tools for parameter estimation of finite mixture models (McLachlan and Krishnan, 2007). However, other methods are advocated in the literature as possible alternatives to EM-type algorithms. For instance, it might be convenient to directly maximize the log-likelihood of the mixture. This is particularly useful when closed-form estimates are not available or when the log-likelihood can be easily evaluated (MacDonald, 2014). To this end, a general-purpose numerical optimizer, such as the quasi-Newton type Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm (Fletcher, 2013), can be used. This approach will be implemented in Chapter 4.

A stochastic-type technique is provided by the Simulated Annealing (SA) algorithm. To optimize a real-valued function $h(\boldsymbol{\Theta})$ on a compact set D, the SA generates an inhomogeneous Markov process on D depending on a positive parameter T. This Markov process has the following Gibbs distribution as its unique stationary distribution

$$p_T(\boldsymbol{\Theta}) = \frac{\exp\left(-h(\boldsymbol{\Theta})/T\right)}{\int_D \exp\left(-h(\boldsymbol{\Theta})/T\right) d\boldsymbol{\Theta}}, \quad \boldsymbol{\Theta} \in D. \tag{2.7}$$

In the limit as T tends to 0, the stationary distribution tends to a distribution concentrated on the points of global optimum of h. In a homogeneous version of the algorithm, a sequence of homogeneous Markov chains is generated at decreasing values of T. The general algorithm requires the following steps:

1. select a starting value for T, say T_0 and for $\boldsymbol{\Theta}$, say $\boldsymbol{\Theta}_0$, with $h(\boldsymbol{\Theta}_0) = h_0$;
2. choose a proposed point $\boldsymbol{\Theta}_1$ at random from a neighborhood of $\boldsymbol{\Theta}_0$ and compute the corresponding h-value;

3. calculate $\Delta_1 = h(\boldsymbol{\Theta}_1) - h(\boldsymbol{\Theta}_0)$. If $\Delta_1 \leq 0$, move to the new point $\boldsymbol{\Theta}_1$, otherwise draw a value u from the uniform distribution over $[0, 1]$. Accept the new point $\boldsymbol{\Theta}_1$ if $u \leq \exp(-\Delta_1/T)$, i.e., $\exp(-\Delta_1/T)$ is the probability of acceptance;

4. repeat steps 2 and 3, after updating the appropriate quantities, until an equilibrium has been reached by the application of a convenient stopping criterion;

5. lower T according to an "annealing schedule" and start at Step 2 with the equilibrium value at the previous T as the initial value. Again a suitable stopping criterion between consecutive T values is used to decide when to stop the algorithm, obtaining the solution to the optimization problem.

As can be easily intuitable, the behavior of the SA algorithm crucially depend on the choice of the stopping rules, the annealing schedule, and the initial and next values of the first two steps. In the context of finite mixture models, Ingrassia (1991, 1992) compared the performances of the SA and EM algorithms. It was found that although the SA algorithm performed more satisfactorily in some cases, it was much slower than the EM algorithm. Additionally, even if in these cases the SA algorithm gave estimates closer to true values, the value of the likelihood was greater at the EM solution. Therefore, neither algorithm overwhelmingly outperforms the other.

2.3.3 Standard errors of the estimates

Let $\widehat{\boldsymbol{\Theta}}$ be the ML estimator of $\boldsymbol{\Theta}$. To assess the precision of the ML estimates, the estimated covariance matrix of $\widehat{\boldsymbol{\Theta}}$, say $\widehat{\text{Cov}}(\widehat{\boldsymbol{\Theta}})$, is typically computed. The square root of the diagonal elements of $\widehat{\text{Cov}}(\widehat{\boldsymbol{\Theta}})$ are then reported as standard errors of the ML estimates.

One criticism of the EM algorithm is that it does not automatically provide an estimate of $\widehat{\text{Cov}}(\widehat{\boldsymbol{\Theta}})$, as do some other approaches, such as Newton-type methods. Nevertheless, in ML theory, $\widehat{\text{Cov}}(\widehat{\boldsymbol{\Theta}})$ can be usually obtained from the information matrix $\mathcal{I}(\boldsymbol{\Theta})$. Under regularity conditions, and if the model is correctly specified, $\mathcal{I}(\boldsymbol{\Theta})$ is given either by the covariance of the score function $\mathbb{E}(S(\boldsymbol{\Theta})S(\boldsymbol{\Theta})')$ or by the negative of the expected value of the Hessian matrix $-\mathbb{E}(H(\boldsymbol{\Theta}))$. However, an analytical evaluation of these expected values is often cumbersome.

A first solution to such problem relies on numerical methods. Specifically, by using some asymptotic results concerning ML estimation (see, e.g. White, 1982), it is possible to obtain the following asymptotic estimators of $\mathcal{I}(\boldsymbol{\Theta})$

$$\mathcal{I}_1 = \sum_{i=1}^{N} S_i(\widehat{\boldsymbol{\Theta}}) S_i \widehat{\boldsymbol{\Theta}}', \quad \mathcal{I}_2 = -\sum_{i=1}^{N} H_i(\widehat{\boldsymbol{\Theta}}),$$

where $S_i(\widehat{\boldsymbol{\Theta}})$ and $H_i(\widehat{\boldsymbol{\Theta}})$ represent the contribution of the ith observation to the score function and Hessian matrix at the ML estimate, respectively. The inverses $\mathcal{I}_1^{-1}$ and $\mathcal{I}_2^{-1}$ are consistent estimators of $\widehat{\text{Cov}}(\widehat{\boldsymbol{\Theta}})$ if the model is correctly specified (Boldea and Magnus, 2009). In general, the so-called "sandwich" (or robust) approach provides a consistent estimator of $\widehat{\text{Cov}}(\widehat{\boldsymbol{\Theta}})$, whether or not the model is not correctly specified. It is computed by

$$\mathcal{I}_3 = \mathcal{I}_2^{-1} \mathcal{I}_1 \mathcal{I}_2^{-1}.$$

A second solution is based on bootstrap techniques. Specifically, it is possible to mention the parametric and the nonparametric bootstrap (see, McLachlan and Peel, 2000 for further details), and the weighted bootstrap (Newton and Raftery, 1994),

which is a version of the nonparametric bootstrap based on scaling the data with weights that are proportional to the number of times an original point occurs in the bootstrap sample (Boldea and Magnus, 2009).

2.4 Model selection and clustering assessment

In many applications, the number of groups K is unknown, and it is commonly selected by using some likelihood-based information criterion. Information criteria are also used for selecting the best fitting model among a set of competitors. The Bayesian information criterion (BIC; Schwarz, 1978) is undoubtedly one of the most commonly used, and for this reason it will be considered in this thesis. In its original formulation, it is defined as

$$\text{BIC} = -2l(\widehat{\boldsymbol{\Theta}}) + \ln(N)\#\text{par},$$

where $l(\widehat{\boldsymbol{\Theta}})$ is the maximized log-likelihood value and #par is the number of parameters of the model. The value of K, and consequently the model, associated to the smallest BIC value is preferred for modeling a given data sets.

To evaluate the clustering performance of a model, when the true classification of the data is known, the adjusted rand index (ARI; Hubert and Arabie, 1985) and the misclassification percentage (ϵ) will be considered. The ARI evaluates the agreement between two partitions, with an upper bound of 1 indicating a perfect classification, whereas ϵ measures the percentage of statistical observations that are misclassified.

2.5 A short introduction on matrix-variate data

In (2.1), the variable X under consideration is univariate, as in the analyses contained in Chapters 3 and 4. In the multivariate literature, X is replaced by a p-dimensional random vector containing p measurements on the phenomenon under study. When p measurements are collected over r different times or situations, this leads to a matrix-variate (or three-way) data structure. This type of data has received an increasing interest by the researchers, especially within the finite mixture model literature (see, e.g., Gallaugher and McNicholas, 2018; Melnykov and Zhu, 2019; Sarkar *et al.*, 2020; Tomarchio *et al.*, 2020 for recent contributions). Typical examples of this data structure include spatial multivariate data, longitudinal data on multiple response variables or spatio-temporal data. In all these cases the data can be arranged in a three-way array characterized by the following dimensions: measurements (rows), situations (columns) and observations (layers). In other terms, each statistical observation is a $p \times r$ matrix $\mathbf{X}$.

In the matrix-variate literature, the matrix-variate normal (MVN) distribution plays the same pivotal role that the multivariate normal distribution has in the multivariate literature. A random $p \times r$ matrix $\mathbf{X}$ is said to follow a MVN distribution if its PDF can be written

$$f_{\text{MVN}}(\mathbf{X}; \mathbf{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi}) = \frac{1}{(2\pi)^{\frac{pr}{2}} |\boldsymbol{\Sigma}|^{\frac{r}{2}} |\boldsymbol{\Psi}|^{\frac{p}{2}}} \exp\left\{-\tfrac{1}{2} \operatorname{tr}\left[\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\mathbf{M})\boldsymbol{\Psi}^{-1}(\mathbf{X}-\mathbf{M})'\right]\right\}, \quad (2.8)$$

where $\mathbf{M}$ is the $p \times r$ mean matrix, $\boldsymbol{\Sigma}$ is the $p \times p$ row covariance matrix and $\boldsymbol{\Psi}$ is the $r \times r$ column covariance matrix. In (2.8), $\operatorname{tr}\left[\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\mathbf{M})\boldsymbol{\Psi}^{-1}(\mathbf{X}-\mathbf{M})'\right]$ is the squared Mahalanobis distance from $\mathbf{X}$ to the center $\mathbf{M}$ with respect to $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$, and

for brevity's sake will be denoted as $\delta(\mathbf{X}; \mathbf{\Omega})$, with $\mathbf{\Omega} = \{\mathbf{M}, \mathbf{\Sigma}, \mathbf{\Psi}\}$. Upon vectorization, the MVN distribution can be reformulated as a pr-multivariate normal distribution with covariance matrix given by $\mathbf{\Psi} \otimes \mathbf{\Sigma}$, where $\otimes$ is the Kronecker product (Gupta and Nagar, 1999). However, a MVN distribution has the desirable feature of simultaneously model and identify the between and within-variables variabilities as well as reducing the number of free parameters from $pr(pr+1)/2$ to $[p(p+1)/2] + [r(r+1)/2] - 1$. The MVN distribution plays a fundamental role in Chapters 5 and 6.

Chapter 3

Modelling the loss given default distribution via a family of zero-and-one inflated mixture models [1]

3.1 Introduction

The Advanced Internal Ratings Based approach, within the Basel II/III regulatory framework, allows banks to calculate the capital requirements on the basis of their internal credit risk models (Basel Committee on Banking Supervision, 2006). Specifically, they need to develop methods for estimating the following three key risk parameters: PD (probability of default), LGD (loss given default) and EAD (exposure at default). The target of this work is the LGD parameter (the equivalent of one minus the recovery rate), which is defined as the percentage of the exposure that is lost in case of default. Other than for regulatory reasons, an accurate LGD estimation is crucial for the correct evaluation of credit derivatives and asset-backed securities, as well as for gaining a competitive advantage in case of models with high predictive power (Grunert and Weber, 2009; Gürtler and Hibbeln, 2013).

As discussed by Baesens *et al.* (2016), the LGD can be measured in several ways. The first method is called "market approach" and looks at the market price of debt securities of the firms soon after their bankruptcy. This market price is then used as a proxy for the recovery rate (Gupton and Stein, 2005). A disadvantage of this method is that it cannot be applied to all types of debts, but only to those traded in the financial markets. A second method is the "implied market approach" (Seidler, 2008). In this case, the LGD is estimated through the analysis of the market price of not defaulted risky securities using asset pricing models. The idea is that prices reflect market's expectation of the loss and hence the LGD can be extracted from there. Another and more widespread method is the "workout approach", and it is based on an economic notion of loss. In detail, all the relevant incoming and outgoing cash flows or costs related to the collection process should be considered and discounted, via a suitable discount rate, to the moment of default to calculate the loss. This approach is adopted within the Basel Accord (Basel Committee on Banking Supervision, 2006), which identifies three types of costs: (1) those associated to the loss of

[1] This work is based on the following publication: Tomarchio S.D., Punzo A. (2019). Modelling the loss given default distribution via a family of zero-and-one inflated mixture models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **182**(4), 1247–1266. The current manuscript is a combined effort of the authors. However, Tomarchio S.D. contributed in conceptualization, implementation, data elaboration and writing–original draft preparation; Punzo A. contributed in conceptualization and supervision.

principal and the foregone interest income, (2) those linked to the recovery process (for example administrative and operating costs), and (3) those related to the time incurring between the emergence of default and the actual recovery (represented by an appropriate discount rate).

However, modeling the LGD has posed serious challenges. The first one is related to the lack or the confidentiality of data, which make difficult for researchers to develop and test their models (Grunert and Weber, 2009; Gürtler and Hibbeln, 2013; Li *et al.*, 2016). A second problem is connected to the peculiarities of its distribution. In fact, the LGD distribution is generally defined between zero and one (both included), often exhibits bi-, and more in general, multi-modality (Schuermann, 2004; Gürtler and Hibbeln, 2013) and it has a high amount of observations at the boundary values 0 and 1 (Friedman and Sandow, 2003; Calabrese, 2010; Tong *et al.*, 2013; de Oliveira Jr *et al.*, 2015). Some studies (see, e.g. Schmit, 2004; Gürtler and Hibbeln, 2013; Miller and Töws, 2018) discuss that when the "workout approach" is used, the LGD can assume values lower than 0 (i.e. the creditor recovers more than the outstanding amount) or greater than 1 (i.e. the creditor loses more than the outstanding amount). However, as pointed out by Gouriéroux and Monfort (2006), this boundary problem has been early noted by the Basel Committee, that imposes to truncate the LGD to the $[0, 1]$ interval, avoiding negative or greater than one values. In the two real data applications considered in this work, the LGD has been computed via the "workout approach", but the data have been previously pre-processed in order to constrain them within the interval $[0, 1]$.

Because of these peculiar characteristics, different approaches have been discussed in the literature. A first class of models is based on regression analyses (see, e.g Huang and Oosterlee, 2008; Sigrist and Stahel, 2011; Bellotti and Crook, 2012). A second set of models make use of machine learning techniques such as artificial neural networks, random forests, regression tree algorithms and many others (see, e.g Bastos, 2010; Qi and Zhao, 2011; Loterman *et al.*, 2012; Tobback *et al.*, 2014; Yao *et al.*, 2017; Nazemi *et al.*, 2017). A third category of models, on which this work focuses, aims to estimate the LGD distribution, either parametrically or nonparametrically. The nonparametric models are mainly based on different kernel density estimators (see, e.g Renault and Scaillet, 2004; Hagmann *et al.*, 2005; Calabrese and Zenga, 2010; Chen and Wang, 2013). Under the parametric model category, we can mention the works of Calabrese (2014b,a); de Oliveira Jr *et al.* (2015), which consider the LGD as a mixed random variable, obtained via the mixture of a Bernoulli random variable (addressing the problem of excess of zeros and ones), and either a single or a mixture of two beta distributions for the continuous part on $(0, 1)$. These latter models are also known as zero-and-one inflated models (Ospina and Ferrari, 2010).

The proposal contained in this chapter is placed in this last branch of the literature. Indeed, the LGD distribution is herein modeled via a family of zero-and-one inflated mixture models, but where the number of mixture components on $(0, 1)$ is not fixed in advance. There is no specific reason for limiting such number of components, since doing this may lead to erroneous and sub-optimal solutions, as will be shown by the real data analyses. As components of the mixture, distributions directly defined on $(0, 1)$ are used, such as the beta and the generalized beta of type 1 (GB1), but also distributions with support $(-\infty, \infty)$ mapped on $(0, 1)$ via an inverse-logit transformation (see Section 3.2 for details). Therefore, by using these transformed distributions, this work overcomes the scarcity of distributions commonly used in the LGD modeling.

Overall, the models herein proposed can be considered as a flexible device for

modeling the LGD distribution, in a similar manner to a nonparametric approach. Hence, an *indirect application* of the finite mixture models, as discussed in Chapter 2, seems to be the more appropriate. Relatedly, in the two real data analyses of Section 3.3, a comparison with several standard semiparametric/nonparametric approaches used in the credit risk literature is presented. Finally, some conclusions are drawn in Section 3.4.

3.2 Methodology

Let X be a random variable taking values in $[0,1]$. Suppose that part of the distribution of X is concentrated at $\{0,1\}$, while the rest of the distribution is continuously spread over $(0,1)$. In such a case, X is said to be a mixed random variable (Bertsekas and Tsitsiklis, 2008); see also Calabrese and Zenga (2010); Calabrese (2010, 2014a,b) within the LGD literature.

To model the distribution of X in a flexible way, a zero-and-one inflated mixture model is herein proposed. Its PDF can be defined as

$$q\left(x;\boldsymbol{\vartheta}\right)=\begin{cases}\alpha_0 & \text{if } x=0\\ \left(1-\alpha_0-\alpha_1\right)g\left(x;\boldsymbol{\Theta}\right) & \text{if } x\in(0,1)\\ \alpha_1 & \text{if } x=1,\end{cases} \tag{3.1}$$

where $\alpha_0=P\left(X=0\right)$ and $\alpha_1=P\left(X=1\right)$ are the parameters of a three level multinomial model over the categories $\{0\}$, $\{1\}$ and $(0,1)$, with $\alpha_0>0$ and $\alpha_1>0$ such that $\alpha_0+\alpha_1<1$. In (3.1), $g(x;\boldsymbol{\Theta})$ is the PDF of the mixture as defined in (2.1) and with support $(0,1)$, while $\boldsymbol{\vartheta}$ contains all the parameters.

As said in Chapter 2, the mixture components in $g(x;\boldsymbol{\Theta})$ are taken to be of the same type. Naturally, given K, the flexibility of the mixture model improves if more flexible distributions, i.e. distributions having a greater number of parameters m, are considered. In this regard, 13 different type of distributions for the mixture components are considered in this work. Conversely, for a given distribution, the flexibility of the mixture model improves when K increases. In this work, K is free to vary in a large enough set of positive integer values, and then selected by the BIC. Thus, two sources of flexibility are investigated.

Possible component distributions with support $(0,1)$ are the beta and the GB1, having $m=2$ and $m=4$ parameters, respectively. Special cases of model (3.1), based on the beta distribution, already exist (see Ospina and Ferrari, 2010 for the cases $K=1$, and de Oliveira Jr *et al.*, 2015 for $K=2$). However, the extension to a generic number of components, as well as the alternative use of the GB1 distribution, are new in the LGD literature.

To further increase the number of parametric distributions on $(0,1)$ to be used as components in $g(x;\boldsymbol{\Theta})$, the approach discussed in Düllmann and Gehde-Trapp (2004); Rösch and Scheule (2006); Stasinopoulos *et al.* (2017a) is followed. Specifically, it consists in transforming classical distributions with support $(-\infty,\infty)$ via the inverse logit transformation

$$X=\frac{1}{1+\exp\left(-Y\right)}, \tag{3.2}$$

where Y is the random variable taking values in $(-\infty,\infty)$. For example, if Y has a normal distribution, and the inverse logit transformation in (3.2) is applied, the

logit-normal distribution for X is obtained (see, e.g., Atchison and Shen, 1980). Following this idea, 11 further candidates are considered for the mixture components, and are obtained via the inverse logit transformation of the following distributions: exponential generalized beta of type 2 (EGB2; $m = 4$), ex-Gaussian ($m = 3$), Gumbel ($m = 2$), Johnson S_u ($m = 4$), logistic ($m = 2$), normal ($m = 2$), reverse Gumbel ($m = 2$), sinh-arcsinh ($m = 4$), skew-normal ($m = 3$), skew-t ($m = 4$), and t ($m = 3$). For details about these distributions, see Rigby *et al.* (2014). Apart from the logit-normal distribution, which is considered to define the zero-and-one inflated mixture of $K = 2$ logit-normal distributions by de Oliveira Jr *et al.* (2015), the logit version of the remaining distributions have never been used to define zero-and-one inflated models.

3.2.1 Parameter estimation

Given a sample $\{X_i\}_{i=1}^N$ of N independent observations from the PDF in (3.1), the log-likelihood function of the model can be decomposed – because of the orthogonality between (α_0, α_1) and $\boldsymbol{\Theta}$ (Stasinopoulos *et al.*, 2017a) – as

$$l(\boldsymbol{\vartheta}) = \underbrace{N_0 \log(\alpha_0)}_{\{0\}} + \underbrace{(N - N_0 - N_1)\log(1 - \alpha_0 - \alpha_1) + l_{(0,1)}(\boldsymbol{\Theta})}_{(0,1)} + \underbrace{N_1 \log(\alpha_1)}_{\{1\}}, \quad (3.3)$$

where $N_0 = \sum_{i=1}^N I_{\{0\}}(x_i)$ and $N_1 = \sum_{i=1}^N I_{\{1\}}(x_i)$, with $I_A(x)$ being the indicator function on the set A, are the number of 0s and 1s in the sample, respectively, and $l_{(0,1)}(\boldsymbol{\Theta})$ is the log-likelihood function of the mixture model in (2.1), with the summation running only over the observations in the interval $(0, 1)$. By looking at (3.3), it is possible to see that the of parameters sets (α_0, α_1) and $\boldsymbol{\Theta}$ can be estimated separately. The ML estimates of α_0 and α_1 correspond to the sample proportions N_0/N and N_1/N, respectively. On the contrary, the ML estimate of $\boldsymbol{\Theta}$ are obtained by means of the EM algorithm. Specifically, it is implemented via the *gamlssMX()* function, contained in the **gamlss.mx** package (Stasinopoulos and Rigby, 2016), for the **R** (Team, 2019) statistical software. Further details about the *gamlssMX()* function can be found in Stasinopoulos *et al.* (2017b).

3.2.2 Some notes on identifiability

Generally, for finite mixtures of distributions, we need to distinguish among different types of non identifiability (see, e.g. Frühwirth-Schnatter, 2006):

1. first of all, non-identifiability is caused by the invariance of a mixture distribution to relabeling the components (*label switching*), as first noted by Redner and Walker (1984): indeed, K kept fixed, all the $K!$ permutations of the K products $\pi_1 f(x; \boldsymbol{\phi}_1), \ldots, \pi_K f(x; \boldsymbol{\phi}_K)$ yield the same mixture model. This is not a serious problem and someone in the literature (see, e.g. Aitkin and Rubin, 1985) handled it by a constraint on the mixing proportions of the form $\pi_1 < \pi_2 < \ldots < \pi_K$.

2. A further identifiability problem, noted by Crawford (1994), is non-identifiability due to potential overfitting. Crawford (1994) showed that any mixture with $K - 1$ components can be equivalently rewritten as a mixture with K components, where either one component is empty (i.e. $\pi_k = 0$) or two components are equal (i.e. $\boldsymbol{\phi}_k = \boldsymbol{\phi}_l$, with $k \neq$l). However, in all the analyses of Section 3.3, this non-identifiability issue has not been encountered.

Nevertheless, finite mixtures may remain unidentifiable even if formal identifiability constraints are considered to rule out the non-identifiability problems described above. Among the models herein considered (beta, GB1, EGB2, ex-Gaussian, Gumbel, Johnson S_u, logistic, normal, reverse Gumbel, sinh-arcsinh, skew-normal, skew-t, and t), in my knowledge the identifiability issue have been discussed only for the beta (Ahmad and Al-Hussaini, 1982), normal (Teicher, 1963), t (Holzmann *et al.*, 2006), skew-normal and skew-t (Otiniano *et al.*, 2015) distributions.

3.3 Real data applications

This section contains the results of the real data analyses. Firstly, the two data sets are accordingly described in Section 3.3.1. Secondly, the zero-and-one inflated models are fitted to the data and discussed in Section 3.3.2. Since some of the parametric models used in the credit risk literature to fit the LGD distribution are special cases of the models herein proposed, a comparison with them is obtained as a by-product. Lastly, in Section 3.3.3, the best fitting models are compared, via a convenient simulation study, to semiparametric and nonparametric density estimation approaches used in the LGD literature.

3.3.1 Data description

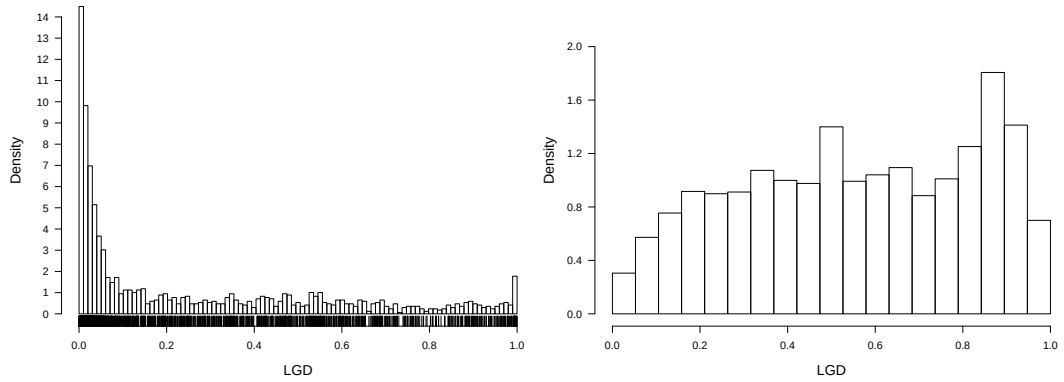
The first data set (Data set A), consists of $N = 2545$ LGDs on loans of a European bank (see Baesens *et al.*, 2016, for details). As discussed in Section 3.1, the data has been preprocessed by the owners in such a way that only values in the interval $[0, 1]$ occur. For their application, Baesens *et al.* (2016) converted 0 and 1 values to 0.00001 and 0.99999, respectively. Therefore, to recover the original data, this transformation is eliminated, obtaining a data set with $N_{(0,1)} = 1674$ observations defined on the interval $(0, 1)$, $N_0 = 728$ observations equal to 0 and $N_1 = 143$ observations equal to 1. Therefore, about the 34.2% of the observations is at the boundaries.

The second data set (Data set B) contains $N = 149378$ loan recovery rates from a comprehensive survey conducted by the Bank of Italy on about 250 banks in the years 2000–2001 (Banca d'Italia, 2001). Considering that this survey deals with loans privately held, the market approach is not feasible, and consequently the Bank of Italy applies the workout approach. Also in this case, the data have been preprocessed following the methodology proposed by Calabrese and Zenga (2008), which permits to compute the recovery rate in the workout approach constraining this variable within $[0, 1]$. Additional details about this data set can be seen in Calabrese and Zenga (2010). The corresponding LGD values are obtained as LGD $= 1 -$ recovery rate. There are $N_{(0,1)} = 103511$ on $(0, 1)$, $N_0 = 11514$ observations equal to 0 and $N_1 = 34353$ equal to 1. Therefore, more than 30% of the LGDs is confined to the boundaries.

Table 3.1 reports some descriptive statistics for the data sets, whereas Figure 3.1 shows the histograms of the LGD values on $(0, 1)$ only; these values are also represented as tick marks below the horizontal LGD-axis for Data set A. Tick marks of the observations are not plotted for Data set B due to the huge number of values covering the whole horizontal axis below the histogram.

TABLE 3.1: Descriptive statistics

Statistic	Value	
	Data set A	Data set B
Number of observations (N)	2545	149 378
Number of zeros (N_0)	728	11 514
Number of observations in $(0,1)$ ($N_{(0,1)}$)	1674	103 511
Number of ones (N_1)	143	34 353
Mean	0.228	0.616
Standard deviation	0.329	0.340
Skewness	1.308	−0.415
Kurtosis (excess)	0.274	−1.176
First quartile	0.000	0.333
Median	0.032	0.667
Third quartile	0.398	0.958



(a) Data set A (b) Data set B

FIGURE 3.1: Histograms of the LGD values on $(0,1)$

3.3.2 Zero-and-one inflated mixture models

3.3.2.1 Computational details

Considering that the likelihood function for mixture models usually has multiple local maxima, the EM algorithm is generally run several times for different starting values. Then, in order to ensure that a global maximum has been reached, the *gamlssMX()* function is run five times, with different starting values randomly determined. In case of different solutions, the one producing the highest log-likelihood is preferred. Furthermore, in the attempt of reducing the cases of overparameterization, we limit to $K = 4$ the maximum number of mixture components to be tried.

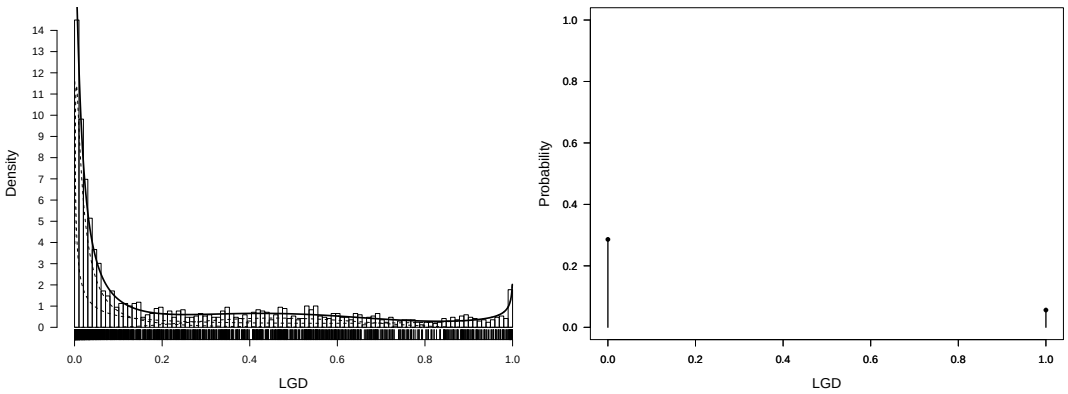
As concerns the component densities of the mixture, the 11 distributions obtained via inverse-logit transformation are generated via the *gen.Family()* function contained in the **gamlss.dist** package (Stasinopoulos and Rigby, 2017). This function offers the possibility to generate the logit version of all the distributions with support $(-\infty, \infty)$ contained in that package. At a preliminary stage of the real data analyses, all the distributions have been considered as components of the mixture. However, the logit version of the power exponential and skew exponential power distributions presented computational issues in at least one of the considered data

sets. Specifically, the EM algorithm fails to converge probably due to the generation of an indeterminate form of the type $0/0$ when the posterior probabilities of group-membership are computed in the E-step. Changing the starting values of the parameters seems not to fix this problem. Given that a considerable number of logit-distributions is still considered, it should not be a great loss to exclude these two models from the analyses. In addition to these 11 transformed distributions, the 2 distributions directly defined on the support $(0,1)$ included in the **gamlss.dist** package are considered, i.e. the beta and GB1.

3.3.2.2 Results

The zero-and-one inflated mixture models are fitted to both data sets for values of $K \in \{1,2,3,4\}$, yielding to a total of $13 \times 4 = 52$ fitted models for each data set. Table 3.2 reports a model comparison in terms of $-2l(\widehat{\boldsymbol{\vartheta}})$ and BIC. For a better comparison, the ranking induced by the BIC is also shown in correspondence of the selected value of K for each of the 13 models. The first and most immediate result is that a model with a single component ($K = 1$) is never the best choice in both data sets, and this confirms the previous conjectures about the need of a more flexible model to better capture the behavior of the LGDs on $(0,1)$.

As concerns Data set A, according to the BIC, $K = 2$ components are selected only for the models whose component PDFs have $m = 4$ parameters as well as for the zero-and-one inflated mixture with logit-t components; for the remaining ones, $K = 3$ components are chosen. This is because the BIC tends to penalize model complexity, and to prefer more parsimonious models. The best models are those with $K = 3$ logit-normal and logit-logistic components, ranked as first and second, respectively. The best zero-and-one inflated mixture with beta components has $K = 3$, and it is ranked third. This confirms how the often used beta distribution is not always the best choice to be used as mixture component in the analysis of LGD data. The graphical representation of the mixture of $K = 3$ logit-normal distributions, i.e. the best fitting model, is given via a solid line superimposed on the histogram, in Figure 3.2(a), with dotted curves showing the component densities multiplied by the corresponding estimated mixture weights $\widehat{\pi}_1$, $\widehat{\pi}_2$ and $\widehat{\pi}_3$. The bar-plot of the discrete part of models is displayed in Figure 3.2(b).



(a) Mixture of 3 logit-normal distributions

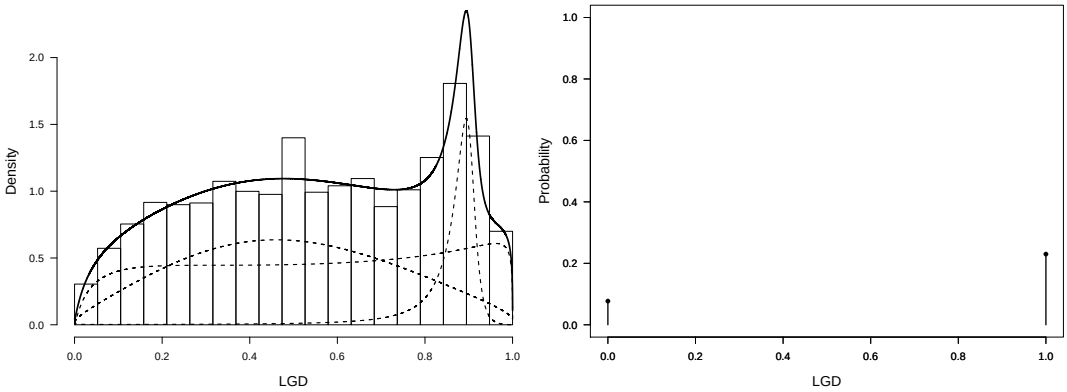
(b) $\widehat{\alpha}_0$ and $\widehat{\alpha}_1$

FIGURE 3.2: Data set A. Histogram with superimposed curves from the mixture model selected by the BIC (panel 3.2(a)), and estimated $\widehat{\alpha}_0$ and $\widehat{\alpha}_1$ (panel 3.2(b)).

TABLE 3.2: Values of $-2l(\widehat{\boldsymbol{\vartheta}})$ and BIC for the zero-and-one inflated mixture models fitted for $K \in \{1, 2, 3, 4\}$. The best BIC value, for each model, is written in bold font; for these models, a ranking is given.

