ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Model-based clustering for count & mixed mode data

by

Fotini Panagou

under the supervision of Prof. Dimitris Karlis

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy
in the
Department of Statistics

September 28, 2022

Abstract

Model based clustering is a common approach for modelling data with the use of finite mixtures of parametric distributions. For count data, the choice of high dimensional multivariate Poisson distribution can lead to increased computational effort. Composite likelihoods concept with the use of bi-variate marginals can offer flexibility in estimations. In order to further reduce the time of estimation of the composite likelihood method associated parameters, we introduce the sampling methods which can offer adequate results, especially for large data samples.
When it comes to mixed data sets, the joint probability is not always easy to be found. Copulas can provide a solution to this problem, and especially Gaussian copula offers flexibility for description of the dependencies between different types of variables. Our aim was to reduce computational effort arisen from the use of Gaussian copula, and the fully parametrized model we assessed, since this approach causes effort from adding different correlation matrices for every component that need to be estimated. So, the main target was to achieve parsimony in estimation with the use of appropriate techniques.

Acknowledgements

I am extremely grateful to my supervisor, Prof. Karlis, D. for his invaluable advice, continuous support, and patience during my PhD study.

Contents

Chapter 1

Introduction

1.1 Count data

The last decade we have seen a tremendous increase on interest about discrete valued mixtures models. While methodologies for univariate integer valued mixtures are now flourishing, literature on multivariate mixtures for counts is a less developed area of research. Such multivariate count data occur in several different disciplines like epidemiology, marketing, criminology and engineering just to name a few. For example in syndromic surveillance systems the number of patients with a given symptom is recorded aiming at being able to discover early an abrupt change on this number, perhaps indicating for a threat for public health. In practice a large number of symptoms are counted creating multiple mixtures that are in fact correlated. Correct evaluation of such multiple series need a model that can take into account the correlation across time but at the same time the cross correlation between the different symptoms.

Another example comes from geophysical research when the number of earthquakes need to be modelled (Boudreault and Charpentier, 2011). In such data the number of earthquakes above a certain magnitude threshold and for a given time period are counted. Different series can be generated from adjacent areas, making an important scientific question the correlation between the two areas. In criminology one counts the number of occurrences of a specific type of crime in successive time periods (let say weeks). Working together with more than one types of crimes generates many count mixtures that can be correlated and need proper models to work with. In finance one wants to model the number of bids and asks for a stock or the number of trades of stocks in a portfolio. Similar examples may be seen for the number of purchases of different but related products

in marketing, the number of claims of different policies in actuarial science and many others.

Literature on multivariate mixtures of counts is less developed. One of the reasons is that even models not being time series are less developed due to analytical and computational problems. However, in recent years, there have been some new models to facilitate modelling approach.

In the literature, methods have been presented for the analysis of count data classified by fixed and crossed factors under the assumptions that this data can be modelled by independent binomial or Poisson distributions. In general, the mean value of these distributions depends on the levels of the classifying factors and a linear model is proposed for the logit transform or the log transform of these mean values. In practice many situations occur which are different, such as:

- The counts are independent, but the observed variation in the data is more than can be explained by e.g. the Poisson distribution;
- The counts are dependent: the factors are not fixed but they are random.

A critical question faced by data analysts while modelling the count data is how to choose a suitable model for a particular study. For modelling the categorical count data with excess zero counts, numerous choices of methodologies have been used by various researchers in literature. Usually Regression models are widely applied for modelling this kind of data. However other data analysis techniques also has been adopted in the recent years, which includes machine learning techniques like artificial neural networks. However the major problem encountered is the selection of most suitable model for analysing the count data, since various methods provides dissimilar results, which also varies from one data to another data. One of the widely accepted and used methods for modelling the count data with excess zero counts is the zero inflated regression models, which supply a broad and rigorous area of research.

1.1.1 Models for multivariate count data

Even ignoring the mixtures correlation one can see that there are not many models for multivariate counts. Inference for multivariate counts is analytically and computationally demanding. Perhaps the case is easier and more developed when dimensions reduce to two but there are several bivariate models that cannot

generalize easy to multivariate ones. This obstacles the development of flexible models to be used also in mixtures context.

We briefly expose some of the issues. Consider for example the simplest extension of the simple Poisson distribution to the bivariate case. As in Kocherlakota and Kocherlakota (1992), the bivariate Poisson has probability mass function (pmf) is given by

$$\begin{aligned} P(y_1, y_2) &= P(Y_1 = y_1, Y_2 = y_2; \boldsymbol{\Theta}) \\ &= e^{-(\theta_1+\theta_2+\theta_0)} \frac{\theta_1^{y_1}}{y_1!} \frac{\theta_2^{y_2}}{y_2!} \sum_{s=0}^{min(y_1,y_2)} \binom{y_1}{s} \binom{y_2}{s} s! \left(\frac{\theta_0}{\theta_1\theta_2}\right)^s, \quad (1.1) \end{aligned}$$

$\theta_1, \theta_2, \theta_0 \geq 0$, $y_1, y_2 = 0, 1, \ldots$, where $\boldsymbol{\Theta} = (\theta_1, \theta_2, \theta_0)$ are the parameters. θ_0 is the covariance while the marginal means and variances are equal to $\theta_1 + \theta_0$ and $\theta_2 + \theta_0$ respectively. The marginal distributions are Poisson. One can easily recognize that this pmf involves a finite summation which can be computational intensive for larger counts. This bivariate Poisson allows only positive correlation. We denote this by $BP(\theta_1, \theta_2, \theta_0)$. For $\theta_0 = 0$ we get two independent Poisson distributions. We may generalize this model in a certain extend by considering mixtures of the bivariate Poisson. There are two ways to do this that have been worked in detail in practice (though other schemes also apply but are less developed). Most of the literature assume a $BP(\alpha\theta_1, \alpha\theta_2, \alpha\theta_0)$ distribution and places a mixing distribution in α. Such a model, depending on the choice of the distribution of α produces overdispersed marginal distributions but with always positive correlation. Correlation comes from two sources, the first is the intrinsic one from θ_0 and the second due to the use of a common α.

A more refined model can be produced by assuming a $BP(\theta_1, \theta_2, 0)$ and letting θ_1, θ_2 jointly vary according to some bivariate continuous distribution, as for example in Chib and Winkelmann (2001) where a bivariate lognormal distribution is assumed. Here all the correlation comes from the correlation of the joint mixing distribution and thus it can be negative as well. Here the obstacle is that we do not have flexible bivariate distributions to use for the mixing or some of them may lead to computational problems. Chib and Winkelmann (2001) used a bivariate lognormal. The resulting bivariate Poisson lognormal does not have closed form pmf and bivariate integration is needed thoroughly. Also note that since there are more than one derivation there is some confusion on the resulting distributions. For example the literature has a large number of distributions under the name bivariate negative binomial.

A different avenue to built multivariate models is to apply copula approach. Copulas (see Nelsen, 2006) have found a remarkable large number of applications in finance, hydrology, biostatistics etc., since they allow the derivation and application of flexible multivariate models with given marginal distributions. The key idea is that the marginal properties can be separated from the association properties leading thus to a wealth of potential models. For the case of discrete data, copula–based modelling is less developed. For example Genest and Nešlehová (2007) provided an excellent review on the topic. Since then there are several attempts to apply copulas to discrete data with quite useful success in practice (Nikoloulopoulos and Karlis, 2009). It is important to keep in mind that some of the desirable properties of copulas are not valid when dealing with count data, as for example dependence properties which are now dependent on the marginal properties. Also calculation of the pmf can be cumbersome in larger dimensions.

To help the exposition we restrict ourselves to the bivariate case, i.e. to bivariate copulas and later we will discuss the problem when going to higher dimensions.

Definition (Nelsen, 2006). A bivariate copula is a function C from $[0,1]^2$ to $[0,1]$ with the following properties: a) For every $\{u,v\} \in [0,1]$, $C(u,0) = 0 = C(0,v)$ and $C(u,1) = u, C(1,v) = v$ and b) For every $\{u_1,u_2,v_1,v_2\} \in [0,1]$ such that $u_1 \leq u_2$ and $v_1 \leq v_2$, $C(u_2,v_2) - C(u_2,v_1) - C(u_1,v_2) + C(u_1,v_1) \geq 0$.

So, copulas are in fact bivariate distributions with uniform marginals. Recall the inversion theorem, central in simulation, where starting from a uniform random variable and applying the inverse transform of a distribution function we can generate whatever distribution we like. Copulas actually extend this idea in the sense that we start from two uniforms that are correlated and hence we end up with variables from whatever distribution we like which are still correlated.

If $F(x), G(y)$ are the *cdf*'s of the univariate random variables X and Y, then $C(F(x), G(y))$ is a bivariate distribution for (X,Y) with marginal distributions F and G respectively. Conversely, if H is a bivariate *cdf* with univariate marginal *cdf*'s F, G, then, according to Sklar (1959)'s theorem there exists a bivariate copula C such that for all (X,Y), $H(x,y) = C(F(x), G(y))$. If F, G are continuous, then C is unique, otherwise, C is uniquely determined on $range\, F \times range\, G$. This lack of uniqueness is not a problem in practical applications as it implies that there may exist two copulas with identical properties.

Actually, copulas provide the joint cumulative function. In order to derive the joint density (for continuous data) or the joint probability function (for discrete

data) we need to take the derivatives or the finite differences of the copula. In the bivariate case we have that for discrete data, the *pmf* is obtained by finite differences of the *cdf* through its copula representation (Genest and Nešlehová, 2007), namely

$$\begin{aligned} h(x,y;\alpha_1,\alpha_2,\theta) &= C(F(x;\alpha_1),G(y;\alpha_2);\theta) - C(F(x-1;\alpha_1),G(y;\alpha_2);\theta) \\ &- C(F(x;\alpha_1),G(y-1;\alpha_2);\theta) + C(F(x-1;\alpha_1),G(y-1;\alpha_2);\theta) \end{aligned}$$

where $F(\cdot)$ and $G(\cdot)$ are the marginal cumulative functions, α_1 and α_2 are the parameters associated with the marginal distributions respectively and θ is the parameter(s) of the copula.

And here the problems occur for more dimensions. Since we need to take differences, for the trivariate case we need to evaluate 8 times the copula and for d dimensions, one needs to evaluate it 2^d times. If the selected copula is not in closed form, recall that it is a cdf and hence for some well known and used one like the Gaussian copula this is a multivariate integral, problems occur as dimensions increase. From the computational point of view one needs to evaluate many time multivariate integrals or at least to add and subtract several number which leads to possible truncation errors.

A relative problem is the lack of many copulas that can easily allow for flexible correlation structure, as for example some multivariate copulas assume the same correlation to all pairs of variables which is too restrictive in practice. Also if one needs to specify both positive and negative correlation more restrictions apply.

Having entered in the realm of more than 2 dimensions similar problems occur for the simple bivariate Poisson and related models. Generalizing the bivariate Poisson to the multivariate Poisson with one correlation parameter for every pair, leads to multiple summation, see the details in Karlis and Meligkotsidou (2005).

There are some more strategies to built flexible models for multivariate counts like models based on conditional distribution; Berkhout and Plug (2004); finite mixtures Karlis and Meligkotsidou (2007). In most cases things are not simple and always they are more complicated with respect to continuous models where the multivariate normal distributions is a cornerstone allowing for great flexibility and feasible calculations.

In the literature other statistical models have been generated to analyze data with count nature DB. (2000), D. (1992). The first model to analyze count outcomes is the Poisson regression model (PRM) Karazsia BT (2008) , Long SJ

(2006), R. (2016). This model is based on Poisson distribution has two restrictive assumptions. First, the variance of the count outcome is equal to the mean. The second assumption is that occurrences of events are independent of each other. However, in practice, these assumptions are usually violated, and count variables tend to have a conditional variance that often exceeds the conditional mean, which is known as "overdispersion". Using the PRM to analyze outcomes in which one of these two assumptions is violated may result in biased data with underestimated standard error.

The second model is the negative binomial regression model (NBRM) that attempted to overcome the abovementioned limitations in the Poisson distribution and has proven to properly represent the observed counts than the Poisson distribution e.g. Karazsia BT (2008), Hausman JA (1984). Accordingly, unlike the PRM, this distribution does not require the mean and variance of the count outcome to be equal. Additionally, the previously mentioned assumption of independence of events required for PRM is no longer mandatory in the NBRM since it assumes that events can be repeated, given the influence of individual differences on the probability of an event to occur.

Two other alternatives count models are the ZeroInflated Count Models: zero-inflated Poisson (ZIP) and zero-inflated negative binomial (ZINB). These models had been developed to overcome circumstances in which the origin of overdispersion is due to excessive zero counts. These kinds of distributions assume that the zero counts originate from two different sources and can be classified into two groups. The zero-inflated model selection, whether ZIP or ZINB, is determined by the sort of overdispersion. If the excessive number of zeros generates the overdispersion, then the ZIP is more appropriate to model count data. On the other hand, if the overdispersion is caused by factors not related to the excessive number of zeros, then the ZINB model is more suitable.

Summarizing this section, there are models for multivariate counts available but they can be demanding in practice which creates problems in their applicability.

1.2 Mixed mode data

Clustering is an important tool in data mining,which has many applications in areas such as bio-informatics, web data analysis, information retrieval, customer relationship management, text mining, and scientific data exploration. It aims to partition a finite, unlabeled dataset into several natural subsets so that data

objects within the same clusters are close to each other and the data objects from different clusters are dissimilar from each other according to the predefined similarity measurement.

Cluster analysis aims to partition unlabeled data into homogeneous groups, such that two instances are similar if they belong to the same cluster, and dissimilar otherwise. Although this unsupervised task is often considered in the context of either continuous or categorical datasets, this task remains challenging when dealing with "heterogeneous" or "mixed" data, i.e. with both types of variables. As previously emphasized, clustering of mixed data is challenging because it is difficult to directly apply mathematical operations to both types of feature variables. One of the main issues arising in the framework of mixed data clustering is thus the choice of the most appropriate distance or model to simultaneously process both data types Preud (2021). Indeed, clinical research usually relies on heterogeneous data: clinical datasets typically include a mix of variables related to clinical history (usually categorical variables), general/anthropometric data (usually continuous variables such as age and body mass index), physical examination (both categorical and ordinal variables) and laboratory or imaging findings (often continuous variables). Such heterogeneity urges for ways to guide users and clinical practitioners in choosing appropriate clustering approaches for heterogeneous clinical datasets in order to achieve efficient phenomapping of patients in various clinical settings.

Discretization and dummy-coding are some of the simple and intuitive solutions to obtain a homogeneous dataset containing only categorical data on which classical techniques can be applied. However, this approach may introduce distortion in the original data and may consequently lead to increased bias. Fortunately, a wide range of clustering algorithms has been specifically developed to deal with mixed data. A detailed taxonomy of available methods has been reported recently by Ahmad and Khan (2019). Nevertheless, the end-user may be bewildered when choosing one of these techniques as there is no clear guidance for choosing the most appropriate technique in a given context. To our knowledge, few benchmark studies have examined the performance of clustering strategies for mixed type variables on both real and simulated data. Moreover, only a few of the available techniques have been tested in previous benchmark attempts. In addition, an external assessment of available techniques, by a group not directly involved in their development, may further strengthen the generalizability of the results. In fact, a better understanding of the strengths and weaknesses of each clustering strategy

may help to clarify the lack of reproducibility and generalization sometimes observed in the setting of mixed data clustering. Most clustering methods fall into one of two classes: distance-based methods and model based methods.

At the very high end of the overall taxonomy, two main categories of distance based clustering, known as partitional clustering and hierarchical clustering, are envisioned in the literature. We will mainly talk about the first two types of methods as applied to the mixed-mode data. The distance-based methods have two sub-branches. One is partitioning algorithms, including K-Means, K-Medians, K-Medoids and K-prototypes; the other is hierarchical algorithms, including agglomerative methods and divisive methods. The probabilistic model-based methods generally assume a specific form of the generative model, like a mixture of Gaussians. The model parameters are estimated (commonly with the EM algorithm) using the maximum likelihood method. Then each data point is assigned to the cluster for which has the highest predicted probability. The density-based methods assume the data space has the granularity between every dense region with arbitrary shape. The most popular density-based clustering method is DBSCAN.

1.2.1 Distance-based methods

One of the more common approaches for clustering mixed-type data involves converting the data set to a single data type, and applying standard distance measures to the transformed data. Dummy coding all categorical variables is one example of such an approach. Dummy coding increases the dimensionality of the data set, which can be problematic when the number of categorical variables and associated categorical levels increase with the size of the data. Further, any semantic similarity that may have been observable in the original data set is lost in the transformed data set. Perhaps most importantly, coding strategies involve a non-trivial choice of numbers or weights that must be used to represent categorical levels. The coding strategy introduced by Hennig and Liao (2013) is one example of such dummy coding strategies. It involves selecting values that control the expected contribution of categorical variables in relation to the quantity $E(X_1 - X_2)^2 = 2$, where X_i denotes independent and identically distributed observations of a continuous variable standardized to unit variance.

Rather than set the expected contribution of the categorical variables equal to this quantity, however, Hennig and Liao (2013) set the expected contribution to half of this quantity, based on a concern that the gaps between the coded categorical dummy variables would unduly influence the resulting clusters. Hennig

and Liao (2013) justify this claim based on results of a Monte Carlo study with 50 simulations per condition, in which they inspect the performance of the k-medoids algorithm CLARA Kaufman (1990) using two weighting schemes.

This family of methods relies exclusively on explicit distances or dissimilarities between individuals. Some algorithms such as Partitioning Around Medoids (PAM) or Hierarchical Ascendant Clustering (HAC) can take any dissimilarity matrix as an input, whereas K-prototypes rather build their own distance. Note that in the present analysis, by misuse of language, the term "distance" sometimes means "dissimilarity" as some measures do not necessarily verify the triangular inequality.

An alternative approach is to use distance measures developed specifically for mixed data sets, e.g., Gower (1971) defined as follows. For two observations x and y , the Gower similarity coefficient is calculated as:

$$S(x,y) = \frac{1}{m}\left(\sum_{j=1}^{q}\left(1 - \frac{|x_j - y_j|}{Range(j)}\right) + \sum_{j=q+1}^{p} s(x_j, y_j)\right) \tag{1.2}$$

where, $s(x_j, y_j)$ equals 1 if $x_j = y_j$ and 0 otherwise, and Range(j) represents the absolute difference between extreme values of the j-th variable. The first term of the right part is the similarity on the continuous variables, while the second term deals with categorical variables. By dividing the difference $|x_j|y_j|$ by the range of variable j, both coefficients for numeric and categorical variables are included in the interval [0, 1] . The dissimilarity matrix is then comprised of the dissimilarity coefficients calculated between each pair of observations.

Huang (1998) proposed the k-prototypes algorithm, a variant of the k-means algorithm that is based on the weighted combination of squared Euclidean distance for continuous variables and matching distance for categorical variables. The k-prototypes algorithm relies on a user-specified weighting factor that determines the relative contribution of continuous and categorical variables, not unlike what is required to use Gower's distance, and thus suffers from the same limitation. The K-prototypes algorithm defines G virtual individuals (or prototypes) as the centers of the groups, built from the means by group for numeric variables, and modes by group for categorical variables. The distance between two subjects X and Y is then defined as:

$$d(x,y) = \sum_{j=1}^{q}(x_i - y_j)^2 + \gamma \sum_{j=q+1}^{p} \delta(x_i - y_j) \tag{1.3}$$

where the first term is the squared Euclidean distance measurement for the continuous variables and the second term is the Hamming distance. The weight γ is used to avoid favoring either type of attribute. It can be specified by the user or estimated via a combined variance of the data. The minimization criteria is the total sum of distances (TSD) between the subjects and the prototype of the class b_g to which they belong:

$$TSD = \sum_{g=1}^{G} \sum_{x \in C_g} \sum_{j=1}^{q} (x_j - b_{jg})^2 + \gamma \sum_{j=q+1}^{p} \delta(x_j - b_{jg}) \tag{1.4}$$

In practice, the algorithm is very similar to the k-means: initial G prototypes are selected as temporary centers of the clusters, then each subject is allocated to the closest prototypes. When all subjects are allocated, the prototypes are updated to represent their optimal class.

Partitioning around medoids (PAM). The PAM method Kaufman (1990) builds a partition by affecting observations to the closest "medoid", i.e. the best representative subject of its cluster. The algorithm is composed of two steps: one for building the current clustering similarly to the K-means (BUILD phase), and another to improve the partition toward a local optimum (SWAP phase). The minimization criteria is the Total Deviation (TD):

$$TD = \sum_{g=1}^{G} \sum_{x \in C_g} d(x_j - m_g) \tag{1.5}$$

where $(m_1, ..., m_G)$ are the medoids, $(C_1, ..., C_G)$ the respective clusters they represent, and $d(x_j - m_g)$ the dissimilarity between the subject x_j and the medoid of the cluster C_g. The BUILD phase finds the first medoid which minimizes the total deviation, i.e. with the smallest dissimilarity to all other subjects. The remaining $G-1$ medoids are then successively found by maximizing the reduction of the TD. The SWAP phase subsequently improves the existing partition by considering all possible swaps of the G medoids with the non-medoids. The swaps which reduce TD the most are applied, and the process is repeated until no further improvement is found.

Ascendant hierarchical clustering (HC) (see e.g. Ward (1963)). This well-known clustering method begins with N clusters (one per subject), then at each step aggregates the two closest clusters until only one remain. The successive fusions are represented on a dendrogram to facilitate the a posteriori choice of an optimal number of clusters. In general, the best partition is the one preceding the

first sizeable increase in intra-cluster variance. Let us suppose that at a particular aggregation step, clusters C_i and C_j are the next to be merged. To determine the distance of the merged cluster with any other cluster C_k , the dissimilarity matrix must be updated by one aggregation method belonging to the Lance-Williams algorithm family:

$$d(C_i \cup Cj, Ck) = \alpha d(C_i, C_k) + \beta d(C_j, C_k) + \eta d(C_i, C_j) \tag{1.6}$$

The coefficients α, β and η are dependent on the aggregation method. These methods for computing distances between clusters are called linkage criteria such as Ward's algorithm, which aims at minimizing the increase in intra-cluster variance at each binary fusion, such that convex and compact clusters are more likely to be formed.

1.2.2 Probabilistic-based methods

Model-based or statistical approaches to clustering mixed-type data typically assume the observations follow a normal-multinomial finite mixture model see e.g. Browne (2012) , Everitt (1988) ,Fraley and Raftery (2002) , Hunt and Jorgensen (2011) and Lawrence and Krzanowski (1996).

When parametric assumptions are met, model-based methods generally perform quite well and are able to effectively use both continuous and categorical variables, while avoiding undue vulnerability to variables with weak association with the identified clusters. Normal-multinomial mixture models can be extended using the location model Krzanowski (1993), which allows a distinct distribution for the continuous variables for each unique combination of categorical levels. While this accounts for any possible dependence structure between continuous and categorical variables, it becomes infeasible when the number of categorical variables or number of levels within each categorical variable is large. For exclusively continuous data, kernel density (KD) methods allow these parametric assumptions to be relaxed however KD methods incur a prohibitively large computational cost with a large number of continuous variables, along with other well-documented problems associated with high-dimensional KD estimation.