			Data set A			Data set B		
Mixture component	m	k	$-2l(\widehat{\boldsymbol{\vartheta}})$	BIC	rank	$-2l(\widehat{\boldsymbol{\vartheta}})$	BIC	rank
beta	2	1	2092.12	2123.49		228502.95	228550.60	
		2	1779.02	1833.91		226465.72	226549.12	
		3	1746.13	**1824.55**	3	224102.27	224221.41	
		4	1738.58	1840.53		224063.10	**224217.99**	6
GB1	4	1	2058.15	2105.20		228497.27	228568.76	
		2	1760.20	**1846.46**	12	226538.30	226669.36	
		3	1744.95	1870.42		224092.90	**224283.52**	8
		4	1741.99	1906.67		224078.87	224329.07	
logit-logistic	2	1	1884.19	1915.55		229064.03	229111.68	
		2	1781.27	1836.16		226038.54	226121.94	
		3	1738.43	**1816.85**	2	224119.95	224239.09	
		4	1738.06	1840.00		224027.16	**224182.04**	3
logit-Gumbel	2	1	2343.96	2375.32		253904.90	253952.50	
		2	1970.95	2025.84		235106.25	235189.65	
		3	1756.67	**1835.09**	6	225902.65	226021.79	
		4	1734.28	1836.22		224821.14	**224976.02**	11
logit-normal	2	1	1863.19	1894.55		227856.91	227904.57	
		2	1835.12	1890.01		227502.65	227586.05	
		3	1738.25	**1816.66**	1	225103.08	225222.22	
		4	1737.65	1839.59		224006.29	**224161.18**	2
logit-rev.Gumbel	2	1	1973.74	2005.11		245260.46	245308.12	
		2	1871.52	1926.41		232438.10	232521.50	
		3	1754.45	**1832.87**	5	229426.51	229545.65	
		4	1748.66	1850.61		224747.63	**224902.52**	10
logit-exGaus	3	1	1847.83	1887.03		227833.25	227892.82	
		2	1834.51	1905.09		227532.84	227640.07	
		3	1738.60	**1840.55**	10	227433.32	227588.20	
		4	1738.34	1871.66		226445.84	**226648.39**	12
logit-skew-normal	3	1	1863.19	1902.40		227856.90	227916.47	
		2	1835.98	1906.55		227565.19	227672.42	
		3	1738.38	**1840.32**	9	227458.73	227613.62	
		4	1737.62	1870.93		227303.12	**227505.66**	13
logit-t	3	1	1857.97	1897.18		227536.57	227596.14	
		2	1765.57	**1836.14**	7	224836.77	224943.99	
		3	1736.51	1838.45		224050.17	**224205.05**	5
		4	1735.59	1868.90		224014.18	224216.72	
logit-Johnson S_u	4	1	1830.49	1877.54		227488.90	227560.39	
		2	1750.64	**1836.90**	8	224631.03	224762.09	
		3	1734.14	1859.61		223908.11	**224098.74**	1
		4	1732.18	1896.86		223862.84	224113.04	
logit-sinh-arcsinh	4	1	1867.28	1914.33		229500.56	229572.04	
		2	1743.13	**1829.39**	4	224505.80	224636.86	
		3	1740.64	1866.11		224348.28	224538.91	
		4	1736.43	1901.11		224099.13	**224349.33**	9
logit-skew-t	4	1	1844.48	1891.54		227490.92	227562.41	
		2	1757.48	**1843.73**	11	227380.36	227511.42	
		3	1735.97	1861.44		224000.32	**224190.95**	4
		4	1734.62	1899.30		223962.61	224212.81	
logit-EGB2	4	1	1861.26	1908.31		228755.06	228826.55	
		2	1768.12	**1854.38**	13	226177.91	226308.96	
		3	1737.97	1863.44		224225.57	224416.20	
		4	1737.16	1901.84		224014.51	**224264.71**	7

Analyzing Data set B, the results are even stronger because there are no cases where a mixture model with $K = 1$ or $K = 2$ components is the best choice. Specifically, the BIC select $K = 4$ components for almost all the competing models, expect for four cases where the it indicates a model with $K = 3$ components. Among them, there is the best one, i.e. the mixture of $K = 3$ logit-Johnson S_u distributions. In comparison with the ranking reported for the previous data set, the zero-and-one inflated mixture based on the beta distribution performs worse, while the one based on the logit-normal distribution is still competitive, although with a greater number ($K = 4$) of mixture components; this remarks the need, for these mixtures fitted on these data, of more than two mixture components. The continuous part of the best zero-and-one inflated mixture model selected by BIC is shown in Figure 3.3(a), while the plot of the discrete part is displayed in Figure 3.3(b).



(a) Mixture of 3 logit-Johnson S_u distributions

(b) $\widehat{\pi}_0$ and $\widehat{\pi}_1$

FIGURE 3.3: Data set B. Histogram with superimposed curves from the mixture model selected by the BIC (panel 3.3(a)), and estimated $\widehat{\alpha}_0$ and $\widehat{\alpha}_1$ (panel 3.3(b)).

3.3.3 A comparison with semiparametric/nonparametric approaches

3.3.3.1 Simulation study

Some semiparametric and nonparametric approaches, used in the LGD literature, are now compared with the zero-and-one inflated models. In detail:

- the semiparametric density considered by Hagmann *et al.* (2005), labeled "H-BK";
- the Gaussian kernel considered by Renault and Scaillet (2004) and Chen and Wang (2013), labeled "GK";
- the beta kernel introduced by Chen (1999), and applied by Renault and Scaillet (2004), labeled "C-BK";
- the beta kernel proposed by Calabrese and Zenga (2010), labeled "CZ-BK".

In this case, the comparison is made challenging because the BIC cannot be used to compare semiparametric and nonparametric approaches, to the author knowledge. To allow such an overall comparison, a simulation-based procedure, similar to the one considered by Renault and Scaillet (2004), is implemented. In their work, the authors assessed the impact of assuming a parametric beta distribution for the LGD,

on the Value at Risk (VaR) computed on the loss distribution of a well diversified portfolio. The VaR represents the maximum loss which can occur with probability c over a specified period of time, and it is defined as

$$\mathrm{VaR}_c(X) = \inf\{x : F(x) \geq c\}, \quad 0 \leq c \leq 1,$$

where F is the cumulative distribution function of X.

With a similar procedure, ten thousand losses $\{L_j\}_{j=1}^{10000}$ are simulated via a default or non-default approach (Bellotti, 2010, 2017), and by considering $\{N_i\}_{i=1}^{50000}$ debtors. In detail:

1. for $j = 1, \ldots, 10000$
   (a) for $i = 1, \ldots, 50000$
      i. generate a latent factor Y_i, which describes the uncertainty on repayment (Vasicek, 2002), as

$$Y_i = T\sqrt{\rho} + T_i\sqrt{1-\rho},$$

      where $T \sim N(0,1)$ represents a common systematic risk factor affecting all the debtors (e.g., the state of the economy), $T_i \sim N(0,1)$ is an idiosyncratic factor independent for each debtor and ρ is the pairwise correlation coefficient which is assumed the same for any two debtors;
   (b) compute the expected loss for that portfolio via

$$L_j = \sum_{i=1}^{N} I_{\left(-\infty, z_{\mathrm{PD}_i}\right)}(Y_i) \times \mathrm{EAD}_i \times \mathrm{LGD}_i, \tag{3.4}$$

   where z_{PD_i} is the quantile of order PD_i from the standard normal distribution and $I_A(x)$ denotes the indicator function, which is equal to 1 when $x \in A$ and 0 otherwise.
2. when $L_1, \ldots, L_{10000}$ are computed, a simulated loss distribution is obtained and the VaR is calculated. The probability level used for the VaR is $c = 0.99$.

In (3.4) all the debtors have the same EAD (which is assumed to be $\mathrm{EAD}_i = 1$) and PD. On the contrary, the LGD values are randomly drawn either from the empirical distribution (that is used as a benchmark) of the available LGD values or from one of the competing models.

Two different risky scenarios are also evaluated, according to the values of PD and ρ shared by all the debtors. The Basel II Accord assumes a decreasing relationship between PD and ρ. However, some studies (see Dietsch and Petey, 2004 and Lee *et al.*, 2009) show that this stylized decreasing relationship seems to have neither theoretical nor empirical support. For this reason, an increasing relationship between PD and ρ is considered. Specifically, the two risky scenarios are: (1) PD $= 0.05$ and $\rho = 0.10$; (2) PD $= 0.10$ and $\rho = 0.20$.

3.3.3.2 Computational details

The following **R** functions and packages are used to implement the semiparametric and nonparametric competitors:

- the function *kdensity()* included in the **kdensity** package (Moss and Tveten, 2018) for the H-BK;
- the function *density()* contained in the **stats** package for the GK;
- the function *chen99Kernel()* included in the **bde** package (Santafe *et al.*, 2015) for the C-BK.

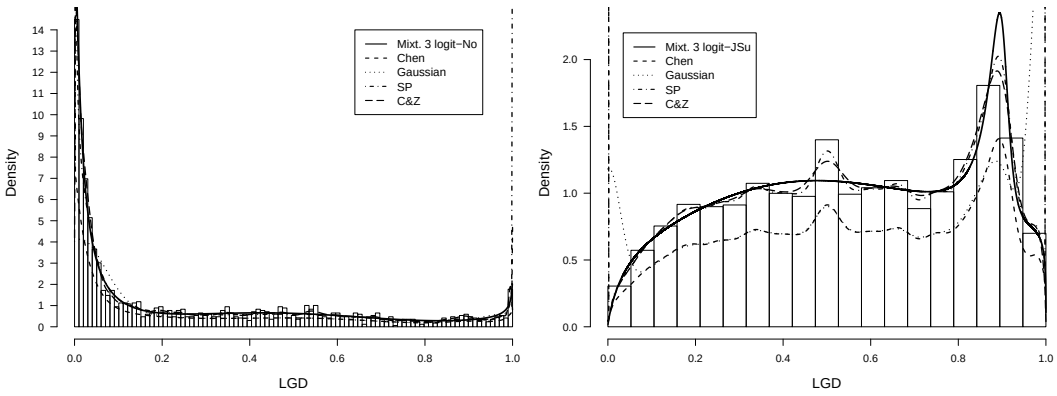
A convenient **R** code has been implemented for the CZ-BK. Bandwidth values are: $\sigma N^{-2/5}$ for H-BK and C-BK (Renault and Scaillet, 2004; Hagmann *et al.*, 2005), where σ is the empirical standard deviation; $0.9AN^{-1/5}$ for GK (Renault and Scaillet, 2004; Silverman, 1986), where $A = \min(\sigma, \text{interquantile range}/1.34)$; estimated by likelihood cross-validation for CZ-BK.

3.3.3.3 Results

In the following, for each of the 13 zero-and-one inflated mixture models, only those with K selected by the BIC are considered. The $\text{VaR}_{99}(X)$ values for both data sets, along with their percentage of variation with respect to the empirical $\text{VaR}_{99}(X)$, are reported in Table 3.3.

The first result is that our models provide estimates that are very close to the empirical ones, in both risky scenarios. Comparing them with the semiparameteric and nonparameteric approaches, only the CZ-BK seems to behave comparably. This is because the C-BK and H-BK, applied as in Renault and Scaillet (2004) and Hagmann *et al.* (2005), respectively, are strongly affected by the structure of the data at the boundaries (Gouriéroux and Monfort, 2006). The GK, instead, performs badly because it suffers from the boundary bias problem (Renault and Scaillet, 2004), i.e. the allocation of probability masses outside the theoretical support of the LGD distribution.

Figure 3.4 shows an overall graphical comparison, for both data sets, between the competing models. For simplicity's sake, only the best overall zero-and-inflated model selected by the BIC, and already depicted in Figures 3.2–3.3, are considered. Specifically, they are superimposed to the competing semiparametric and nonparametric densities in Figure 3.4(a) for Data set A, and Figure 3.4(b) for Data set B.



(a) Data set A (b) Data set B

FIGURE 3.4: Comparison between parametric, semiparametric, and nonparametric densities for Data set A, in panel 3.4(a), and Data set B, in panel 3.4(b). Among the parametric models, only the best selected by the BIC is depicted.

TABLE 3.3: Estimated $\text{VaR}_{99}(X)$, and difference (in percentage) with respect to the empirical $\text{VaR}_{99}(X)$, for the best zero-and-one inflated mixture models according to the BIC and the semiparametric/nonparametric densities.

Model	m	K	PD $= 0.05$, $\rho = 0.10$		PD $= 0.10$, $\rho = 0.20$	
			$\text{VaR}_{99}(X)$	Difference %	$\text{VaR}_{99}(X)$	Difference %
Data set A						
empirical			1905.79		4445.26	
beta	2	3	1886.59	−1.00	4371.00	−1.67
GB1	4	2	1888.52	−0.91	4555.93	2.49
logit-logistic	2	3	1894.61	−0.59	4542.75	2.19
logit-Gumbel	2	3	1919.75	0.73	4471.65	0.59
logit-normal	2	3	1875.60	−1.58	4415.49	−0.67
logit-rev.Gumbel	2	3	1882.55	−1.22	4511.99	1.50
logit-exGaus	3	3	1908.00	0.12	4274.67	−3.84
logit-skew-normal	3	3	1935.73	1.57	4398.44	−1.05
logit-t	3	2	1892.31	−0.71	4356.53	−2.00
logit-Johnson S_u	4	2	1843.19	−3.28	4384.85	−1.36
logit-sinh-arcsinh	4	2	1938.08	1.69	4373.87	−1.61
logit-skew-t	4	2	1946.75	2.15	4531.06	1.93
logit-EGB2	4	2	1904.94	−0.04	4383.17	−1.40
C-BK			2202.29	15.56	5376.00	20.94
H-BK			2184.86	14.64	5226.58	17.58
CZ-BK			1898.44	−0.39	4589.08	3.24
GK			1673.63	−12.18	4042.99	−9.05
Data set B						
empirical			5225.72		12475.91	
beta	2	4	5191.68	−0.65	12200.98	−2.20
GB1	4	3	5185.73	−0.77	12019.52	−3.66
logit-logistic	2	4	5040.42	−3.55	12295.00	−1.45
logit-Gumbel	2	4	5077.97	−2.83	12230.92	−1.96
logit-normal	2	4	5116.21	−2.10	12102.48	−3.00
logit-rev.Gumbel	2	4	5129.31	−1.84	12698.18	1.78
logit-exGaus	3	4	5158.77	−1.28	12245.80	−1.84
logit-skew-normal	3	4	5124.58	−1.94	12453.30	−0.18
logit-t	3	3	5168.55	−1.09	12045.76	−3.45
logit-Johnson S_u	4	3	5206.82	−0.36	12121.95	−2.84
logit-sinh-arcsinh	4	4	5143.43	−1.57	12233.89	−1.94
logit-skew-t	4	3	5060.07	−3.17	12298.01	−1.43
logit-EGB2	4	4	5172.43	−1.02	12307.16	−1.35
C-BK			5026.75	−3.81	11780.16	−5.58
H-BK			5434.86	4.00	13101.11	5.01
CZ-BK			5142.40	−1.59	12097.50	−3.03
GK			4149.65	−20.59	9445.05	−24.29

To sum up, the results of this simulation study remark the necessity of taking into account the multilevel dimension of the LGD. Indeed, among the considered semiparametric and nonparametric approaches, only the one based on a zero-and-one inflated approach shows a competitive performance.

3.4 Conclusions

Modelling the loss given default is an important aspect, both from a regulatory and a risk management point of view. Unfortunately, the distinctive characteristics of its distribution makes this task difficult. In this work, zero-and-one inflated mixture models, in which a three level multinomial model is considered for the membership of the LGD values to the sets $\{0\}$, $(0,1)$ and $\{1\}$, and a finite mixture of distributions is used on $(0,1)$, are proposed. Differently from de Oliveira Jr *et al.* (2015), where the number of mixture components is limited to two, this number is herein left free and selected by the BIC. Moreover, the family of candidate distributions on $(0,1)$ to be used as mixture components is extended by applying the inverse-logit transformation to some classical distributions with support $(-\infty,\infty)$.

The real banking loans data applications suggested that limiting the number of mixture components to one or two is too restrictive. In fact, according to the BIC, quite often all the estimated models had at least 3 mixture components. This family of models showed its effectiveness also when compared to other well-established semiparametric and nonparametric approaches used in the credit risk literature. A further main finding from the empirical analysis was that there is not a specific model that works universally better than the others. So, almost all the proposed zero-and-one inflated mixture models were reasonably good candidates for fitting the LGD distribution, and the suggestion is to fit all of them and to choose the best one *a posteriori*. Given the flexibility, interpretability, and tractability of these models, they should be closer considered in credit risk modelling.

Chapter 4

Dichotomous unimodal compound models: Application to the distribution of insurance losses [1]

4.1 Introduction

It is pivotal in the insurance industry to find adequate models for loss data, in order to correctly compute premiums, risk measures and the required reserves. This necessity has been accelerated over the last ten years by a revised regulatory framework such as Solvency II and Basel II/III (Brazauskas and Kleefeld, 2016). However, modeling insurance losses is not an easy task because of the distinctive characteristics of their distribution. As widely documented, the loss distribution is unimodal hump-shaped, highly positively skewed and with a heavy right tail (Furman, 2008; Ahn *et al.*, 2012; Jeon and Kim, 2013; Abu Bakar *et al.*, 2015).

Among the different approaches, the parametric one has been the most followed in the actuarial literature. The flexibility of a parametric distribution is a desirable feature, but usually multi-parameter distributions can present several computational challenges. This prompted researchers to seek parsimonious yet sufficiently flexible and interpretable models for insurance losses. Some authors argue that observed losses can be described by a single probability distribution, such as the log-normal (Bickerstaff, 1972; Burnecki *et al.*, 2000), or the Pareto distribution (Packová and Brebera, 2015; Burnecki *et al.*, 2005). However, as pointed out by Cooray and Ananda (2005), the Pareto distribution, due to the monotonically decreasing shape of the density, does not provide a reasonable fit when the density of data is hump-shaped. In these cases the log-normal distribution is typically used, but it fades away to zero more quickly than the Pareto distribution. This implies that the log-normal model fails to cover the higher losses. Some models have been proposed to solve this issue in the actuarial literature (see, e.g., Cooray and Ananda, 2005; Pigeon and Denuit, 2011; Abu Bakar *et al.*, 2015; Punzo *et al.*, 2018). Alternative models are based on the skew-normal, skew-*t* or skew-logistic Adcock *et al.* (2015); Eling *et al.* (2010); Kazemi and Noorizadeh (2015). However, these distributions defined on the whole real line are not adequate to the positive support of the losses, because of the boundary bias problem mentioned in Section 3.3.3.3.

In Section 4.2 a compound approach is proposed, accommodating all the peculiarities of the loss distribution until here discussed. Starting from the scale mixture

[1] This work is based on the following publication: Tomarchio S.D., Punzo A. (2020). Dichotomous unimodal compound models: application to the distribution of insurance losses. *Journal of Applied Statistics*, **47**(13–15), 2328–2353. The current manuscript is a combined effort of the authors. However, Tomarchio S.D. contributed in conceptualization, implementation, data elaboration and writing–original draft preparation; Punzo A. contributed in conceptualization and supervision.

model illustrated in Section 2.2, a 2-parameter unimodal hump-shaped distribution, defined on a positive support and reparameterized with respect to the mode $\theta > 0$ and to another parameter $\gamma > 0$ related to the distribution variability, is considered. The γ parameter is then scaled by a dichotomous mixing variable that depends on a vector of parameters $\boldsymbol{\nu}$ governing the tails behavior. The resulting model can be seen as a 2-component contaminated model (Punzo and McNicholas, 2017a; Punzo *et al.*, 2019; Mazza and Punzo, 2020) in which one component, often called "contaminant", is an inflated version of the other, herein called "conditional", and allows a more flexible accommodation of the outlying observations. Additionally, since both components have the same mode, the model guarantees unimodality in θ.

The proposed model can also allow for an automatic detection of atypical losses via a simple procedure based on maximum *a posteriori* probabilities. Specifically, and in the fashion of Aitkin and Wilson (1980), atypical losses are defined with respect to the conditional distribution as points producing an overall distribution that is too heavy-tailed in order to be modeled by the conditional distribution only. Furthermore, such a detection rule allows the partition of the positive real line in two regions (see, Duda *et al.*, 2012; Ingrassia and Punzo, 2016) that could be used to identify different categories of losses. Indeed, their classification can be of interest for insurance companies in tuning premiums and credit scores (Yeo *et al.*, 2001; Kellison and Brockett, 2003).

A drawback of the proposed model is that when extremely large losses need to be accounted for, this makes also heavier its left tail, rising the probability of losses close to zero. Nevertheless, this is a minor problem for at least two reasons: 1) because of its distinctive characteristics, the loss distribution has a very short left tail that could be considered negligible; 2) risk managers are mainly interested in a good description of the right tail, because large losses, though rare in frequency, are the ones that have the most impact on the financial stability of insurance companies Berkowitz (2001).

Two examples of unimodal hump-shaped distributions are examined in Section 4.2. Parameter estimation via the ML approach is discussed in Section 4.3, while computational aspects are analyzed in Section 4.4. A sensitivity analysis is described in Section 4.5, where the robustness of the ML estimator for the proposed models is investigated. These models are then applied to two real insurance loss data sets, along with other well-known competitors, in Section 4.6. Lastly, some conclusions are commented in Section 4.7.

4.2 Methodology

4.2.1 Dichotomous unimodal compound models

Let X be a positive random loss. Requiring that the PDF of X is unimodal hump-shaped and positively skewed, the scale mixture model introduced in Section 2.2 and proposed by Punzo *et al.* (2018) has PDF

$$f_{\text{SM}}(x;\theta,\gamma,\boldsymbol{\nu}) = \int_0^\infty f(x;\theta,\gamma/w)h(w;\boldsymbol{\nu})dw, \quad x > 0, \tag{4.1}$$

where $f(x;\theta,\gamma)$ is the PDF of a unimodal hump-shaped distribution, with mode $\theta > 0$ and variability parameter $\gamma > 0$. If W is degenerate in 1 (i.e. $W \equiv 1$), then $f(x;\theta,\gamma)$ is obtained.

An interesting special case of model (4.1), is obtained if

$$W = \begin{cases} 1 & \text{with probability } \pi \\ 1/\eta & \text{with probability } 1-\pi, \end{cases} \tag{4.2}$$

where $\pi \in (0,1)$ and $\eta > 1$. The PMF of W in (4.2) is

$$h(w;\pi,\eta) = \pi^{\frac{w-1/\eta}{1-1/\eta}} (1-\pi)^{\frac{1-w}{1-1/\eta}}, \quad w \in \{1, 1/\eta\}.$$

Then, model (4.1) can be written as a contaminated model with PDF

$$g(x;\theta,\gamma,\eta,\pi) = \pi f(x;\theta,\gamma) + (1-\pi) f(x;\theta,\eta\gamma), \quad x > 0, \tag{4.3}$$

in which the contaminant distribution $f(x;\theta,\eta\gamma)$ is an inflated version of the conditional one $f(x;\theta,\gamma)$. Consequently, atypical losses can be modeled in a better way. As discussed in Chapter 2, Titterington *et al.* (1985) identifies such type of model as an example of *indirect application* of finite mixture models.

As often happens in robust statistics half of the losses are assumed to be typical (Punzo and McNicholas, 2016; Templ *et al.*, 2019; Cerioli *et al.*, 2019); this is the reason why in this work $\pi \in (0.5,1)$. It is also important to notice that, because both components have their maximum in θ, $g(x)$ will have mode θ. Furthermore, considering that $W \in \{1, 1/\eta\}$, $g(x)$ will have heavier tails with respect to $f(x;\theta,\gamma)$ (or at the limit they are equal when $\pi \to 1^-$ and $\eta \to 1^+$).

Differently from Punzo *et al.* (2018), the additional parameters π and η have an interpretation of practical interest:

- π is the proportion of points from the conditional distribution; in other words, it represents the proportion of typical losses.
- η is the degree of contamination and, since $\eta > 1$, it can be meant as the increase in variability due to the points which do not come from the conditional distribution, i.e. due to the presence of either an excessive number of losses close to zero or to excessively large losses. Therefore, it is an inflation parameter.

Another interesting characteristic of model (4.3) is that, once the parameters are estimated (marked with a "hat"), it is possible to determine whether a generic loss x is typical via the *a posteriori* probability

$$v\left(x;\hat{\theta},\hat{\gamma},\hat{\eta},\hat{\pi}\right) = \frac{\hat{\pi} f\left(x;\hat{\theta},\hat{\gamma}\right)}{g\left(x;\hat{\theta},\hat{\gamma},\hat{\eta},\hat{\pi}\right)}. \tag{4.4}$$

Specifically, x is considered typical if $v(x;\hat{\theta},\hat{\gamma},\hat{\eta},\hat{\pi}) > 0.5$, while it is considered atypical otherwise. Such a decision rule can be equivalently defined in terms of the discriminant functions

$$D_{\text{typical}}\left(x;\hat{\theta},\hat{\gamma},\hat{\pi}\right) = \hat{\pi} f\left(x;\hat{\theta},\hat{\gamma}\right)$$

and

$$D_{\text{atypical}}\left(x;\hat{\theta},\hat{\gamma},\hat{\eta},\hat{\pi}\right) = (1-\hat{\pi}) f\left(x;\hat{\theta},\hat{\eta}\hat{\gamma}\right),$$

such that x is classified as typical if

$$D_{\text{typical}}\left(x;\hat{\theta},\hat{\gamma},\hat{\pi}\right) > D_{\text{atypical}}\left(x;\hat{\theta},\hat{\gamma},\hat{\eta},\hat{\pi}\right), \tag{4.5}$$

and atypical otherwise (Duda *et al.*, 2012; Ingrassia and Punzo, 2016). By solving (4.5) as a function of x, the positive real line is partitioned in two regions of typical and atypical data, delimited by the intersection points between the two discriminant functions. Indeed, these points represent the situation of maximum assignment uncertainty, where the probabilities to be a typical or an atypical point coincide. As better shown in Section 4.6, the region of atypical data involves the two tails of the model, whereas the area between them entails the typical region. This might be useful to identify different categories of losses that could be classified as atypically low (left tail), typical (center) and atypically high (right tail) with respect to the conditional distribution.

Among the existing 2-parameter unimodal hump-shaped distributions that can be used for f, log-normal and unimodal gamma are considered in the next paragraphs. Furthermore, by using and extending the notation of Punzo *et al.* (2018), the model in (4.3) will be referred as dichotomous unimodal compound model in the following.

4.2.2 Specific cases

4.2.2.1 Mode-parametrized log-normal distribution

The PDF of a log-normal (LN) distribution with the standard parameterization is given by

$$f(x;\mu,\sigma) = \frac{e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma x}, \quad x > 0,$$

where $\mu \in$ and $\sigma > 0$ are the mean and the standard deviation of the variable's natural logarithm, respectively.

With the purpose of having a distribution that can be inserted in model (4.3), a reparameterization is needed. Imposing

$$\begin{cases} \mu &= \ln(\theta) + \gamma \\ \sigma^2 &= \gamma \end{cases} \quad \Rightarrow \quad \begin{cases} \theta &= e^{\mu-\sigma^2} \\ \gamma &= \sigma^2 \end{cases},$$

the PDF becomes

$$f(x;\theta,\gamma) = \frac{e^{-\frac{(\ln(x)-\ln(\theta)-\gamma)^2}{2\gamma}}}{\sqrt{2\pi\gamma}x}, \quad x > 0, \tag{4.6}$$

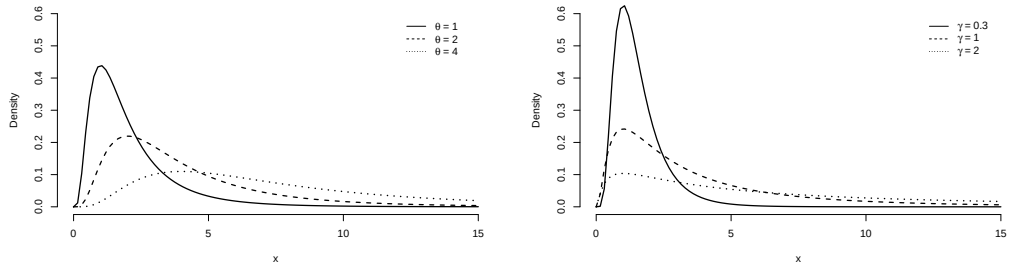
with $\theta > 0$ and $\gamma > 0$.

The effect of varying the mode θ, keeping fixed γ, is shown in Figure 4.1(a). The variance of a random variable with density function (4.6) is

$$(e^{\gamma} - 1)\,\theta^2 e^{3\gamma}. \tag{4.7}$$

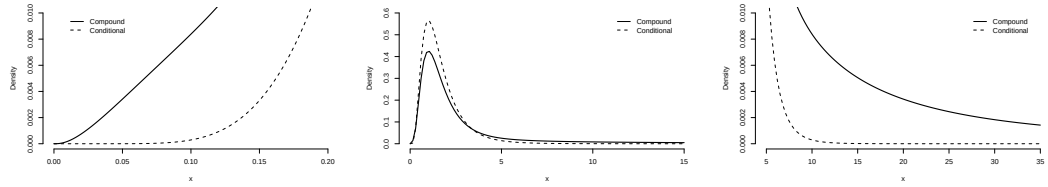
For a fixed θ in (4.7), the variance rises if γ increases, confirming that γ governs the variability of the distribution. This is illustrated in Figure 4.1(b).

When the LN distribution is chosen in (4.3), the LN dichotomous unimodal compound (LN-DUC) model is obtained. An example of LN-DUC model is illustrated in Figure 4.2. It is possible to see that the tails of the compound model (4.3) are heavier than those of the simple conditional distribution.



(a) $\gamma = 0.5$ (b) $\theta = 1$

FIGURE 4.1: Mode-parameterized log-normal densities (4.6) in (a) and (b).



(a) (b) (c)

FIGURE 4.2: A LN-DUC compared to a LN distribution in (b) with a specific zoom in the left (a) and right (c) tails, respectively.

Finally, when the LN-DUC model is considered, the intersection points between the discriminant functions, delimiting the typical and the atypical regions, are

$$x_1 = \theta e^{-\frac{\sqrt{\gamma(\eta-1)\eta[\gamma(\eta-1)-2\ln(1-\alpha)+2\ln(\alpha)+\ln(\eta)]}}{\eta-1}},$$

with $x_1 \in (0, \theta)$, and

$$x_2 = \theta e^{\frac{\sqrt{\gamma(\eta-1)\eta[\gamma(\eta-1)-2\ln(1-\alpha)+2\ln(\alpha)+\ln(\eta)]}}{\eta-1}},$$

with $x_2 \in (\theta, \infty)$.

4.2.2.2 Mode-parametrized unimodal gamma distribution

The PDF of a unimodal hump-shaped gamma (UG) distribution with the standard parameterization is

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1} e^{-\frac{x}{\beta}}}{\beta^{\alpha} \Gamma(\alpha)}, \quad x > 0,$$

with shape parameter $\alpha > 1$ and scale parameter $\beta > 0$. In order to have a distribution that can be inserted in model (4.3), a reparameterization is needed. Setting

$$\begin{cases} \alpha &= \frac{\theta}{\gamma} + 1 \\ \beta &= \gamma \end{cases} \quad \Rightarrow \quad \begin{cases} \theta &= \beta(\alpha - 1) \\ \gamma &= \beta \end{cases},$$

the PDF becomes

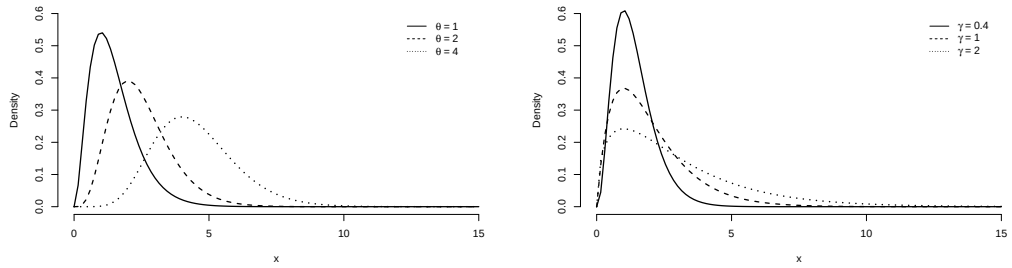
$$f(x;\theta,\gamma) = \frac{x^{\frac{\theta}{\gamma}} e^{-\frac{x}{\gamma}}}{\gamma^{\frac{\theta}{\gamma}+1}\Gamma\left(\frac{\theta}{\gamma}+1\right)}, \quad x > 0, \tag{4.8}$$

with $\theta > 0$ and $\gamma > 0$.

The effect of varying the mode θ, keeping fixed γ, is shown in Figure 4.3(a). The variance of a random variable X with density function (4.8) is

$$\gamma^2 + \theta\gamma. \tag{4.9}$$

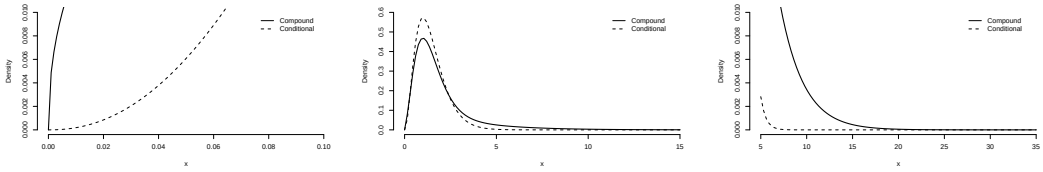
Fixing θ in (4.9), the variance increases if γ increases, confirming that γ governs the variability of the distribution. The effect of varying γ, keeping fixed θ is shown in Figure 4.3(b).



(a) $\gamma = 0.5$ (b) $\theta = 1$

FIGURE 4.3: Mode-parameterized unimodal gamma densities (4.8).

When the UG distribution is chosen in (4.3), the UG dichotomous unimodal compound (UG-DUC) model is obtained. An example of UG-DUC model is presented in Figure 4.4. Also in this case, it is possible to notice how the tails of the UG-DUC model are heavier than those of the UG distribution.



(a) (b) (c)

FIGURE 4.4: A UG-DUC model compared to a UG distribution in (b) with a specific zoom in the left (a) and right (c) tails, respectively.

Lastly, when the UG-DUC model is considered, recovering a closed-form expression for the intersection points between the discriminant functions is analytically cumbersome. However, they can be easily obtained numerically by using, for instance, the *uniroot.all()* function of the **rootSolve** package Soetaert (2009).

4.3 Parameter estimation

As mentioned in Section 2.3.2, the direct maximization approach is used to estimate the parameters of model (4.3). This is because closed-form expressions are not available for most of the parameters involved in the M-step; for these parameters, numerical methods, such as the BFGS algorithm, must be used. On this regard, the *optim()* function, included in the **stats** package, is used for the maximization of the log-likelihood function. The BFGS algorithm, is passed to this function via the argument *method*.

To start the optimization algorithm, the initial values for the parameters must be specified. A system of two equations is solved to initialize θ and γ. The first equation matches the empirical and the theoretical modes. The second equation is model-dependent. In detail, it matches the empirical and the theoretical variances, for the UG-DUC model, and the empirical and the theoretical means for the LN-DUC model. As discussed in Section 4.2, when $\pi \to 1^-$ and $\eta \to 1^+$, the conditional distribution $f(x;\theta,\gamma)$ is obtained. For this reason, the starting values for π and η are set to $\pi_0 = 0.99$ and $\eta_0 = 1.01$. From an operative point of view, thanks to the monotonicity of the BFGS algorithm, this ensures that the log-likelihood of model (4.3) will be always greater than, or equal to, the log-likelihood of the conditional distribution. This is an important consideration for choosing between the conditional distribution and its corresponding dichotomous unimodal compound version, when using likelihood-based model selection criteria.