The Kamila algorithm Foss (2016) is a model-based adaptation of the k-means for managing heterogeneous datasets. The sample of continuous variables is assumed to follow a mixture distribution with arbitrary spherical clusters (where the density of the data is only dependent on the distance to the center of the

distribution). This assumption is less restrictive than those from Mixmod or LCM (see below). Categorical variables are supposed to be sampled from a mixture of multinomial variables. Factors are also assumed to be conditionally independent given the clusters to which they belong. The Kamila algorithm begins with a set of centroids for the continuous variables and a set of parameters for the categorical variables. For continuous variables, the Euclidean distance with the closest centroid is computed. This set of N minimal distances is used to estimate the mixture distribution of continuous variables. For categorical variables, the probabilities of observing the data given the cluster are computed. The log-likelihood of the sum of these two components is then used to find the most appropriate cluster for each subject. Based on this temporary partition, the centroids and the parameters are updated to best represent the clusters. These steps are repeated until the clusters are stable. Finally, multiple runs of this process are performed with different initializations, and the partition maximizing the sum of the best final likelihoods is retained. The KAMILA clustering algorithm is a scalable version of k-means well suited to handle mixed-type data sets. It overcomes the challenges inherent in the various extant methods for clustering mixed continuous and categorical data, i.e., either they require strong parametric assumptions (e.g., the normal–multinomial mixture model), they are unable to minimize the contribution of individual variables (e.g.Modha–Spangler weighting), or they require an arbitrary choice of weights determining the relative contribution of continuous and categorical variables (e.g., dummy/simplex coding and Gower's distance). The KAMILA algorithm combines the best features of two of the most popular clustering algorithms, the k-means algorithm and Gaussian-multinomial mixture models both of which have been adapted successfully to very large data sets. Like k-means, KAMILA does not make strong parametric assumptions about the continuous variables. Like Gaussian-multinomial mixture models, KAMILA can successfully balance the contribution of continuous and categorical variables without specifying weights, but KAMILA is based on an appropriate density estimator computed from the data, effectively relaxing the Gaussian assumption.

Clustering by mixture modeling (Mixmod). Clustering by mixture modeling was proposed a number of years ago Everitt (1988) , although powerful computers are needed to realize its full potential. Nowadays, many R packages implement mixture models such as clustMD or fpc. Mixture models assume that continuous variables follow a multivariate normal distribution whereas categorical variables

follow a multivariate multinomial distribution. For an observation x_i , the probability distribution function is defined as:

$$f(x_i, \theta) = \sum_{g=1}^{G} \tau_g h(x_i|\alpha_g) \tag{1.7}$$

where $h(x_i|\alpha_g)$ is the distribution function for cluster g , with parameters α_g. For example, if h is defined as a multivariate normal distribution, α_g would be the mean vector $\mu_g g$ and the variance–covariance matrix Σ_g . The mixing proportions $\tau_g \in [0, 1]$, describe the expected size of each cluster. The set of parameters to be determined is $\theta = (\tau_1, ..., \tau_g, \alpha_1, ..., \alpha_g)$. Following an Expectation–Maximization (EM) framework, the set of parameters θ is computed such that the log-likelihood is maximized. The $\tau_{ig}(\theta)$ are then updated and so forth until convergence is reached. The crucial portion of this process relies on the choice of the model for the data within a specific cluster, i.e. the distribution function h . Several models are available with different levels of constraints. For continuous variables, the variance–covariance matrices are assumed to be diagonal. The user can decide to set all cluster volumes equal, and/or all intra-variances equal, which yields 4 possible models. With regard to categorical variables, a re-parametrization allows an interpretation similar to the center and the variance matrix used for continuous data. The dispersion parameter can be chosen to be the same across clusters and/or across variables, or across levels, thereby yielding 5 possibilities.

Latent class model (LCM). This method Marbac (2018) is another type of mixture modeling quite similar to Mixmod but, in addition, it can also determine whether a variable is useful for clustering, as well as the optimal number of clusters. If the j-th variable is relevant (i.e., its distribution differs significantly across clusters), it is labeled with $\omega_j = 1$ and belongs to Ω . If j is irrelevant (i.e. its distribution is similar across clusters), it is labeled with $\omega_j = 0$ and belongs to Ω complementary. Let $\omega = \omega_1, ..., \omega_p$ be the binary vector of the role of the p variables, and let $m = (G, \omega)$ be the resulting model. For an observation x_i , the probability density function of the mixture distribution is:

$$f(x_i|m, \theta) = \prod_{j \in \Omega^c} h_j(x_{ij}|a_{1j}) \sum_{g=1}^{G} \tau_g \prod_{j \in \Omega} h_j(x_{ij}|a_{gj}) \tag{1.8}$$

In LCM, the variables are assumed to be independent within clusters. Similarly to Mixmod, an EM algorithm is used to determine the optimal partition. When the selection of relevant variables is enabled, a penalization on the Bayesian

Information Criterion (BIC) or the Maximum Integrated Complete-data Likelihood (MICL) is applied at the maximization step. The selection of the number of clusters is achieved by running the algorithm for each number of clusters in a specified range, and selecting the one which yields the best value of the selected criterion.

Latent class analysis (LCA). This clustering technique Bandeen-roche (1997) is derived from the Latent Class Regression. LCA has the particularity of being applied to categorical data only, implying that continuous variables must be discretized. This transformation can be achieved based on percentiles in order to obtain balanced level counts, or based on practitioner knowledge such that the categories are clinically relevant. Each categorical variable is supposed to be sampled from a mixture of multinomial distributions, depending to which latent cluster the subjects belong to. Similarly to mixture modeling methods, the overall density function is used:

$$f(x_i, \theta) = \sum_{g=1}^{G} \tau_g h(x_i|\alpha_g) \tag{1.9}$$

In this instance, the α_g are the sets of probabilities for each level of each categorical variable if the subject belongs to the latent cluster C_g . Initially, the τ_g are uniform (equal cluster sizes), and the α_g are randomly sampled. As in Mixmod, the $\tau_{ig}(\theta)$ are computed and used to update the α_g according to the Bayes theorem and the observed data. With the new probabilities of the multinomial mixture, the τ_g are updated. Finally, the new parameters allow computing the log-likelihood of the present iteration:

$$\ell(x|\theta) = \sum_{i=1}^{N} \log \left(\sum_{g=1}^{G} \tau_g h(x|\alpha_g) \right) \tag{1.10}$$

The parameter update is repeated until the maximum number of iterations is achieved, or the difference between two successive log-likelihoods is too small. Several runs are subsequently performed to avoid finding a local optimum, and the run with the best final log-likelihood returns the resulting partition.

Chapter 2

Composite Likelihoods

2.1 Introduction

There is an increasing interest the last years about an inferential approach named Composite Likelihood (CL hereafter). The method is applicable when the standard likelihood is hard to be derived because, for example, the underlying model is too complicated involving multivariate integrals, and hence it maximization is almost impossible. In such cases CL can be the basis for inference in the sense that it replaces the computationally impossible likelihood with some other estimating function which while it captures part of the model it is computationally less intensive.

The CL approach, was originally described in Lindsay (1988) and further developed on the last decade, see e.g. the review paper of Varin et al. (2011). The idea behind composite likelihood is the following: In models with complex interdependencies, the joint distribution of the data may be difficult to evaluate, or even to specify. Typical problems arise from the need to invert large matrices and/or from approximation of intractable integrals/sums. To avoid evaluation of the full likelihood, one may approximate/replace the full likelihood with some other function which while retains a certain amount of information about the quantities of interest are easier to work with. This is also valid when the multivariate model is hard to be determined with full detail while an approximate model is possible.

In general CL is an inference function derived by multiplying a collection of component likelihoods; the particular collection used is often determined by the context. Typically they are of smaller dimension and hence easier to work with. Because each individual component is a conditional or marginal density, the resulting estimating equation obtained from the derivative of the composite

log-likelihood is an unbiased estimating equation. Because the components are multiplied, whether or not they are independent, the inference function has the properties of likelihood from a misspecified model.

More Formally, consider an m-dimensional vector random variable Y with probability density $f(y;\theta)$ for some unknown p-dimensional parameter $\theta \in \Theta$. Denote by $\{\mathcal{A}_1, \ldots, \mathcal{A}_k\}$ a set of marginal or conditional events with associated likelihoods $\mathcal{L}_k(\theta; y)$ then a composite likelihood is the weighted product

$$CL(\theta; y) = \prod_{k=1}^{K} \mathcal{L}_k(\theta; y)^{w_k}$$

where w_k are nonnegative weights used in certain cases to improve efficiency.

The CL is also useful when the multivariate model is difficult to fully determine yet an approximate model is possible. Such an approach is of particular interest in the case of multivariate counts. As an example, although it would be challenging to specify an 8-variate Poisson distribution (as the probabilities would be difficult to obtain), we could approximate it using products of lower-dimensional (e.g., bivariate Poisson) probability functions, corresponding to a pseudo-likelihood that in this case is a pairwise likelihood. In composite likelihood approaches, the key ingredient is to identify an approximation that retains as much information as possible; the corresponding price to pay for the model misspecification relates to the efficiency of the estimates and more complicated asymptotics. Also note that the model is now misspecified which can help on the robustness properties of the procedure. The derived CL estimators are asymptotically unbiased and normally distributed with variance the inverse of the Godambe Information. It is typical in practice to construct a surrogate function for the true likelihood using either marginal of conditional densities. More details can be found in Lindsay (1988) and Varin et al. (2011). Note also that the approach has been used under certain other names so far.

Among composite likelihood methods, pairwise likelihood (PL) methods, which use bivariate likelihoods to approximate the multivariate likelihood, have been used in many applications. For example, Kuk and Nott (2000) used such an approach for correlated binary data, and pairwise approaches have been used in mixtures models (see, e.g., Davis and Yau (2011)), spatial models (e.g., Varin et al. (2005)), and image models (e.g., Nott and Rydén (1999)). Several improvements for such pairwise methods have also been proposed. For example, Fieuws and Verbeke (2006) suggested the use of pairwise likelihood, where relevant estimators were

derived by averaging estimators obtained from each pair; an improved version of this approach with optimal weighting is described in Vasdekis et al. (2014). Joe and Lee (2009) examined optimally weighted pairwise likelihoods to improve the efficiency of the estimators, while Kuk (2007) proposed replacing the pairwise likelihood score equations by the optimal linear combinations of the marginal score functions. More recently, Papageorgiou and Moustaki (2018) proposed a sampling approach for a factor model based on pairwise likelihoods, where pairs of variables are sampled rather than using all pairs. In the context of multivariate mixed models, Hui et al. (2018) proposed constructing a quadratic approximation to each term in the pairwise likelihood function, which is then augmented with a penalty to encourage both individual and group coefficient sparsity.

2.2 Composite Likelihoods Concept

Let $(X_{i1}, \ldots, X_{id})$, $i = 1, \ldots, n$, be independent random vectors with a common density f_0 in $L_2(\mathbf{R}^d)$, the space of functions f on $\mathbf{R}^d$ such that $\int |f|^p < \infty$. The underlying density f_0 is assumed to lie in an identifiable parametric family $\{f(\cdot;\theta), \theta \in \Theta\}$ for some compact subset Θ of an Euclidean space $\mathbf{R}^q$, where q is some natural number greater or equal to one. Let θ_0 denote the element of Θ such that $f_0 = f_{\theta_0}$. It is convenient to use the following notation. Let $\mathcal{A}$ be the set of all pairs of variables so that the cardinal of $\mathcal{A}$ is $d(d-1)/2$. (Both (X_{i1}, X_{i2}) and (X_{i2}, X_{i1}) count for the same pair.) The pairs in $\mathcal{A}$ are ordered in the lexicographical order. For a in $\mathcal{A}$, denote by $a(1)$ the index of the first variable and by $a(2)$ the index of the second variable. Denote by f_a the marginal density of $(X_{ia(1)}, X_{ia(2)})$.

The PL function is given by

$$L_n^{\mathrm{PL}}(\theta) = \sum_{a \in \mathcal{A}} \sum_{i=1}^{n} \log f_a(X_{ia(1)}, X_{ia(2)}; \theta), \quad \theta \in \Theta. \tag{2.1}$$

A maximizer of (2.1) over a compact subset of Θ that contains θ_0 is called a maximum pairwise likelihood estimator (MPLE).

The randomized pairwise likelihood (hereafter RPL) approach is similar to Dillon and Lebanon (2010), developed for conditional composite likelihood approach, and corresponds to binary weights in the weighting case of Joe and Lee (2009). Although the proposed approach reduces both the number of pairs and the number of observations to be used in the calculations, both the sampling ratio and

the sampling scheme are important for the efficiency of the approach. A consistency result was given and also a discussion of the properties of the method; in particular, given the family of models considered, the underlying model structure may be reduced, and the sampling approach can thus be improved using suitable assumptions. Based on the asymptotic normality of the proposed estimator one can base inference on the estimator, something not possible for other approaches like, for example, variational approaches.

The use of finite mixture models in clustering is finding a large number of applications, mainly because it allows st1andard statistical modelling tools to be used in order to assess and evaluate the clustering. The density or probability mass function of a finite mixture model is defined as

$$h(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\pi}) = \sum_{j=1}^{k} \pi_j f_j(\boldsymbol{x};\boldsymbol{\theta}_j) \quad (\boldsymbol{x} \in \Re^p)\,, \tag{2.2}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_k^\top)^\top \in \Theta_1 \times \ldots \times \Theta_k$, and $\pi_j \in (0,1)$ with $\sum_{j=1}^{k} \pi_j = 1$. Appropriate choices of $f_j(\boldsymbol{x};\boldsymbol{\theta}_j)$ can result in flexible models of small complexity. Banfield and Raftery (1993) and the book of McNicholas (2016) provide a detailed treatment of the framework of finite mixture modelling for clustering and classification.

For continuous data, a common choice for the component densities $f_j(\boldsymbol{x};\boldsymbol{\theta}_j)$ $(j = 1, \ldots, k)$ is the density of the multivariate Gaussian distribution, also known as Gausian Mixture Model (GMM). This is mainly because of the convenience it offers in estimation (closed-form maximization steps in the EM algorithm) and interpretation (easy marginalization for visualising fitted components and the mixture density), see the `R` package `mclust` for an implementation (Fraley et al. (2012)).

CL can be useful in this setting for the following reasons:

- For cases with $p > n$ we would like to avoid implementing very large covariance matrices
- For several cases model specification in the high dimension is hard and almost impossible, e.g. multivariate discrete data

CL can be used to circumvent the issues above. This idea has been used for finite mixtures in Ranalli and Rocci (2017) at the past.

2.2.1 Composite Likelihoods Concept for mixtures

Let's assume n observations of $X = (X_1, X_2, ..., X_p) \sim f_p(X, \mathbf{\Theta})$, from $k = 1, ..., K$ clusters and the complete likelihood to be optimized is written in the form:

$$\mathcal{L}(X, \mathbf{\Theta}) = \prod_{i=1}^{n} \sum_{k=1}^{K} p_k f_p(X, \mathbf{\Theta})$$

where p_k the mixing probabilities for the k-th component. The augmented log-likelihood is defined as follows:

$$\ell = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \Big\{ \log p_k + \log f_p(X, \mathbf{\Theta}) \Big\}$$

. With the use of the bi-variate composite likelihoods we can reduce the p dimensions of the full dataset by using the bivariate p.d.f.s of all combinations $\binom{p}{2}$ of X_i's . The complete composite likelihood and log likelihood is now written in the below forms:

$$\mathcal{L}_{CL}(X, \mathbf{\Theta}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \prod_{s<t} p_k f_{st}(X_s, X_t, \mathbf{\Theta_{stk}}) \tag{2.3}$$

$$\ell_{CL} = \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{s<t} z_{ik} \Big\{ \log p_k + \log f(X_{isk}, X_{itk} | \mathbf{\Theta_{stk}}) \Big\} \tag{2.4}$$

where, $s < t$, $t = 1, ..., p$,and $\mathbf{\Theta_{stk}}$ is the parameter space of the joint probabilities of X_s, X_t for the k-th cluster. The 2.4 can be maximized with the use of EM algorithm described in the next section.

2.2.2 EM algorithm

With the use of EM algorithm we maximize the complete-data log likelihood of the model for the r^{th} iteration as follows:

E-step: Calculate for $i = 1, ..., n$ and $k = 1, 2, ..., K$ and for $s < t$ for all $m = \binom{p}{2}$ combinations of the p attributes

$$w_{ikst}^{(r+1)} = \frac{p_k^{(r)} f(X_{si}, X_{ti} | \Theta_{stk}^{(r)}))}{\sum_{k=1}^{K} p_k^{(r)} f(X_{si}, X_{ti} | \Theta_{stk}^{(r)}))} \tag{2.5}$$

therefore,

$$w_{ik}^{(r+1)} = \frac{\sum_{s<t} w_{ikst}^{(r+1)}}{\binom{p}{2}}$$

M-step : Update $p_k^{(r+1)} = \sum_{i=1}^{n} w_{ik}^{(r+1)}/n$ for all $k = 1, 2, ..., K$ and then maximize the quantity

$$Q_k = \sum_{i=1}^{n} \sum_{s<t} \left\{ w_{ikst}^{(r+1)} f(X_{si}, X_{ti}|\Theta_{stk}^{(r)}) \right\}$$

to get updated values for all parameters of the parametric space $\boldsymbol{\Theta_k}$ associated with the k-th component $k = 1, 2, ..., K$ as defined above.

2.2.3 Model Selection

Composite likelihood inference based on low-dimensional marginal or conditional distributions is common when the full likelihood is computationally too difficult. It is expected that CLM can also be more robust under possible miss specification of the higher order dimensional distributions and they can allow a less complex structure on the parameter space that might lead to a smoother likelihood surface. The central limit theorem for the composite likelihood score statistic implies that the distribution of θ_{CL} can be approximated by the Normal with mean θ and variance-covariance matrix $G^{-1}(\theta_{CL})$ where $G(\theta_{CL})$ is the Godambe information matrix (also known as sandwich information). For model selection with composite likelihood, the question is if the use of limited or reduced information leads to different decisions. To understand this, an asymptotic theory based on the theory of a sequence of contiguous local alternatives is developed to compare Akaike information criterion (AIC) and Bayesian information criterion (BIC) in their full likelihood and composite marginal likelihood versions. Consider the composite likelihood versions of Akaike information criterion (AIC) and Bayesian information criterion (BIC) described in Varin and Vidoni (2005). They are defined as:

$$\text{CLAIC} = -2L_{CL}(\hat{\theta_{CL}}) + 2tr\{J(\hat{\theta}_{CL})H^{-1}(\hat{\theta}_{CL})\} \tag{2.6}$$

and

$$\text{CLBIC} = -2L_{CL}(\hat{\theta_{CL}}) + (\log n)\, tr\{J(\hat{\theta}_{CL})H^{-1}(\hat{\theta}_{CL})\} \tag{2.7}$$

Here, $\hat{\theta_{CL}}$ is the composite likelihood estimator that maximizes the log composite likelihood. The matrices $H(\hat{\theta}_{CL})$ and $J(\hat{\theta}_{CL})$ are the Hessian matrix and the covariance matrix of the score function, respectively,

$$H(\hat{\theta}) = -\lim_{n\to\infty} \frac{\partial^2 L_{CL}}{\partial\theta\partial\theta^T} \tag{2.8}$$

and

$$J(\hat{\theta}) = Cov\Big\{n^{-1}\frac{\partial L_{CL}}{\partial\theta}\Big\} \tag{2.9}$$

The sample estimators for matrices $H(\hat{\theta}_{CL})$ and $J(\hat{\theta}_{CL})$ are:

$$H(\hat{\theta}) = \frac{\partial^2 L_{CL}}{\partial\theta\partial\theta^T} \tag{2.10}$$

and

$$J(\hat{\theta}) = n^{-1}\left(\frac{\partial L_{CL}}{\partial\theta}\right)\left(\frac{\partial L_{CL}}{\partial\theta}\right)^T \tag{2.11}$$

2.3 Multivariate Poisson mixtures

Mixed multivariate Poisson models, using some mixing distribution for the parameters as, for instance, multivariate negative binomial models, can solve the problem of overdispersion. If the mixing distribution is multivariate and allows for negative correlations, then the resulting model also allows for negative correlations like, for example, the multivariate Poisson-lognormal model of Aitchinson (1989). However, in this case the computational burden required for parameter estimation is quite large Chib (2001).

Karlis and Meligkotsidou (2007) propose finite multivariate Poisson mixtures as an alternative class of models for multivariate count data. These models are defined by assuming a finite step mixing distribution and they allow for both negative correlations and overdispersion while being computationally tractable. In general, step mixing distributions are quite flexible and the resulting finite mixture models have become a well established approach for modelling non-standard distributions. Non-parametric maximum likelihood estimation of the mixing distribution can be performed along the lines of Lindsay (1995) and Bohning (2000). Several inferential procedures for finite mixtures are also available and can be applied. Moreover, the proposed models can be considered as the basis for model based clustering for multivariate count data, offering a great potential for real data applications.

Let $\boldsymbol{X} = (X_1, X_2, ..., X_m)^T$ be a vector of discrete random variables. The definition of the multivariate Poisson model is based on the existence of a mapping $u : \mathbb{N}^q \to \mathbb{N}^m$, $q \geq m$, such that $\boldsymbol{X} = u(\boldsymbol{Y}) = \boldsymbol{AY}$, where $\boldsymbol{Y} = (Y_1, \ldots, Y_q)^T$, Y_r, $r = 1, ..., q$, are independent univariate Poisson random variables with parameters θ_r respectively, (denoted as $Y_r \sim Poisson(\theta_r)$), and $\boldsymbol{A}$ is an $m \times q$ binary matrix with no duplicate columns. Then, the vector $\boldsymbol{X}$ is said to follow a multivariate Poisson distribution with parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)^T$. The mean and the variance covariance matrix of $\boldsymbol{X}$ are given by

$$E(\boldsymbol{X} \mid \boldsymbol{\theta}) = \boldsymbol{A\theta} \quad \text{and} \quad Var(\boldsymbol{X} \mid \boldsymbol{\theta}) = \boldsymbol{A\Sigma A}^T,$$

where $\boldsymbol{\Sigma} = diag(\theta_1, \theta_2, \ldots, \theta_q)$ is the variance covariance matrix of $\boldsymbol{Y}$ ($\boldsymbol{\Sigma}$ is diagonal because of the independence of the Y_r's). Each element of $\boldsymbol{X}$ marginally follows a univariate Poisson distribution. (For further details for the model, see Karlis and Meligkotsidou (2005)).