All the parameters involved are subject to constraints, and to make the maximization of the log-likelihood unconstrained, as required by the BFGS algorithm, a transformation/back-transformation approach has been implemented (Zucchini *et al.*, 2017; Bagnato and Punzo, 2019). In detail, the original constrained parameters are mapped to unconstrained real values (marked with a "tilde") and, after the log-likelihood is maximized with respect to the unconstrained parameters, a back-transformation is applied to obtain the constrained parameter estimates. The following transformations and back-transformations are used:

$$\tilde{\theta} = \ln(\theta) \quad \leftrightarrow \quad \theta = \exp(\tilde{\theta}), \qquad \tilde{\gamma} = \ln(\gamma) \quad \leftrightarrow \quad \gamma = \exp(\tilde{\gamma}),$$

$$\tilde{\eta} = \ln(\eta - 1) \quad \leftrightarrow \quad \eta = \exp(\tilde{\eta}) + 1,$$

$$\tilde{\pi} = \ln\left(\frac{\phi}{1-\phi}\right) \quad \leftrightarrow \quad \pi = \frac{0.5 + \exp(\tilde{\pi})}{1 + \exp(\tilde{\pi})},$$

where $\phi = (\pi - 0.5)/0.5$.

4.4 Computational and operative aspects

4.4.1 Model comparison

Several measures are used to compare the fitted models. Specifically, in Section 4.4.1.1 a likelihood-ratio (LR) test is discussed, whereas in Section 4.4.1.2 a specific analysis on the right tail goodness of fit is conducted.

4.4.1.1 Global fit evaluation

A LR test is often used to compare the goodness of fit of two competing models, one of which (the null model) is a special case of the other (the alternative model). In

this work, when the null model is the UG distribution, then the alternative model is the UG-DUC, while when the null model is the LN distribution, then the alternative model is the LN-DUC. Under the null hypothesis of no improvement, the test statistic is

$$\mathrm{LR} = 2\left[l(\hat{\boldsymbol{\delta}}_1) - l(\hat{\boldsymbol{\delta}}_0)\right],$$

where $\hat{\boldsymbol{\delta}}_1$ and $\hat{\boldsymbol{\delta}}_0$ are the parameter vectors of the alternative and null models, respectively, and $l(\hat{\boldsymbol{\delta}}_1)$ and $l(\hat{\boldsymbol{\delta}}_0)$ are their maximized log-likelihood values, respectively. However, regularity conditions do not hold for mixture-based models, and the LR statistic has not its usual asymptotic null distribution of a χ^2 random variable with m degrees of freedom, where $m \in \mathbb{N}_+$ is the difference between the number of estimated parameters of the alternative and the null models. To overcome this issue, under the same null and alternative hypotheses, the following parametric double bootstrap procedure is implemented (McLachlan and Peel, 2000; MacKinnon, 2009):

1. Fit the null and alternative models to the sample and compute the LR statistic, say $\mathrm{LR}_{\mathrm{obs}}$;
2. Generate B_1 bootstrap samples, of size N, from the model fitted under the null;
3. For each of the B_1 bootstrap samples, fit the null and the alternative models, and compute the first-level bootstrap LR statistic, say LR_j^*, with $j = 1, \ldots, B_1$;
4. Calculate the first-level bootstrap p^*-value as $\frac{1}{B_1}\sum_{j=1}^{B_1} I\left(\mathrm{LR}_j^* > \mathrm{LR}_{\mathrm{obs}}\right)$;
5. For every B_1 bootstrap sample, generate B_2 bootstrap samples, of size N, from the model fitted under the null to B_1;
6. For each of the B_2 bootstrap samples, fit the null and the alternative models, and compute the second-level bootstrap LR statistic, say LR_{jl}^{**}, with $l = 1, \ldots, B_2$;
7. For every B_1 bootstrap sample, compute the second-level bootstrap p_j^{**}-value as $\frac{1}{B_2}\sum_{l=1}^{B_2} I\left(\mathrm{LR}_{jl}^{**} > \mathrm{LR}_j^*\right)$;
8. Calculate the double bootstrap p-value as the proportion of the p_j^{**} that are more extreme than p^*, i.e. $p = \frac{1}{B_1}\sum_{j=1}^{B_1} I\left(p_j^{**} < p^*\right)$.

The double bootstrap procedure reduces the bias in the bootstrap estimates obtained from the first level, but is computationally demanding, since $1 + B_1 + (B_1 \times B_2)$ test statistics must be calculated. In this work, $B_1 = 500$ and $B_2 = 250$, yielding to a total of 125501 estimates. The double bootstrap p-value is compared with the 0.05 significance level.

Besides comparing the dichotomous unimodal compound models with their corresponding conditional distributions, the BIC is used to make comparisons with some benchmark distributions, making possible to draw up an overall goodness fit ranking.

4.4.1.2 Right tail fit evaluation

A standard procedure in the insurance literature consists in comparing the empirical value of some risk measures, with those estimated by the fitted models (see, e.g. Eling, 2012; Bernardi *et al.*, 2012; Kazemi and Noorizadeh, 2015; Abu Bakar *et al.*, 2015; Punzo *et al.*, 2018). This is useful for assessing the estimated tail behavior, since it is

of particular interest for risk managers. Specifically, two well-known risk measures are considered: the value at risk (VaR) and the tail-value at risk (TVaR). Being related to the quantiles of a distribution, the closer the estimated risk measures are to the empirical ones, the better the fitting on the tail is.

The VaR has been already defined in Section 3.3.3. The TVaR quantifies the expected value of the loss given that an event outside a given probability level has occurred. It is defined as

$$\text{TVaR}_c(X) = E\left[X | X \geq \text{VaR}_c(X)\right], \quad 0 \leq c \leq 1.$$

If the underlying distribution for X is continuous, then the TVaR is the same as the expected shortfall. An alternative formulation, that will be useful in the following, expresses the TVaR in terms of the VaR as

$$\text{TVaR}_c(X) = \frac{1}{1-c}\int_c^1 \text{VaR}_u(X)du. \tag{4.10}$$

In this work, the probability levels used for both risk measures are $c = 0.95$ and $c = 0.99$.

To evaluate the goodness of the VaR and TVaR estimates produced by the competing models, two backtesting procedures are also implemented. For the VaR, a binomial test examines, under the null hypothesis, if the proportion of violations $\hat{\rho}$ obtained using the estimates of the VaR ($\hat{\rho} = y/n$, where y is the number of losses exceeding the estimated VaR and N is the sample size), is compatible with the one expected $\rho = (1-c)$ Kupiec (1995). The test is performed via the *VaRTest()* function of the **rugarch** package Ghalanos (2015).

For the TVaR, the backtest suggested by Emmer *et al.* (2015) is implemented. Specifically, it relies on a simple approximation of the TVaR representation in (4.10). Given a $\text{TVaR}_c(X)$, they suggest to compute and backtest the VaR at the following four levels: $\text{VaR}_c(X)$, $\text{VaR}_{0.75c+0.25}(X)$, $\text{VaR}_{0.5c+0.5}(X)$ and $\text{VaR}_{0.25c+0.75}(X)$. If all the four backtests are not rejected, then the estimate of TVaR can be considered acceptable. As a consequence, the minimum among the four p-values is enough to decide whether the TVaR estimate has to be discarded; such a minimum will be reported in the analyses of Section 4.6.

In this work, both backtesting procedures are compared with the 0.05 significance level.

4.4.2 Competing models and approaches

The proposed models are compared to several standard distributions used in the actuarial literature, and whose parameters are estimated by using the ML approach. Specifically, they are listed in Table 4.1, along with the **R** functions and packages used to fit them to the data. About the UG and LN distributions, a convenient code is implemented to find the ML estimates of the parameters of these distributions, as done for the proposed models. Indeed, the θ and γ parameters are estimated via the *optim()* function, by using the same strategy explained in Section 4.3.

In addition to the ML approach, two further methodologies are considered: the t-score and the the PORT-MO$_p$. Specific details about them are provided in Section 4.4.2.1 and Section 4.4.2.2, respectively.

TABLE 4.1: **R** functions and packages used for the ML-based competitors.

Distribution	Function	Package
Exponential	*fitdistr()*	**MASS** (Venables and Ripley, 2002)
Weibull	*fitdist()*	**fitdistrplus** (Delignette-Muller and Dutang, 2015)
Normal	*fitdist()*	**fitdistrplus**
Logistic	*fitdist()*	**fitdistrplus**
Skew-logistic	*glogisfit()*	**glogis** (Zeileis and Windberger, 2014)
Skew-Normal	*snormFit()*	**fGarch** (Wuertz and Chalabi, 2016)
Skew-*t*	*sstdFit()*	**fGarch**
Hyperbolic	*hyperbFit()*	**HyperbolicDist** (Scott, 2009)

4.4.2.1 The t-score approach

The t-score moment estimator has been proposed and discussed by Fabián (2006, 2007, 2010). In the insurance literature this estimator has been used by Stehlík *et al.* (2008, 2010). Specifically, the authors discussed the t-score moment estimator for the Pareto distribution, both in its "American" and "European" parametrization. The main difference between these two parametrizations is that the former starts in zero, whereas the latter begins in a threshold that either should be carefully estimated (Stehlík *et al.*, 2008, 2010) or is assumed to be known (Rytgaard, 1990). Both versions have been fitted in the two real data applications, but the risk measures obtained with the "European" parametrization were very far from their empirical counterparts, suggesting that this parametrization is not adequate for modeling these data. On the contrary, the estimated risk measures obtained with the "American" parametrization are close to the empirical values, and for this reason only this version of the Pareto distribution is considered hereafter. Specifically, it has PDF $\alpha\lambda^{\alpha}/(x+\lambda)^{\alpha+1}$, with $\alpha > 0$ and $\lambda > 0$. According to Fabián (2007), its parameters are estimated by solving the following system of two equations:

$$\sum_{i=1}^{n} \frac{\alpha x_i - \lambda}{x_i + \lambda} = 0 \tag{4.11}$$

$$\frac{1}{n}\sum_{i=1}^{n} \left(\frac{\alpha x_i - \lambda}{x_i + \lambda}\right)^2 = \frac{\alpha}{\alpha + 2} \tag{4.12}$$

Therefore, by using a generalized method of the moments approach, (4.11) and (4.12) match the theoretical t-score moments to their empirical counterparts.

4.4.2.2 The PORT-MO$_p$ approach

The peaks over a random threshold-mean of order p (PORT-MO$_p$) estimator has been recently introduced by Gomes *et al.* (2016) and used for VaR estimation by Figueiredo *et al.* (2017). Specifically, given a sample of ascending order statistics $x_{1:N} \leq \cdots \leq x_{N:N}$, the PORT-MO$_p$VaR is

$$\text{PORT-MO}_p\text{VaR}\,(k; p, s) = (x_{n-k:n} - x_{n_s:n}) \left(\frac{k}{nc}\right)^{\text{H}_k(p,s)} + x_{n_s:n},$$

where $N_s := \lfloor Ns \rfloor + 1$, with $\lfloor \cdot \rfloor$ denoting the floor function, $\mathrm{H}_k(p, s)$ is the PORT-MO$_p$ estimator, c is the probability level for compute the VaR and $\{k, p, s\}$ are the tuning parameters to be estimated. Relatedly, the tuning parameters are estimated by combining the bootstrap scheme in Longin (2016) with the double bootstrap algorithm discussed by Brilhante *et al.* (2013). This results in a reasonably sophisticated and time-consuming procedure. Due to the specificity of the underlying concepts and the length of the entire procedure, its detailed explanation is here avoided. In the following, the references and the execution order of the steps are mentioned, with a comment only on those where an arbitrary choice is required:

1. Steps 1 to 2.1 of Longin (2016), page 130;
   (a) In step 1, we set Q to be the sequence of values from 0 to 1, with increments of 0.1.
2. Steps 1 to 16 of Brilhante *et al.* (2013), pages 527–528;
   (a) In step 5, we set b to be the sequence of values from 0.925 to 0.995, with increments of 0.01.
3. Steps 3 to 5 of Longin (2016), page 130.

4.5 Simulation study

The aim of this section is to investigate how atypical observations affect the ML estimator for the θ and γ parameters of the UG and LN distributions, and how its robustness is increased when the corresponding UG-DUC and LN-DUC models are considered. A similar simulation study can be found in Stehlík *et al.* (2010), where the authors show how the ML estimator for the shape parameter of the Pareto distribution is affected when atypical observations are added to the data. For similar purposes, two sensitivity analyses are conducted, that differ depending on how the data are contaminated. The first analysis in Section 4.5.1 considers the following scenario:

1. generate πn typical observations from a UG distribution;
2. generate $(1 - \pi)n$ atypical observations from the same UG distribution, with the only difference that the γ-parameter is multiplied by an inflation factor η;
3. fit both the UG and UG-DUC to the merged data.

The same procedure is repeated by changing the UG with the LN distribution and the UG-DUC model with the LN-DUC. Therefore, in the first analysis, typical and atypical points come from a distribution of the same type, as assumed in model (4.3). Furthermore, only the UG and the LN distribution misspecify the data.

The second analysis in Section 4.5.2 considers the following scenario:

1. generate πn typical observations from a UG distribution;
2. generate $(1 - \pi)n$ atypical observations from a LN distribution, with the same $\theta^* = \theta$ of the UG distribution and with variability parameter γ^*;
3. fit the UG and UG-DUC models to the merged data.

Also in this case, the procedure is repeated for the complementary situation where, in steps 1 and 3 the UG and UG-DUC models are substituted by the LN and LN-DUC models, respectively, and the UG replaces the LN distribution in step 2. In this second study, the atypical points come from a distribution of a different type with respect to the one generating the typical points. Therefore, the estimation performances are evaluated in a context of misspecification also for model (4.3).

In each analysis, the sample size is $N = 1000$, and three different proportions of typical points π are combined with three levels of contamination, governed by η in the first analysis and by the variability parameter γ^* of the contaminant distribution in the second one. This yields a total of 9 different contamination cases. For each of them, 10000 replications are considered; then, a total of $2 \times 10000 \times 9 \times 2 = 360000$ samples are generated. The mean and the standard deviation (in brackets) of the estimated θ and γ, over these replications, are reported. For the sake of simplicity, each simulation scenario is identified according to the data generating process (DGP), labeled by matching with a "+" the name of the distributions generating the typical and the atypical observations, respectively.

4.5.1 Sensitivity analysis I

The parameters used for the UG+UG DGP are $\theta = 1$ and $\gamma = 1$, while those for the LN+LN DGP are $\theta = 1$ and $\gamma = 0.5$. The average estimated parameters, along with their standard deviation, under each scenario, are shown in Table 4.2. Specifically, each subtable displays the ML estimates of θ and γ obtained by fitting the UG and UG-DUC models for the UG+UG DGP, and by the LN and LN-DUC models for the LN+LN DGP.

It is easy to see that, for a fixed π, the more η increases, the more the differences between the estimates produced by the competing models become. The same occurs keeping fixed η and decreasing π. Furthermore, in presence of atypical observations, the estimates produced by the UG-DUC and LN-DUC models are always closer to the true values, indicating the increased robustness of the ML estimator.

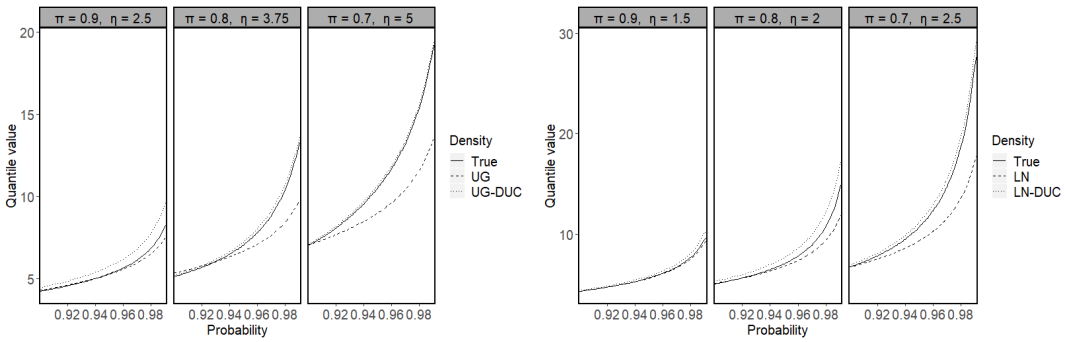
Since the VaR and the TVaR are based on quantiles, it is interesting to investigate how the differences between the parameter estimates produced by the competing models have an effect on the corresponding quantile estimates. For illustrative purposes, only the (π, η) combinations that are highlighted with a gray background in Table 4.2 are considered. They represent situations with growing levels of contamination in the data, within each DGP. Figure 4.5 illustrates the quantile values of the conditional distributions and their dichotomous unimodal compound versions, obtained by using the average estimated parameters in the diagonals of Table 4.2, against the true quantiles of the corresponding DGPs, for growing probability levels. It is worth to notice that, within each DGP and for low levels of contamination, the estimated quantiles of the competing models are close enough to true ones. However, when the contamination in the data starts to increase, the estimated quantiles of the conditional distributions start to diverge from the true ones, whereas the corresponding dichotomous compound models fit always better.

4.5.2 Sensitivity analysis II

The parameters used for the UG+LN DGP are $\theta = \theta^* = 1$ and $\gamma = 1$, whereas those for the LN+UG DGP are $\theta = \theta^* = 1$ and $\gamma = 0.5$. The average estimated parameters, along with their standard deviation, under each scenario, are shown in Table 4.3. Also in this case, each subtable displays the ML estimates of θ and γ obtained by

TABLE 4.2: Average $\hat{\theta}$ and $\hat{\gamma}$ values, with standard deviations in brackets, estimated over 10000 replications by the UG and UG-DUC models for the UG+UG DGP, and by the LN and LN-DUC models for the LN+LN DGP.

		UG+UG DGP			LN+LN DGP		
		η			η		
		2.5	3.75	5	1.5	2	2.5
	UG				***LN***		
π	0.9	$\hat{\theta} = 0.97$ (0.05) $\hat{\gamma} = 1.18$ (0.06)	$\hat{\theta} = 0.91$ (0.06) $\hat{\gamma} = 1.37$ (0.08)	$\hat{\theta} = 0.83$ (0.07) $\hat{\gamma} = 1.57$ (0.11)	$\hat{\theta} = 0.99$ (0.03) $\hat{\gamma} = 0.53$ (0.02)	$\hat{\theta} = 0.97$ (0.03) $\hat{\gamma} = 0.57$ (0.03)	$\hat{\theta} = 0.95$ (0.04) $\hat{\gamma} = 0.62$ (0.03)
	0.8	$\hat{\theta} = 0.94$ (0.06) $\hat{\gamma} = 1.36$ (0.07)	$\hat{\theta} = 0.83$ (0.07) $\hat{\gamma} = 1.72$ (0.10)	$\hat{\theta} = 0.67$ (0.09) $\hat{\gamma} = 2.13$ (0.14)	$\hat{\theta} = 0.99$ (0.03) $\hat{\gamma} = 0.56$ (0.03)	$\hat{\theta} = 0.96$ (0.04) $\hat{\gamma} = 0.64$ (0.03)	$\hat{\theta} = 0.92$ (0.04) $\hat{\gamma} = 0.74$ (0.04)
	0.7	$\hat{\theta} = 0.92$ (0.06) $\hat{\gamma} = 1.53$ (0.08)	$\hat{\theta} = 0.76$ (0.08) $\hat{\gamma} = 2.06$ (0.12)	$\hat{\theta} = 0.54$ (0.10) $\hat{\gamma} = 2.66$ (0.17)	$\hat{\theta} = 0.98$ (0.03) $\hat{\gamma} = 0.59$ (0.03)	$\hat{\theta} = 0.95$ (0.04) $\hat{\gamma} = 0.70$ (0.03)	$\hat{\theta} = 0.89$ (0.04) $\hat{\gamma} = 0.84$ (0.04)
	UG-DUC				***LN-DUC***		
π	0.9	$\hat{\theta} = 1.00$ (0.05) $\hat{\gamma} = 0.95$ (0.13)	$\hat{\theta} = 1.00$ (0.05) $\hat{\gamma} = 0.98$ (0.10)	$\hat{\theta} = 1.00$ (0.05) $\hat{\gamma} = 0.99$ (0.08)	$\hat{\theta} = 1.00$ (0.03) $\hat{\gamma} = 0.48$ (0.06)	$\hat{\theta} = 1.00$ (0.04) $\hat{\gamma} = 0.48$ (0.06)	$\hat{\theta} = 1.00$ (0.03) $\hat{\gamma} = 0.49$ (0.05)
	0.8	$\hat{\theta} = 1.00$ (0.05) $\hat{\gamma} = 0.96$ (0.14)	$\hat{\theta} = 1.00$ (0.06) $\hat{\gamma} = 0.99$ (0.11)	$\hat{\theta} = 1.00$ (0.06) $\hat{\gamma} = 0.99$ (0.09)	$\hat{\theta} = 1.00$ (0.03) $\hat{\gamma} = 0.49$ (0.06)	$\hat{\theta} = 1.00$ (0.04) $\hat{\gamma} = 0.48$ (0.06)	$\hat{\theta} = 1.00$ (0.04) $\hat{\gamma} = 0.49$ (0.05)
	0.7	$\hat{\theta} = 1.00$ (0.06) $\hat{\gamma} = 0.98$ (0.16)	$\hat{\theta} = 1.00$ (0.06) $\hat{\gamma} = 0.99$ (0.13)	$\hat{\theta} = 1.00$ (0.06) $\hat{\gamma} = 0.99$ (0.12)	$\hat{\theta} = 1.00$ (0.04) $\hat{\gamma} = 0.50$ (0.07)	$\hat{\theta} = 1.00$ (0.04) $\hat{\gamma} = 0.49$ (0.07)	$\hat{\theta} = 1.00$ (0.04) $\hat{\gamma} = 0.49$ (0.06)



(a) UG+UG DGP (b) LN+LN DGP

FIGURE 4.5: Quantile values from the conditional distributions, their dichotomous unimodal compound versions, and the true DGPs.

fitting the UG and UG-DUC models for the UG+LN DGP, and by the LN and LN-DUC models for the LN+UG DGP.

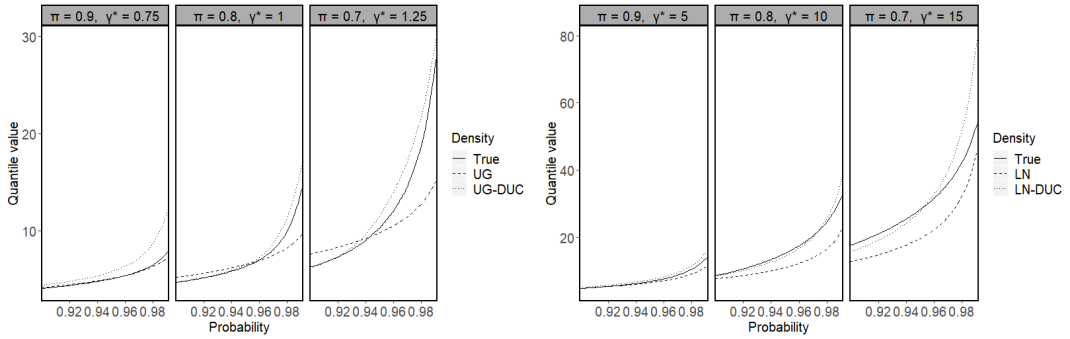
Similar conclusions to those previously discussed can be drawn, in terms of differences between the estimates produced by the competing models. Even in presence of a misspecification effect, the ML estimation carried out by the UG-DUC and LN-DUC models is more robust than the one obtained by fitting the corresponding conditional distributions.

Again, a comparison in terms of quantile values, for the (π, η) combinations that are highlighted with a gray background in Table 4.3, is illustrated in Figure 4.6 for

TABLE 4.3: Average $\hat{\theta}$ and $\hat{\gamma}$ values, with standard deviations in brackets, estimated over 10000 replications by the UG and UG-DUC models for the UG+LN DGP, and by the LN and LN-DUC models for the LN+UG DGP.

		UG+LN DGP			LN+UG DGP		
		η			η		
		0.75	1	1.25	5	10	15
	UG				***LN***		
	0.9	$\hat{\theta} = 0.98$ (0.05) $\hat{\gamma} = 1.12$ (0.07)	$\hat{\theta} = 0.90$ (0.07) $\hat{\gamma} = 1.35$ (0.12)	$\hat{\theta} = 0.73$ (0.13) $\hat{\gamma} = 1.72$ (0.22)	$\hat{\theta} = 0.95$ (0.04) $\hat{\gamma} = 0.64$ (0.03)	$\hat{\theta} = 0.88$ (0.04) $\hat{\gamma} = 0.76$ (0.04)	$\hat{\theta} = 0.82$ (0.04) $\hat{\gamma} = 0.87$ (0.04)
π	0.8	$\hat{\theta} = 0.96$ (0.06) $\hat{\gamma} = 1.25$ (0.08)	$\hat{\theta} = 0.80$ (0.09) $\hat{\gamma} = 1.69$ (0.16)	$\hat{\theta} = 0.46$ (0.17) $\hat{\gamma} = 2.44$ (0.31)	$\hat{\theta} = 0.91$ (0.04) $\hat{\gamma} = 0.76$ (0.04)	$\hat{\theta} = 0.81$ (0.04) $\hat{\gamma} = 0.99$ (0.04)	$\hat{\theta} = 0.72$ (0.04) $\hat{\gamma} = 1.18$ (0.05)
	0.7	$\hat{\theta} = 0.95$ (0.06) $\hat{\gamma} = 1.37$ (0.10)	$\hat{\theta} = 0.71$ (0.11) $\hat{\gamma} = 2.03$ (0.19)	$\hat{\theta} = 0.22$ (0.17) $\hat{\gamma} = 3.13$ (0.33)	$\hat{\theta} = 0.88$ (0.05) $\hat{\gamma} = 0.87$ (0.04)	$\hat{\theta} = 0.77$ (0.04) $\hat{\gamma} = 1.18$ (0.05)	$\hat{\theta} = 0.67$ (0.04) $\hat{\gamma} = 1.43$ (0.06)
	UG-DUC				***LN-DUC***		
	0.9	$\hat{\theta} = 1.02$ (0.05) $\hat{\gamma} = 0.95$ (0.10)	$\hat{\theta} = 1.02$ (0.05) $\hat{\gamma} = 1.01$ (0.08)	$\hat{\theta} = 1.01$ (0.05) $\hat{\gamma} = 1.04$ (0.07)	$\hat{\theta} = 0.99$ (0.04) $\hat{\gamma} = 0.44$ (0.06)	$\hat{\theta} = 1.00$ (0.04) $\hat{\gamma} = 0.44$ (0.04)	$\hat{\theta} = 1.00$ (0.03) $\hat{\gamma} = 0.45$ (0.03)
π	0.8	$\hat{\theta} = 1.04$ (0.05) $\hat{\gamma} = 0.96$ (0.11)	$\hat{\theta} = 1.03$ (0.05) $\hat{\gamma} = 1.03$ (0.09)	$\hat{\theta} = 0.91$ (0.24) $\hat{\gamma} = 1.08$ (0.10)	$\hat{\theta} = 0.98$ (0.04) $\hat{\gamma} = 0.45$ (0.07)	$\hat{\theta} = 1.00$ (0.04) $\hat{\gamma} = 0.41$ (0.04)	$\hat{\theta} = 1.00$ (0.04) $\hat{\gamma} = 0.41$ (0.03)
	0.7	$\hat{\theta} = 1.06$ (0.05) $\hat{\gamma} = 0.96$ (0.11)	$\hat{\theta} = 1.05$ (0.05) $\hat{\gamma} = 1.05$ (0.10)	$\hat{\theta} = 1.03$ (0.09) $\hat{\gamma} = 1.13$ (0.14)	$\hat{\theta} = 0.96$ (0.05) $\hat{\gamma} = 0.55$ (0.11)	$\hat{\theta} = 1.00$ (0.04) $\hat{\gamma} = 0.41$ (0.04)	$\hat{\theta} = 1.01$ (0.04) $\hat{\gamma} = 0.38$ (0.03)

growing probability values. With the exclusion of the combination $(\pi = 0.9, \gamma^* = 0.75)$ for the UG+LN DGP case, representing a situation of low contamination and where the right tail of the UG-DUC model is heavier than necessary, in all the other situations the dichotomous unimodal compound models provide a better fit then their conditional counterparts. This is due to their greater flexibility, that allows the accommodation of the contaminating points in a better way, regardless of whether they are generated by a distribution of the same type or not.



(a) UG+LN DGP (b) LN+UG DGP

FIGURE 4.6: Quantile values from the conditional distributions, their dichotomous unimodal compound versions, and the true DGPs.

4.6 Real data applications

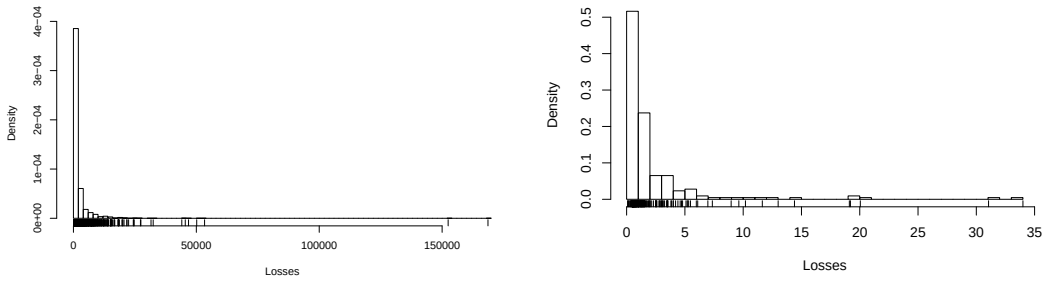
In this section, the dichotomous unimodal compound models, along with the nested conditional distributions, are applied to two real insurance loss data sets.

4.6.1 Data description

The first data set (labeled Frebiloss hereafter) consists of $N = 2387$ French business interruption losses (in French francs; FF), over the period 1985 to 2000, and it is contained in the **CASdatasets** package (Dutang and Charpentier, 2016). For each observation, the total cost in FF is considered. For scaling purposes, payments amount is divided by 1000; thus, thousand of French francs (TFF), instead of FF, are considered.

The second data set (called Swefire hereafter) contains 218 Swedish fire losses (in millions of Swedish krona; SEK) collected in 1982, and it can be extracted from Embrechts and Schmidli (1994). The claims equal to zero are removed from the analysis, implying a final number of $N = 215$ observations.

The histograms of both data sets are displayed in Figure 4.7, whereas their summary statistics are reported in Table 4.4. As it is possible to see, both data sets share the classic characteristics of the insurance losses described in Section 4.1. Indeed, the mean is considerably higher than the median and the third quartile, suggesting extreme right skewness and a heavy right tail, as also confirmed by some large losses that lay quite far from the bulk of the data.



(a) (b)

FIGURE 4.7: Histograms of the (a) Frebiloss and (b) Swefire data sets.

TABLE 4.4: Summary statistics of the Frebiloss and Swefire data sets.

	Frebiloss	Swefire
Min	100.29	0.09
1st Quart.	304.90	0.63
Median	762.25	1.00
Mean	2027.74	2.31
3rd Quart.	1829.39	2.00
Max	168654.35	34.00
St. Dev	5938.06	4.18
Skewness	17.62	4.81
Kurtosis	438.89	27.48

4.6.2 Global results

The comparison between the likelihood-estimated models is presented in Table 4.5, for both data sets. In detail, the $l(\hat{\delta})$ and the BIC values, with corresponding ranking, are provided. In both data sets, the LN-DUC model is the best, while UG-DUC model is ranked third and second, respectively. Importantly, they provide an improvement compared to their conditional distributions, since the p-values of the double bootstrap LR test are 0.00. As highlighted by Kazemi and Noorizadeh (2015), the skew-logistic seems to work slightly better than the skew-normal, and the normal distribution is the worst in any case. Globally, the dichotomous unimodal compound models appear to be very competitive.

TABLE 4.5: Values of $l(\hat{\delta})$ and BIC for the competing models. A ranking is also provided.

Model	Frebiloss			Swefire		
	$l(\hat{\delta})$	BIC	rank	$l(\hat{\delta})$	BIC	rank
UG-DUC	-19983.29	39997.69	3	-328.93	679.34	2
LN-DUC	-19842.98	39717.08	1	-320.30	662.08	1
UG	-20563.23	41142.01	8	-395.35	801.45	8
LN	-19893.29	39802.14	2	-345.68	702.10	4
Exponential	-20563.23	41134.23	7	-395.35	796.08	7
Weibull	-20254.73	40525.02	5	-389.98	790.70	6
Normal	-24127.48	48270.51	12	-612.16	1235.07	12
Logistic	-22261.65	44538.85	10	-521.02	1052.78	11
Skew-logistic	-21270.53	43464.40	9	-462.27	940.66	9
Skew-normal	-22592.65	45208.64	11	-490.06	996.24	10
Skew-t	-20039.17	40109.44	4	-339.10	699.69	3
Hyperbolic	-20445.57	40922.26	6	-377.72	776.93	5

It may be interesting to notice that the estimated value of the degree of contamination parameter, in the Frebiloss data set, is $\hat{\eta} = 2.64$ for the LN-DUC model and $\hat{\eta} = 11.36$ for the UG-DUC one. Similarly, in the Swefire data set, it is $\hat{\eta} = 8.73$ for the LN-DUC model and $\hat{\eta} = 27.95$ for the UG-DUC one. This means that in both data sets there are atypical values that contaminate the conditional distribution, but a higher level of contamination is detected in the Swefire data set.

4.6.3 Risk measures analysis

4.6.3.1 Frebiloss data set

Table 4.6 reports the empirical and the estimated VaR values of all the competing models and approaches. Again, a ranking is introduced to simplify the reading, but this time it is based on the absolute value of the percentage of variation with respect to the empirical VaR; the lower the difference, the better the position in the ranking. The corresponding backtesting results are also provided in the last two columns.

When $c = 0.95$, the LN-DUC model is ranked first, extremely close to the empirical value. On the contrary, some models seem to provide better VaR values than the UG-DUC model. However, if the corresponding backtesting p-values are examined, only the LN-DUC model does not lead to the rejection of the null hypothesis.

In contrast with the previous case, when the $c = 0.99$, the best model is the UG-DUC, providing a good estimate of the VaR. The LN-DUC model falls in sixth position, exceeding the empirical value. Even if the proposed models have acceptable p-values, four benchmark models have good values too, among which there are the Pareto t-estimator and the PORT-MO$_p$ (with estimated tuning parameters $k = 47$, $p = 0.19$ and $s = 0.2$).