The multivariate Poisson model for m variables derived by setting $\boldsymbol{A} = [\boldsymbol{A}_1, \boldsymbol{1}_m]$, where $\boldsymbol{A}_1$ is the identity matrix of size $m \times m$ and $\boldsymbol{1}_m$ is the m-column vector of 1's, is frequently used in the literature. The resulting distribution is commonly referred to as the multivariate Poisson distribution (see, e.g. Karlis (2003)). This model assumes that all the pairwise covariances are equal. However, this assumption is often not realistic in practice.

Karlis and Meligkotsidou (2007) focus on the case where the matrix $\boldsymbol{A}$ takes the form

$$\boldsymbol{A} = [\boldsymbol{A}_1, \boldsymbol{A}_2], \tag{2.12}$$

where $\boldsymbol{A}_1$ is again the identity matrix of size $m \times m$ and $\boldsymbol{A}_2$ is an $m \times \frac{m(m-1)}{2}$ binary matrix; each column of $\boldsymbol{A}_2$ has exactly 2 ones and $(m-2)$ zeros and no duplicate columns exist. The columns of $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ can be interpreted as main effects and two-way covariance effects, respectively, in an ANOVA like fashion. This model, which will be referred to as the multivariate Poisson model with two-way covariance structure, allows for different pairwise covariances. Therefore, it can be considered as a counterpart of the multivariate normal distribution and is suitable for multivariate count data.

In this case the parameter vector $\boldsymbol{\theta}$ can be split into a vector $\boldsymbol{\theta}^{(1)} = (\theta_1, \ldots, \theta_m)^T$ containing the main effects (hereafter the mean parameters since they appear only in the means and the variances of the X_j's but not in the covariances)

and a vector $\boldsymbol{\theta}^{(2)} = (\theta_{m+1}, \ldots, \theta_q)^T$ containing the pairwise covariances (hereafter the covariance parameters). We will denote by λ_{ij}, $i = 1, \ldots, m-1$, $j = i+1, \ldots, m$, the covariance between the pair of variables X_i and X_j. Note that $\boldsymbol{\theta}^{(2)} = (\lambda_{12}, \ldots, \lambda_{1m}, \lambda_{23}, \ldots, \lambda_{2m}, \ldots, \lambda_{m-1,m})$.

The main problem, which limits the use of the multivariate Poisson distribution, is the difficulty in calculating the probability mass function (pmf). Recalling the definition of an m-dimensional Poisson model through the mapping u, the joint probability of an m-dimensional vector $\boldsymbol{x} = (x_1, x_2, ..., x_m)^T$ is given by the sum of the joint probabilities of all the q-dimensional vectors $\boldsymbol{y} = (y_1, y_2, ..., y_q)^T$ such that $u(\boldsymbol{y}) = \boldsymbol{x}$. If $\boldsymbol{x} \in \mathbb{N}^m$, let the set $u^{-1}(\boldsymbol{x}) \subset \mathbb{N}^q$ denote the inverse image of $\boldsymbol{x}$ under u. The pmf of $\boldsymbol{X}$ is then defined as

$$MP_m(\boldsymbol{x} \mid \boldsymbol{\theta}) = \Pr(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{\theta}) = \sum_{\boldsymbol{y} \in u^{-1}(\boldsymbol{x})} \Pr(\boldsymbol{Y} = \boldsymbol{y} \mid \boldsymbol{\theta}).$$

Since the elements of $\boldsymbol{Y}$ follow independent univariate Poisson distributions, one obtain that

$$MP_m(\boldsymbol{x} \mid \boldsymbol{\theta}) = \sum_{\boldsymbol{y} \in u^{-1}(\boldsymbol{x})} \prod_{r=1}^{q} Po(y_r \mid \theta_r), \tag{2.13}$$

where $Po(y \mid \theta) = e^{-\theta}\theta^y / y!$, $y = 0, 1, \ldots, \theta \geq 0$, i.e. the probability function of the univariate Poisson distribution with parameter θ. At least one of the elements of $\boldsymbol{\theta}$ is assumed to be non-zero to avoid degenerate cases. In the sequel the m-variate Poisson model with two-way covariance structure will be denoted by $MP_m(\boldsymbol{\theta})$ and $MP_m(. \mid \boldsymbol{\theta})$, with parameter vector $\boldsymbol{\theta}$ and its joint probability function, respectively.

The calculation of these multivariate Poisson probabilities can be computationally expensive. Fortunately, computation of the probabilities can be accomplished via simple recursive schemes. Kano (1991) provided a general scheme for constructing recurrence relations for multivariate Poisson distributions. Even those recursive relations must be used efficiently in order to lead to feasible calculations.

In the case of model (2.12), each row of $\boldsymbol{A}$ contains exactly m ones, hence each recurrence relationship for the calculation of $\Pr(\boldsymbol{X} = \boldsymbol{x} \mid \boldsymbol{\theta})$ requires the computation of m previous probabilities. Obviously, as m increases, the complexity of the model and hence the computational effort also increase. Efficient use of the recurrence relationships is described in Tsiamyrtzis (2004).

2.3.1 Multivariate Mixed Poisson Distributions

Let $f(x \mid \theta) \underset{\theta}{\Lambda} g(\theta)$ be a general mixture of the density $f(x \mid \cdot)$ with respect to its parameter $\theta \in \Theta$, where $g(\theta)$ is the mixing distribution. Clearly, θ can be vector-valued. The density of the mixed distribution is given by $f(x) = \int_{\Theta} f(x \mid \theta) dG(\theta)$, where $G(\theta)$ is the cumulative function of the mixing distribution.

Mixtures of multivariate Poisson distributions are rather rare in the literature mainly due to their complicated form. The existing models can be gathered in two general groups. The first kind of multivariate Poisson mixtures has the form $MP_m(\alpha\boldsymbol{\theta}) \underset{\alpha}{\Lambda} g(\alpha)$, where α is a scalar, i.e. all the parameters have a common element α, which is distributed according to the mixing distribution $g(\alpha)$ and $\boldsymbol{\theta}$ is fixed. This type of mixing always leads to positive correlation between any pair of variables. The marginal distributions are mixtures of the Poisson distribution of the form $Poisson(\alpha\theta) \underset{\alpha}{\Lambda} g(\alpha)$.

The second kind of multivariate Poisson mixtures assumes a multivariate mixing distribution, i.e. the parameters are jointly distributed according to a joint probability distribution function $g(\boldsymbol{\theta})$. Formally, it has the form $MP_m(\boldsymbol{\theta}) \underset{\boldsymbol{\theta}}{\Lambda} g(\boldsymbol{\theta})$. If $\boldsymbol{X}$ is a random vector which follows a mixed multivariate Poisson distribution $MP_m(\boldsymbol{\theta}) \underset{\boldsymbol{\theta}}{\Lambda} g(\boldsymbol{\theta})$, then the unconditional expectation of $\boldsymbol{X}$ is given by

$$E(\boldsymbol{X}) = E(\boldsymbol{\theta})$$

while the unconditional variance covariance matrix is given by

$$Var(\boldsymbol{X}) = \boldsymbol{A}\boldsymbol{D}(\boldsymbol{\theta})\boldsymbol{A}^T, \tag{2.14}$$

where

$$\boldsymbol{D}(\boldsymbol{\theta}) = \begin{bmatrix} Var(\theta_1) + E(\theta_1) & Cov(\theta_1, \theta_2) & \dots & Cov(\theta_1, \theta_q) \\ Cov(\theta_1, \theta_2) & Var(\theta_2) + E(\theta_2) & \dots & Cov(\theta_2, \theta_q) \\ & \dots & & \\ Cov(\theta_1, \theta_q) & \dots & & Var(\theta_q) + E(\theta_q) \end{bmatrix}.$$

Here are some interesting points derived from this result.

Remark 1: Equation (2.14) implies that if the mixing distribution allows for covariances between the $\boldsymbol{\theta}$'s then the resulting unconditional variables are correlated. Even if one starts with independent Poisson variables, i.e. the covariance terms are zero, the mixing operation, as expected, leads to correlated variables.

Remark 2: More importantly, if $Cov(\theta_i, \theta_j) < 0$, then the unconditional variables may exhibit negative correlation. Although the multivariate Poisson distribution cannot incorporate negative correlations, this is not true for mixtures which offer a wide range of models for real data applications.

Remark 3: The covariances of the unconditional random variables are simple expressions of the covariances of the mixing parameters and hence of the mixing distribution's moments. Having fitted a multivariate Poisson mixture model one is then able to estimate consistently the reproduced covariance structure of the data. This is true since the moments of the multivariate Poisson distribution are simple polynomials with respect to the mixing parameters. Comparing the estimated covariance matrix to its observed counterpart may serve as a goodness of fit index.

Note that the first kind of multivariate Poisson mixtures can be viewed as a special case of the second kind. More details and references for both models can be found in Karlis (2005).

A particular case of the second kind of multivariate Poisson mixtures arises if a step mixing distribution is used. This gives rise to the class of finite mixtures of multivariate Poisson distributions.

2.3.2 The Finite Mixture Model

The p.m.f. of a finite mixture of K multivariate Poisson distributions is given by

$$P(x|\psi) = \sum_{k=1}^{K} p_k MP_m(\boldsymbol{x} \mid \boldsymbol{\theta}_k), \quad x \in \mathrm{IN}^m, \tag{2.15}$$

where $MP_m(\boldsymbol{x} \mid \boldsymbol{\theta})$ is defined in (2.13), $\boldsymbol{\psi} = (p_1, \ldots, p_{K-1}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$, $\boldsymbol{\theta}_k = (\theta_{1k}, \ldots, \theta_{qk})^T$ is the specific parameter vector of the k-th component, and the mixing proportions satisfy $0 < p_k \leq 1$, $k = 1, \ldots, K$, and $\sum_{k=1}^{K} p_k = 1$. Each mixture component is defined through a different vector of latent variables $\boldsymbol{Y}_k$, associated with $\boldsymbol{\theta}_k$, $k = 1, \ldots, K$. Without loss of generality, it is assumed that all of the mixture components are defined through the same matrix $\boldsymbol{A}$ of the form (2.12).

Under this mixture model, the unconditional expectation of $\boldsymbol{X}$ is given by

$$E(\boldsymbol{X}) = \sum_{k=1}^{K} p_k \boldsymbol{A}\boldsymbol{\theta}_k,$$

while its variance covariance matrix is given by

$$Var(\boldsymbol{X}) = \boldsymbol{A}\left[\sum_{k=1}^{K} p_k\left(\boldsymbol{\Sigma}_k + \boldsymbol{\theta}_k\boldsymbol{\theta}_k^T\right) - \left(\sum_{k=1}^{K} p_k\boldsymbol{\theta}_k\right)\left(\sum_{k=1}^{K} p_k\boldsymbol{\theta}_k\right)^T\right]\boldsymbol{A}^T,$$

where $\boldsymbol{\Sigma}_k = diag(\theta_{1k}, \ldots, \theta_{qk})$.

This is a typical finite mixture model and hence several properties and inferential procedures are applicable, as described in McLachlan and Peel (2000). Such procedures include ML estimation, selection of the optimal number of components K^*, non-parametric ML estimation of the mixing distribution etc.

2.3.2.1 Identifiability

Definition (Teicher, 1961): Mixtures of the density $f(x \mid \theta)$ are identifiable if and only if $\int_\Theta f(x \mid \theta)dG_1(\theta) = \int_\Theta f(x \mid \theta)dG_2(\theta)$ implies that $G_1(\cdot) = G_2(\cdot)$.

Following Al-Hussaini (1981), let

$$\mathcal{F}_{m,q} = \{F(x \mid \theta) : x \in \mathbb{R}^m, \theta \in \mathbb{R}_1^q\} \tag{2.16}$$

be a family of distribution functions and

$$\mathcal{K} = \{H : H(x) = \int_{\mathbb{R}_1^q} F(x \mid \theta)dG(\theta), G \in \mathcal{R}\}$$

be the class of mixtures H generated by $\mathcal{F}_{m,q}$, where $m, q \in \mathbb{N}$ and $m, q \geq 1$, $\mathbb{R}_1^q$ is a Borel subset of the Euclidean q-space $\mathbb{R}^q$, $F(x \mid \theta)$ is measurable in $\mathbb{R}^m \times \mathbb{R}_1^q$ and $\mathcal{R}$ is the class of q-dimensional distributions G whose induced measures μ_G assigns measure one to $\mathbb{R}_1^q$. The following theorem holds:

Theorem 1 (Al-Hussaini and Ahmad, 1981): For integers $m, q \geq 1$, let $\mathcal{F}_{m,q}$ be defined as in (2.16) with transforms $\phi(\boldsymbol{t})$ where $\boldsymbol{t} = (t_1, \ldots, t_m)^T$ defined for the domain S_ϕ of definition of ϕ such that the mapping $M : F \to \phi$ is linear and one-to-one. Suppose that there exists a total ordering $(\preceq)$ of $\mathcal{F}_{m,q}$ such that $F_1 \prec F_2$, $(F_j(x) = F(x \mid \theta_j))$ implies that (i) $S_{\phi_1} \subseteq S_{\phi_2}$ and (ii) there exists some $\boldsymbol{T} \in \bar{S}_{\phi_1}$ independent of ϕ_2 such that

$$\lim_{\boldsymbol{t}\to\mathbf{T}}\left(\frac{\phi_2(\boldsymbol{t})}{\phi_1(\boldsymbol{t})}\right) = 0.$$

Then the class of all finite mixtures of $\mathcal{F}_{m,q}$ is identifiable.

Theorem 2: Finite mixtures of the multivariate Poisson distribution with two-way covariance structure are identifiable.

Proof: Write the covariance parameters in matrix form, namely define

$$\mathbf{\Lambda} = \begin{bmatrix} 0 & \lambda_{12} & \dots & \lambda_{1m} \\ \lambda_{12} & 0 & \dots & \lambda_{2m} \\ \dots & \dots & \dots & \dots \\ \lambda_{1m} & \lambda_{2m} & \dots & 0 \end{bmatrix},$$

i.e. $\mathbf{\Lambda}$ has zeros at the diagonal and the covariance λ_{ij} as its ij-th element. For a general multivariate distribution the probability generating function (pgf) of the m-variate random vector $\boldsymbol{X}$ is defined as $\mathcal{G}(\boldsymbol{t}) = E\left(\prod_{j=1}^{m} t_j^{x_j}\right)$. In our multivariate Poisson case, the pgf is written in the form

$$\mathcal{G}(\boldsymbol{t}) = \exp\left(\boldsymbol{t}^T\boldsymbol{\theta}^{(1)} + \frac{1}{2}\boldsymbol{t}^T\mathbf{\Lambda}\boldsymbol{t} - \mathbf{1}_q^T\boldsymbol{\theta}\right), \tag{2.17}$$

where $\boldsymbol{\theta}^{(1)} = (\theta_1, \dots, \theta_m)^T$ and $\boldsymbol{t} \in \mathbb{R}^m$ (see, Johnson et al. (1997)).

We use the pgf given in (2.17) as the transform needed in Theorem 1 above. Then,

$$\lim_{\boldsymbol{t}\to\boldsymbol{T}} \frac{\mathcal{G}_2(\boldsymbol{t})}{\mathcal{G}_1(\boldsymbol{t})} = 0,$$

where $\boldsymbol{t} \to \boldsymbol{T}$ implies that every term of the vector $\boldsymbol{t}$ tends point-wise to the corresponding term of $\boldsymbol{T}$, if

a) $\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} \lambda_{1ij} > \sum_{i=1}^{m-1}\sum_{j=i+1}^{m} \lambda_{2ij}$, with $\boldsymbol{T} = (\infty, \dots, \infty)$,

b) $\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} \lambda_{1ij} = \sum_{i=1}^{m-1}\sum_{j=i+1}^{m} \lambda_{2ij}$ and $\sum_{i=1}^{m} \theta_{1i} > \sum_{i=1}^{m} \theta_{2i}$, with $\boldsymbol{T} = (\infty, \dots, \infty)$,

c) $\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} \lambda_{1ij} = \sum_{i=1}^{m-1}\sum_{j=i+1}^{m} \lambda_{2ij}$, $\sum_{i=1}^{m} \theta_{1i} = \sum_{i=1}^{m} \theta_{2i}$ and for some index $j \in \{1, \dots, m\}$, $\theta_{1j} > \theta_{2j}$, with $\boldsymbol{T} = (T_1, \dots, T_m)$ and elements $T_j = \infty$ and $T_r = c$ for all $r \neq j$, where c is a positive constant, and

d) $\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} \lambda_{1ij} = \sum_{i=1}^{m-1}\sum_{j=i+1}^{m} \lambda_{2ij}$, $\sum_{i=1}^{m} \theta_{1i} = \sum_{i=1}^{m} \theta_{2i}$ and for some indices $i \in \{1, \dots, m-1\}$, $j \in \{i+1, \dots, m\}$, $\lambda_{1ij} > \lambda_{2ij}$, with $\boldsymbol{T} = (T_1, \dots, T_m)$ and elements $T_j = \infty$ and $T_r = c$ for all $r \neq j$, where c is a positive constant.

The subscripts on $\mathcal{G}$ and their parameters denote the ordered components. Hence, Theorem 1 implies the identifiability of finite multivariate Poisson mixtures.

2.3.2.2 Marginal and Conditional Distributions

Under the finite multivariate Poisson mixture model the marginal distributions of the elements of $\boldsymbol{X}$ are finite univariate Poisson mixtures. Specifically, if we denote by $\boldsymbol{a}_j$ the j-th row of matrix $\boldsymbol{A}$, then each X_j follows a finite mixture of K Poisson distributions with mixing proportions $p_1, \ldots, p_K$ and parameters $\boldsymbol{a}_j\boldsymbol{\theta}_k$, $k = 1, \ldots, K$. Clearly, the marginal distributions are overdispersed.

The joint marginal distributions are again multivariate Poisson mixture distributions. Let $\boldsymbol{X}^{m'} = (X_{j_1}, \ldots, X_{j_{m'}})^T$ be a vector consisting of m' out of the m components of $\boldsymbol{X}$, where $m' < m$ and j_i, $i = 1, \ldots, m'$, are distinct indices with $j_i \in \{1, \ldots, m\}$. Then, $\boldsymbol{X}^{m'}$ follows a finite mixture of K multivariate Poisson distributions with mixing proportions p_k, $k = 1, \ldots, K$; each distribution has parameter vector $\boldsymbol{A}^{(m')}\boldsymbol{\theta}_k$, where $\boldsymbol{A}^{(m')}$ is the submatrix of $\boldsymbol{A}$ containing the rows indexed by $j_1, \ldots, j_{m'}$.

The conditional distributions are much more cumbersome and not of standard form. In the simplest case, the one without covariance terms, the conditional probability function takes the form

$$P(x_j \mid \boldsymbol{x}_{-j}) = \frac{\sum_{k=1}^{K} p_k \prod_{r=1}^{m} Po(x_r \mid \theta_{rk})}{\sum_{k=1}^{K} p_k \prod_{r \neq j} Po(x_r \mid \theta_{rk})}, \quad j = 1, \ldots, m,$$

where $\boldsymbol{x}_{-j}$ is the vector which contains all the elements of $\boldsymbol{x}$ apart from the j-th. This can be written as a finite Poisson mixture of the form

$$P(x_j \mid \boldsymbol{x}_{-j}) = \sum_{k=1}^{K} \pi_k Po(x_j \mid \theta_{jk}) \quad j = 1, \ldots, m,$$

where

$$\pi_k = \frac{p_k \prod_{r \neq j} Po(x_r \mid \theta_{rk})}{\sum_{\ell=1}^{K} p_\ell \prod_{r \neq j} Po(x_r \mid \theta_{r\ell})}, \quad k = 1, \ldots, K.$$

Each π_k corresponds to the posterior probability of the vector $\boldsymbol{x}_{-j}$ belonging to the $k-th$ component. Thus, the conditional distribution is again a finite Poisson mixture with updated mixing proportions.

2.3.2.3 Inference

Assume that we have observed m-variate random vectors $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{im})^T$, $i = 1, \ldots, n$, which, under the finite multivariate Poisson mixture model, belong to a superpopulation consisting of K distinct subpopulations in some proportions $\boldsymbol{p} = (p_1, \ldots, p_K)$, with corresponding multivariate Poisson probability distributions $MP_m(\boldsymbol{x} \mid \boldsymbol{\theta}_k t)$, $k = 1, ..., K$, where t is an offset. The probability of each observation $\boldsymbol{x}_i$ can then be represented in the finite mixture form defined in (2.15).

The offset t_i, $i = 1, \ldots, n$, may represent time, area, population etc. related to the i-th observation. In practice we may have different offsets for each variable. For example, if the vector of counts refers to the number of incidents in different age groups, naturally the offset for each variable will be the population of each group. Hence t_i can be a vector. For ease of exposition, in what follows we will treat t_i as a scalar, having in mind that the possibility of vector offsets slightly complicates the approach.

2.3.2.4 ML Estimation via an EM Algorithm

This section focuses on ML estimation of the parameter vector of the finite multivariate Poisson mixture model, $\boldsymbol{\psi} = (p_1, \ldots, p_{K-1}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$. Estimation of $\boldsymbol{\psi}$ can be obtained as a solution of the likelihood equation. Since this is quite cumbersome,a standard approach is followed for finite mixtures; using an EM type algorithm.

For $i = 1, \ldots, n$ we define the vector of indicator variables $\boldsymbol{Z}_i = (Z_{i1}, Z_{i2}, \ldots, Z_{iK})^T$ with $Z_{ik} = 1$ if $\boldsymbol{x}_i$ belongs to the k-th subpopulation and 0 otherwise. $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ are independent and identically distributed random variables following a multinomial distribution with parameters 1 and $\boldsymbol{p}$.

The derivation of the multivariate Poisson distribution via multivariate reduction implies that an EM scheme is also needed for the estimation of the parameter $\boldsymbol{\theta}_k = (\theta_{1k}, \ldots, \theta_{qk})^T$ of the distribution of each component. For the i-th observation and for the k-th component we define a vector of latent variables $\boldsymbol{Y}_{ik} = (Y_{i1k}, \ldots, Y_{iqk})^T$, $i = 1, \ldots, n$, $k = 1, \ldots, K$. According to the model specification, for each random variable we have that $Y_{irk} \sim Poisson(\theta_{rk} t_i)$, $r = 1, \ldots, q$. The complete data consist of the latent variables $\boldsymbol{Y}_{ik}$ and the indicator variables $\boldsymbol{Z}_i$. Using this data augmentation scheme the complete data loglikelihood is given by

$$\ell(\boldsymbol{\psi}) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\left[\log p_k + \sum_{r=1}^{q}\log(Po(y_{irk} \mid \theta_{rk}t_i))\right].$$

The EM algorithm proceeds as follows. Starting from some initial values for the model parameters, at the E-step the conditional expectations of the indicator variables $\boldsymbol{Z}_i$, $i = 1, \ldots, n$, given the $\boldsymbol{x}_i$'s and the current values of the estimates $\boldsymbol{\psi}^{(b-1)} = (p_1^{(b-1)}, \ldots, p_{K-1}^{(b-1)}, \boldsymbol{\theta}^{(b-1)})$, i.e. $\boldsymbol{w}_i = E(\boldsymbol{Z}_i \mid \boldsymbol{x}_i, \boldsymbol{\psi}^{(b-1)})$, are calculated first. Then, the conditional expectations of the latent variables $\boldsymbol{Y}_{ik}$, given the observed data and the current values of the estimates, i.e. the vectors $\boldsymbol{s}_{ik} = E(\boldsymbol{Y}_{ik} \mid \boldsymbol{x}_i, \boldsymbol{\psi}^{(b-1)})$ are computed for $i = 1, \ldots, n$ and $k = 1, \ldots, K$. At the M-Step the complete likelihood is maximized using $\boldsymbol{w}_i, \boldsymbol{s}_{ik}$ and the parameter estimates are updated.