TABLE 4.6: Frebiloss: $\text{VaR}_{95}(X)$ and $\text{VaR}_{99}(X)$ with corresponding ranking and backtesting p-values.

Model	Value at Risk				p-value	
	$\text{VaR}_{95}(X)$	Rank	$\text{VaR}_{99}(X)$	Rank	$\text{VaR}_{95}(X)$	$\text{VaR}_{99}(X)$
Empirical	7675.81		18293.88			
UG-DUC	9454.65	8	18931.39	1	0.00	0.55
LN-DUC	7787.84	1	21810.33	6	0.90	0.05
UG	6074.52	7	9338.02	11	0.00	0.00
LN	6189.57	5	14304.85	7	0.00	0.00
Exponential	6074.55	6	9338.07	10	0.00	0.00
Weibull	6893.35	2	12358.04	9	0.01	0.00
Normal	11792.92	13	15838.83	4	0.00	0.16
Logistic	5097.50	11	7257.37	13	0.00	0.00
Skew-Logistic	4930.70	12	7093.86	14	0.00	0.00
Skew-Normal	12385.83	14	16246.27	3	0.00	0.31
Skew-t	5809.08	10	12804.28	8	0.00	0.00
Hyperbolic	5879.79	9	8986.22	12	0.00	0.00
Pareto(t-score)	6327.10	4	15284.05	5	0.00	0.08
PORT-MO$_p$	8772.87	3	19041.06	2	0.03	0.55

Table 4.7 shows the empirical and estimated values of the TVaR along with the corresponding backtesting results. The rankings are computed as in Table 4.6. According to the backtesting procedure, when $c = 0.95$, only the LN-DUC provides a p-value that does not lead to the rejection of the null hypothesis. On the contrary, when $c = 0.99$, both the LN-DUC and UG-DUC model pass the backtest, along with the Pareto t-estimator. Overall, the dichotomous unimodal compound models seem to suggest a good description of the tail behavior of the empirical distribution.

4.6.3.2 Swefire data set

Table 4.8 reports the empirical VaR as well as the VaR for all the considered models and approaches, along with the corresponding backtesting results. When $c = 0.95$, about the dichotomous unimodal compound models, only the LN-DUC performs well, even if by the analysis of the p-values, all models except the skew-t seem able to produce acceptable estimates of the empirical VaR (the estimated tuning parameters for the PORT-MO$_p$ are $k = 11$, $p = 4.24$ and $s = 0.9$). It is possible to appreciate the real difference between the dichotomous unimodal compound models and the benchmark ones, moving deeper in the right tail of the empirical distribution. In fact, when $c = 0.99$ the UG-DUC and LN-DUC models are the best, as also confirmed by the p-values of the backtest.

Table 4.9 reports the empirical and estimated values of the TVaR, with always the same ranking mechanism. Also in this case, the LN-DUC is the best model, as also

TABLE 4.7: Frebiloss: $\text{TVaR}_{95}(X)$ and $\text{TVaR}_{99}(X)$ with corresponding ranking and backtesting p-values.

Model	Tail Value at Risk				min p-value	
	$\text{VaR}_{95}(X)$	Rank	$\text{VaR}_{99}(X)$	Rank	$\text{VaR}_{95}(X)$	$\text{VaR}_{99}(X)$
Empirical	17062.59		38135.05			
UG-DUC	15341.69	2	24789.96	3	0.00	0.07
LN-DUC	17872.90	1	40499.24	1	0.09	0.05
UG	8102.25	9	11365.75	9	0.00	0.00
LN	11822.42	6	23777.22	5	0.00	0.00
Exponential	8102.29	10	11365.80	10	0.00	0.00
Weibull	10335.17	8	16253.92	8	0.00	0.00
Normal	14273.68	4	17850.61	7	0.00	0.00
Logistic	6393.65	12	8508.83	12	0.00	0.00
Skew-logistic	6274.92	13	8422.52	13	0.00	0.00
Skew-normal	14754.25	3	18227.81	6	0.00	0.00
Skew-t	11297.15	7	24449.78	4	0.00	0.00
Hyperbolic	7809.57	11	10914.98	11	0.00	0.00
Pareto(t-score)	13176.71	5	29258.68	2	0.00	0.05

TABLE 4.8: Swefire: $\text{VaR}_{95}(X)$ and $\text{VaR}_{99}(X)$ with corresponding ranking and backtesting p-values.

Model	Value at Risk				p-value	
	$\text{VaR}_{95}(X)$	Rank	$\text{VaR}_{99}(X)$	Rank	$\text{VaR}_{95}(X)$	$\text{VaR}_{99}(X)$
Empirical	7.84		19.93			
UG-DUC	9.45	10	16.47	2	0.80	0.10
LN-DUC	8.43	3	19.40	1	0.94	0.58
UG	6.93	6	10.65	9	0.50	0.00
LN	6.14	11	11.90	7	0.50	0.01
Exponential	6.93	5	10.65	8	0.50	0.00
Weibull	7.43	1	12.20	5	0.94	0.01
Normal	9.18	7	12.02	6	0.81	0.01
Logistic	5.38	13	7.54	14	0.12	0.00
Skew-logistic	5.44	12	7.78	13	0.12	0.00
Skew-normal	9.36	8	12.27	4	0.81	0.01
Skew-t	4.33	14	9.33	11	0.00	0.00
Hyperbolic	6.26	9	9.44	10	0.50	0.00
Pareto(t-score)	7.01	4	13.62	3	0.70	0.03
PORT-MO_p	7.37	2	8.60	12	0.94	0.00

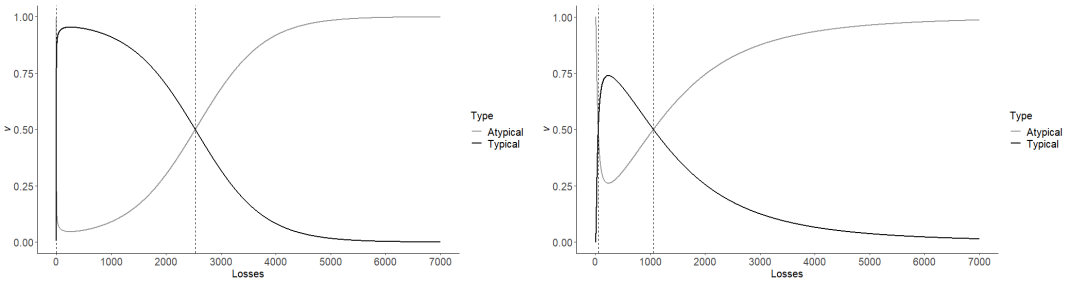
confirmed by the backtest results. In an opposite way with respect to the previous application, the UG-DUC seems to provide a good estimate of the TVaR only when $c = 0.95$. Overall, the TVaR values from both LN-DUC and UG-DUC models are the only ones being not rejected for at least one probability level c.

TABLE 4.9: Swefire: $\text{TVaR}_{95}(X)$ and $\text{TVaR}_{99}(X)$ with corresponding ranking and backtesting p-values.

Model	Tail Value at Risk				min p-value	
	$\text{TVaR}_{95}(X)$	Rank	$\text{TVaR}_{99}(X)$	Rank	$\text{TVaR}_{95}(X)$	$\text{TVaR}_{99}(X)$
Empirical	17.39		28.37			
UG-DUC	13.81	2	20.77	2	0.20	0.03
LN-DUC	15.90	1	31.23	1	0.20	0.30
UG	9.25	9	12.97	10	0.00	0.00
LN	9.97	7	17.45	5	0.01	0.00
Exponential	9.25	8	12.97	9	0.00	0.00
Weibull	10.41	6	15.34	6	0.01	0.00
Normal	10.92	5	13.43	8	0.03	0.00
Logistic	6.73	13	8.86	13	0.00	0.00
Skew-logistic	6.90	12	9.22	12	0.00	0.00
Skew-normal	11.14	4	13.76	7	0.03	0.00
Skew-t	8.35	10	18.20	4	0.00	0.00
Hyperbolic	8.23	11	11.41	11	0.00	0.00
Pareto(t-score)	11.46	3	20.27	3	0.03	0.01

4.6.4 Comments on typical and atypical losses

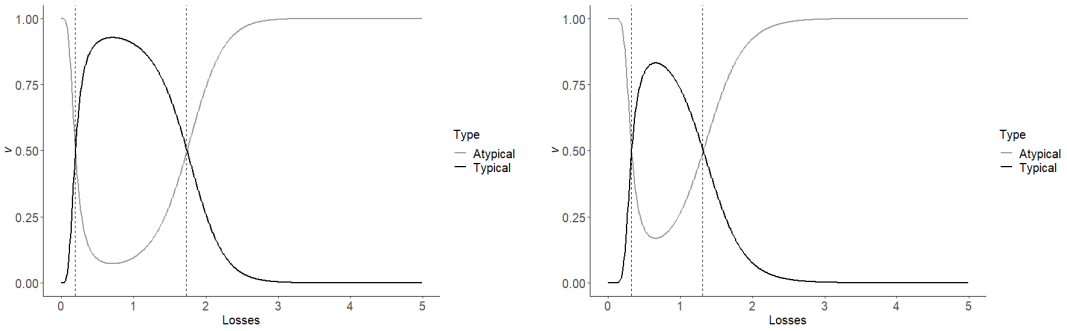
Finally, Figure 4.8 illustrates the estimated probabilities to be typical or atypical points, as defined in (4.4), according to the UG-DUC and LN-DUC models for the Frebiloss data set. The same is done in Figure 4.9 for the Swefire data set. For a better graphical visualization, the attention is focused to the $(0, 7000]$ and $(0, 5]$ intervals for the two data sets, respectively. As discussed in Section 4.2.1, these probabilities coincide in the intersection points between the discriminant functions, delimiting the typical and the atypical regions (marked with the vertical dashed lines in Figure 4.8 and Figure 4.9).



(a) UG-DUC (b) LN-DUC

FIGURE 4.8: Frebiloss: estimated probabilities to be typical or atypical points by the UG-DUC (a) and LN-DUC models (b). The corresponding typical and atypical regions are separated by the vertical dashed lines.

In detail, when the Frebiloss data set is considered, the intersection points are in $x_1 = 0.12$ and $x_2 = 2531.82$ for the UG-DUC model and in $x_1 = 52.21$ and $x_2 = 1057.17$ for the LN-DUC model. Since x_1 is lower than the minimum observed (100.29; see Table 4.4) for both models, none of the points is considered atypically



(a) UG-DUC (b) LN-DUC

FIGURE 4.9: Swefire: estimated probabilities to be typical or atypical points by the UG-DUC (a) and LN-DUC models (b). The corresponding typical and atypical regions are separated by the vertical dashed lines.

low. However, a different classification of typical and atypically high data is produced in the interval $(1057.17, 2531.82)$, containing the 21.49% of the data. When a classification of the data is of interest, a possible way to choose among the two models could be to rely on the model producing the best overall fitting according to the information criteria, i.e. in this case the LN-DUC one. If it is of particular interest the set of atypically high data, another option could be to choose the classification produced by the model with the best fit in the right tail. Again the LN-DUC model seems to be the best option even if, when $\text{VaR}_{99}(X)$ is selected, it performs worse than the UG-DUC one.

When the Swefire data set is analyzed, the intersection points delimiting the typical and the atypical regions are in $x_1 = 0.19$ and $x_2 = 1.73$ for the UG-DUC model and in $x_1 = 0.33$ and $x_2 = 1.31$ for the LN-DUC one. The differences between the partitions produced by the two models concern the interval $(0.19, 0.33)$, containing the 1.39% of the data, and the interval $(1.31, 1.73)$, containing the 10.23% of the data. Therefore, they are more similar to each other than they are in the Frebiloss data set. Here, both from a global and a right tail goodness of fit point of view, the best model is the LN-DUC one, and then its partition of the data should be preferred.

In any case, regarding of the data set considered, it is possible to see both models agree always more in defining the losses as atypically high, as their values become larger and larger.

4.7 Conclusions

Several models have been suggested in the actuarial literature for insurance loss data. However, losses distributions show characteristics that are hardly compatible with the choice of fitting a single parametric distribution, calling for more flexible models.

In this work, a general dichotomous unimodal compound model has been introduced by compounding a unimodal hump-shaped conditional distribution on a positive support with a convenient mixing dichotomous distribution. As a result, the density of the proposed model is: defined on positive support, unimodal hump-shaped, positively skewed, and with tails heavier than those of the conditional distribution. For illustrative purposes, two hump-shaped distributions have been considered, i.e. the log-normal and the unimodal gamma. The resulting models have

been firstly analyzed in a sensitivity study, and therefore fitted to two real data sets, along with several benchmark competitors (among which there are the t-score estimator and the PORT-MO$_p$ approach). A double-bootstrap likelihood-ratio test, the BIC and two well-known risk measures, VaR and TVaR, which typically focus on the right tail of the distribution, have been used for comparisons. The main findings are: 1) by using a dichotomous unimodal compound model the ML estimator of the parameters of the conditional distribution is more robust compared to the case in which the simple conditional distribution is fitted; 2) the proposed models behave very well both in terms of global and right tail fit of the data; 3) by using a dichotomous unimodal compound model it is possible to detect typical and atypical losses, with respect to the conditional distribution, and to define the corresponding typical and atypical regions on the x-axis.

Chapter 5

Two new matrix-variate distributions with application in model-based clustering [1]

5.1 Introduction

As mentioned in Section 2.5, over the last years there has been an increased interest in the analysis of matrix-variate data via mixture models. Originally proposed by Viroli (2011a), the matrix-variate normal mixtures (MVN-Ms) are the first example of model-based clustering in the context of matrix-variate data. The same author generalizes the MVN-Ms in a Bayesian framework in Viroli (2011b). However, for many real phenomena, the tails of the MVN distribution are lighter than required, with a direct effect on the corresponding mixture model. This is often due to the presence of mild outliers (Ritter, 2015), i.e. observations that produce an overall distribution that is too heavy-tailed to be modeled by MVN-Ms. The most commonly used solution for managing this type of situations consists in relaxing the normality assumption of the mixture components. Unfortunately, and differently from the multivariate literature (see, e.g. Peel and McLachlan, 2000 and Dang *et al.*, 2015), only finite mixtures of matrix-variate t (MVt-Ms) distributions have been proposed, as three-way elliptical heavy-tailed model, to cope with this issue (Doğru *et al.*, 2016). Therefore, in this work two new heavy-tailed matrix-variate distributions are introduced, namely the matrix-variate shifted exponential normal (MVSEN) and the matrix-variate tail-inflated normal (MVTIN), generalization of the corresponding multivariate distributions recently presented in Punzo and Bagnato (2020a) and Punzo and Bagnato (2020b), respectively. As explained in Section 5.2, to define the MVSEN and MVTIN distributions, the well-known normal scale mixture model is used. Then, these distributions are used for model-based clustering via mixture models. Because of their heavier-than-normal tails, these models are able to cope with clusters having potential mild outliers in a proper way. This implies a better fit of the data, and may avoid the disruption of the true underlying grouping structure.

Section 5.3 examines parameter estimation, that is carried out by means of different extensions of the EM algorithm. Simulated data analyses involving computational time and parameter recovery of the aforementioned algorithms are discussed in Section 5.4. Additionally, the fitting and clustering performances of the proposed

[1] This work is based on the following publication: 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. The current manuscript is a combined effort of the authors. However, Tomarchio S.D. contributed in conceptualization, implementation, data elaboration and writing–original draft preparation; Punzo A. and Bagnato, L. contributed in conceptualization and supervision.

and some competing models, in presence of outlying matrices, are also therein evaluated. For similar comparative purposes, two real data applications are analyzed in Section 5.5. Lastly, Section 5.6 concludes.

5.2 Methodology

5.2.1 The matrix-variate normal scale mixture model

In a matrix-variate framework, the normal scale mixture model introduced in Section 2.2 has PDF

$$f_{\text{MVNSM}}(\mathbf{X};\boldsymbol{\Omega},\boldsymbol{\nu}) = \int_0^\infty f_{\text{MVN}}(\mathbf{X};\mathbf{M},\boldsymbol{\Sigma}/w,\boldsymbol{\Psi})h(w;\boldsymbol{\nu})dw. \tag{5.1}$$

where MVNSM stands for matrix-variate normal scale mixture. By looking at the moments of the MVNSM, it is possible to see the impact of the mixing distribution in terms of deviation from normality. Indeed, the mean, the covariance matrix and the kurtosis of the MVNSM are

$$\text{E}(\mathbf{X}) = \mathbf{M}, \tag{5.2}$$

$$\text{Var}(\mathbf{X}) = a(\boldsymbol{\nu})\,\boldsymbol{\Psi}\otimes\boldsymbol{\Sigma}, \tag{5.3}$$

$$\text{Kurt}(\mathbf{X}) = b(\boldsymbol{\nu})\,pr(pr+2), \tag{5.4}$$

where $pr(pr+2)$ is the kurtosis of the nested MVN distribution, with p and r indicating the number of rows and columns of $\mathbf{X}$, respectively, $a(\boldsymbol{\nu}) = \text{E}(1/W)$ and $b(\boldsymbol{\nu}) = \text{E}\left[(1/W)^2\right] / \left[\text{E}(1/W)\right]^2$, with W having PDF $h(w;\boldsymbol{\nu})$, are the multiplicative factors governing the deviation from the nested MVN distribution. Notice that, in the matrix-variate literature, $\text{Var}(\mathbf{X})$ and $\text{Kurt}(\mathbf{X})$ are computed on the vectorized data (Sánchez-Manzano *et al.*, 2002; Gupta *et al.*, 2013). Results in (5.2) and (5.3) can be found in Gupta *et al.* (2013), whereas (5.4) is a generalization of the results given in Punzo and Bagnato (2020b), provided that the well-known Mardia's measure of kurtosis (Mardia, 1970) is considered.

Since $\text{E}\left[(1/W)^2\right] \geq \left[\text{E}(1/W)\right]^2$, the excess of kurtosis (with respect to the MVN distribution) is non-negative, with the equality holding only when $W \equiv 1$. Then, apart from this limit case, the resulting distribution is leptokurtic and $\boldsymbol{\nu}$ can be meant as the tailedness parameter of the MVNSM model. An example of distribution belonging to the MVNSM family is the matrix-variate t distribution, that is obtained by considering a convenient gamma as mixing distribution (Doğru *et al.*, 2016).

5.2.2 The matrix-variate shifted exponential normal distribution

Definition 5.2.1. A $p \times r$ random matrix $\mathbf{X}$ is said to have a matrix-variate shifted exponential normal (MVSEN) distribution with $p \times r$ mean matrix $\mathbf{M}$, $p \times p$ and $r \times r$ scale matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$, and tailedness parameter $\theta > 0$, if its PDF is given by

$$f_{\text{MVSEN}}(\mathbf{X};\boldsymbol{\phi}) = \frac{\theta\exp(\theta)}{(2\pi)^{\frac{pr}{2}}|\boldsymbol{\Sigma}|^{\frac{r}{2}}|\boldsymbol{\Psi}|^{\frac{p}{2}}}\varphi_{\frac{pr}{2}}\left(\frac{\delta(\mathbf{X};\boldsymbol{\Omega})}{2}+\theta\right), \tag{5.5}$$

where $\varphi_m(z)$ is the Misra function (Misra, 1940), generalization of the generalized exponential integral function (Abramowitz and Stegun, 1965), and $\boldsymbol{\phi}$ contains all the parameters of the PDF.

The PDF in (5.5) can be obtained from the PDF in (5.1) by considering W as a shifted exponential random variable defined on the $(1,\infty)$ interval, with PDF $h_{\text{SE}}(w;\nu) = \theta \exp\left[-\theta\left(w-1\right)\right]$, as mixing distribution. Indeed, the PDF in (5.5) can be written as

$$f_{\text{MVSEN}}(\mathbf{X};\boldsymbol{\phi}) = \frac{\theta\exp(\theta)}{(2\pi)^{\frac{pr}{2}}|\boldsymbol{\Sigma}|^{\frac{r}{2}}|\boldsymbol{\Psi}|^{\frac{p}{2}}}\int_1^\infty w^{\frac{pr}{2}}\exp\left\{-w\left[\frac{\delta(\mathbf{X};\boldsymbol{\Omega})}{2}+\theta\right]\right\}dw.$$

By noting that

$$\begin{aligned}&\int_1^\infty w^{\frac{pr}{2}}\exp\left\{-w\left[\frac{\delta(\mathbf{X};\boldsymbol{\Omega})}{2}+\theta\right]\right\}dw = \\ &= \left[\frac{\delta(\mathbf{X};\boldsymbol{\Omega})}{2}+\theta\right]^{-\left(\frac{pr}{2}+1\right)}\Gamma\left(\frac{pr}{2}+1,\frac{\delta(\mathbf{X};\boldsymbol{\Omega})}{2}+\theta\right) = \varphi_{pr}\left(\frac{\delta(\mathbf{X};\boldsymbol{\Omega})}{2}+\theta\right),\end{aligned}$$

where $\Gamma(c,z)$ denotes the upper incomplete gamma function (Abramowitz and Stegun, 1965), the PDF in (5.5) is obtained.

A hierarchical representation of the MVSEN distribution can be given as

1. $W \sim \mathcal{SE}\left(\theta\right)$,
2. $\mathbf{X}|W = w \sim \mathcal{N}_{p\times r}(\mathbf{M},\boldsymbol{\Sigma}/w,\boldsymbol{\Psi})$,

where $\mathcal{SE}\left(\theta\right)$ is a shifted exponential distribution on $(1,\infty)$, and $\mathcal{N}_{p\times r}(\mathbf{M},\boldsymbol{\Sigma}/w,\boldsymbol{\Psi})$ is a matrix-variate normal distribution. This alternative way to see the MVSEN distribution is useful for random data generation and for the implementation of the EM-based algorithm discussed in Section 5.3.

When the considered shifted exponential is chosen as mixing distribution, the multiplicative factors to be inserted in (5.3) and (5.4) are

$$a\left(\theta\right) = \theta\exp\left(\theta\right)\varphi_{-1}(\theta), \tag{5.6}$$

$$b\left(\theta\right) = \frac{\theta\left[1-a\left(\theta\right)\right]}{a\left(\theta\right)^2}, \tag{5.7}$$

and their graphical representation, for growing values of θ, is drawn in Figure 5.1 (Punzo and Bagnato, 2020a). It is easy to see that $a\left(\theta\right)$ is a smoothly increasing function of θ and, because of its values, $\text{Var}(\mathbf{X})$ is a deflated version of $\boldsymbol{\Psi}\otimes\boldsymbol{\Sigma}$. However, when $\theta\to\infty$, then $a\left(\theta\right)\to 1$ and $\text{Var}(\mathbf{X})\to\boldsymbol{\Psi}\otimes\boldsymbol{\Sigma}$. On the contrary $b\left(\theta\right)$ is decreasing in θ and, because of its values, $\text{Kurt}(\mathbf{X})$ is greater than $pr\left(pr+2\right)$, implying that the MVSEN distribution is leptokurtic. However, when $\theta\to\infty$, then $b\left(\theta\right)\to 1$ and $\text{Kurt}(\mathbf{X})\to pr\left(pr+2\right)$. Notice that, the decrease of $b\left(\theta\right)$ is sudden when θ is close to 0, meaning that only low values of θ allow to reach high levels of kurtosis.

5.2.3 The matrix-variate tail-inflated normal distribution

Definition 5.2.2. A $p\times r$ random matrix $\mathbf{X}$ is said to have a matrix-variate tail-inflated normal (MVTIN) distribution with $p\times r$ mean matrix $\mathbf{M}$, $p\times p$ and $r\times r$ scale matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$, tailedness parameter $\theta\in(0,1)$, if its PDF is given by

$$f_{\text{MVTIN}}(\mathbf{X};\boldsymbol{\phi}) = \frac{2(\pi)^{-\frac{pr}{2}}|\boldsymbol{\Sigma}|^{-\frac{r}{2}}|\boldsymbol{\Psi}|^{-\frac{p}{2}}}{\theta\delta(\mathbf{X};\boldsymbol{\Omega})^{\frac{pr}{2}+1}}\left[\Gamma\left(\frac{pr}{2}+1,(1-\theta)\frac{\delta(\mathbf{X};\boldsymbol{\Omega})}{2}\right)-\Gamma\left(\frac{pr}{2}+1,\frac{\delta(\mathbf{X};\boldsymbol{\Omega})}{2}\right)\right], \tag{5.8}$$

where $\boldsymbol{\phi}$ contains all the parameters of the density.

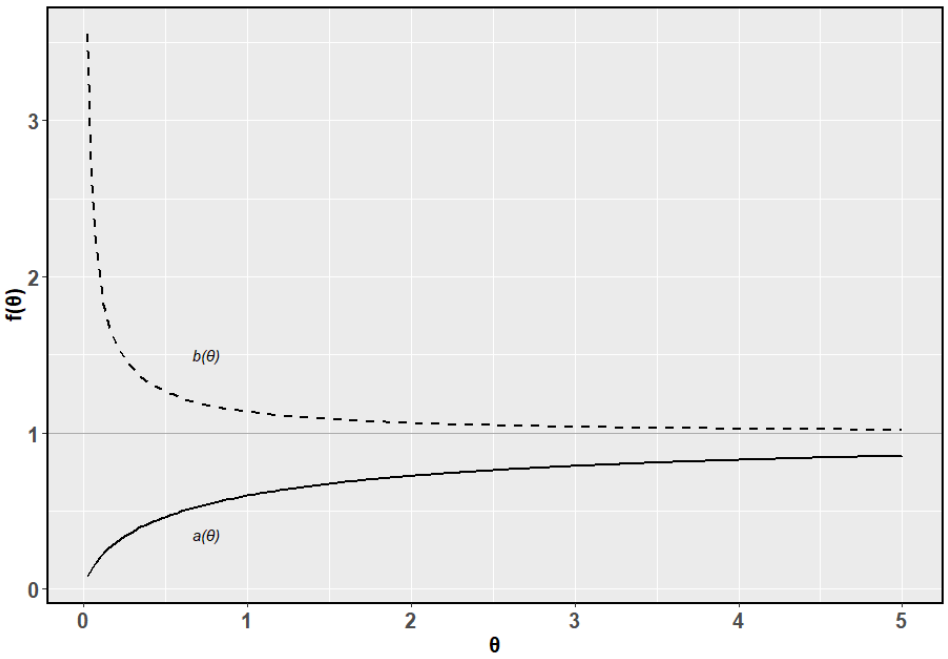


FIGURE 5.1: Multiplicative factors $a(\theta)$ (solid line) and $b(\theta)$ (dashed line) for the MVSEN distribution.

The PDF in (5.8) can be obtained from the PDF in (5.1) by considering W as a uniform mixing random variable defined on the $(1-\theta, 1)$ interval, with PDF $h_{\mathrm{U}}(w;\boldsymbol{\nu}) = 1/\theta$. Indeed, the PDF in (5.8) can be written as

$$f_{\mathrm{MVTIN}}(\mathbf{X};\boldsymbol{\phi}) = \frac{(2\pi)^{-\frac{pr}{2}}|\boldsymbol{\Sigma}|^{-\frac{r}{2}}|\boldsymbol{\Psi}|^{-\frac{p}{2}}}{\theta}\int_{1-\theta}^{1} w^{\frac{pr}{2}}\exp\left\{-\frac{w}{2}\delta(\mathbf{X};\boldsymbol{\Omega})\right\}dw.$$

By noting that

$$\begin{aligned}&\int_{1-\theta}^{1} w^{\frac{pr}{2}}\exp\left\{-\frac{w}{2}\delta(\mathbf{X};\boldsymbol{\Omega})\right\}dw = \\ &= \left[\frac{2}{\delta(\mathbf{X};\boldsymbol{\Omega})}\right]^{\left(\frac{pr}{2}+1\right)}\left[\Gamma\left(\frac{pr}{2}+1, (1-\theta)\frac{\delta(\mathbf{X};\boldsymbol{\Omega})}{2}\right) - \Gamma\left(\frac{pr}{2}+1, \frac{\delta(\mathbf{X};\boldsymbol{\Omega})}{2}\right)\right],\end{aligned}$$

the PDF in (5.8) is obtained.

For the same purposes of Section 5.2.2, a hierarchical representation of the MVTIN distribution is given. Specifically,

1. $W \sim \mathcal{U}(1-\theta, 1)$,
2. $\mathbf{X}|W = w \sim \mathcal{N}_{p\times r}(\mathbf{M}, \boldsymbol{\Sigma}/w, \boldsymbol{\Psi})$,

where $\mathcal{U}(1-\theta, 1)$ denotes a uniform distribution on $(1-\theta, 1)$.

When the considered uniform is chosen as mixing distribution, the multiplicative factors to be inserted in (5.3) and (5.4) are

$$a(\theta) = -\frac{\ln(1-\theta)}{\theta}, \tag{5.9}$$

$$b(\theta) = \frac{\theta^2}{(1-\theta)\ln^2(1-\theta)}, \tag{5.10}$$

and their graphical representation, for growing values of θ, is drawn in Figure 5.2 (Punzo and Bagnato, 2020b). Similarly to the MVSEN distribution, $a(\theta)$ is a smoothly

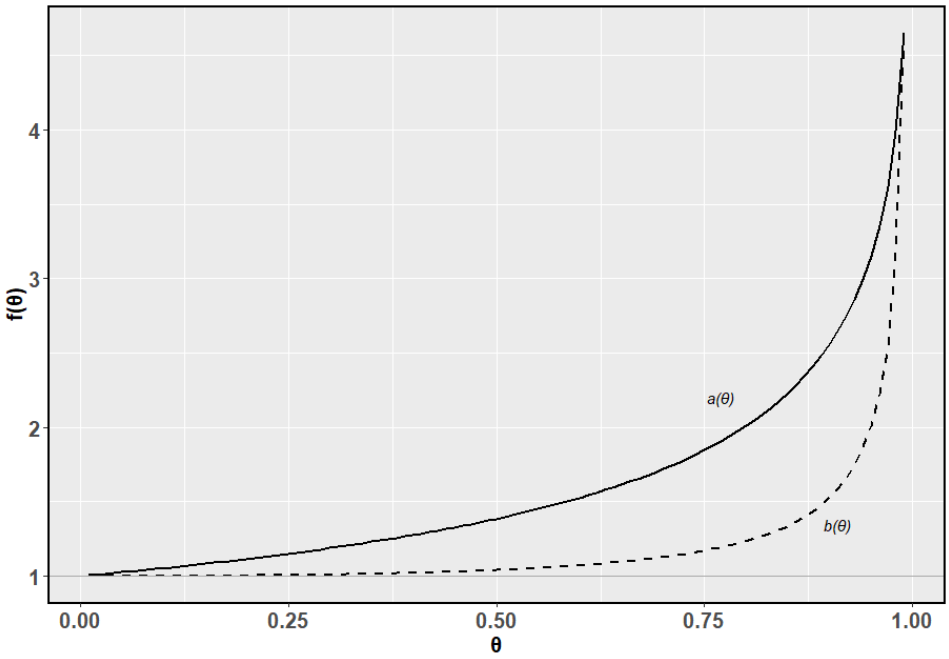


FIGURE 5.2: Multiplicative factors $a(\theta)$ (solid line) and $b(\theta)$ (dashed line) for the MVTIN distribution.

increasing function of θ and, because of its values, $\text{Var}(\mathbf{X})$ is an inflated version of $\boldsymbol{\Psi} \otimes \boldsymbol{\Sigma}$. However, when $\theta \to 0$, then $a(\theta) \to 1$ and $\text{Var}(\mathbf{X}) \to \boldsymbol{\Psi} \otimes \boldsymbol{\Sigma}$. Here, $b(\theta)$ is increasing in θ and, because of its values, as for the MVSEN distribution, $\text{Kurt}(\mathbf{X})$ is greater than $pr(pr+2)$, implying leptokurtosis. However, when $\theta \to 0$, then $b(\theta) \to 1$ and $\text{Kurt}(\mathbf{X}) \to pr(pr+2)$. Notice that $b(\theta)$ suddenly increases when θ is close to 1, meaning that we need high θ-values to reach relevant levels of kurtosis.

5.3 Parameter estimation

When the PDFs (5.5) or (5.8) are chosen as component in model in (2.1), the corresponding mixtures are obtained; these models will be denoted herein as MVSEN-Ms and MVTIN-Ms, respectively. Here, the complete-data are $\mathcal{S}_c = \{\mathbf{X}_i, \mathbf{z}_i, w_i\}_{i=1}^N$. By saying that $\mathbf{Z}_i$, random counterpart of $\mathbf{z}_i$, is distributed according to a multinomial distribution consisting of one draw from K categories with probabilities $\boldsymbol{\pi} = \{\pi_k\}_{k=1}^K$, say $\mathcal{M}_K(\boldsymbol{\pi})$, the hierarchical representations introduced in Section 5.2 for the single distributions can be extended, to the mixture modeling framework, as

$$\begin{aligned} \mathbf{Z}_i &\sim \mathcal{M}_K(\boldsymbol{\pi}) \\ W_i|z_{ik}=1 &\sim \begin{cases} \mathcal{SE}(\theta_k) & \text{for MVSEN-Ms} \\ \mathcal{U}(1-\theta_k, 1) & \text{for MVTIN-Ms} \end{cases} \\ \mathbf{X}_i|W_i=w_i, z_{ik}=1 &\sim \mathcal{N}_{p\times r}(\mathbf{M}_k, \boldsymbol{\Sigma}_k/w_i, \boldsymbol{\Psi}_k). \end{aligned} \tag{5.11}$$

Unfortunately, the EM algorithm cannot be directly implemented for MVSEN-Ms and MVTIN-Ms, since the M-steps of both models present some issues. In detail, there is no closed form solution for the covariance matrices, considering that one

of the two depends on the value of the other at the previous iteration, inheriting this characteristic from the MVN distribution (Dutilleul, 1999). Therefore, this is not a specific problem of MVSEN-Ms and MVTIN-Ms, but of all the distributions that can be obtained via the MVNSM model in (4.1). Additionally, for MVTIN-Ms, the equation involving the tailedness parameter is not well-defined, since the update for θ tends to 0 as the number of iterations of the algorithm grows. This implies that regardless of its true but unknown value, the EM algorithm fails to converge. For these reasons, the modified versions of the EM algorithm mentioned in Section 2.3 are herein implemented. Specifically, an ECM algorithm is implemented for MVSEN-Ms in Section 5.3.1. On the contrary, in Section 5.3.2 two alternatives are discussed for MVTIN-Ms: an ECME algorithm and an AECM algorithm.