A full description of the EM algorithm is the following. At the b-th iteration of the algorithm we have

- *E-Step:* Using the observed data and the current estimates $\boldsymbol{\psi}^{(b-1)}$, calculate the pseudo values

$$w_{ik} = \frac{p_k^{(b-1)} MP_m(\boldsymbol{x}_i \mid \boldsymbol{\theta}_k^{(b-1)} t_i)}{\sum_{\ell=1}^{K} p_\ell^{(b-1)} MP_m(\boldsymbol{x}_i \mid \boldsymbol{\theta}_\ell^{(b-1)} t_i)}, \quad i = 1, \ldots, n, \quad k = 1, \ldots, K,$$

  and

$$\boldsymbol{s}_{ik} = \frac{\sum_{\boldsymbol{y}_i \in g^{-1}(\boldsymbol{x}_i)} \boldsymbol{y}_{ik} \prod_{r=1}^{q} Po(y_{irk} \mid \theta_{rk}^{(b-1)} t_i)}{MP_m(\boldsymbol{x}_i \mid \boldsymbol{\theta}_k^{(b-1)} t_i)}, \quad i = 1, \ldots, n, \quad k = 1, \ldots, K,$$

  where $\boldsymbol{s}_{ik} = (s_{i1k}, \ldots, s_{iqk})^T$, and $\boldsymbol{\theta}_k^{(b)} = (\theta_{1k}^{(b)}, \ldots, \theta_{qk}^{(b)})$.

- *M-Step:* Update the estimates by

$$p_k^{(b)} = \frac{\sum_{i=1}^{n} w_{ik}}{n} \quad k = 1, \ldots, K.$$

  and in vector form

$$\boldsymbol{\theta}_k^{(b)} = \frac{\sum_{i=1}^{n} w_{ik}\boldsymbol{s}_{ik}}{\sum_{i=1}^{n} w_{ik}t_i} \quad k = 1, \ldots, K.$$

- If some convergence criterion is satisfied stop iterating, else go back to the E-step for one more iteration.

The above described EM algorithm has the pros and cons of every EM type algorithm. Since the size of missing information (latent variables) is large the algorithm is slow. Moreover, since in every iteration several probabilities are needed, it is important to use efficient algorithms to calculate the required probabilities through recursive schemes. On the other hand the algorithm is easily programmable in standard statistical packages. Calculations are tremendously simplified if the covariance parameters are set equal to zero.

Another feature to mention is that one may choose to work with the frequency table instead of the original counts, since some values are expected to appear more than once in the data. In the special case that all the offsets are equal this speeds up the estimation process substantially. The procedure is particularly suitable for data mining purposes where the datasets are very large but with some cells occurring at high frequencies.

2.4 Composite Likelihoods for Gaussian mixtures

Composite Likelihoods can apply in various data types further to the Poisson mixtures for count data we described in previous section. For the case of Multivariate Gaussian mixtures the estimators of the parameters associated to the probability mass function and the marginal bi-variate distributions are easy to calculate. For that case we explore the mathematical background of the Composite Likelihood concept and we propose an alternative composite likelihood method. For simplicity we will consider a 3-dimensional example of Gaussian mixture.

Example 2.4:

Let us consider $X = (X_1, X_2, X_3)^T$ be a vector of random variables which follows a 3-variate Normal distribution with parameters $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)$ and covariance matrix $\boldsymbol{\Sigma}$, where:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3 \end{bmatrix}$$

The marginal distributions for all pairs of X_i's are defined as follows:
$(X_1, X_2) \sim N_2(\boldsymbol{\mu_{12}} = (\mu_1, \mu_2), \boldsymbol{\Sigma_{12}})$, where

$$\boldsymbol{\Sigma_{12}} = \begin{bmatrix} \sigma_1 & \sigma_{12} \\ \sigma_{12} & \sigma_2 \end{bmatrix}$$

$(X_1, X_3) \sim N_2(\boldsymbol{\mu_{13}} = (\mu_1, \mu_3), \boldsymbol{\Sigma_{13}})$, where

$$\boldsymbol{\Sigma_{13}} = \begin{bmatrix} \sigma_1 & \sigma_{13} \\ \sigma_{13} & \sigma_3 \end{bmatrix}$$

$(X_2, X_3) \sim N_2(\boldsymbol{\mu_{23}} = (\mu_2, \mu_3), \boldsymbol{\Sigma_{23}})$, where

$$\boldsymbol{\Sigma_{23}} = \begin{bmatrix} \sigma_2 & \sigma_{23} \\ \sigma_{23} & \sigma_3 \end{bmatrix}$$

2.4.1 Full model evaluation

For the case of the illustrative example 2.4 the complete & the augmented likelihood for the full model evaluation is written in the below form for the above example and for a dataset of size n:

$$\mathcal{L}(X, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{n} \sum_{k=1}^{K} p_k N_3(\boldsymbol{X}_i | \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left\{ p_k N_3(\boldsymbol{X}_i | \boldsymbol{\mu}, \boldsymbol{\Sigma}) \right\}^{z_{ik}}$$

where $\boldsymbol{X}_i = (X_{1i}, X_{2i}, X_{3i})$ and $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ as described in example 2.4. The log-likelihood is defined as follows:

$$\ell = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \left\{ \log p_k + \log N_3(\boldsymbol{X}_i | \boldsymbol{\mu}, \boldsymbol{\Sigma}) \right\}$$

With the use of EM algorithm we maximize the complete-data log likelihood of the model for the r^{th} iteration as follows:

E-step: Calculate for $i = 1, ..., n$ and $k = 1, 2, ..., K$

$$w_{ik}^{(r+1)} = \frac{p_k^{(r)} N_3(\boldsymbol{X}_i|\boldsymbol{\mu_k}^{(r)}, \boldsymbol{\Sigma_k}^{(r)})}{\sum_{k=1}^{K} p_k^{(r)} N_3(\boldsymbol{X}_i|\boldsymbol{\mu_k}^{(r)}, \boldsymbol{\Sigma_k}^{(r)})}$$

M-step : Update $p_k^{(r+1)} = \sum_{i=1}^{n} w_{ik}^{(r+1)}/n$ for all $k = 1, 2, ..., K$ to estimate the mixing probabilities and then maximize the quantity

$$Q_k = E(\ell(x; z)) = \sum_{i=1}^{n} \Big\{ w_{ik}^{(r+1)} N_3(\boldsymbol{X}_i|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}) \Big\}$$

to get updated values for $\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}$ parameters associated with the k-th cluster $k = 1, 2, ..., K$.

For the k-th component the above optimization is solved as shown below:

$$\frac{\partial Q_k}{\partial \mu_k} = \frac{\partial}{\partial \mu_k} \sum_{i=1}^{n} w_{ik}^{(r+1)} \Big\{ -\log(\sqrt{(2\pi)^3}|\Sigma|) - \frac{1}{2}(X_i - \mu_k)^T \Sigma^{-1} (X_i - \mu_k) \Big\} = 0$$

$$\frac{1}{2} \sum_{i=1}^{n} w_{ik}^{(r+1)} \Big\{ (X_i - \mu_k)\Sigma^{-1} \Big\} = 0$$

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{k}}^{(\boldsymbol{r})} = \frac{\sum_{i=1}^{n} w_{ik}^{(r+1)} X_i}{n} = \bar{x}_k \tag{2.18}$$

and

$$\frac{\partial Q_k}{\partial \Sigma_k} = \frac{\partial}{\partial \Sigma_k} \sum_{i=1}^{n} w_{ik}^{(r+1)} \Big\{ \frac{1}{|\Sigma_k|} - \frac{1}{2}(X_i - \mu_k)^T (X_i - \mu_k) \frac{1}{\Sigma_k^T \Sigma_k} \Big\} = 0$$

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{k}}^{(\boldsymbol{r})} = \frac{1}{n} \sum_{i=1}^{n} w_{ik}^{(r+1)} (X_i - \bar{x}_k)^T (X_i - \bar{x}_k) \tag{2.19}$$

where 2.18 and 2.19 provide the estimators of the parameters of the 3-variate Normal distribution of the k^{th} component, $k = 1, 2, ..., K$.

2.4.2 Composite Likelihood evaluation

The complete likelihood for the composite likelihood method model evaluation is written in the below form for the above example and for a dataset of size n:

$$\begin{aligned}
\mathcal{L}(X, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \prod_{i=1}^{n} \prod_{s<t} \sum_{k=1}^{K} p_k N_2(X_s, X_t | \boldsymbol{\mu_{st}}, \boldsymbol{\Sigma_{st}}) \\
&= \prod_{i=1}^{n} \left\{ \sum_{k=1}^{K} p_k N_2(X_{1i}, X_{2i} | \boldsymbol{\mu_{12}}, \boldsymbol{\Sigma_{12}}) \right\} \left\{ \sum_{k=1}^{K} p_k N_2(X_{1i}, X_{3i} | \boldsymbol{\mu_{13}}, \boldsymbol{\Sigma_{13}}) \right\} \\
&\times \left\{ \sum_{k=1}^{K} p_k N_2(X_{2i}, X_{3i} | \boldsymbol{\mu_{23}}, \boldsymbol{\Sigma_{23}}) \right\} \\
&= \prod_{i=1}^{n} \prod_{k=1}^{K} \left\{ p_k N_2(X_{1k}, X_{2k} | \boldsymbol{\mu_{12}}, \boldsymbol{\Sigma_{12}}) \right\}^{z_{ik}} \prod_{k=1}^{K} \left\{ p_k N_2(X_{1k}, X_{3k} | \boldsymbol{\mu_{13}}, \boldsymbol{\Sigma_{13}}) \right\}^{z_{ik}} \\
&\times \prod_{k=1}^{K} \left\{ p_k N_2(X_{2k}, X_{3k} | \boldsymbol{\mu_{23}}, \boldsymbol{\Sigma_{23}}) \right\}^{z_{ik}}
\end{aligned} \tag{2.20}$$

The log-likelihood is defined as follows:

$$\begin{aligned}
\ell &= \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log p_k + \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log N_2(X_{1i}, X_{2i} | \boldsymbol{\mu_{12}}, \boldsymbol{\Sigma_{12}}) \\
&+ \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log p_k + \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log N_2(X_{1i}, X_{3i} | \boldsymbol{\mu_{13}}, \boldsymbol{\Sigma_{13}}) \\
&+ \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log p_k + \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log N_2(X_{2i}, X_{3i} | \boldsymbol{\mu_{23}}, \boldsymbol{\Sigma_{23}}) \\
&= 3 \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log p_k + \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log N_2(X_{1i}, X_{2i} | \boldsymbol{\mu_{12}}, \boldsymbol{\Sigma_{12}}) \\
&+ \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log N_2(X_{1i}, X_{3i} | \boldsymbol{\mu_{13}}, \boldsymbol{\Sigma_{13}}) \\
&+ \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log N_2(X_{2i}, X_{3i} | \boldsymbol{\mu_{23}}, \boldsymbol{\Sigma_{23}})
\end{aligned} \tag{2.21}$$