In any case, for the implementation of all these algorithms, it is convenient to use the hierarchical representation in (5.11). Indeed, the complete-data likelihood function $\mathcal{L}_c(\boldsymbol{\Theta}; \mathcal{S}_c)$ can be factored as

$$\mathcal{L}_c(\boldsymbol{\Theta}; \mathcal{S}_c) = \prod_{i=1}^{N}\prod_{k=1}^{K}\left[\pi_k f_{\text{MVN}}(\mathbf{X}_i; \mathbf{M}_k, \boldsymbol{\Sigma}_k/w_i, \boldsymbol{\Psi}_k) h(w_i; \boldsymbol{\nu}_k)\right]^{z_{ik}},$$

where $h(w_i; \boldsymbol{\nu}_k)$ is model dependent. Accordingly, the complete-data log-likelihood function can be written as

$$\ell_c(\boldsymbol{\Theta}; \mathcal{S}_c) = \ln\left[\ell_c(\boldsymbol{\Theta}; \mathcal{S}_c)\right] = \ell_{1c}(\boldsymbol{\pi}; \mathcal{S}_c) + \ell_{2c}(\boldsymbol{\Xi}; \mathcal{S}_c) + \ell_{3c}(\boldsymbol{\nu}; \mathcal{S}_c), \tag{5.12}$$

where

$$\ell_{1c}(\boldsymbol{\pi}; \mathcal{S}_c) = \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik} \ln(\pi_k), \tag{5.13}$$

$$\ell_{2c}(\boldsymbol{\Xi}; \mathcal{S}_c) = \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}\left[-\frac{pr}{2}\ln(2\pi_k) + \frac{pr}{2}\ln(w_{ik}) - \frac{r}{2}\ln|\boldsymbol{\Sigma}_k| - \frac{p}{2}\ln|\boldsymbol{\Psi}_k| - \frac{w_{ik}\delta(\mathbf{X};\boldsymbol{\Omega}_k)}{2}\right], \tag{5.14}$$

with $\boldsymbol{\Xi} = \{\mathbf{M}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\Psi}_k\}_{k=1}^{K}$ and $\ell_{3c}(\boldsymbol{\nu}; \mathcal{S}_c)$, with $\boldsymbol{\nu} = \{\theta_k\}_{k=1}^{K}$, is different according to the considered MVNSM model; see Section 5.3.1 for MSEN-Ms and Section 5.3.2 for MTIN-Ms.

5.3.1 An ECM-algorithm for MVSEN-Ms

When MVSEN-Ms are considered, the last term in (5.12) is equal to

$$\ell_{3c}(\boldsymbol{\nu}; \mathcal{S}_c) = \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}\left[\ln(\theta_k) - \theta_g(w_{ik} - 1)\right]. \tag{5.15}$$

Then, the ECM algorithm proceeds as follows (we recall that the dots refer to the algorithm iterations, as explained in Section 2.3).

E-Step At the E-step, it is necessary to compute

$$\ddot{z}_{ik} := \mathrm{E}_{\dot{\boldsymbol{\Theta}}}(Z_{ik}|\mathbf{X}_i) = \frac{\dot{\pi}_k f_{\text{MVSEN}}(\mathbf{X}_i; \dot{\boldsymbol{\phi}}_k)}{\sum_{h=1}^{K}\dot{\pi}_h f_{\text{MVSEN}}(\mathbf{X}_i; \dot{\boldsymbol{\phi}}_h)}, \tag{5.16}$$

which is the posterior probability that the unlabeled observation $\mathbf{X}_i$ belongs to the kth component of the mixture, and

$$\ddot{w}_{ik} := \mathrm{E}_{\dot{\boldsymbol{\Theta}}}\left(W_{ik}|\mathbf{X}_i, \mathbf{z}_i\right) = \frac{\varphi_{\frac{pr}{2}+1}\left(\frac{\delta(\mathbf{X}_i;\dot{\boldsymbol{\Omega}}_k)}{2} + \dot{\theta}_k\right)}{\varphi_{\frac{pr}{2}}\left(\frac{\delta(\mathbf{X}_i;\dot{\boldsymbol{\Omega}}_k)}{2} + \dot{\theta}_k\right)}, \tag{5.17}$$

which corresponds to the expected value of a left-truncated gamma distribution with parameters $(pr/2)+1$ and $[\delta(\mathbf{X}_i;\boldsymbol{\Omega}_k)/2]+\theta_k$, on the interval $(1,\infty)$; for details see Punzo and Bagnato, 2020a. Finally, there is no need to compute the expectation of $\ln(w_{ik})$, since it is not related to the parameters.

Now, consider $\boldsymbol{\Theta}_1 = \{\pi_k, \mathbf{M}_k, \boldsymbol{\Sigma}_k, \theta_k\}_{k=1}^K$ and $\boldsymbol{\Theta}_2 = \{\boldsymbol{\Psi}_k\}_{k=1}^K$.

CM-Step 1 At the first CM-step, by fixing $\boldsymbol{\Theta}_2$ at $\dot{\boldsymbol{\Theta}}_2$, it is possible to obtain

$$\ddot{\pi}_k = \frac{\sum_{i=1}^N \ddot{z}_{ik}}{N}, \tag{5.18}$$

$$\ddot{\mathbf{M}}_k = \frac{\sum_{i=1}^N \ddot{z}_{ik}\ddot{w}_{ik}\mathbf{X}_i}{\sum_{i=1}^N \ddot{z}_{ik}\ddot{w}_{ik}}, \tag{5.19}$$

$$\ddot{\boldsymbol{\Sigma}}_k = \frac{\sum_{i=1}^N \ddot{z}_{ik}\ddot{w}_{ik}\left(\mathbf{X}_i - \ddot{\mathbf{M}}_k\right)\left(\dot{\boldsymbol{\Psi}}_k\right)^{-1}\left(\mathbf{X}_i - \ddot{\mathbf{M}}_k\right)'}{r\sum_{i=1}^N \ddot{z}_{ik}}, \tag{5.20}$$

$$\ddot{\theta}_k = \frac{\sum_{i=1}^N \ddot{z}_{ik}}{\sum_{i=1}^N \ddot{z}_{ik}\left(\ddot{w}_{ik} - 1\right)}. \tag{5.21}$$

CM-Step 2 At the second CM-step, keeping fixed $\boldsymbol{\Theta}_1$ at $\ddot{\boldsymbol{\Theta}}_1$, it is possible to obtain

$$\ddot{\boldsymbol{\Psi}}_k = \frac{\sum_{i=1}^N \ddot{z}_{ik}\ddot{w}_{ik}\left(\mathbf{X}_i - \ddot{\mathbf{M}}_k\right)'\left(\ddot{\boldsymbol{\Sigma}}_k\right)^{-1}\left(\mathbf{X}_i - \ddot{\mathbf{M}}_k\right)}{p\sum_{i=1}^N \ddot{z}_{ik}}. \tag{5.22}$$

5.3.2 EM-based algorithms for MVTIN-Ms

When MVTIN-Ms are considered, the last term in (5.12) is equal to

$$\ell_{3c}\left(\boldsymbol{\nu}; \mathcal{S}_c\right) = \sum_{i=1}^N \sum_{k=1}^K z_{ik}\left\{-\ln\left(\theta_k\right) + \ln\left[\mathbb{1}_{(1-\theta_k,1)}\left(w_{ik}\right)\right]\right\}, \tag{5.23}$$

where $\mathbb{1}_A(\cdot)$ is the indicator function on the set A. The ECME and AECM algorithms have in common the E-step and the first two CM-steps, whereas the last CM-step presents two options depending on which algorithm is chosen. In detail, both algorithms proceed as follows.

E-Step The E-step requires the calculation of

$$\ddot{z}_{ik} := \mathrm{E}_{\dot{\boldsymbol{\Theta}}}\left(Z_{ik}|\mathbf{X}_i\right) = \frac{\dot{\pi}_k f_{\text{MVTIN}}\left(\mathbf{X}_i; \dot{\boldsymbol{\phi}}_k\right)}{\sum_{h=1}^K \dot{\pi}_h f_{\text{MVTIN}}\left(\mathbf{X}_i; \dot{\boldsymbol{\phi}}_h\right)}, \tag{5.24}$$

which, also in this case, corresponds to the posterior probability that the unlabeled observation $\mathbf{X}_i$ belongs to the kth component of the mixture, and

$$\begin{aligned}\ddot{w}_{ik} &:= \mathrm{E}_{\dot{\boldsymbol{\Theta}}}\left(W_{ik}|\mathbf{X}_i, \mathbf{z}_i\right) \\ &= \frac{2}{\delta(\mathbf{X}_i;\dot{\boldsymbol{\Omega}}_k)} \frac{\left[\Gamma\left(\frac{pr}{2}+2, \left(1-\dot{\theta}_k\right)\frac{\delta\left(\mathbf{X}_i;\dot{\boldsymbol{\Omega}}_k\right)}{2}\right) - \Gamma\left(\frac{pr}{2}+2, \frac{\delta\left(\mathbf{X}_i;\dot{\boldsymbol{\Omega}}_k\right)}{2}\right)\right]}{\left[\Gamma\left(\frac{pr}{2}+1, \left(1-\dot{\theta}_k\right)\frac{\delta\left(\mathbf{X}_i;\dot{\boldsymbol{\Omega}}_k\right)}{2}\right) - \Gamma\left(\frac{pr}{2}+1, \frac{\delta\left(\mathbf{X}_i;\dot{\boldsymbol{\Omega}}_k\right)}{2}\right)\right]},\end{aligned} \tag{5.25}$$

which is the expected value of a doubly-truncated gamma distribution with parameters $(pr/2)+1$ and $\delta(\mathbf{X}_i;\boldsymbol{\Omega}_k)/2$, on the interval $(1-\theta_k, 1)$; for details see Punzo and Bagnato, 2020b. Lastly, there is no need to compute the expectation of $\ln(w_{ik})$, since it is not related to the parameters, and the expectation of $\ln\left[\mathbb{1}_{(1-\theta_k,1)}(w_{ik})\right]$, since the conditional expectation of (5.23) is not used to update θ.

Now, consider the following parameter sets $\boldsymbol{\Theta}_1 = \{\pi_k, \mathbf{M}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$, $\boldsymbol{\Theta}_2 = \{\boldsymbol{\Psi}_k\}_{k=1}^K$ and $\boldsymbol{\Theta}_3 = \{\theta_k\}_{k=1}^K$.

CM-Step 1-2 The first two CM-steps involve the updates $\ddot{\pi}_k$, $\ddot{\mathbf{M}}_k$, $\ddot{\boldsymbol{\Sigma}}_k$ and $\ddot{\boldsymbol{\Psi}}_k$, which are the same as in (5.18), (5.19), (5.20) and (5.22), respectively.

CM-Step 3 The third CM-step depends on the selected algorithm.

- If an ECME algorithm is chosen, the following incomplete-data log-likelihood function is maximized

$$\ell\left(\boldsymbol{\Theta};\mathcal{S}\right) = \sum_{i=1}^{N} \ln\left[\sum_{k=1}^{K} \pi_k f_{\text{MVTIN}}(\mathbf{X};\mathbf{M}_k,\boldsymbol{\Sigma}_k,\boldsymbol{\Psi}_k,\theta_k)\right], \tag{5.26}$$

  with respect to $\boldsymbol{\Theta}_3$, keeping fixed $\boldsymbol{\Theta}_1$ at $\ddot{\boldsymbol{\Theta}}_1$ and $\boldsymbol{\Theta}_2$ at $\ddot{\boldsymbol{\Theta}}_2$. Operationally, the *optim()* function, in the **stats** package is used to perform a numerical search of $\ddot{\boldsymbol{\nu}} = \{\theta_k\}_{k=1}^K$.

- If an AECM algorithm is chosen, the following specification of the complete-data log-likelihood function is maximized

$$\ell_c\left(\boldsymbol{\Theta};\mathcal{S}_c\right) = \ell_{1c}\left(\boldsymbol{\pi};\mathcal{S}_c\right) + \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik} \ln f_{\text{MVTIN}}(\mathbf{X};\mathbf{M}_k,\boldsymbol{\Sigma}_k,\boldsymbol{\Psi}_k,\theta_k), \tag{5.27}$$

  with respect to $\boldsymbol{\Theta}_3$, keeping fixed $\boldsymbol{\Theta}_1$ at $\ddot{\boldsymbol{\Theta}}_1$ and $\boldsymbol{\Theta}_2$ at $\ddot{\boldsymbol{\Theta}}_2$. In (5.27), $\ell_{1c}\left(\boldsymbol{\pi};\mathcal{S}_c\right)$ is equal to (5.13). Notice that in this case the complete data are $\mathcal{S}_c = \{\mathbf{X}_i, \mathbf{z}_i\}_{i=1}^N$. Operationally, the *optimize()* function, in the **stats** package, is used to perform a numerical search of the maximum $\ddot{\theta}_k$ of (5.27).

5.3.3 A note on the initialization strategy

As discussed in Section 2.3.1.1, the initialization of EM-based algorithms is an important aspect. Here, the short-EM procedure suggested by Biernacki *et al.* (2003) is generalized in the matrix-variate framework. The reasons why such initialization is chosen, with respect to an approach that starts from the M-step, derives from the fact that initialize only the z_{ik} in (5.16) and (5.24) is not sufficient here. Indeed, starting values should be provided also for the w_{ik} in (5.17) and (5.25), and this is not an

easy task. On the contrary, the proposed strategy has shown stable results in both simulated and real data analyses. The parameter estimates and the classification produced have displayed excellent performances, as will be shown in Section 5.4.

About the short-EM procedure, by recalling that H is the number of short runs of the considered algorithm, and s is the (small) number of iterations of these short runs, in this work $H = 100$ and $s = 1$. Then, the parameter sets producing the ten largest log-likelihood values are used to initialize ten complete runs of the considered algorithm. The solution providing the highest log-likelihood value is finally reported. For better comparability purposes, the common parameters $\{\pi_k, \mathbf{M}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\Psi}_k\}_{k=1}^{K}$ of the different competing models considered in this work, are initialized with the same values in the H runs.

5.4 Simulated data analyses

5.4.1 Comparison between ECME and AECM algorithms for MVTIN-Ms

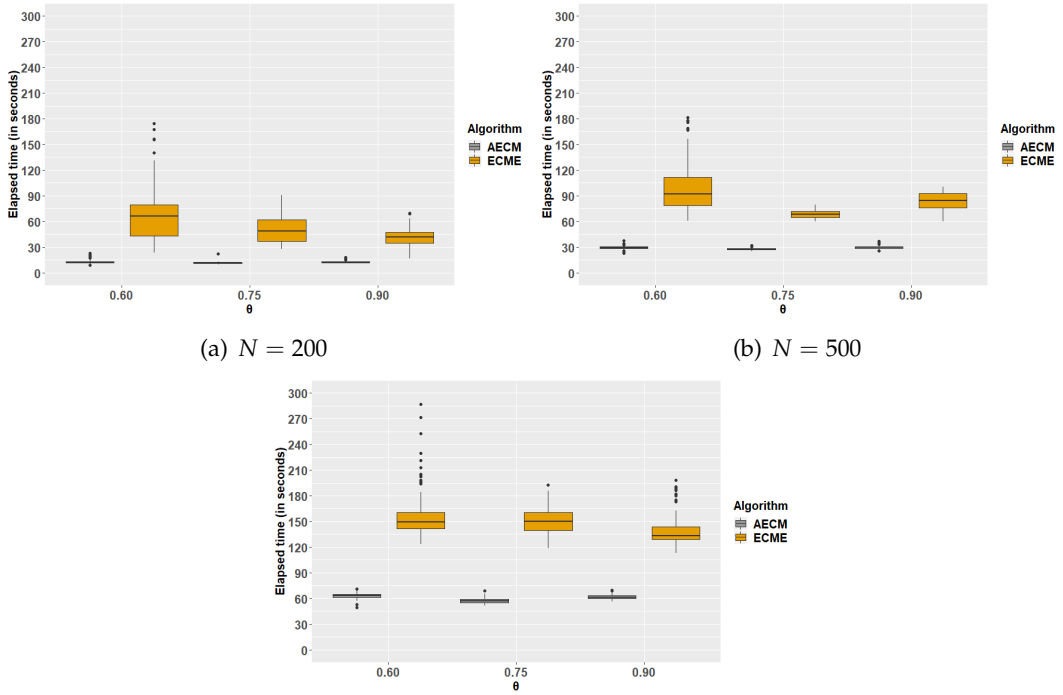
The performance of the ECME and AECM algorithms for MVTIN-Ms are compared in terms of computational times and average log-likelihood at convergence. Two experimental factors are considered: the sample size $N \in \{200, 500, 1000\}$ and the inflation parameter $\theta \in \{0.60, 0.75, 0.90\}$. The values of θ are unbalanced on the right to have scenarios with more kurtosis (see Section 5.2.3). The dimension of the matrices are $p = 3$ and $r = 4$. For each pair (N, θ), one hundred data sets are sampled from a MVTIN-M with $K = 2$ by using the hierarchical representation in (5.11). Overall, a total of $3 \times 3 \times 100 = 900$ data sets are generated. For easiness, it is assumed that all the mixture components have the same θ, whereas the other parameters used to generate the data are displayed in Table 5.1.

TABLE 5.1: Parameters used in the MVTIN-M to generate the data of Section 5.4.1.

Parameters	Group 1	Group 2
π_k	0.35	0.65
$\mathbf{M}_k$	$\begin{pmatrix} -5.00 & -4.00 & -4.00 & -3.00 \\ 4.00 & 4.00 & 5.00 & 5.00 \\ 5.00 & 6.00 & 7.00 & 6.00 \end{pmatrix}$	$\begin{pmatrix} 5.00 & 4.00 & 4.00 & 3.00 \\ -5.00 & -5.00 & -6.00 & -5.00 \\ -2.00 & -2.00 & -3.00 & -3.00 \end{pmatrix}$
$\boldsymbol{\Sigma}_k$	$\begin{pmatrix} 2.00 & 1.00 & 0.20 \\ 1.00 & 2.00 & 1.00 \\ 0.20 & 1.00 & 2.00 \end{pmatrix}$	$\begin{pmatrix} 2.50 & 0.20 & 0.20 \\ 0.20 & 2.50 & 0.20 \\ 0.20 & 0.20 & 2.50 \end{pmatrix}$
$\boldsymbol{\Psi}_k$	$\begin{pmatrix} 1.00 & 0.50 & 0.20 & 0.10 \\ 0.50 & 1.00 & 0.50 & 0.20 \\ 0.20 & 0.50 & 1.00 & 0.50 \\ 0.10 & 0.20 & 0.50 & 1.00 \end{pmatrix}$	$\begin{pmatrix} 1.50 & 0.80 & 0.30 & 0.20 \\ 0.80 & 1.50 & 0.80 & 0.30 \\ 0.30 & 0.80 & 1.50 & 0.80 \\ 0.20 & 0.30 & 0.80 & 1.50 \end{pmatrix}$

On each generated data set, MVTIN-Ms with $K = 2$ are fitted by using both the ECME and AECM algorithms. The computational time is calculated via the *system.time()* function of the **base** package and refers to the execution of ten runs in parallel of the selected algorithm. Specifically, each core executes one of the ten runs discussed in Section 5.3.3 for the considered algorithm. The computation is performed on a Windows 10 PC, with AMD Ryzen 7 3700x CPU, 16.0 GB RAM.

Figure 5.3 illustrates the box-plots of the elapsed time for each combination of θ and the chosen algorithm, when varying the sample size N. It easy to see that, the AECM algorithm has always a much lower computational time than the ECME algorithm. As it is reasonable to expect, the more the sample size grows, the greater the computational time becomes, which in the case of the AECM algorithm seems to double with each increase. Furthermore, it is interesting to notice that, the elapsed time is approximately a decreasing function of θ, with the only exception of the ECME algorithm when $N = 500$ and θ moves from 0.75 to 0.90.



(a) $N = 200$ (b) $N = 500$

(c) $N = 1000$

FIGURE 5.3: Elapsed time (in seconds) for each combination of θ and the chosen algorithm, when varying $N \in \{200, 500, 1000\}$. Each box plot refers to 100 replications.

Let ℓ_j^{AECM} and ℓ_j^{ECME} be the log-likelihood values obtained by the AECM and ECME algorithms, respectively, for the jth generated data set, $j = 1, \ldots, 100$. Based on these quantities, the performance of the algorithms is also evaluated in terms of: average log-likelihood value over the 100 replications, say $\overline{\ell}^{\text{AECM}}$ for the AECM and $\overline{\ell}^{\text{ECME}}$ for the ECME, and number of times that each algorithm reaches $\max\left\{\ell_j^{\text{AECM}}, \ell_j^{\text{ECME}}\right\}$, over the 100 replications, say $\#\text{max}^{\text{AECM}}$ for the AECM and $\#\text{max}^{\text{ECME}}$ for the ECME. Table 5.2 reports these summary measures for each pair (N, θ). It is possible to notice that the average log-likelihoods obtained via the AECM algorithm are constantly slightly better than those produced by the ECME one. Furthermore, the differences between these averages seem to increase for growing values of the sample size. By looking at the last two columns of Table 5.2, and excluding three occasions in the case $(N = 200, \theta = 0.60)$, the AECM algorithm produces a log-likelihood that is always greater than, or equal to, the log-likelihood from the ECME algorithm. Therefore, considering the results of this section, the AECM algorithm will be used in the rest of the paper for MVTIN-Ms.

TABLE 5.2: Average log-likelihood values and number of times the best log-likelihood is reached, for the AECM and ECME algorithms of the MVTIN-Ms, over 100 replications.

N	θ	$\overline{\ell}^{\text{AECM}}$	$\overline{\ell}^{\text{ECME}}$	#max$^{\text{AECM}}$	#max$^{\text{ECME}}$
200	0.60	-4946.89	-4947.06	97	85
	0.75	-5141.09	-5141.11	100	97
	0.90	-5441.68	-5443.06	100	95
500	0.60	-12394.10	-12394.11	100	98
	0.75	-12909.98	-12910.19	100	98
	0.90	-13647.17	-13647.91	100	99
1000	0.60	-24807.21	-24807.66	100	96
	0.75	-25852.09	-25853.29	100	96
	0.90	-27312.12	-27315.32	100	97

5.4.2 Parameter recovery

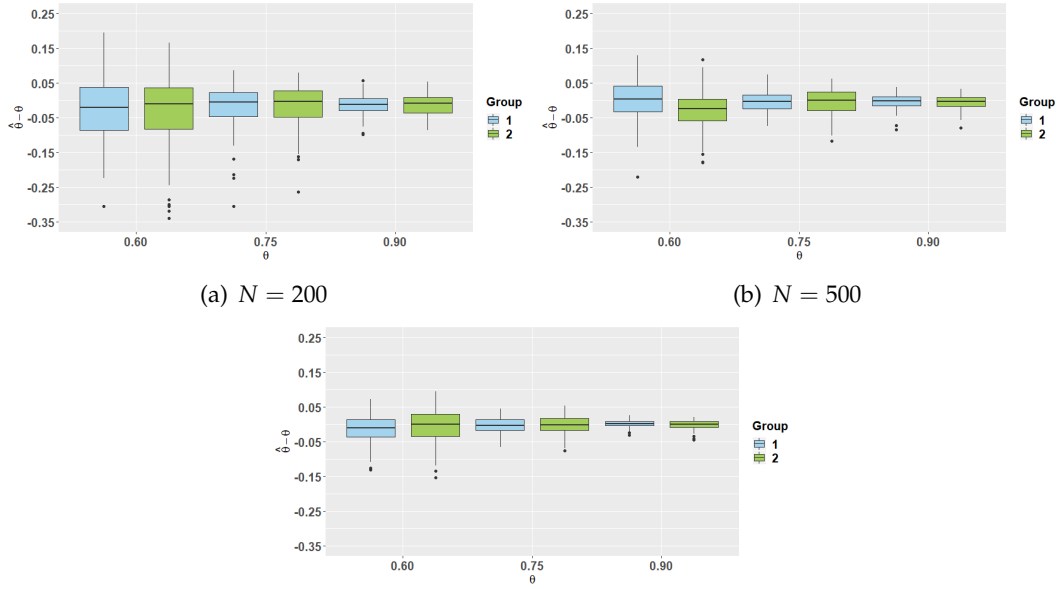
In this section, the parameter recovery of the AECM algorithm for MVTIN-Ms and of the ECM algorithm for MVSEN-Ms is investigated. In detail, the analysis is focused only on θ, since in the normal scale mixtures literature the parameter(s) governing the tail-weight is (are) generally the most difficult to be estimated.

For MVTIN-Ms, the data analyzed in Section 5.4.1 are used. The parameter set and the experimental factors used for MVTIN-Ms are also adopted for MVSEN-Ms, with the only exception of the inflation parameter that assumes the values $\theta \in \{0.15, 0.30, 0.60\}$, in order to have scenarios with more kurtosis (see Section 5.2.2). Similarly to Section 5.4.1, the hierarchical representation in (5.11) is used for generate data from MVSEN-Ms.

Before showing the results, it is important to underline the well-known label switching issue, caused by the invariance of the likelihood function under relabeling the components of a mixture model (Frühwirth-Schnatter, 2006). There is no generally accepted labeling method, and considering the substantial separation between the two groups, the label 1 is attached to the component with the lowest estimated values in the first row of $\mathbf{M}_k$.

In the case of MVTIN-Ms, Figure 5.4 and Figure 5.5 illustrate the box-plots of the differences $(\hat{\theta}_k - \theta_k)$ between the estimated and the true parameters for bias evaluation, and the squared differences $(\hat{\theta}_k - \theta_k)^2$ for mean square error (MSE) evaluation, respectively. The same quantities are displayed in Figure 5.6 and Figure 5.7 for MVSEN-Ms. In any case, each plot refers to a specific sample size and shows the box-plots of both latent groups for the different values of θ_k. Specifically, each box-plot summarizes the behavior of the considered differences with respect to the available 100 replications. The first and immediate result is that the differences under evaluation improve as N increases, for both models. Furthermore, the higher is the kurtosis in the data, the more precise are the estimates for both models. Indeed, the average estimates of θ look quite unstable when the kurtosis is low. To understand this result consider for example the MVSEN distribution and Figure 5.1. In this case the instability arises from the fact that, although θ can take all positive values, those larger than (more or less) 1 lead to a mesokurtic distribution that is practically the same as the matrix-variate normal. The same reasoning can be done for the MVTIN distribution, since in Figure 5.2 it is illustrated that θ values approximately lower than 0.5

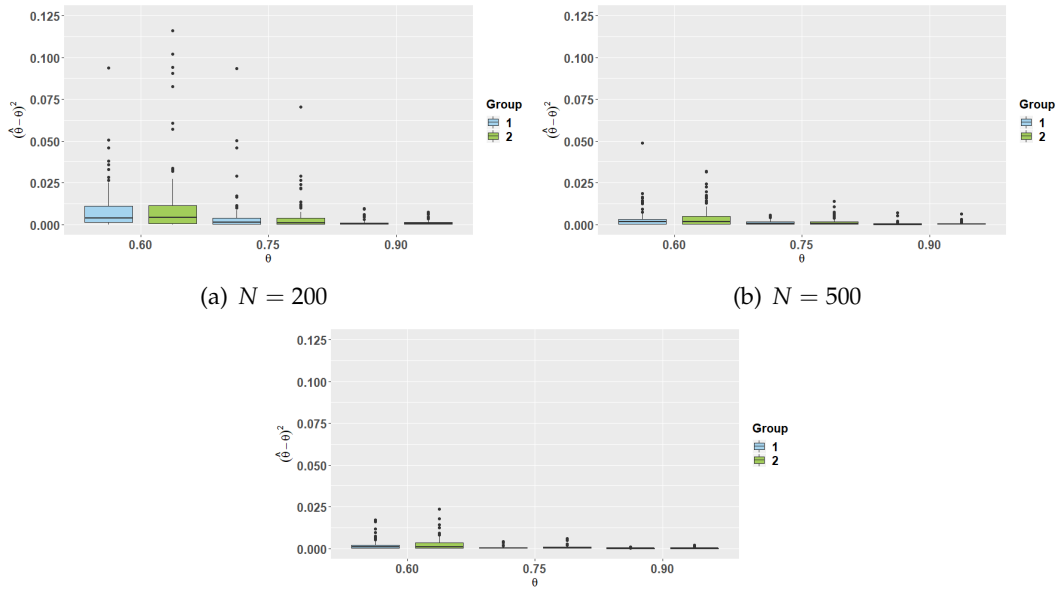
lead to situations close to the normality. Overall, both algorithms seem to properly estimate the tailedness parameters.



(a) $N = 200$ (b) $N = 500$

(c) $N = 1000$

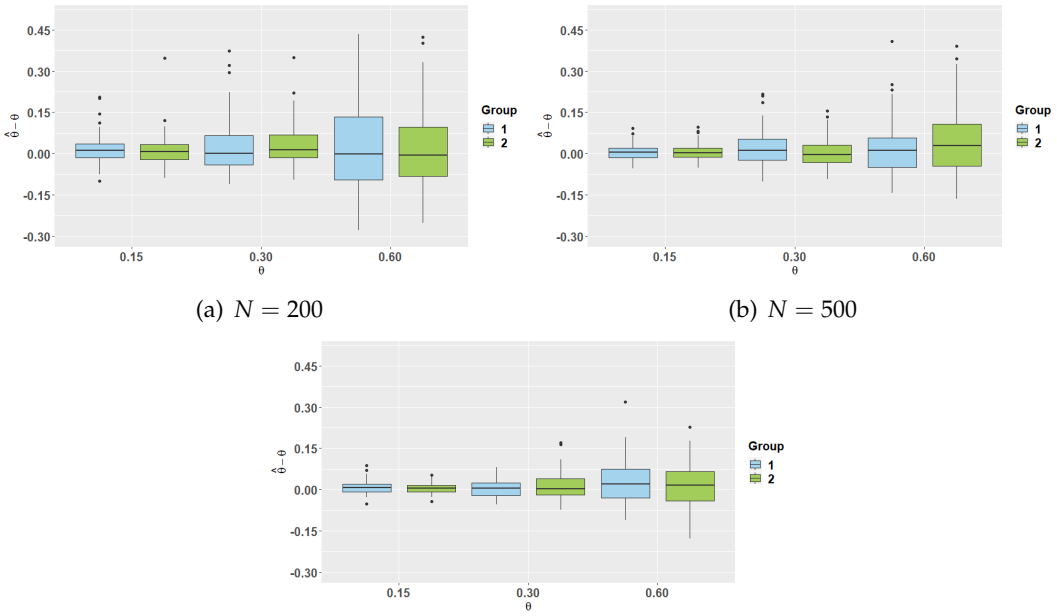
FIGURE 5.4: Box-plots of $(\hat{\theta}_k - \theta_k)$, in the case of the MVTIN distribution, for each latent group and pair (N, θ). Each box-plot summarizes the results over 100 replications.



(a) $N = 200$ (b) $N = 500$

(c) $N = 1000$

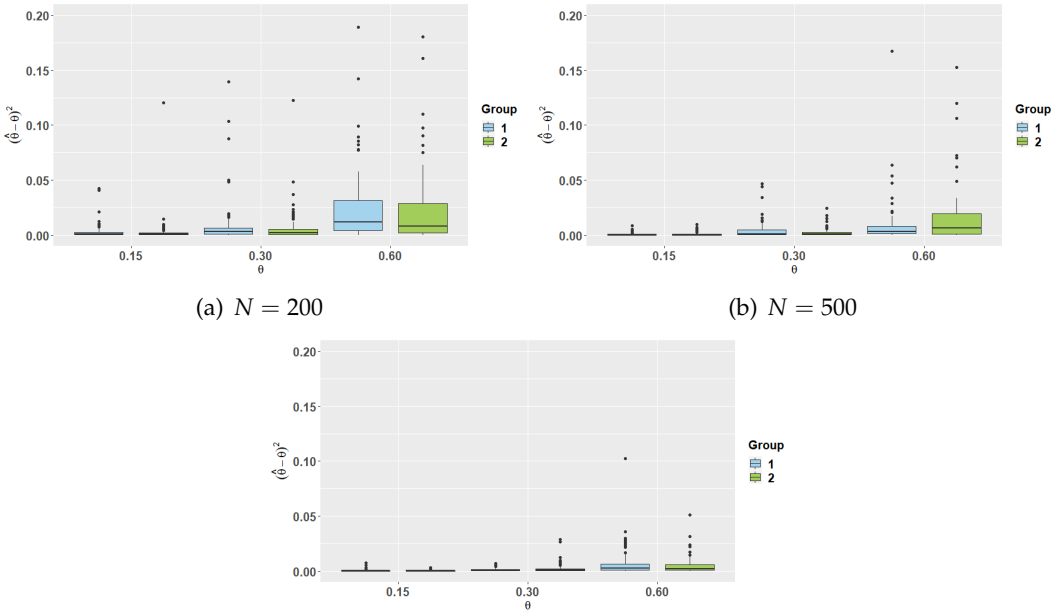
FIGURE 5.5: Box-plots of $(\hat{\theta}_k - \theta_k)^2$, in the case of the MVTIN distribution, for each latent group and pair (N, θ). Each box-plot summarizes the results over 100 replications.



(a) $N = 200$ (b) $N = 500$

(c) $N = 1000$

FIGURE 5.6: Box-plots of $(\hat{\theta}_k - \theta_k)$, in the case of the MVSEN distribution, for each latent group and pair (N, θ). Each box-plot summarizes the results over 100 replications.



(a) $N = 200$ (b) $N = 500$

(c) $N = 1000$

FIGURE 5.7: Box-plots of $(\hat{\theta}_k - \theta_k)^2$, in the case of the MVSEN distribution, for each latent group and pair (N, θ). Each box-plot summarizes the results over 100 replications.

5.4.3 Assessing the impact of outlying matrices

Here, two scenarios are considered to evaluate the effect of outlying matrices on model selection and clustering performance. On each scenario, $p = 3$, $r = 3$, $N = 200$ and one hundred data sets are generated from a MVN-M with $G = 2$

and parameters in Table 5.3.

TABLE 5.3: Parameters used to generate the data of Section 5.4.3.

Parameters	Group 1	Group 2
π_g	0.50	0.50
$\mathbf{M}_g$	$\begin{pmatrix} -3.00 & -2.00 & -2.00 \\ -1.00 & 0.00 & 0.00 \\ -2.00 & -1.00 & -1.00 \end{pmatrix}$	$\begin{pmatrix} 0.00 & 1.00 & 0.00 \\ 2.00 & 2.00 & 3.00 \\ 0.00 & 2.00 & 1.00 \end{pmatrix}$
$\mathbf{\Sigma}_g$	$\begin{pmatrix} 0.70 & -0.05 & -0.05 \\ -0.05 & 0.70 & -0.01 \\ -0.05 & -0.01 & 0.70 \end{pmatrix}$	$\begin{pmatrix} 0.60 & 0.10 & 0.01 \\ 0.10 & 0.60 & -0.05 \\ 0.01 & -0.05 & 0.60 \end{pmatrix}$
$\mathbf{\Psi}_g$	$\begin{pmatrix} 0.60 & 0.10 & -0.05 \\ 0.10 & 0.80 & 0.10 \\ -0.05 & 0.10 & 0.70 \end{pmatrix}$	$\begin{pmatrix} 0.60 & -0.10 & 0.05 \\ -0.10 & 0.90 & 0.01 \\ 0.05 & 0.01 & 1.00 \end{pmatrix}$

Then, a certain percentage T of matrices is randomly selected and, in turn, for each of them one column, say the jth, $j = 1, \ldots, r$, is chosen at random. The values in this column are substituted with random numbers lying on the surface of a 3-dimensional sphere (since $p = 3$), with ray y and center in the corresponding jth column of $\mathbf{M}_k$. In this way, these T matrices have atypical values on one of their columns. In the first scenario (Scenario A) $T = 30\%$ and $y = 2.7$, while in the second one (Scenario B) $T = 15\%$ and $y = 4.0$. This means that, in Scenario A there are a large number of matrices that slightly deviate from the bulk of the data, while in Scenario B there are few but more distant outlying matrices.

MVSEN-Ms and MVTIN-Ms are fitted to the data for $K \in \{1, 2, 3\}$, along with MVN-Ms and MVt-Ms for comparison purposes. Table 5.4 and Table 5.5 report the number of times each K is selected by the BIC, the average classification results and the average BIC computed over the best BIC models on the 100 replications ($\overline{\text{BIC}}$). In Scenario A, because of the relative proximity of the outlying matrices to the bulk of the data, the BIC correctly recognize the number of groups in the data and a perfect classification is always obtained for all the competing models. However, in terms of $\overline{\text{BIC}}$, the MVN-Ms have the worst fitting performance, whereas the MVSEN-Ms are the best.

In Scenario B, the BIC selects almost always three groups for MVN-Ms; additionally, the $\overline{\text{BIC}}$ is clearly the worst among the competing models. This is due to the tails of the mixture components, that are not sufficiently heavy to model data with such atypical observations. In the attempt of modeling the outlying matrices, the true underlying group structure is disrupted, and an additional third component is detected. This has an implication on the resulting classification, as shown by the $\overline{\text{ARI}}$ and $\overline{e}$ values. On the contrary, for MVTIN-Ms and MVt-Ms the number of groups selected is always equal to two, with an almost perfect classification. However, the MVTIN-Ms show the best fitting performance as indicated by the lowest $\overline{\text{BIC}}$ value.

5.5 Real data applications

5.5.1 Data description

The first data set (herein called "Education data") contains career indicators from the Italian National Agency for the Evaluation of Universities and Research Institutes.

TABLE 5.4: Scenario A: number of times each K is selected by the BIC along with the average ARI, ϵ and BIC values computed over the best BIC models with respect to the 100 replications.

Model	$G=1$	$G=2$	$G=3$	$\overline{\text{ARI}}$	$\bar{\epsilon}$	$\overline{\text{BIC}}$
MVN-Ms	0	100	0	1.00	0.00	4857.09
MVSEN-Ms	0	100	0	1.00	0.00	4846.38
MVTIN-Ms	0	100	0	1.00	0.00	4851.42
MVt-Ms	0	100	0	1.00	0.00	4852.36

TABLE 5.5: Scenario B: number of times each K is selected by the BIC along with the average ARI, ϵ and BIC values computed over the best BIC models with respect to the 100 replications.

Model	$G=1$	$G=2$	$G=3$	$\overline{\text{ARI}}$	$\bar{\epsilon}$	$\overline{\text{BIC}}$
MVN-Ms	0	2	98	0.85	8.52	4930.48
MVSEN-Ms	0	99	1	0.99	0.17	4839.73
MVTIN-Ms	0	100	0	1.00	0.03	4813.76
MVt-Ms	0	100	0	1.00	0.04	4825.83

It consists of $p = 3$ numerical indicators, collected over $r = 3$ years, for $N = 75$ study programs in the universities of southern Italy. Specifically, they measure (i) the percentage of course credits earned in the first year over the total to be achieved (V1), (ii) the percentage of students that continued in the second year of the same study program (V2) and (iii) the percentage of students that completed their studies within the normal duration of the course (V3). Therefore, each data point is a 3×3 matrix and reports the average value of all the study programs of the same type across southern Italy. There are $K = 2$ groups in the data, i.e. $N_1 = 33$ bachelor's degrees and $N_2 = 42$ master's degrees.

The second data set (herein called "R&D data"), contains $p = 3$ variables measuring the level of (i) sales, (ii) employment and (iii) capital for $N = 509$ R&D-performing US manufacturing companies in the years 1982-89 ($r = 8$). Therefore, each data point consists of a 3×8 matrix. It is contained in the **pder** package (Croissant and Millo, 2019) under the name *RDPerfComp*. Differently from the previous data set, there is neither a classification of the data, nor any information about a possible underlying group structure. However, useful insights can be gained by looking at Figure 5.8, where the histogram of each variable is displayed, for the full 8-years of data. A similar way of thinking can be find in Melnykov and Zhu (2019). The multimodality in all these histograms seems to suggest the existence of groups in the data.

5.5.2 Results

5.5.2.1 Education data

The competing models are fitted to the data for $K \in \{1,2,3\}$ and the results are reported in Table 5.6. The BIC selects $K = 3$ groups for the MVN-Ms, with a third group that is strongly overlapped to the second one. Additionally, they show the worst fitting and classification performances. Notice that, even if MVN-Ms had been fitted directly with $K = 2$ groups, the classification produced would be the same as for MVt-Ms, but it would have anyway the worst performance in terms of

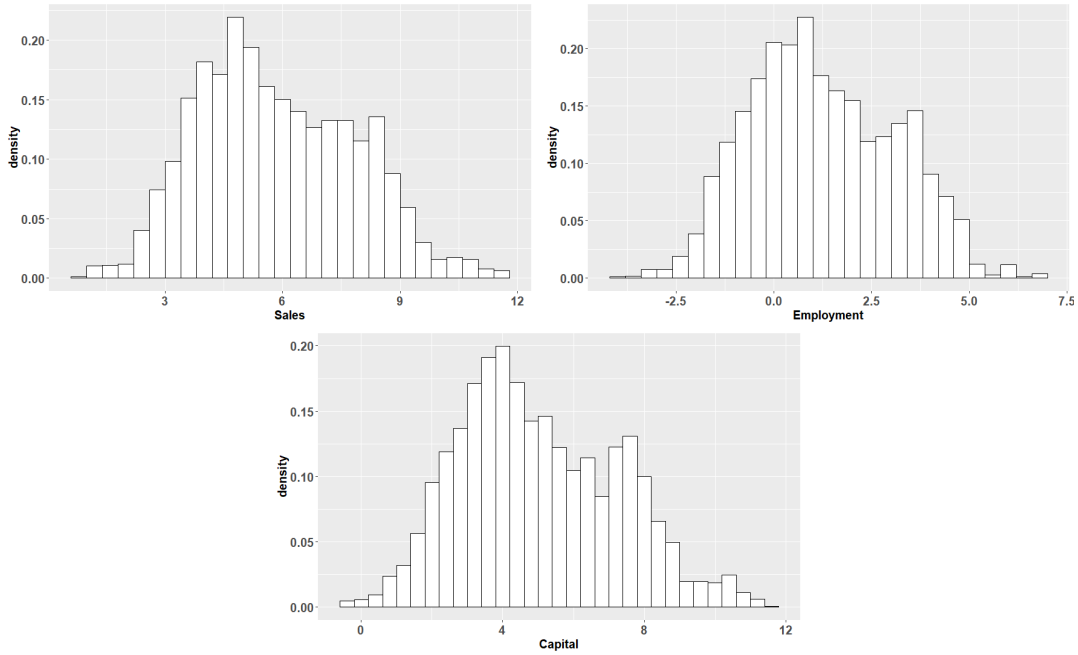


FIGURE 5.8: Histograms of the distribution of each variable for the R&D data.

TABLE 5.6: Number of selected groups (K), BIC values, and classification performance (ARI and ϵ) of the competing models on the education data.

Model	K	BIC	ARI	ϵ
MVN-Ms	3	-2212.05	0.62	16.00
MVSEN-Ms	2	-2262.12	0.84	4.00
MVTIN-Ms	2	-2265.37	0.84	4.00
MVt-Ms	2	-2264.07	0.79	5.33

BIC, i.e. $-2201,79$. For all the other competing models $K = 2$ is chosen. In terms of overall fitting, the best model according to the BIC is the MVTIN-M, whereas in terms of classification both MVSEN-Ms and MVTIN-Ms produce the same partition of the data, that yields a high value of the ARI and the lowest percentage of misclassified observations. This data partition is depicted in Figure 5.9 by using parallel coordinate plots, and where group 1 corresponds to the bachelor's degrees, while group 2 represents the master's degrees. The dashed line in each subfigure represents to the estimated mean for that variable, across the time, in that group. It is possible to see that the two degree typologies are quite separated and, as it is reasonable to expect, the performances of the master's group are better than those of the bachelor's one. This might be due to the difficulties that bachelor's students have in the transition from high schools to universities, while master's students have already overcome these problems. Anyway, there are study programs that seem far from to the bulk of the data, with respect to the considered variable, in both groups. These considerations might suggest the presence of outlying matrices. As a confirmation of this, the estimated tailedness parameters $\hat{\theta}_k$ and kurtoses by the competing models are reported in Table 5.7. The sample weighted kurtoses, with weights

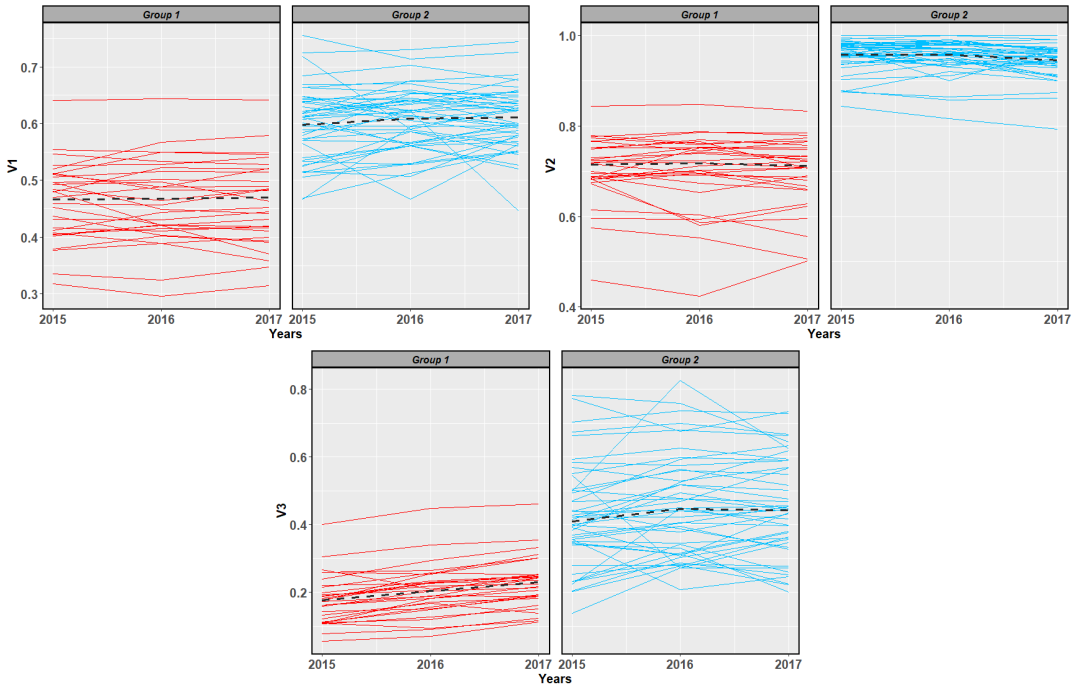


FIGURE 5.9: Education data: parallel coordinate plots constructed for the best BIC model (MVTIN-Ms with $K = 3$).

given by the posterior probabilities produced by each model for the detected partition, are also reported for comparative purposes. The estimated values of $\hat{\theta}_k$ indicate a strong deviation from the normality assumption in terms of tail weight (see Figure 5.1 and Figure 5.2). Additionally, the mixture component of the master's degrees has heavier tails than the bachelor's degrees one. With the exclusion of the MVN-Ms that, as already said, overfit the data, all the models agree in detecting a considerable level of kurtosis. Compared to their sample counterparts, MVSEN-Ms and MVt-Ms moderately overestimate the sample kurtoses, but in any case provide better values than MVt-Ms.

TABLE 5.7: Estimated tailedness parameters and kurtoses by the competing models, along with the sample weighted kurtoses of the soft groups, on the education data.

Model	Group	$\hat{\theta}_k$	Sample kurtosis	Estimated kurtosis
MVN-Ms	1	-	100.86	99.00
	2	-	94.04	99.00
	3	-	119.51	99.00
MVSEN-Ms	1	0.34	102.30	135.42
	2	0.16	124.28	166.09
MVTIN-Ms	1	0.89	101.98	146.48
	2	0.95	124.39	192.27
MVt-Ms	1	6.23	102.63	188.01
	2	4.79	124.47	348.38

5.5.2.2 R&D data

The competing models are fitted to the data for $K \in \{1, \ldots, 5\}$ and the results in terms of BIC are reported in Table 5.8. Considering that the true group memberships are unknown, the ARI and ϵ cannot be used to compare the models. The best

TABLE 5.8: BIC values, and corresponding number of groups selected, for the competing models under the R&D data.

Model	K	BIC
MVN-Ms	5	-9046.91
MVSEN-Ms	3	-10019.55
MVTIN-Ms	3	-9897.02
MVt-Ms	3	-9985.52

model in terms of BIC is the MVSEN-M with $K = 3$ components, and it is illustrated in Figure 5.10 via parallel coordinate plots. Again, the dashed line in each subfigure are the estimated mean for that variable, across the time, in that group. The third group appears to be quite separated from the other two, that show a certain degree of overlap instead. The classification produced by the MVSEN-Ms seems to partition the manufacturing companies according to their resources and productive capacities, that are high for Group 2, medium for Group 1 and scarce for Group 3. It is interesting to note that all the models with heavy tailed component distribu-

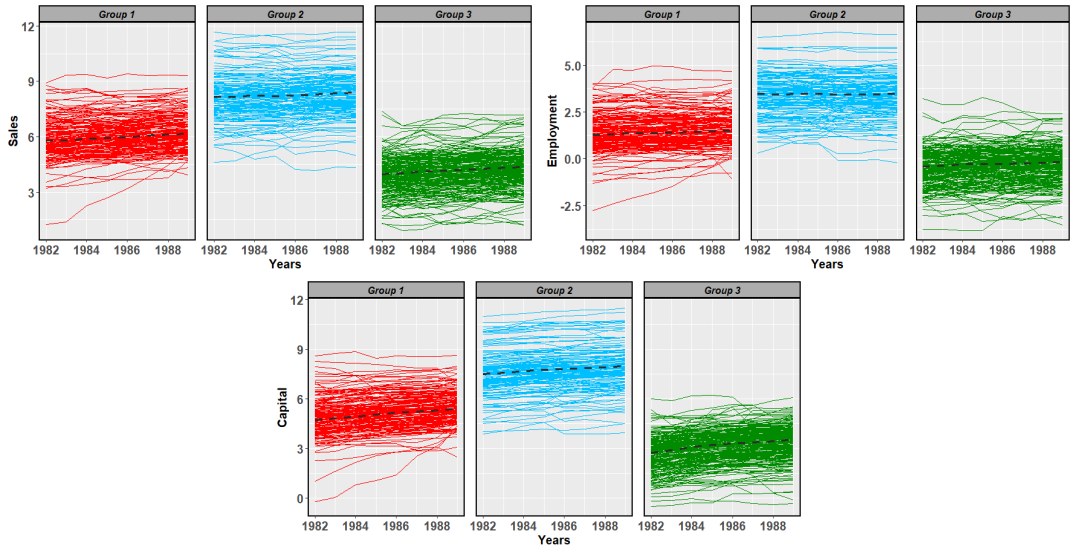


FIGURE 5.10: R&D data: parallel coordinate plots constructed for the best BIC model (MVSEN-M with $K = 3$).

tions agree in detecting three groups in the data, whereas for MVN-Ms the best BIC is obtained when $K = 5$. Also in this data set, MVN-Ms seem to overfit the data because, from the analyses of its mean matrices (not shown here for brevity's sake), the two additional detected groups are strongly overlapped to two of the other three groups. Furthermore, they show the worst fitting performance. Similarly to the previous section, this may be an indication that the components of MVN-Ms are not heavy tailed enough to adequately model these data. On a related note, Table 5.9 reports the estimated tailedness parameters $\hat{\theta}_k$ and kurtoses by the competing models, along with the sample weighted kurtoses, as done in Section 5.5.2.1. The $\hat{\theta}_k$ values

are more extreme than those analyzed in the previous real data application, highlighting the presence of clusters with very high levels of kurtosis. Also in this case there is a moderate overestimation of the sample kurtoses by our models, whereas the kurtosis for MVt-Ms is undefined because the estimated $\hat{\theta}_k$ are lower than 4. Indeed, compared to MVt-Ms, MVSEN-Ms and MVTIN-Ms have the advantage that the kurtosis always exists and can be computed.

TABLE 5.9: Estimated tailedness parameters and kurtoses by the competing models, along with the sample weighted kurtoses of the soft groups, on the R&D performing companies data.

Model	Group	$\hat{\theta}_k$	Sample kurtosis	Estimated kurtosis
MVN-Ms	1	-	645.23	624.00
	2	-	654.12	624.00
	3	-	661.63	624.00
	4	-	684.67	624.00
	5	-	652.08	624.00
MVSEN-Ms	1	0.10	947.05	1214.89
	2	0.11	1023.40	1202.44
	3	0.07	1005.65	1386.42
MVTIN-Ms	1	0.97	967.73	1492.39
	2	0.97	1121.82	1513.12
	3	0.97	928.89	1527.91
MVt-Ms	1	2.96	1193.50	-
	2	3.47	1208.43	-
	3	3.03	948.22	-

5.6 Conclusions

In this work two new matrix-variate distributions have been introduced, both belonging to the normal scale mixture family of models. In detail, when a convenient shifted exponential is chosen as mixing distribution, the matrix-variate shifted exponential normal distribution is obtained. Instead, by choosing a convenient uniform as mixing distribution, the matrix-variate tail-inflated normal distribution is defined. Both distributions have a closed-form characterization of the probability density function and heavier tails than the (nested) matrix-variate normal distribution, implying that they are able to model data with mild outliers in a better way. The application of both distributions in model-based clustering has been also discussed. Specifically, each of the two distributions has been chosen as component distribution of the respective finite mixture model. Different EM-based algorithms for maximum likelihood parameter estimation have been considered and tested, in terms of computational time and parameter recovery, via simulated analyses. For MVTIN-Ms, the AECM algorithm has shown better performances with respect to the ECME algorithm. It has been also illustrated that, because of their greater flexibility with respect to matrix-variate normal mixtures, the proposed mixture models may avoid the disruption of the true underlying group structure, and provide a better fit both in simulated and real data sets.

Chapter 6

The Matrix Normal Cluster-Weighted Model [1]

6.1 Introduction

In the general finite mixture model illustrated in Section (2.1), no exogenous variables explain the location and the variability parameters of each mixture component. However, when there is a linear relationship between some variables, important insight can be gained by accounting for functional dependencies between them. For this reason, finite mixtures of regression models with fixed covariates (FMR) have been proposed in the literature (see DeSarbo and Cron, 1988; Frühwirth-Schnatter, 2006, for examples). An extension of FMR are the so-called finite mixtures of regression models with concomitant variables (FMRC; Dayton and Macready, 1988), in which the mixing weights depend on some variables (which are often the same covariates) and are generally modeled by a multinomial logistic model (see Ingrassia and Punzo, 2016, 2019, for details). Unfortunately, none of these methodologies explicitly use the distribution of the covariates for clustering, i.e. the assignment of data points to clusters does not directly utilize any information from the distribution of the covariates. Differently from these approaches, finite mixtures of regressions with random covariates (Gershenfeld, 1997; Gershenfeld *et al.*, 1999), also known as cluster-weighted models (CWMs), allow for such functional dependency. This occurs because, for each mixture component, the CWMs decompose the joint distribution of responses and covariates into the product between the marginal distribution of the covariates and the conditional distribution of the responses given the covariates.

Several CWMs have been introduced in the univariate and multivariate literature. Most of them consider a univariate response variable, along with a set of covariates, modeled by a univariate and a multivariate distribution, respectively (see Ingrassia *et al.*, 2012, 2014; Punzo, 2014, for examples). Fewer CWMs exist in which several responses and covariates are both modeled by multivariate distributions (see Punzo and McNicholas, 2017b; Dang *et al.*, 2017).

However, as discussed in Section (2.5), over the years there has been an increasing interest in applications involving matrix-variate data. Nevertheless, there exists a limited number of contributions involving matrix-variate regressions. Firstly introduced by Viroli (2012), this model has been recently considered in the mixture framework by Melnykov and Zhu (2019), where matrix-variate finite mixtures of

[1] This work is based on the following unpublished manuscript: Tomarchio S.D., McNicholas P.D., Punzo A.. Matrix Normal Cluster-Weighted Models. It is currently under review at the *Journal of Classification*. The current manuscript is a combined effort of the authors. However, Tomarchio S.D. contributed in conceptualization, implementation, data elaboration and writing–original draft preparation; McNicholas P.D. and Punzo A. contributed in conceptualization and supervision.

regressions with fixed covariates (MVN-FMR) are introduced. There are no matrix-variate CWMs presently in the literature and this work aims to fill this gap by introducing and discussing a matrix-variate CWM, in which the cluster-specific marginal distribution of the covariates, and the cluster-specific conditional distribution of the responses given the covariates, are assumed to be matrix normal.

In Section 6.2, the matrix normal cluster-weighted model (MVN-CWM) is introduced. In Section 6.3, an ECM algorithm for maximum likelihood parameter estimation is presented, along with some computational and operational aspects. In the simulation studies outlined in Section 6.4, the parameter recovery, the classification performance and the initialization strategy for the MVN-CWM are investigated as well as the capability of the BIC to detect the underlying group structure. A comparison with the MVN-FMR is also therein done. The application of the MVN-CWM to two real data sets concerning educational and non-life Italian insurance data is then analyzed in Section 6.5. Lastly, some conclusions are drawn in Section 6.6.

6.2 Methodology

6.2.1 Background

Initially, the matrix-variate regression model is recalled, since it can be considered the fist building block for the MVN-CWM. Let $\boldsymbol{Y}$ be a continuous random matrix of dimension $p \times r$, containing p responses measured in r occasions. Suppose that a set of q covariates is observed for each r, and inserted in a matrix $\mathbf{X}$ of dimension $q \times r$. A generic matrix-variate regression model for $\boldsymbol{Y}$ has the form

$$\boldsymbol{Y} = \boldsymbol{\beta}\boldsymbol{w}^{\top} + \boldsymbol{B}\mathbf{X} + \boldsymbol{U}, \tag{6.1}$$

where $\boldsymbol{\beta}$ is the $p \times 1$ vector consisting in the parameters related with the intercept, $\boldsymbol{w}$ is a $r \times 1$ vector of ones, $\boldsymbol{B}$ is the $p \times q$ matrix containing the parameters related to the q covariates and $\boldsymbol{U}$ is the $p \times r$ error term matrix. Model (6.1) can be expressed in compact notation as

$$\boldsymbol{Y} = \boldsymbol{B}^{*}\mathbf{X}^{*} + \boldsymbol{U}, \tag{6.2}$$

where $\boldsymbol{B}^{*}$ is the $p \times (q+1)$ matrix involving all the parameters to be estimated and $\mathbf{X}^{*}$ is the $(q+1) \times r$ matrix containing the information about the intercept and q covariates (Viroli, 2012; Anderlucci *et al.*, 2014). If $\boldsymbol{U} \sim \mathcal{N}_{p\times r}(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{\Psi})$, then $\boldsymbol{Y}|\mathbf{X}^{*} \sim \mathcal{N}_{p\times r}\left(\boldsymbol{B}^{*}\mathbf{X}^{*}, \boldsymbol{\Sigma}, \boldsymbol{\Psi}\right)$.

The second building block for the construction of the MNV-CWM consists in extending the matrix-variate regression model to a mixture setting. Specifically, within a fixed covariates framework (FMR), we have

$$g\left(\boldsymbol{Y}\right) = \sum_{k=1}^{K} \pi_k f_k\left(\boldsymbol{Y}|\mathbf{X}^{*}\right), \tag{6.3}$$

where π_k is the mixing weight and $f_k\left(\boldsymbol{Y}|\mathbf{X}^{*}\right)$ is the cluster-specific conditional distribution of the responses.

6.2.2 The matrix normal CWM

Now, let $(\mathbf{X}, \mathbf{Y})$ be a pair of random matrices, having the same meaning as in Section 6.2.1, with joint distribution $g(\mathbf{X}, \mathbf{Y})$. Then, a general matrix CWM has the following joint distribution

$$g(\mathbf{X}, \mathbf{Y}) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{Y}|\mathbf{X}^*) f_k(\mathbf{X}), \tag{6.4}$$

where $f_k(\mathbf{Y}|\mathbf{X}^*)$ is the cluster-specific conditional distribution of the responses, $f_k(\mathbf{X})$ is the cluster-specific marginal distribution of the covariates, and π_k is the mixing weight. Furthermore, in each group the conditional expectation $\mathbb{E}(\mathbf{Y}|\mathbf{X}^*)$ is assumed to be a linear function of $\mathbf{X}^*$ depending on some parameters.

Herein, it is assumed that in model (6.4) both $f_k(\mathbf{Y}|\mathbf{X}^*)$ and $f_k(\mathbf{X})$ are MVN distributions, and $\mathbb{E}(\mathbf{Y}|\mathbf{X}^*) = \boldsymbol{B}^*\mathbf{X}^*$, as described in Section 6.2.1. Thus, model (6.4) can be rewritten as

$$g(\mathbf{X}, \mathbf{Y}; \boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_k f_{\text{MVN}}\left(\mathbf{Y}|\mathbf{X}^*; \boldsymbol{B}_k^* \mathbf{X}^*, \boldsymbol{\Sigma}_{\mathbf{Y}_k}, \boldsymbol{\Psi}_{\mathbf{Y}_k}\right) f_{\text{MVN}}\left(\mathbf{X}; \mathbf{M}_k, \boldsymbol{\Sigma}_{\mathbf{X}_k}, \boldsymbol{\Psi}_{\mathbf{X}_k}\right). \tag{6.5}$$

For ease of understanding, a subscript with the variable name is added to the respective covariance matrices.

Notice that there is a lack of model identifiability related to the properties of the Kronecker product. Indeed, for a MVN distribution $\boldsymbol{\Psi} \otimes \boldsymbol{\Sigma} = \boldsymbol{\Psi}^* \otimes \boldsymbol{\Sigma}^*$ if $\boldsymbol{\Sigma}^* = a\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}^* = a^{-1}\boldsymbol{\Psi}$. As a result, matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$ are identifiable up to a multiplicative constant a (Melnykov and Zhu, 2019). According to Gallaugher and McNicholas (2018), and as adopted in this thesis, a way to obtain a unique solution is to set the first diagonal element of the row covariance matrix to 1. In detail, all the elements of the estimated $\boldsymbol{\Sigma}_{\mathbf{Y}_k}$ and $\boldsymbol{\Sigma}_{\mathbf{X}_k}$ are divided by their first diagonal element. Therefore, from an interpretative point of view, all the elements of these matrices are proportionally scaled with respect to the first one.

6.3 Parameter estimation

Parameter estimation is carried out by means of the ECM algorithm. The EM algorithm cannot be directly implemented because of the characteristics of the MVN distribution (see Section 5.3). Let $\mathcal{S}_c = \{\mathbf{X}_i, \mathbf{Y}_i, \mathbf{z}_i\}_{i=1}^N$ be the complete-data. Then, the complete-data likelihood is

$$L_c(\boldsymbol{\Theta}) = \prod_{i=1}^{N} \prod_{k=1}^{K} \left[\pi_k f_{\text{MVN}}\left(\mathbf{Y}_i|\mathbf{X}_i^*; \boldsymbol{B}_k^* \mathbf{X}_i^*, \boldsymbol{\Sigma}_{\mathbf{Y}_k}, \boldsymbol{\Psi}_{\mathbf{Y}_k}\right) f_{\text{MVN}}\left(\mathbf{X}_i; \mathbf{M}_k, \boldsymbol{\Sigma}_{\mathbf{X}_k}, \boldsymbol{\Psi}_{\mathbf{X}_k}\right)\right]^{z_{ik}}. \tag{6.6}$$

Therefore, the corresponding complete-data log-likelihood can be written as

$$\begin{aligned} l_c(\boldsymbol{\Theta}) = &\sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \ln(\pi_k) + \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \ln\left[f_{\text{MVN}}\left(\mathbf{Y}_i|\mathbf{X}_i^*; \boldsymbol{B}_k^* \mathbf{X}_i^*, \boldsymbol{\Sigma}_{\mathbf{Y}_k}, \boldsymbol{\Psi}_{\mathbf{Y}_k}\right)\right] \\ &+ \sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \ln\left[f_{\text{MVN}}\left(\mathbf{X}_i; \mathbf{M}_k, \boldsymbol{\Sigma}_{\mathbf{X}_k}, \boldsymbol{\Psi}_{\mathbf{X}_k}\right)\right]. \end{aligned} \tag{6.7}$$

E-Step The E-step requires the calculation of

$$\begin{aligned}\ddot{z}_{ik} &:= \mathrm{E}_{\dot{\boldsymbol{\Theta}}}\left(Z_{ik}|\mathbf{X}_i, \boldsymbol{Y}_i\right) \\ &= \frac{\dot{\pi}_k f_{\mathrm{MVN}}(\boldsymbol{Y}_i|\mathbf{X}_i^*; \dot{\boldsymbol{B}}_k^*\mathbf{X}_i^*, \dot{\boldsymbol{\Sigma}}_{\boldsymbol{Y}_k}, \dot{\boldsymbol{\Psi}}_{\boldsymbol{Y}_k}) f_{\mathrm{MVN}}(\mathbf{X}_i; \dot{\mathbf{M}}_k, \dot{\boldsymbol{\Sigma}}_{\mathbf{X}_k}, \dot{\boldsymbol{\Psi}}_{\mathbf{X}_k})}{\sum_{j=1}^{K} \dot{\pi}_j f_{\mathrm{MVN}}(\boldsymbol{Y}_i|\mathbf{X}_i^*; \dot{\boldsymbol{B}}_j^*\mathbf{X}_i^*, \dot{\boldsymbol{\Sigma}}_{\boldsymbol{Y}_j}, \dot{\boldsymbol{\Psi}}_{\boldsymbol{Y}_j}) f_{\mathrm{MVN}}(\mathbf{X}_i; \dot{\mathbf{M}}_j, \dot{\boldsymbol{\Sigma}}_{\mathbf{X}_j}, \dot{\boldsymbol{\Psi}}_{\mathbf{X}_j})},\end{aligned} \tag{6.8}$$

which corresponds to the posterior probability that the unlabeled observation $(\mathbf{X}_i, \boldsymbol{Y}_i)$ belongs to the kth component of the mixture. Because the posterior probability in (6.8) depends on the density of $\mathbf{X}_i$, the accuracy of clustering can be increased by using the covariates as well (Zarei *et al.*, 2018).

Now, consider $\boldsymbol{\Theta}_1 = \{\pi_k, \mathbf{M}_k, \boldsymbol{\Sigma}_{\mathbf{X}_k}, \boldsymbol{B}_k, \boldsymbol{\Sigma}_{\boldsymbol{Y}_k}\}_{k=1}^K$, and $\boldsymbol{\Theta}_2 = \{\boldsymbol{\Psi}_{\mathbf{X}_k}, \boldsymbol{\Psi}_{\boldsymbol{Y}_k}\}_{k=1}^K$.

CM-Step 1 At the first CM-step, by fixing $\boldsymbol{\Theta}_2$ at $\dot{\boldsymbol{\Theta}}_2$, it is possible to obtain

$$\ddot{\pi}_k = \frac{\sum_{i=1}^N \ddot{z}_{ik}}{N}, \tag{6.9}$$

$$\ddot{\mathbf{M}}_k = \frac{\sum_{i=1}^N \ddot{z}_{ik}\mathbf{X}_i}{\sum_{i=1}^N \ddot{z}_{ik}}, \tag{6.10}$$

$$\ddot{\boldsymbol{B}}_k^* = \left[\sum_{i=1}^N \ddot{z}_{ik}\boldsymbol{Y}_i(\dot{\boldsymbol{\Psi}}_{\boldsymbol{Y}_k})^{-1}\mathbf{X}_i^{*\prime}\right]\left[\sum_{i=1}^N \ddot{z}_{ik}\mathbf{X}_i^*(\dot{\boldsymbol{\Psi}}_{\boldsymbol{Y}_k})^{-1}\mathbf{X}_i^{*\prime}\right]^{-1}, \tag{6.11}$$

$$\ddot{\boldsymbol{\Sigma}}_{\mathbf{X}_k} = \frac{\sum_{i=1}^N \ddot{z}_{ik}\left(\mathbf{X}_i - \ddot{\mathbf{M}}_k\right)\left(\dot{\boldsymbol{\Psi}}_{\mathbf{X}_k}\right)^{-1}\left(\mathbf{X}_i - \ddot{\mathbf{M}}_k\right)'}{r\sum_{i=1}^N \ddot{z}_{ik}}, \tag{6.12}$$

$$\ddot{\boldsymbol{\Sigma}}_{\boldsymbol{Y}_k} = \frac{\sum_{i=1}^N \ddot{z}_{ik}\left(\boldsymbol{Y}_i - \ddot{\boldsymbol{B}}_k^*\mathbf{X}_i^*\right)\left(\dot{\boldsymbol{\Psi}}_{\boldsymbol{Y}_k}\right)^{-1}\left(\boldsymbol{Y}_i - \ddot{\boldsymbol{B}}_k^*\mathbf{X}_i^*\right)'}{r\sum_{i=1}^N \ddot{z}_{ik}}. \tag{6.13}$$

CM-Step 2 At the second CM-step, keeping fixed $\boldsymbol{\Theta}_1$ at $\ddot{\boldsymbol{\Theta}}_1$, it is possible to obtain

$$\ddot{\boldsymbol{\Psi}}_{\mathbf{X}_k} = \frac{\sum_{i=1}^N \ddot{z}_{ik}\left(\mathbf{X}_i - \ddot{\mathbf{M}}_k\right)'\left(\ddot{\boldsymbol{\Sigma}}_{\mathbf{X}_k}\right)^{-1}\left(\mathbf{X}_i - \ddot{\mathbf{M}}_k\right)}{q\sum_{i=1}^N \ddot{z}_{ik}}, \tag{6.14}$$

$$\ddot{\boldsymbol{\Psi}}_{\boldsymbol{Y}_k} = \frac{\sum_{i=1}^N \ddot{z}_{ik}\left(\boldsymbol{Y}_i - \ddot{\boldsymbol{B}}_k^*\mathbf{X}_i^*\right)'\left(\ddot{\boldsymbol{\Sigma}}_{\boldsymbol{Y}_k}\right)^{-1}\left(\boldsymbol{Y}_i - \ddot{\boldsymbol{B}}_k^*\mathbf{X}_i^*\right)}{p\sum_{i=1}^N \ddot{z}_{ik}}. \tag{6.15}$$

6.3.1 ECM initialization

The ECM algorithm is initialized by specifying the initial quantities in (6.8). Indeed, with respect to the approach implemented in Section 5.3.3, it is more convenient to start from the E-step, since it only requires the initialization of the z_{ik}. Furthermore, starting from the M-step would require a random starting value for the $\boldsymbol{B}_k^*$, and this is not a straightforward task. In any case, an initial value also for $\boldsymbol{\Psi}_{\mathbf{X}_k}$ and $\boldsymbol{\Psi}_{\boldsymbol{Y}_k}$ must be provided. Therefore, the following initialization strategy is implemented:

1. generate K random positive definite matrices for $\boldsymbol{\Psi}_{\mathbf{X}_k}$ and $\boldsymbol{\Psi}_{\boldsymbol{Y}_k}$. This can be done via the *genPositiveDefMat()* function of the **clusterGeneration** package (Qiu and Joe., 2015).

2. generate N random vectors $z_i^{(1)} = \left(z_{i1}^{(1)}, \ldots, z_{iK}^{(1)}\right)^\top$, $i = 1, \ldots, N$. This is done with the following three approaches:

   (a) in a "soft" way, by generating K positive random values from a uniform distribution in [0,1] for each observation, that are subsequently divided by their total to sum 1. Being purely random, this procedure is repeated 15 times, and the solution maximizing the observed-data log-likelihood among these runs is considered.

   (b) in a "hard" way, by using the classification produced by the k-means algorithm, computed on the vectorized data;

   (c) in a "hard" way, by using the classification produced by mixtures of matrix-normal distributions, computed on the merged data.

The approach producing the highest observed-data log-likelihood is finally selected.

6.4 Simulation studies

6.4.1 Simulation 1: A focus on the matrix-normal CWM

In this study, several aspects concerning the matrix-normal CWM are discussed. First and foremost, since the ECM algorithm is used to fit the model, it is useful to evaluate its parameter recovery, i.e. whether it can correctly recover the generating parameters. For this reason, simulated data are generated from a MVN-CWM with $p = q = r = 3$ and $K = 4$. Two scenarios are then considered, according to the different level of overlap of the mixture components. In the first scenario (called "Scenario A_1"), the mixture components are well-separated both in $\mathbf{X}$, by assuming relatively distant mean matrices, and in $\boldsymbol{Y}|\mathbf{X}^*$, by using different intercepts and slopes. Conversely, in the second scenario (called "Scenario B_1"), there is a certain degree of overlap, since the intercepts are all equal among the mixture components, and their mean matrices and slopes have nearly the same values. The parameters used for Scenario A_1 are displayed in Table 6.1. In Scenario B_1, the set of parame-

TABLE 6.1: Parameters used to generate the simulated data sets under Scenario A_1.

Parameter	Group 1	Group 2	Group 3	Group 4
π_k	0.30	0.30	0.20	0.20
$\mathbf{M}_k$	$\begin{pmatrix} 1.00 & 2.00 & 0.00 \\ -4.00 & -3.00 & -3.00 \\ 1.00 & 2.00 & 1.00 \end{pmatrix}$	$\begin{pmatrix} 6.00 & 8.00 & 6.00 \\ 2.00 & 1.00 & 3.00 \\ 5.00 & 6.00 & 6.00 \end{pmatrix}$	$\begin{pmatrix} -4.00 & -3.00 & -4.00 \\ -9.00 & -9.00 & -7.00 \\ -4.00 & -3.00 & -5.00 \end{pmatrix}$	$\begin{pmatrix} 12.00 & 12.00 & 11.00 \\ 6.00 & 7.00 & 7.00 \\ 10.00 & 11.00 & 11.00 \end{pmatrix}$
$\boldsymbol{\Sigma}_{\mathbf{X}_k}$	$\begin{pmatrix} 1.00 & 0.50 & 0.25 \\ 0.50 & 1.00 & 0.50 \\ 0.25 & 0.50 & 1.00 \end{pmatrix}$	$\begin{pmatrix} 2.00 & 0.40 & 0.08 \\ 0.40 & 0.20 & 0.40 \\ 0.08 & 0.40 & 2.00 \end{pmatrix}$	$\begin{pmatrix} 1.50 & 0.75 & 0.38 \\ 0.75 & 1.50 & 0.75 \\ 0.38 & 0.75 & 1.50 \end{pmatrix}$	$\begin{pmatrix} 1.20 & 0.60 & 0.30 \\ 0.60 & 1.20 & 0.60 \\ 0.30 & 0.60 & 1.20 \end{pmatrix}$
$\boldsymbol{\Psi}_{\mathbf{X}_k}$	$\begin{pmatrix} 1.20 & 0.60 & 0.30 \\ 0.60 & 1.20 & 0.60 \\ 0.30 & 0.60 & 1.20 \end{pmatrix}$	$\begin{pmatrix} 1.40 & 0.70 & 0.35 \\ 0.70 & 1.40 & 0.70 \\ 0.35 & 0.70 & 1.40 \end{pmatrix}$	$\begin{pmatrix} 0.80 & 0.40 & 0.20 \\ 0.40 & 0.80 & 0.40 \\ 0.20 & 0.40 & 0.80 \end{pmatrix}$	$\begin{pmatrix} 1.60 & 0.80 & 0.40 \\ 0.80 & 1.60 & 0.80 \\ 0.40 & 0.80 & 1.60 \end{pmatrix}$
$\boldsymbol{B}_k^*$	$\begin{pmatrix} 0.00 & 1.00 & 1.00 & 1.00 \\ -2.00 & 1.00 & 1.50 & 1.00 \\ 1.00 & 1.50 & 1.50 & 1.00 \end{pmatrix}$	$\begin{pmatrix} 6.00 & -1.00 & -1.50 & -1.00 \\ 4.00 & -1.00 & -1.50 & -1.00 \\ 8.00 & -1.50 & -1.50 & -1.00 \end{pmatrix}$	$\begin{pmatrix} -5.00 & 1.00 & 1.00 & 1.00 \\ -3.00 & 1.50 & 1.00 & 1.00 \\ -6.00 & 1.50 & 1.50 & 1.00 \end{pmatrix}$	$\begin{pmatrix} 1.00 & -1.00 & -1.00 & -1.00 \\ -5.00 & -1.00 & -1.50 & -1.50 \\ 0.00 & -1.50 & -1.00 & -1.50 \end{pmatrix}$
$\boldsymbol{\Sigma}_{\mathbf{Y}_k}$	$\begin{pmatrix} 1.40 & 0.84 & 0.50 \\ 0.84 & 1.40 & 0.84 \\ 0.50 & 0.84 & 1.40 \end{pmatrix}$	$\begin{pmatrix} 1.80 & 1.26 & 0.88 \\ 1.26 & 1.80 & 1.26 \\ 0.88 & 1.26 & 1.80 \end{pmatrix}$	$\begin{pmatrix} 1.20 & 0.84 & 0.59 \\ 0.84 & 1.20 & 0.84 \\ 0.59 & 0.84 & 1.20 \end{pmatrix}$	$\begin{pmatrix} 1.60 & 0.96 & 0.58 \\ 0.96 & 1.60 & 0.96 \\ 0.58 & 0.96 & 1.60 \end{pmatrix}$
$\boldsymbol{\Psi}_{\mathbf{Y}_k}$	$\begin{pmatrix} 2.00 & 0.60 & 0.18 \\ 0.60 & 0.20 & 0.60 \\ 0.18 & 0.60 & 2.00 \end{pmatrix}$	$\begin{pmatrix} 1.10 & 0.55 & 0.28 \\ 0.55 & 1.10 & 0.55 \\ 0.28 & 0.55 & 1.10 \end{pmatrix}$	$\begin{pmatrix} 1.90 & 1.71 & 1.54 \\ 1.71 & 1.90 & 1.71 \\ 1.54 & 1.71 & 1.90 \end{pmatrix}$	$\begin{pmatrix} 1.40 & 1.26 & 1.13 \\ 1.26 & 1.40 & 1.26 \\ 1.13 & 1.26 & 1.40 \end{pmatrix}$

ters $\{\pi_k, \boldsymbol{\Sigma}_{\mathbf{X}_k}, \boldsymbol{\Psi}_{\mathbf{X}_k}, \boldsymbol{\Sigma}_{\mathbf{Y}_k}, \boldsymbol{\Psi}_{\mathbf{Y}_k}\}_{k=1}^4$ and $\mathbf{M}_1$, as well as the slopes in $\boldsymbol{B}_1^*$ and $\boldsymbol{B}_3^*$, are the same of Scenario A. The other mean matrices are obtained by adding a number c to

each element of the corresponding mean matrices used for Scenario A_1. In detail, c is equal to -5, 5 and -10 for $\mathbf{M}_2$, $\mathbf{M}_3$ and $\mathbf{M}_4$, respectively. The intercept column of all the mixture components is equal to $(7, 2, 5)^\top$, while the slopes in $\boldsymbol{B}_2^*$ and $\boldsymbol{B}_4^*$ are all multiplied by -1, with respect to those used in Scenario A_1. Finally, two sample sizes, i.e. $N = 200$ and $N = 500$, are considered within each scenario.

The MVN-CWM is fitted directly to each generated data set with $G = 4$, and the bias and the mean squared error (MSE) of the parameter estimates are computed. For the sake of brevity, and as also supported by the existing CWM literature (see, e.g. Punzo, 2014; Ingrassia *et al.*, 2015; Punzo and McNicholas, 2017a), the attention is only focused on the $\left\{\boldsymbol{B}_k^*\right\}_{k=1}^K$. Similarly to Section 5.4.2, the label switching issue is controlled by analyzing the overall estimated parameters on each generated data set, in order to properly identify each mixture component.

Table 6.2 shows the estimated bias and MSE of the parameter estimates for Scenario A_1, over one hundred replications for each sample size N, after fitting the MVN-CWM with $G = 4$. The same is illustrated in Table 6.3 for Scenario B_1. The first and most immediate result is that the biases and the MSE assume very small values in both scenarios. This is particularly important for Scenario B_1, because of the presence of overlap between the mixture components. Additionally, within in each scenario, an increase in the sample size leads to a rough improvement of the parameter estimates, and it systematically reduces the MSE.

TABLE 6.2: Estimated bias and MSE of the $\left\{\boldsymbol{B}_k^*\right\}_{k=1}^K$, over one hundred replications for each sample size N, under Scenario A_1.

		$N = 200$	$N = 500$
Group 1	Bias	$\begin{pmatrix} 0.032 & -0.001 & 0.005 & 0.002 \\ -0.025 & -0.010 & -0.008 & -0.004 \\ -0.028 & 0.006 & -0.014 & -0.004 \end{pmatrix}$	$\begin{pmatrix} 0.001 & -0.003 & -0.002 & 0.007 \\ -0.033 & 0.002 & -0.007 & 0.004 \\ 0.003 & -0.004 & -0.002 & 0.005 \end{pmatrix}$
	MSE	$\begin{pmatrix} 0.343 & 0.011 & 0.018 & 0.014 \\ 0.337 & 0.010 & 0.017 & 0.018 \\ 0.365 & 0.011 & 0.020 & 0.016 \end{pmatrix}$	$\begin{pmatrix} 0.111 & 0.004 & 0.006 & 0.007 \\ 0.114 & 0.004 & 0.006 & 0.006 \\ 0.104 & 0.004 & 0.006 & 0.005 \end{pmatrix}$
Group 2	Bias	$\begin{pmatrix} 0.039 & -0.004 & 0.001 & -0.001 \\ -0.005 & -0.004 & 0.002 & 0.006 \\ -0.004 & -0.008 & -0.004 & 0.009 \end{pmatrix}$	$\begin{pmatrix} 0.001 & -0.003 & -0.002 & -0.001 \\ 0.019 & -0.000 & -0.000 & -0.000 \\ 0.042 & -0.004 & -0.002 & -0.001 \end{pmatrix}$
	MSE	$\begin{pmatrix} 0.252 & 0.003 & 0.003 & 0.004 \\ 0.204 & 0.002 & 0.003 & 0.004 \\ 0.170 & 0.002 & 0.003 & 0.005 \end{pmatrix}$	$\begin{pmatrix} 0.084 & 0.001 & 0.001 & 0.001 \\ 0.089 & 0.001 & 0.001 & 0.001 \\ 0.099 & 0.001 & 0.001 & 0.001 \end{pmatrix}$
Group 3	Bias	$\begin{pmatrix} 0.005 & -0.008 & 0.005 & 0.001 \\ -0.020 & -0.010 & -0.002 & -0.000 \\ -0.051 & -0.008 & -0.003 & -0.001 \end{pmatrix}$	$\begin{pmatrix} 0.002 & -0.000 & -0.001 & -0.000 \\ 0.055 & 0.001 & 0.004 & 0.001 \\ 0.018 & 0.003 & 0.002 & -0.001 \end{pmatrix}$
	MSE	$\begin{pmatrix} 0.229 & 0.005 & 0.002 & 0.003 \\ 0.244 & 0.005 & 0.002 & 0.003 \\ 0.235 & 0.006 & 0.002 & 0.004 \end{pmatrix}$	$\begin{pmatrix} 0.104 & 0.002 & 0.001 & 0.001 \\ 0.122 & 0.002 & 0.001 & 0.001 \\ 0.111 & 0.002 & 0.001 & 0.001 \end{pmatrix}$
Group 4	Bias	$\begin{pmatrix} 0.097 & -0.008 & 0.011 & -0.005 \\ 0.027 & -0.006 & 0.006 & 0.002 \\ -0.017 & -0.005 & 0.006 & 0.005 \end{pmatrix}$	$\begin{pmatrix} -0.041 & 0.003 & 0.001 & -0.002 \\ -0.045 & 0.003 & 0.005 & -0.002 \\ -0.006 & 0.002 & 0.003 & -0.004 \end{pmatrix}$
	MSE	$\begin{pmatrix} 0.412 & 0.003 & 0.005 & 0.004 \\ 0.412 & 0.003 & 0.007 & 0.004 \\ 0.397 & 0.003 & 0.006 & 0.004 \end{pmatrix}$	$\begin{pmatrix} 0.242 & 0.001 & 0.001 & 0.001 \\ 0.200 & 0.001 & 0.001 & 0.001 \\ 0.209 & 0.001 & 0.002 & 0.001 \end{pmatrix}$

Other aspects that are evaluated consist in the analysis of the classification produced by the MVN-CWM, as well as the capability of the BIC in identifying the true number of groups in the data. For this reason, under each scenario, the MVN-CWM

TABLE 6.3: Estimated bias and MSE of the $\{\boldsymbol{B}_k^*\}_{k=1}^K$, over one hundred replications for each sample size N, under Scenario B_1.

		$N = 200$	$N = 500$
Group 1	Bias	$\begin{pmatrix} -0.058 & -0.011 & -0.015 & 0.015 \\ -0.037 & -0.008 & -0.011 & 0.018 \\ -0.082 & -0.002 & -0.027 & 0.010 \end{pmatrix}$	$\begin{pmatrix} 0.052 & -0.001 & 0.008 & -0.002 \\ 0.034 & -0.001 & 0.004 & 0.002 \\ -0.018 & 0.000 & -0.006 & 0.001 \end{pmatrix}$
	MSE	$\begin{pmatrix} 0.368 & 0.014 & 0.018 & 0.021 \\ 0.410 & 0.014 & 0.021 & 0.022 \\ 0.361 & 0.011 & 0.019 & 0.022 \end{pmatrix}$	$\begin{pmatrix} 0.118 & 0.004 & 0.007 & 0.006 \\ 0.117 & 0.004 & 0.007 & 0.007 \\ 0.124 & 0.004 & 0.006 & 0.007 \end{pmatrix}$
Group 2	Bias	$\begin{pmatrix} -0.046 & 0.002 & -0.013 & -0.001 \\ -0.037 & 0.005 & 0.001 & 0.001 \\ -0.013 & 0.008 & 0.004 & 0.005 \end{pmatrix}$	$\begin{pmatrix} -0.014 & 0.006 & -0.002 & 0.005 \\ -0.030 & 0.004 & -0.006 & 0.008 \\ -0.008 & 0.000 & -0.002 & 0.009 \end{pmatrix}$
	MSE	$\begin{pmatrix} 0.046 & 0.003 & 0.006 & 0.004 \\ 0.046 & 0.004 & 0.003 & 0.005 \\ 0.043 & 0.004 & 0.004 & 0.005 \end{pmatrix}$	$\begin{pmatrix} 0.015 & 0.001 & 0.001 & 0.002 \\ 0.013 & 0.001 & 0.001 & 0.001 \\ 0.013 & 0.001 & 0.001 & 0.002 \end{pmatrix}$
Group 3	Bias	$\begin{pmatrix} 0.035 & -0.017 & 0.011 & 0.016 \\ 0.011 & -0.006 & 0.012 & 0.008 \\ 0.023 & -0.010 & 0.005 & 0.010 \end{pmatrix}$	$\begin{pmatrix} 0.007 & -0.004 & 0.002 & -0.000 \\ 0.015 & -0.004 & 0.001 & 0.000 \\ 0.028 & -0.005 & 0.004 & -0.000 \end{pmatrix}$
	MSE	$\begin{pmatrix} 0.078 & 0.006 & 0.003 & 0.003 \\ 0.073 & 0.005 & 0.003 & 0.003 \\ 0.080 & 0.005 & 0.002 & 0.003 \end{pmatrix}$	$\begin{pmatrix} 0.027 & 0.002 & 0.001 & 0.001 \\ 0.025 & 0.002 & 0.001 & 0.001 \\ 0.030 & 0.002 & 0.001 & 0.001 \end{pmatrix}$
Group 4	Bias	$\begin{pmatrix} 0.039 & 0.002 & 0.002 & -0.003 \\ 0.005 & -0.001 & -0.004 & -0.007 \\ -0.051 & -0.003 & 0.004 & 0.004 \end{pmatrix}$	$\begin{pmatrix} -0.043 & 0.004 & 0.001 & -0.003 \\ -0.014 & -0.000 & 0.003 & 0.002 \\ -0.008 & -0.002 & 0.002 & 0.003 \end{pmatrix}$
	MSE	$\begin{pmatrix} 0.147 & 0.003 & 0.005 & 0.004 \\ 0.160 & 0.006 & 0.007 & 0.006 \\ 0.132 & 0.003 & 0.007 & 0.006 \end{pmatrix}$	$\begin{pmatrix} 0.061 & 0.001 & 0.002 & 0.002 \\ 0.060 & 0.001 & 0.002 & 0.002 \\ 0.069 & 0.001 & 0.002 & 0.002 \end{pmatrix}$

is fitted to the generated data sets for $K \in \{1, 2, 3, 4, 5\}$, and the results are reported in Table 6.4. It is easy to see that in scenario A_1, a perfect classification is always

TABLE 6.4: Average $\overline{\text{ARI}}$ and $\bar{\epsilon}$, along with the number of times in which the correct K is selected by the BIC, over one hundred replications for each sample size N, under both scenarios.

	$N = 200$			$N = 500$		
	$\overline{\text{ARI}}$	$\bar{\epsilon}$	Correct K	$\overline{\text{ARI}}$	$\bar{\epsilon}$	Correct K
Scenario A_1	1.00	0.00%	100	1.00	0.00%	100
Scenario B_1	0.91	3.04%	99	0.92	2.71%	100

obtained, despite of the considered sample size. Furthermore, the BIC regularly detects the true number of groups in the data. In scenario B_1, because of the presence of overlap, the $\overline{\text{ARI}}$ assumes lower but in any case good values. Relatedly, the percentage of misclassified units stands at around the 3% for both sample sizes. About the BIC, also in this case it correctly identifies the underlying group structure, with only one exception when $N = 200$.

The last aspect evaluated in this study deals with the initialization strategy. In detail, Table 6.5 shows the number of times each strategy for the $\left\{z_i^{(1)}\right\}_{i=1}^N$ produces the highest log-likelihood at convergence, within each scenario and for both sample sizes. If multiple strategies lead to the same optimal solution, they are all counted. The initial K random matrices for $\boldsymbol{\Psi}_{\mathbf{X}_k}^{(1)}$ and $\boldsymbol{\Psi}_{\mathbf{Y}_k}^{(1)}$ are assumed to be the same. The

first result highlights the importance of considering multiple initialization strategies, since none of them are preferred in all the generated data sets. However, the random strategy is quite close to this target, since it only fails in 3 data sets under scenario B_1. Very similar performances are obtained when the mixture initialization is used. Conversely, the k-means strategy has the worst performances, even if it produces the best solution in approximately the 80% of the data sets.

TABLE 6.5: Number of times indicating which of the initializations for the $\left\{z_i^{(1)}\right\}_{i=1}^{N}$ produced the highest log-likelihood at convergence, over one hundred replications for each sample size N, under both scenarios.

	$N = 200$			$N = 500$		
	Random	k-means	Mixture	Random	k-means	Mixture
Scenario A_1	100	77	98	100	79	95
Scenario B_1	97	74	87	100	83	100

6.4.2 Simulation 2: A comparison between the MVN-CWM and the MVN-FMR

In this study, the MVN-CWM is compared to the MVN-FMR. In detail, three scenarios with $N = 200$, $p = 2$, $q = 3$ and $r = 4$ are considered, and in each of them thirty data sets from a MVN-CWM with $K = 2$ are generated. The first scenario (called "Scenario A_2") consists of two groups that differ only for the intercepts and the covariance matrices. This means that they have totally overlapped mean matrices, which should make the distribution of the covariates $p_k(\mathbf{X})$ not important for clustering. The parameters used to generate the data sets are displayed in Table 6.6.

In the second scenario (called "Scenario B_2") the two groups have the same $\boldsymbol{B}_k^*$ and π_k. The parameters used to generate the data sets are the same of "Scenario A_2", with the only difference that a value of 5 is added to each element of $\mathbf{M}_2$, and that $\boldsymbol{B}_2^*$ assumes the same values of $\boldsymbol{B}_1^*$. Finally, in the third scenario (called "Scenario C_2"), the two groups have only the same slopes and π_k. Here, with respect to the parameters used under Scenario B_2, the only difference is in the intercepts vectors which are $(-3, -4)^\top$ and $(-7, -8)^\top$, for the first and the second group, respectively.

The MVN-CWM and the MVN-FMR are then fitted to the data sets of each scenario for $K \in \{1, 2, 3\}$, and the results in terms of model selection and clustering are reported in Table 6.7. It is possible to see that in Scenario A_2, the BIC properly select two groups for both models and the classifications produced are on average perfect. Therefore, even if the two groups have the same means and are strongly overlapped, the MVN-CWM seems able to correctly identify the true underlying grouping structure. However, under such scenario the MVN-FMR should be preferred, since the distribution of the covariates $p_k(\mathbf{X})$ is not useful for clustering, and it is more parsimonious than the MVN-CWM. Conversely, Scenarios B_2 and C_2 represent typical examples of the usefulness of $p_k(\mathbf{X})$. In detail, the BIC always identifies just one group under both scenarios for the MVN-FMR, with clear consequences in terms of the classification produced. Notice that, even if the MVN-FMR had been fitted directly with $K = 2$, the resulting classifications would lead to almost identical $\overline{\text{ARI}}$ and $\bar{e}$ for Scenario B_2, and slightly better performance for Scenario C_2, since $\overline{\text{ARI}} = 0.15$ and

TABLE 6.6: Parameters used to generate the simulated data sets under Scenario A_2.

Parameter	Group 1	Group 2
π_g	0.50	0.50
$\mathbf{M}_k$	$\begin{pmatrix} 1.00 & 2.00 & 2.00 & 0.00 \\ -1.00 & 1.00 & 1.00 & 2.00 \\ 0.00 & 2.00 & 2.00 & 1.00 \end{pmatrix}$	$\begin{pmatrix} 1.00 & 2.00 & 2.00 & 0.00 \\ -1.00 & 1.00 & 1.00 & 2.00 \\ 0.00 & 2.00 & 2.00 & 1.00 \end{pmatrix}$
$\mathbf{\Sigma}_{\mathbf{X}_k}$	$\begin{pmatrix} 1.00 & 0.50 & 0.25 \\ 0.50 & 1.00 & 0.50 \\ 0.25 & 0.50 & 1.00 \end{pmatrix}$	$\begin{pmatrix} 2.00 & 0.40 & 0.08 \\ 0.40 & 0.20 & 0.40 \\ 0.08 & 0.40 & 2.00 \end{pmatrix}$
$\mathbf{\Psi}_{\mathbf{X}_k}$	$\begin{pmatrix} 1.70 & 0.85 & 0.42 & 0.21 \\ 0.85 & 1.70 & 0.85 & 0.42 \\ 0.42 & 0.85 & 1.70 & 0.85 \\ 0.21 & 0.42 & 0.85 & 1.70 \end{pmatrix}$	$\begin{pmatrix} 1.00 & 0.50 & 0.25 & 0.12 \\ 0.50 & 1.00 & 0.50 & 0.25 \\ 0.25 & 0.50 & 1.00 & 0.50 \\ 0.12 & 0.25 & 0.50 & 1.00 \end{pmatrix}$
$\boldsymbol{B}_k^*$	$\begin{pmatrix} 2.00 & 1.00 & 1.00 & -1.00 \\ 3.00 & 1.00 & -1.00 & 1.00 \end{pmatrix}$	$\begin{pmatrix} -7.00 & 1.00 & 1.00 & -1.00 \\ -8.00 & 1.00 & -1.00 & 1.00 \end{pmatrix}$
$\mathbf{\Sigma}_{\mathbf{Y}_k}$	$\begin{pmatrix} 1.00 & 0.50 \\ 0.50 & 1.00 \end{pmatrix}$	$\begin{pmatrix} 2.00 & 1.20 \\ 1.20 & 2.00 \end{pmatrix}$
$\mathbf{\Psi}_{\mathbf{Y}_k}$	$\begin{pmatrix} 2.00 & 1.00 & 0.50 & 0.25 \\ 1.00 & 2.00 & 1.00 & 0.50 \\ 0.50 & 1.00 & 2.00 & 1.00 \\ 0.25 & 0.50 & 1.00 & 2.00 \end{pmatrix}$	$\begin{pmatrix} 1.70 & 0.75 & 0.38 & 0.19 \\ 0.75 & 1.50 & 0.75 & 0.38 \\ 0.38 & 0.75 & 1.50 & 0.75 \\ 0.19 & 0.39 & 0.75 & 1.50 \end{pmatrix}$

$\overline{\epsilon} = 32.48\%$. This highlights that regardless of the BIC, the MVN-FMR is not able to properly model such data structures.

TABLE 6.7: Average $\overline{\text{ARI}}$ and $\overline{\eta}$, along with the number of times in which the correct G is selected by the BIC, over thirty replications in each scenario, for the MVN-CWM and MVN-FMR.

	MVN-CWM			MVN-FMR		
	$\overline{\text{ARI}}$	$\overline{\epsilon}$	Correct K	$\overline{\text{ARI}}$	$\overline{\epsilon}$	Correct K
Scenario A_2	1.00	0.00%	100	1.00	0.00%	100
Scenario B_2	0.99	0.03%	100	0.00	47.22%	0
Scenario C_2	1.00	0.01%	100	0.00	47.18%	0

6.5 Real data applications

6.5.1 Data description

The first data set (herein called "Education data") contains career indicators from the Italian National Agency for the Evaluation of Universities and Research Institutes. For this application, the following $p = 2$ responses, that measure the level of completion of studies by students, are considered: (i) the percentage of students that graduate within $T + 1$ years ($\boldsymbol{Y1}$) and (ii) the percentage of students that drop after $T + 1$ years ($\boldsymbol{Y2}$), where T is the normal duration of the study program. Moreover, the following $q = 2$ covariates, that may be helpful in explaining this progress, are taken into account: (i) the percentage of course credits earned in the first year over

the total to be achieved (**X1**) and (ii) the percentage of students that have earned at least 40 course credits during the solar year (**X2**).

For sake of simplicity, hereafter these variables will be referred by using the corresponding name in brackets. All the measurements refer to $N = 75$ study programs in the non-telematic Italian universities, over $r = 3$ years. Each study program is measured at national level, i.e. it is the average value of all the study programs of the same type across all the country, for the reference period. There are $K = 2$ groups in the data, i.e. $N_1 = 33$ bachelor's degrees and $N_2 = 42$ master's degrees.

The second data set (herein called "Insurance data") contains variables related to the non-life insurance market in Italy. It is contained in the **splm** package (Millo and Piras, 2012) under the name *Insurance*. This data set was introduced by Millo and Carmeci (2011) and refers to $N = 103$ Italian provinces in the years 1998–2002 ($r = 5$). In this application, the following $p = 2$ responses, that are related to the consumption and the presence of insurance products in the market, are considered: (i) the real per-capita non-life premiums in 2000 euros (PPCD) and (ii) the density of insurance agencies per 1000 inhabitants (AGEN). Then, the following $q = 3$ financial covariates are selected: (i) the real per-capita GDP (RGDP), (ii) the real per-capita bank deposits (BANK) and (iii) the real interest rate on lending to families and small enterprises (RIRS). The focus is on this subset of covariates because: (1) they are almost regularly used in the literature, and their relevant effects on the consumption or development of insurance products has been widely discussed (see the references in Millo and Carmeci, 2011, for further details). Indeed, they are commonly used as proxies for income and general level of economic activity (RGDP), stock of wealth (BANK) and opportunity cost of allocate funds in insurance policies (RIRS); (2) avoid an excessive parametrization of the models.

Differently from the previous data set, there is not a classification of the data. Hoverer, the findings of Millo and Carmeci (2011) seem to suggest the presence of two groups in the data. Specifically, they underline the existence of two macro areas, namely the Central-Northern Italy, characterized by an insurance penetration level relatively close to the European averages, and the Southern-Insular Italy, where a general economic underdevelopment has long been standing as a fundamental social and political problem. The presence of groups in the data can be confirmed by looking at the sampling distribution of each covariate, since $p(\mathbf{X})$ is used in (6.5). In detail, they are reported in Figure 6.1, for the full 5-year data, as done by Melnykov and Zhu (2019). The bimodality in all these histograms seems to confirm the existence of two groups in the data, validating the conclusions of Millo and Carmeci (2011).

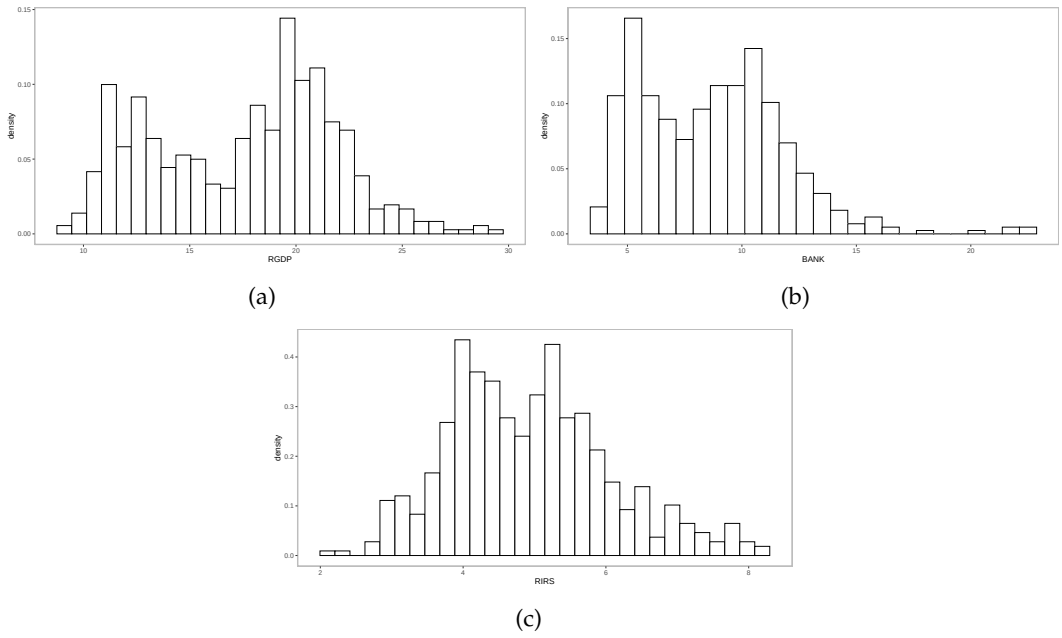
6.5.2 Results

In both data sets, the MVN-CWM and the MVN-FMR are fitted to for $K \in \{1, 2, 3\}$. When the Education data are considered, the results are reported in Table 6.8. The

TABLE 6.8: Education data: ARI and η for the MVN-CWM and MVN-FMR selected by the BIC.

Model	K	ARI	η
MVN-CWM	2	1.00	0.00%
MVN-FMR	3	0.88	6.67%

BIC selects a two-component MVN-CWM that yields a perfect classification of the



(a) (b) (c)

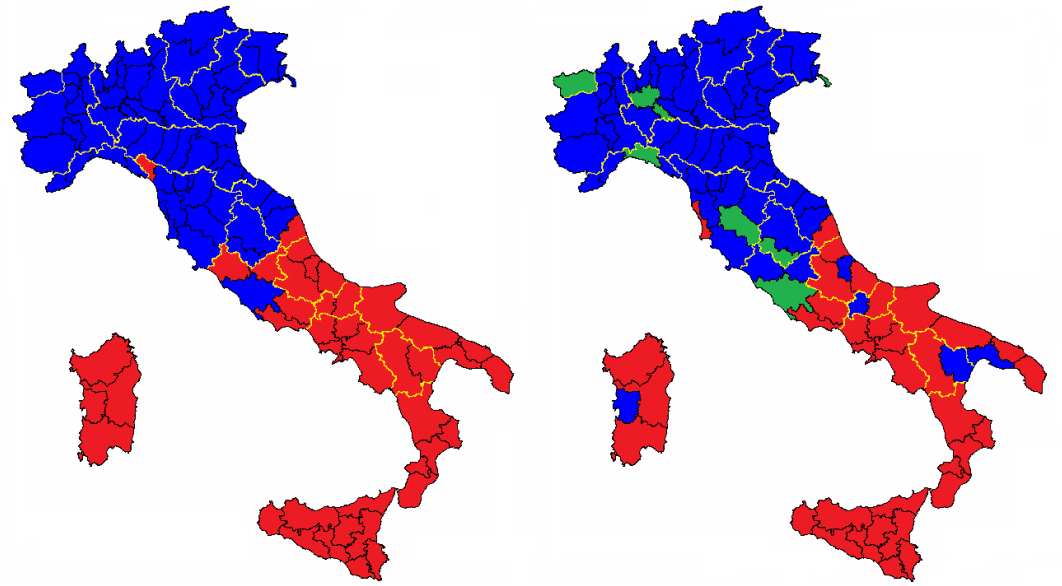
FIGURE 6.1: Sampling distributions of the covariates.

data. On the contrary, a three-component MVN-FMR is chosen by the BIC, with a 6.67% of misclassified units. Note that, even if the MVN-FMR was directly fitted with $K = 2$, this will produce a similar classification with an ARI of 0.89. Therefore, this results highlight how the CWM is able to completely recognize the underlying group structure, differently from the MVN-FMR.

When the Insurance data are analyzed, the BIC selects a two-component MVN-CWM and a three-component MVN-FMR, respectively. Considering that we do not have a classification of the data, we cannot compute the ARI or η. To give a representation of the classifications produced by the competing models, they are illustrated in Figure 6.2 by using the Italian political map. Specifically, the Italian regions are bordered in yellow (islands excluded), while the internal provinces are delimited with the black lines and colored according to the estimated group memberships, both for the MVN-CWM and the MVN-FMR.
Here, it is possible to see that the classification produced by the MVN-CWM appears in line with the findings of Millo and Carmeci (2011), with a clear separation between the Central-Northern Italy and the Southern-Insular Italy. Furthermore, with the exclusion of three cases, all the provinces belonging to the same region are clustered together. The only exceptions concern the province of Rome (in the Lazio region), which due to its social-economic development is reasonably assigned to the Central-Northern Italy group, the province of Ascoli-Piceno (in the Marche region) and the province of Massa-Carrara (in the Toscana region). On the contrary, the three groups detected by the MVN-FRM are not supported by the literature and are difficult to interpret, even because they put together provinces spanning all over the country without a straightforward and reasonable justification.

6.6 Conclusion

The first matrix-variate CWM has been introduced in this work. In the MVN-CWM framework, the sets of responses and covariates may be simultaneously observed at



(a) (b)

FIGURE 6.2: Partitions produced by the MVN-CWM (a) and MVN-FMR (b).

different time points or locations. The matrix normal was used for both the cluster-specific distributions of covariates and responses given covariates, and an ECM algorithm for parameter estimation was presented. A simulation study analyzed parameter recovery, the classification performance and the initialization strategy for the proposed model as well as the capability of the BIC to detect the underlying group structure of the data. Furthermore, the scenarios illustrated in the simulated analyses have highlighted how the covariates' distribution can affect the clustering structure and how important can be to take it into account when defining the model. Similar conclusions are produced in the first real data application analyzed, where the MVN-CWM produced a perfect classification, differently from the MVN-FMR model. Lastly, in the second real data analysis, the MVN-CWM seemed to provide a more reliable partition of the Italian provinces, according to the existing literature, than the MVN-FMR model.

Chapter 7

Conclusions and future developments

In this thesis several new finite mixtures models have been proposed and applied to economic, financial and education data sets. With the attempt of summarizing in a nutshell the main results of each chapter, it is possible to say that:

- the thirteen zero-and-one inflated mixture models introduced in Chapter 3 have shown the importance of not using a single distribution (or not fixing *a priori* the number of mixture components to 2) when modeling the loss given default. Additionally, the use of logit-transformed distributions have provided better performances with respect to the traditionally used beta distribution. The family of models proposed have also produced better estimates of the value at risk when compared to other well-established semiparametric and nonparametric approaches.

- the two dichotomous unimodal compound models introduced in Chapter 4 have shown several characteristics that make them attractive for the modelization of the insurance losses distribution. Indeed, they provide a robust alternative compared to the corresponding simple conditional distribution, allow the detection of typical/atypical losses and the definition of the typical/atypical regions. Furthermore, the proposed models have also displayed better fitting and risk measures estimates when compared to other distributions and approaches, among which there are the t-score estimator and the PORT-MO$_p$ method.

- the two matrix-variate mixture models introduced in Chapter 5, based on the two new distributions therein proposed, have provided an alternative to the matrix-variate t mixture models, which can be considered the only matrix-variate model, having elliptical and heavy-tailed mixture components, used for clustering. Because of their heavier-than-normal tails, these models are able to cope with clusters having potential mild outliers in a proper way and may avoid the disruption of the true underlying grouping structure, as shown both in the simulated and real data analyses.

- the matrix normal cluster-weighted model introduced in Chapter 6 has represented the first finite mixture of regression model with random covariates in the matrix-variate literature. The results of the simulated and real data applications have highlighted how the covariates' distribution can affect the clustering structure and how important is to take it into account when defining the model. This is of particular importance since the competing approach (the finite mixture of regression model with fixed covariates) has performed poorly when fitted to the considered data.

All the models introduced in this thesis can be be extended in different ways. Specifically, future works will concern:

- for the zero-and-one inflated mixture models of Chapter 3, it might be interesting to consider them in a nonparametric context by using smoothers for the LGD values on (0,1). Furthermore, covariates are often available along with LGD in real data. Then, another extension would deal with the regression framework, in which the response variable is conditionally distributed according to one of the zero-and-one inflated mixture models. Relatedly, by following the approach of Centoni *et al.* (2019), a concomitant-variable latent-class model could also be implemented. In their work, the authors considered a latent class of beta inflated distributions to assess students' performance of a private Italian university. Therefore, it might be also interesting to apply in that context the other distributions used in the chapter of this thesis.

- for the dichotomous unimodal compound models of Chapter 4, different unimodal hump-shaped distributions (defined on a positive support) could be considered. Additionally, their extension to the multivariate context, where the insurance losses are jointly modeled with the allocated loss adjustment expenses (ALAE; see e.g. Abu Bakar *et al.*, 2015) could be studied. Finally, also in this case, an extension in a regression framework might be implemented, in which the response variable is conditionally distributed according to one of the two dichotomous unimodal compound models.

- for the matrix-variate mixture models of Chapter 5, it might be useful to consider constrained parametrizations of the means and covariance matrices, in order to reduce the number of parameters to be estimated, and introduce parsimony in the models (Sarkar *et al.*, 2020). Furthermore, to accommodate skewness in the data, the two distributions herein defined could also be generalized in order to include of a skewness parameter (Gallaugher and McNicholas, 2018).

- for the CWM in Chapter 6, a first extension could consists in using such model in further application areas. In this work, education and insurance frameworks have been considered, but other possible applications might involve: 1) environmental data, where economic and political variables can be used as regressors, whereas variables measuring the concept of "environmental concern" can be considered as dependent variables. Each variable is measured for a set of countries across several years (Dunlap and Michelson, 2002; Hao, 2016); 2) financial data, where firm related variables are used are regressors, whereas performance measures such as the ROA, ROI and ROE can be used as dependent variables. Also in this case, all the variables are measured over the time for a set of firms (Bou and Satorra, 2018).

  A second extension could concern the development of a procedure for variable selection. Specifically, it would be interesting to evaluate which variables contribute to the clustering and which are not required. This would lead to a parsimonious model and might avoid that noisy variables are inserted in the model. While variable selection procedures in model-based clustering have been widely discussed (see, e.g. Fop *et al.*, 2018), to my knowledge, such task has not yet been treated for any of the cluster weighted model present in the literature. Indeed, we should understand which responses, regressors or both are not useful for clustering, and the approaches mentioned so far cannot be straightforward applied.

References

Abramowitz, M. and Stegun, I. A. (1965). Handbook of mathematical functions with formulas, graphs, and mathematical table. In *US Department of Commerce*. National Bureau of Standards Applied Mathematics series 55.

Abu Bakar, S. A., Hamzah, N. A., Maghsoudi, M., and Nadarajah, S. (2015). Modeling loss data using composite models. *Insurance: Mathematics and Economics*, **61**, 146–154.

Adcock, C., Eling, M., and Loperfido, N. (2015). Skewed distributions in finance and actuarial science: a review. *The European Journal of Finance*, **21**(13-14), 1253–1281.

Ahmad, K. E. and Al-Hussaini, E. K. (1982). Remarks on the non-identifiability of mixtures of distributions. *Annals of the Institute of Statistical Mathematics*, **34**(3), 543–544.

Ahn, S., Kim, J. H. T., and Ramaswami, V. (2012). A new class of models for heavy tailed distributions in finance and insurance risk. *Insurance: Mathematics and Economics*, **51**, 43–52.

Aitkin, M. and Rubin, D. B. (1985). Estimation and hypothesis testing in finite mixture models. *Journal of the Royal Statistical Society: Series B (Methodological)*, **47**(1), 67–75.

Aitkin, M. and Wilson, G. T. (1980). Mixture models, outliers, and the EM algorithm. *Technometrics*, **22**(3), 325–331.

Anderlucci, L., Montanari, A., and Viroli, C. (2014). A matrix-variate regression model with canonical states: An application to elderly danish twins. *Statistica*, **74**(4), 367–381.

Atchison, J. and Shen, S. M. (1980). Logistic-normal distributions: Some properties and uses. *Biometrika*, **67**(2), 261–272.

Baesens, B., Roesch, D., and Scheule, H. (2016). *Credit Risk Analytics: Measurement Techniques, Applications, and Examples in SAS*. John Wiley & Sons.

Bagnato, L. and Punzo, A. (2019). Unconstrained representation of orthogonal matrices with application to common principle components. *arXiv preprint arXiv:1906.00587*.

Banca d'Italia (2001). Principali Risultati della Rilevazione sull'Attività di Recupero dei Crediti. Bollettino di Vigilanza 12.

Basel Committee on Banking Supervision (2006). *International convergence of capital measurement and capital standards: a revised framework*. Bank for International Settlements.

Bastos, J. A. (2010). Forecasting bank loans loss-given-default. *Journal of Banking & Finance*, **34**(10), 2510–2517.

Bellotti, A. (2017). Estimating unbiassed expected loss, with application to consumer credit. Available at SSRN: https://ssrn.com/abstract=2916145 or http://dx.doi.org/10.2139/ssrn.2916145.

Bellotti, T. (2010). A simulation study of basel II expected loss distributions for a portfolio of credit cards. *Journal of Financial Services Marketing*, **14**(4), 268–277.

Bellotti, T. and Crook, J. (2012). Loss given default models incorporating macroeconomic variables for credit cards. *International Journal of Forecasting*, **28**(1), 171–182.

Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business & Economic Statistics*, **19**(4), 465–474.

Bernardi, M., Maruotti, A., and Petrella, L. (2012). Skew mixture models for loss distributions: a bayesian approach. *Insurance: Mathematics and Economics*, **51**(3), 617–623.

Bertsekas, D. P. and Tsitsiklis, J. N. (2008). *Introduction to Probability*, volume 1 of *Athena Scientific optimization and computation series*. Athena Scientific.

Bickerstaff, D. R. (1972). Automobile collision deductibles and repair cost groups: The lognormal model. *PCAS LIX*, pages 68–84.

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

Boldea, O. and Magnus, J. R. (2009). Maximum likelihood estimation of the multivariate normal mixture model. *Journal of the American Statistical Association*, **104**(488), 1539–1549.

Boris Choy, S. and Chan, J. S. (2008). Scale mixtures distributions in statistical modelling. *Australian & New Zealand Journal of Statistics*, **50**(2), 135–146.

Bou, J. C. and Satorra, A. (2018). Univariate versus multivariate modeling of panel data: Model specification and goodness-of-fit testing. *Organizational Research Methods*, **21**(1), 150–196.

Bouguila, N. and Fan, W. (2020). *Mixture Models and Applications*. Springer, Cham.

Brazauskas, V. and Kleefeld, A. (2016). Modeling severity and measuring tail risk of norwegian fire claims. *North American Actuarial Journal*, **20**(1), 1–16.

Brilhante, M. F., Gomes, M. I., and Pestana, D. (2013). A simple generalisation of the hill estimator. *Computational Statistics & Data Analysis*, **57**(1), 518–535.

Burnecki, K., Kukla, G., and Weron, R. (2000). Property insurance loss distributions. *Physica A: Statistical Mechanics and its Applications*, **287**(1), 269–278.

Burnecki, K., Misiorek, A., and Weron, R. (2005). Loss distributions. In *Statistical Tools for Finance and Insurance*, pages 289–317. Springer.

Calabrese, R. (2010). Regression for recovery rates with both continuous and discrete characteristics. In *Proceedings of the Italian Statistical Society Conference, 2010, Padua*.

Calabrese, R. (2014a). Downturn loss given default: Mixture distribution estimation. *European Journal of Operational Research*, **237**(1), 271–277.

Calabrese, R. (2014b). Predicting bank loan recovery rates with a mixed continuous-discrete model. *Applied Stochastic Models in Business and Industry*, **30**(2), 99–114.

Calabrese, R. and Zenga, M. (2008). Measuring loan recovery rate: methodology and empirical evidence. *Statistica & Applicazioni*, **6**(2), 193–214.

Calabrese, R. and Zenga, M. (2010). Bank loan recovery rates: Measuring and non-parametric density estimation. *Journal of Banking & Finance*, **34**(5), 903–911.

Centoni, M., Del Panta, V., Maruotti, A., and Raponi, V. (2019). Concomitant-variable latent-class beta inflated models to assess students' performance: An italian case study. *Social Indicators Research*, **146**(1-2), 7–18.

Cerioli, A., Farcomeni, A., and Riani, M. (2019). Wild adaptive trimming for robust estimation and cluster analysis. *Scandinavian Journal of Statistics*, **46**(1), 235–256.

Chen, R. and Wang, Z. (2013). Curve fitting of the corporate recovery rates: The comparison of beta distribution estimation and kernel density estimation. *PloS one*, **8**(7), e68238.

Chen, S. X. (1999). Beta kernel estimators for density functions. *Computational Statistics & Data Analysis*, **31**(2), 131–145.

Cooray, K. and Ananda, M. M. A. (2005). Modeling actuarial data with a composite lognormal-Pareto model. *Scandinavian Actuarial Journal*, **2005**(5), 321–334.

Crawford, S. L. (1994). An application of the laplace method to finite mixture distributions. *Journal of the American Statistical Association*, **89**(425), 259–267.

Croissant, Y. and Millo, G. (2019). ***pder**: Panel Data Econometrics with* . R package version 1.0-1.

Dang, U. J., Browne, R. P., and McNicholas, P. D. (2015). Mixtures of multivariate power exponential distributions. *Biometrics*, **71**(4), 1081–1089.

Dang, U. J., Punzo, A., McNicholas, P. D., Ingrassia, S., and Browne, R. P. (2017). Multivariate response and parsimony for gaussian cluster-weighted models. *Journal of Classification*, **34**(1), 4–34.

Dayton, C. M. and Macready, G. B. (1988). Concomitant-variable latent-class models. *Journal of the american statistical association*, **83**(401), 173–178.

de Oliveira Jr, M. R., Louzada, F., de Araujo Pereira, G. H., Moreira, F. F., and Calabrese, R. (2015). Inflated mixture models: Applications to multimodality in loss given default. Available at SSRN: https://ssrn.com/abstract=2634919 or http://dx.doi.org/10.2139/ssrn.2634919.

Delignette-Muller, M. L. and Dutang, C. (2015). fitdistrplus: An R package for fitting distributions. *Journal of Statistical Software*, **64**(4), 1–34.

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

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

Dietsch, M. and Petey, J. (2004). Should SME exposures be treated as retail or corporate exposures? a comparative analysis of default probabilities and asset correlations in French and German SMEs. *Journal of Banking & Finance*, **28**(4), 773–788.

Doğru, F. Z., Bulut, Y. M., and Arslan, O. (2016). Finite mixtures of matrix variate t distributions. *Gazi University Journal of Science*, **29**(2), 335–341.

Duda, R. O., Hart, P. E., and Stork, D. G. (2012). *Pattern classification*. John Wiley & Sons.

Düllmann, K. and Gehde-Trapp, M. (2004). Systematic risk in recovery rates: an empirical analysis of us corporate credit exposures. Bundesbank Series 2 Discussion Paper No. 2004,02. Available at SSRN: https://ssrn.com/abstract=2793954.

Dunlap, R. E. and Michelson, W. (2002). *Handbook of environmental sociology*. Greenwood Publishing Group.

Dutang, C. and Charpentier, A. (2016). ***CASdatasets****: Insurance datasets (Official website)*. Version 1.0-6 (2016-05-28).

Dutilleul, P. (1999). The mle algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation*, **64**(2), 105–123.

Eling, M. (2012). Fitting insurance claims to skewed distributions: Are the skew-normal and skew-student good models? *Insurance: Mathematics and Economics*, **51**, 239–248.

Eling, M., Farinelli, S., Rossello, D., and Tibiletti, L. (2010). Skewness in hedge funds returns: classical skewness coefficients vs azzalini's skewness parameter. *International Journal of Managerial Finance*, **6**(4), 290–304.

Embrechts, P. and Schmidli, H. (1994). Modelling of extremal events in insurance and finance. *Zeitschrift für Operations Research*, **39**(1), 1–34.

Emmer, S., Kratz, M., and Tasche, D. (2015). What is the best risk measure in practice? a comparison of standard measures. *Journal of Risk*, **18**(2).

Fabián, Z. (2006). Johnson point and johnson variance. *Proc. Prague Stochastics 2006*, pages 354–363.

Fabián, Z. (2007). Parametric estimation using generalized moment method. Technical report, Research report 1014, Inst. of Computer Sciences.

Fabián, Z. (2010). Scalar score function and score correlation. Technical report, Research report 1077, Inst. of Computer Sciences.

Figueiredo, F., Gomes, M. I., and Henriques-Rodrigues, L. (2017). Value-at-risk estimation and the port mean-of-order-p methodology. *Revstat*, **15**(2), 187–204.

Fletcher, R. (2013). *Practical methods of optimization*. John Wiley & Sons.

Fop, M., Murphy, T. B., *et al.* (2018). Variable selection methods for model-based clustering. *Statistics Surveys*, **12**, 18–65.

Friedman, C. and Sandow, S. (2003). Recovery rates: Ultimate recoveries. *Risk-London-Risk Magazine Limited*, **16**(8), 69–73.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer Science & Business Media, New York.

Furman, E. (2008). On a multivariate gamma distribution. *Statistics & Probability Letters*, **78**(15), 2353–2360.

Gallaugher, M. P. B. and McNicholas, P. D. (2018). Finite mixtures of skewed matrix variate distributions. *Pattern Recognition*, **80**, 83–93.

Gershenfeld, N. (1997). Nonlinear inference and cluster-weighted modeling. *Annals of the New York Academy of Sciences*, **808**(1), 18–24.

Gershenfeld, N., Schoner, B., and Metois, E. (1999). Cluster-weighted modelling for time-series analysis. *Nature*, **397**(6717), 329.

Ghalanos, A. (2015). ***rugarch**: Univariate GARCH Models*. Version 1.3-6 (2015-08-16).

Gomes, M. I., Henriques-Rodrigues, L., and Manjunath, B. (2016). Mean-of-order-p location-invariant extreme value index estimation. *Revstat*, **14**(3), 273–296.

Gouriéroux, C. and Monfort, A. (2006). (Non) consistency of the beta kernel estimator for recovery rate distribution. *CREST-DP*, **31**, 1–27.

Grunert, J. and Weber, M. (2009). Recovery rates of commercial lending: Empirical evidence for german companies. *Journal of Banking & Finance*, **33**(3), 505–513.

Gupta, A. K. and Nagar, D. K. (1999). *Matrix variate distributions*, volume 104. CRC Press.

Gupta, A. K., Varga, T., and Bodnar, T. (2013). *Elliptically contoured models in statistics and portfolio theory*. Springer, New York.

Gupton, G. M. and Stein, R. M. (2005). LossCalc v2: Dynamic prediction of LGD. *Moodys KMV Investors Services*.

Gürtler, M. and Hibbeln, M. (2013). Improvements in loss given default forecasts for bank loans. *Journal of Banking & Finance*, **37**(7), 2354–2366.

Hagmann, M., Renault, O., Scaillet, O., *et al.* (2005). Estimation of recovery rate densities: non-parametric and semi-parametric approaches versus industry practice. *The Next Challenge in Credit Risk Management*, pages 323–346.

Hao, F. (2016). A panel regression study on multiple predictors of environmental concern for 82 countries across seven years. *Social Science Quarterly*, **97**(5), 991–1004.

Holzmann, H., Munk, A., and Gneiting, T. (2006). Identifiability of finite mixtures of elliptical distributions. *Scandinavian journal of statistics*, **33**(4), 753–763.

Huang, X. and Oosterlee, C. W. (2008). *Generalized beta regression models for random loss-given-default*. Delft University of Technology.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, **2**(1), 193–218.

Ingrassia, S. (1991). Mixture decomposition via the simulated annealing algorithm. *Applied stochastic models and data analysis*, **7**(4), 317–325.

Ingrassia, S. (1992). A comparison between the simulated annealing and the em algorithms in normal mixture decompositions. *Statistics and Computing*, **2**(4), 203–211.

Ingrassia, S. and Punzo, A. (2016). Decision boundaries for mixtures of regressions. *Journal of the Korean Statistical Society*, **45**(2), 295–306.

Ingrassia, S. and Punzo, A. (2019). Cluster validation for mixtures of regressions via the total sum of squares decomposition. *Journal of Classification*, pages 1–22.

Ingrassia, S., Minotti, S. C., and Vittadini, G. (2012). Local statistical modeling via a cluster-weighted approach with elliptical distributions. *Journal of classification*, **29**(3), 363–401.

Ingrassia, S., Minotti, S. C., and Punzo, A. (2014). Model-based clustering via linear cluster-weighted models. *Computational Statistics & Data Analysis*, **71**, 159–182.

Ingrassia, S., Punzo, A., Vittadini, G., and Minotti, S. C. (2015). The generalized linear mixed cluster-weighted model. *Journal of Classification*, **32**(1), 85–113.

Jeon, Y. and Kim, J. H. (2013). A gamma kernel density estimation for insurance loss data. *Insurance: Mathematics and Economics*, **53**(3), 569–579.

Kazemi, R. and Noorizadeh, M. (2015). A comparison between skew-logistic and skew-normal distributions. *Matematika*, **31**(1), 15–24.

Kellison, J. B. and Brockett, P. (2003). Check the score: Credit scoring and insurance losses: Is there a connection? *Texas Business Review*.

Kupiec, P. (1995). Techniques for verifying the accuracy of risk measurement models. *The J. of Derivatives*, **3**(2).

Lee, J., Wang, J., and Zhang, J. (2009). The relationship between average asset correlation and default probability. *Moody's KMV Company*.

Lee, S. X. and McLachlan, G. J. (2019). *Scale Mixture Distribution*, pages 1–16. American Cancer Society.

Li, P., Qi, M., Zhang, X., and Zhao, X. (2016). Further investigation of parametric loss given default modeling. *Journal of Credit Risk*, **12**(4), 17–47.

Liu, C. and Rubin, D. B. (1994). The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. *Biometrika*, **81**(4), 633–648.

Longin, F. (2016). *Extreme events in finance: A handbook of extreme value theory and its applications*. John Wiley & Sons.

Loterman, G., Brown, I., Martens, D., Mues, C., and Baesens, B. (2012). Benchmarking regression algorithms for loss given default modeling. *International Journal of Forecasting*, **28**(1), 161–170.

MacDonald, I. L. (2014). Numerical maximisation of likelihood: A neglected alternative to em? *International Statistical Review*, **82**(2), 296–308.

MacKinnon, J. G. (2009). Bootstrap hypothesis testing. *Handbook of computational econometrics*, **183**, 213.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, **57**(3), 519–530.

Mazza, A. and Punzo, A. (2020). Mixtures of multivariate contaminated normal regression models. *Statistical Papers*, **61**(2), 787–822.

Mazza, A., Punzo, A., and Ingrassia, S. (2018). flexcwm: a flexible framework for cluster-weighted models. *J Stat Softw*, **86**(2), 1–30.

McLachlan, G. J. and Krishnan, T. (2007). *The EM Algorithm and Extensions*. Wiley, New York, 2 edition.

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

McNicholas, P. D. (2016). *Mixture Model-Based Classification*. Chapman & Hall/CRC Press, Boca Raton.

Melnykov, V. and Melnykov, I. (2012). Initializing the em algorithm in gaussian mixture models with an unknown number of components. *Computational Statistics & Data Analysis*, **56**(6), 1381–1395.

Melnykov, V. and Zhu, X. (2019). Studying crime trends in the USA over the years 2000–2012. *Advances in Data Analysis and Classification*, **13**(1), 325–341.

Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, **80**, 267–278.

Meng, X.-L. and Van Dyk, D. (1997). The EM algorithm: an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **59**(3), 511–567.

Miller, P. and Töws, E. (2018). Loss given default adjusted workout processes for leases. *Journal of Banking & Finance*, **91**, 189–201.

Millo, G. and Carmeci, G. (2011). Non-life insurance consumption in italy: a sub-regional panel data analysis. *Journal of Geographical Systems*, **13**(3), 273–298.

Millo, G. and Piras, G. (2012). splm: Spatial panel data models in R. *Journal of Statistical Software*, **47**(1), 1–38.

Misra, R. D. (1940). On the stability of crystal lattices. II. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 36, pages 173–182. Cambridge University Press.

Moss, J. and Tveten, M. (2018). *kdensity: Kernel Density Estimation with Parametric Starts and Asymmetric Kernels*. R package version 1.0.0.

Nazemi, A., Pour, F. F., Heidenreich, K., and Fabozzi, F. J. (2017). Fuzzy decision fusion approach for loss-given-default modeling. *European Journal of Operational Research*, **262**(2), 780–791.

Newton, M. A. and Raftery, A. E. (1994). Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society: Series B (Methodological)*, **56**(1), 3–26.

Ospina, R. and Ferrari, S. L. P. (2010). Inflated beta distributions. *Statistical Papers*, **51**(1), 111–126.

Otiniano, C., Rathie, P., and Ozelim, L. (2015). On the identifiability of finite mixture of skew-normal and skew-t distributions. *Statistics & Probability Letters*, **106**, 103–108.

Packová, V. and Brebera, D. (2015). Loss distributions in insurance risk management. *Business Administration*, pages 17–22.

Peel, D. and McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing*, **10**(4), 339–348.

Pigeon, M. and Denuit, M. (2011). Composite lognormal-Pareto model with random threshold. *Scandinavian Actuarial Journal*, **2011**(3), 177–192.

Punzo, A. (2014). Flexible mixture modelling with the polynomial gaussian cluster-weighted model. *Statistical Modelling*, **14**(3), 257–291.

Punzo, A. and Bagnato, L. (2020a). Allometric analysis using the multivariate shifted exponential normal distribution. *Biometrical Journal*.

Punzo, A. and Bagnato, L. (2020b). The multivariate tail-inflated normal distribution and its application in finance. *Journal of Statistical Computation and Simulation*, pages 1–36.

Punzo, A. and McNicholas, P. D. (2016). Parsimonious mixtures of multivariate contaminated normal distributions. *Biometrical Journal*, **58**(6), 1506–1537.

Punzo, A. and McNicholas, P. D. (2017a). Robust clustering in regression analysis via the contaminated Gaussian cluster-weighted model. *Journal of Classification*, **34**(2), 249–293.

Punzo, A. and McNicholas, P. D. (2017b). Robust clustering in regression analysis via the contaminated gaussian cluster-weighted model. *Journal of Classification*, **34**(2), 249–293.

Punzo, A., Bagnato, L., and Maruotti, A. (2018). Compound unimodal distributions for insurance losses. *Insurance: Mathematics and Economics*, **81**, 95–107.

Punzo, A., Blostein, M., and McNicholas, P. D. (2019). High-dimensional unsupervised classification via parsimonious contaminated mixtures. *Pattern Recognition*, **98**, 107031.

Qi, M. and Zhao, X. (2011). Comparison of modeling methods for loss given default. *Journal of Banking & Finance*, **35**(11), 2842–2855.

Qiu, W. and Joe., H. (2015). *clusterGeneration: Random Cluster Generation (with Specified Degree of Separation)*. R package version 1.3.4.

Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the em algorithm. *SIAM review*, **26**(2), 195–239.

Renault, O. and Scaillet, O. (2004). On the way to recovery: A nonparametric bias free estimation of recovery rate densities. *Journal of Banking & Finance*, **28**(12), 2915–2931.

Rigby, B., Stasinopoulos, M., Heller, G., and Voudouris, V. (2014). The distribution toolbox of gamlss. *The GAMLSS Team*.

Ritter, G. (2015). *Robust Cluster Analysis and Variable Selection*, volume 137 of *Chapman & Hall/CRC Monographs on Statistics & Applied Probability*. Chapman & Hall/CRC Press, Boca Raton.

Rösch, D. and Scheule, H. (2006). A multi-factor approach for systematic default and recovery risk. *The Basel II Risk Parameters*, pages 105–125.

Rytgaard, M. (1990). Estimation in the pareto distribution. *ASTIN Bulletin: The Journal of the IAA*, **20**(2), 201–216.

Sánchez-Manzano, E. G., Gomez-Villegas, M. A., and Marín-Diazaraque, J.-M. (2002). A matrix variate generalization of the power exponential family of distributions. *Communications in Statistics-Theory and Methods*, **31**(12), 2167–2182.

Santafe, G., Calvo, B., Perez, A., and Lozano, J. A. (2015). *bde: Bounded Density Estimation*. R package version 1.0.1.

Sarkar, S., Zhu, X., Melnykov, V., and Ingrassia, S. (2020). On parsimonious models for modeling matrix data. *Computational Statistics & Data Analysis*, **142**, 106822.

Schlattmann, P. (2009). *Medical applications of finite mixture models*. Springer, Berlin.

Schmit, M. (2004). Credit risk in the leasing industry. *Journal of Banking & Finance*, **28**(4), 811–833.

Schuermann, T. (2004). What do we know about loss given default? Technical report, Wharton Financial Institutions Center. Working Paper No. 04-01.

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

Scott, D. (2009). *HyperbolicDist: The hyperbolic distribution*. R package version 0.6-2.

Seidler, J. (2008). Implied market loss given default: Structural-model approach. Technical report, IES Working Paper.

Sigrist, F. and Stahel, W. A. (2011). Using the censored gamma distribution for modeling fractional response variables with an application to loss given default. *ASTIN Bulletin: The Journal of the IAA*, **41**(2), 673–710.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall \CRC.

Soetaert, K. (2009). ***rootSolve**: Nonlinear root finding, equilibrium and steady-state analysis of ordinary differential equations*.

Stasinopoulos, M. and Rigby, B. (2016). ***gamlss.mx**: Fitting Mixture Distributions with GAMLSS*. package Version 4.3-5 (2016-05-18).

Stasinopoulos, M. and Rigby, B. (2017). ***gamlss.dist**: Distributions for Generalized Additive Models for Location Scale and Shape*. package Version 5.0-4 (2017-12-11).

Stasinopoulos, M., Enea, M., Rigby, R. A., and Hossain, A. (2017a). *Inflated distributions on the interval* $[0, 1]$.

Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V., and De Bastiani, F. (2017b). *Flexible Regression and Smoothing: Using GAMLSS in* . CRC Press.

Stehlík, M., Potocký, R., Waldl, H., and Fabián, Z. (2008). Some notes on the favourable estimation of fitting heavy tailed data. Technical Report 32, IFAS Research Paper Series.

Stehlík, M., Potocký, R., Waldl, H., and Fabián, Z. (2010). On the favorable estimation for fitting heavy tailed data. *Computational Statistics*, **25**(3), 485–503.

Team, R. C. (2019). *R: A Language and Environment for Statistical Computing*. Foundation for Statistical Computing, Vienna, Austria.

Teicher, H. (1963). Identifiability of finite mixtures. *The annals of Mathematical statistics*, pages 1265–1269.

Templ, M., Gussenbauer, J., and Filzmoser, P. (2019). Evaluation of robust outlier detection methods for zero-inflated complex data. *Journal of Applied Statistics*, pages 1–24.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons, Chichester.

Tobback, E., Martens, D., Van Gestel, T., and Baesens, B. (2014). Forecasting loss given default models: impact of account characteristics and the macroeconomic state. *Journal of the Operational Research Society*, **65**(3), 376–392.

Tomarchio, S. D. and Punzo, A. (2019). Modelling the loss given default distribution via a family of zero-and-one inflated mixture models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **182**(4), 1247–1266.

Tomarchio, S. D. and Punzo, A. (2020). Dichotomous unimodal compound models: application to the distribution of insurance losses. *Journal of Applied Statistics*, **47**(13–15), 2328–2353.

Tomarchio, S. D., Punzo, A., and Bagnato, L. (2020). Two new matrix-variate distributions with application in model-based clustering. *Computational Statistics & Data Analysis*, **152**, 107050.

Tong, E. N., Mues, C., and Thomas, L. (2013). A zero-adjusted gamma model for mortgage loan loss given default. *International Journal of Forecasting*, **29**(4), 548–562.

Vasicek, O. (2002). The distribution of loan portfolio value. *Risk*, **15**(12), 160–162.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Viroli, C. (2011a). Finite mixtures of matrix normal distributions for classifying three-way data. *Statistics and Computing*, **21**(4), 511–522.

Viroli, C. (2011b). Model based clustering for three-way data structures. *Bayesian Analysis*, **6**(4), 573–602.

Viroli, C. (2012). On matrix-variate regression analysis. *Journal of Multivariate Analysis*, **111**, 296–309.

Wedel, M. and Kamakura, W. A. (2012). *Market segmentation: Conceptual and methodological foundations*, volume 8. Springer Science & Business Media, Norwell.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, pages 1–25.

Wuertz, D. and Chalabi, Y. (2016). ***fGarch**: Rmetrics - Autoregressive Conditional Heteroskedastic Modelling*. Version 3010.82.1 (2016-08-15).

Yao, X., Crook, J., and Andreeva, G. (2017). Enhancing two-stage modelling methodology for loss given default with support vector machines. *European Journal of Operational Research*, **263**(2), 679–689.

Yeo, A. C., Smith, K. A., Willis, R. J., and Brooks, M. (2001). Clustering technique for risk classification and prediction of claim costs in the automobile insurance industry. *Intelligent Systems in Accounting, Finance & Management*, **10**(1), 39–50.

Zarei, S., Mohammadpour, A., Ingrassia, S., and Punzo, A. (2018). On the use of the sub-gaussian α-stable distribution in the cluster-weighted model. *Iranian Journal of Science and Technology, Transactions A: Science*, pages 1–11.

Zeileis, A. and Windberger, T. (2014). ***glogis**: Fitting and Testing Generalized Logistic Distributions*. R package version 1.0-0.

Zucchini, W., MacDonald, I. L., and Langrock, R. (2017). *Hidden Markov Models for Time Series: An Introduction Using* . Chapman & Hall/CRC Monographs on Statistics and Applied Probability. CRC Press.