With the use of EM algorithm we maximize the augmented log likelihood of the model for the r^{th} iteration as follows:

E-step: Calculate for $i = 1, ..., n$ and $k = 1, 2, ..., K$

$$w_{ik12}^{(r+1)} = \frac{p_k^{(r)} N_2(X_{1i}, X_{2i}|\boldsymbol{\mu}_{\mathbf{12k}}^{(r)}, \boldsymbol{\Sigma}_{\mathbf{12k}}^{(r)})}{\sum_{k=1}^{K} p_k^{(r)} N_2(X_{1i}, X_{2i}|\boldsymbol{\mu}_{\mathbf{12k}}^{(r)}, \boldsymbol{\Sigma}_{\mathbf{12k}}^{(r)})}$$

$$w_{ik13}^{(r+1)} = \frac{p_k^{(r)} N_2(X_{1i}, X_{3i}|\boldsymbol{\mu}_{\mathbf{13k}}^{(r)}, \boldsymbol{\Sigma}_{\mathbf{13k}}^{(r)})}{\sum_{k=1}^{K} p_k^{(r)} N_2(X_{1i}, X_{3i}|\boldsymbol{\mu}_{\mathbf{13k}}^{(r)}, \boldsymbol{\Sigma}_{\mathbf{13k}}^{(r)})}$$

$$w_{ik23}^{(r+1)} = \frac{p_k^{(r)} N_2(X_{2i}, X_{3i}|\boldsymbol{\mu}_{\mathbf{23k}}^{(r)}, \boldsymbol{\Sigma}_{\mathbf{23k}}^{(r)})}{\sum_{k=1}^{K} p_k^{(r)} N_2(X_{2i}, X_{3i}|\boldsymbol{\mu}_{\mathbf{23k}}^{(r)}, \boldsymbol{\Sigma}_{\mathbf{23k}}^{(r)})}$$

and so

$$w_{ik}^{(r+1)} = \frac{w_{ik12}^{(r+1)} + w_{ik13}^{(r+1)} + w_{ik23}^{(r+1)}}{3}$$

M-step : Update $p_k^{(r+1)} = \sum_{i=1}^{n} w_{ik}^{(r+1)}/n$ for all $k = 1, 2, ..., K$ and with the use of updated values of the quantities w_{ikst}, $s < t$ maximize the quantity:

$$\begin{aligned} Q_k = E(\ell(x; z)) = &\sum_{i=1}^{n} w_{ik12}^{(r+1)} \log N_2(X_{1i}, X_{2i}|\boldsymbol{\mu}_{\mathbf{12}}, \boldsymbol{\Sigma}_{\mathbf{12}}) \\ &+ \sum_{i=1}^{n} w_{ik13}^{(r+1)} \log N_2(X_{1i}, X_{3i}|\boldsymbol{\mu}_{\mathbf{13}}, \boldsymbol{\Sigma}_{\mathbf{13}}) \\ &+ \sum_{i=1}^{n} w_{ik23}^{(r+1)} \log N_2(X_{2i}, X_{3i}|\boldsymbol{\mu}_{\mathbf{23}}, \boldsymbol{\Sigma}_{\mathbf{23}}) \end{aligned} \tag{2.22}$$

to get updated values for $\boldsymbol{\mu}_{\mathbf{12k}}, \boldsymbol{\mu}_{\mathbf{13k}}, \boldsymbol{\mu}_{\mathbf{23k}}, \boldsymbol{\Sigma}_{\mathbf{12k}}, \boldsymbol{\Sigma}_{\mathbf{13k}}, \boldsymbol{\Sigma}_{\mathbf{23k}}$ parameters associated with the k-th component $k = 1, 2, ..., K$.

For the k-th component the above maximization is performed as shown below:

$$\frac{\partial Q_k}{\partial \mu_{12k}} = 0, \qquad \frac{\partial Q_k}{\partial \mu_{13k}} = 0, \qquad \frac{\partial Q_k}{\partial \mu_{23k}} = 0$$

and

$$\frac{\partial Q_k}{\partial \Sigma_{12k}} = 0, \qquad \frac{\partial Q_k}{\partial \Sigma_{13k}} = 0, \qquad \frac{\partial Q_k}{\partial \Sigma_{23k}} = 0$$

The estimators provided by equations 2.18 & 2.19 for all clusters $k = 1, 2, ..., K$

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{k}}^{(\boldsymbol{r})} = \frac{\sum_{i=1}^{n} w_{ik}^{(r+1)} X_i}{n}$$

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{k}}^{(\boldsymbol{r})} = \frac{1}{n} \sum_{i=1}^{n} w_{ik}^{(r+1)} (X_i - \bar{x}_k)^T (X_i - \bar{x}_k)$$

are solving the above equations and so we can observe that the methodology works and provides adequate estimations.

2.4.3 An alternative approach

At this section we consider an alternative method of composite likelihood which in many cases is less computational complicated. For the case of multivariate Normal mixtures, where the estimators of composite likehood are provided by the equations 2.18 & 2.19, the alternative proposed method is not adequate, though it can provide solution to other mixtures which needs an optimization process. The complete & the augmented likelihood for the composite likelihood method model is written in the below form for the above example and for a dataset of size n:

$$\begin{aligned}
\mathcal{L}(X, \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \prod_{i=1}^{n} \sum_{k=1}^{K} p_k \prod_{s<t} N_2(X_s, X_t | \boldsymbol{\mu_{st}}, \boldsymbol{\Sigma_{st}}) \\
&= \prod_{i=1}^{n} \sum_{k=1}^{K} p_k N_2(X_{1i}, X_{2i} | \boldsymbol{\mu_{12}}, \boldsymbol{\Sigma_{12}}) N_2(X_{1i}, X_{3i} | \boldsymbol{\mu_{13}}, \boldsymbol{\Sigma_{13}}) \\
&\times N_2(X_{2i}, X_{3i} | \boldsymbol{\mu_{23}}, \boldsymbol{\Sigma_{23}}) \\
&= \prod_{i=1}^{n} \prod_{k=1}^{K} \Big\{ p_k N_2(X_{1k}, X_{2k} | \boldsymbol{\mu_{12}}, \boldsymbol{\Sigma_{12}}) N_2(X_{1k}, X_{3k} | \boldsymbol{\mu_{13}}, \boldsymbol{\Sigma_{13}}) \\
&\times N_2(X_{2k}, X_{3k} | \boldsymbol{\mu_{23}}, \boldsymbol{\Sigma_{23}}) \Big\}^{z_{ik}}
\end{aligned} \tag{2.23}$$

The log-likelihood is defined as follows:

$$\begin{aligned}
\ell &= \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log p_k + \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log N_2(X_{1i}, X_{2i} | \boldsymbol{\mu_{12}}, \boldsymbol{\Sigma_{12}}) \\
&+ \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log N_2(X_{1i}, X_{3i} | \boldsymbol{\mu_{13}}, \boldsymbol{\Sigma_{13}}) \\
&+ \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log N_2(X_{2i}, X_{3i} | \boldsymbol{\mu_{23}}, \boldsymbol{\Sigma_{23}})
\end{aligned} \tag{2.24}$$

As we can observe from equations 2.21 & 2.24 Difference of this alternative method compared to the composite likelihood method is that the factor $\sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log p_k$ appears only one time in the equation of the alternative method. As a result, the expectation maximization algorithm can be performed via the following steps:

With the use of EM algorithm we maximize the augmented log likelihood of the model for the r^{th} iteration as follows:

E-step: Calculate for $i = 1, ..., n$ and $k = 1, 2, ..., K$

$$w_{ik}^{(r+1)} = \frac{p_k^{(r)} N_2(X_{1i}, X_{2i}|\boldsymbol{\mu}_{\mathbf{12}\boldsymbol{k}}^{(r)}, \boldsymbol{\Sigma}_{\mathbf{12}\boldsymbol{k}}^{(r)}) N_2(X_{1i}, X_{3i}|\boldsymbol{\mu}_{\mathbf{13}\boldsymbol{k}}^{(r)}, \boldsymbol{\Sigma}_{\mathbf{13}\boldsymbol{k}}^{(r)}) N_2(X_{2i}, X_{3i}|\boldsymbol{\mu}_{\mathbf{23}\boldsymbol{k}}^{(r)}, \boldsymbol{\Sigma}_{\mathbf{23}\boldsymbol{k}}^{(r)})}{\sum_{k=1}^{K} p_k^{(r)} N_2(X_{1i}, X_{2i}|\boldsymbol{\mu}_{\mathbf{12}\boldsymbol{k}}^{(r)}, \boldsymbol{\Sigma}_{\mathbf{12}\boldsymbol{k}}^{(r)}) N_2(X_{1i}, X_{3i}|\boldsymbol{\mu}_{\mathbf{13}\boldsymbol{k}}^{(r)}, \boldsymbol{\Sigma}_{\mathbf{13}\boldsymbol{k}}^{(r)}) N_2(X_{2i}, X_{3i}|\boldsymbol{\mu}_{\mathbf{23}\boldsymbol{k}}^{(r)}, \boldsymbol{\Sigma}_{\mathbf{23}\boldsymbol{k}}^{(r)})}$$

M-step : Update $p_k^{(r+1)} = \sum_{i=1}^{n} w_{ik}^{(r+1)}/n$ for all $k = 1, 2, ..., K$ and with the use of updated values of the quantities w_{ik}, maximize the quantity:

$$\begin{aligned} Q_k = E(\ell(x; z)) = &\sum_{i=1}^{n} w_{ik}^{(r+1)} \log N_2(X_{1i}, X_{2i}|\boldsymbol{\mu}_{\mathbf{12}}, \boldsymbol{\Sigma}_{\mathbf{12}}) \\ &+ \sum_{i=1}^{n} w_{ik}^{(r+1)} \log N_2(X_{1i}, X_{3i}|\boldsymbol{\mu}_{\mathbf{13}}, \boldsymbol{\Sigma}_{\mathbf{13}}) \\ &+ \sum_{i=1}^{n} w_{ik}^{(r+1)} \log N_2(X_{2i}, X_{3i}|\boldsymbol{\mu}_{\mathbf{23}}, \boldsymbol{\Sigma}_{\mathbf{23}}) \end{aligned}$$

to get updated values for $\boldsymbol{\mu}_{\mathbf{12}\boldsymbol{k}}, \boldsymbol{\mu}_{\mathbf{13}\boldsymbol{k}}, \boldsymbol{\mu}_{\mathbf{23}\boldsymbol{k}}, \boldsymbol{\Sigma}_{\mathbf{12}\boldsymbol{k}}, \boldsymbol{\Sigma}_{\mathbf{13}\boldsymbol{k}}, \boldsymbol{\Sigma}_{\mathbf{23}\boldsymbol{k}}$ parameters associated with the k-th component $k = 1, 2, ..., K$.

The EM algorithm has properties which are violated for the specific approach. Reason for is that the calculation for the E-step step and the $p_k \prod_{s<t} N_2(X_s, X_t, m_{st}, \Sigma_{st})$, $k = 1, 2, ..., K$ does not correspond to a distribution function. Once this function is transformed into a distribution function with a normalized parameter α, so as

$$\alpha \int\int\int_{(X_1, X_2, X_3)} p_k \prod_{s<t} N_2(X_{si}, X_{ti}|\boldsymbol{\mu}_{\boldsymbol{stk}}, \boldsymbol{\Sigma}_{\boldsymbol{stk}}) dX_1 dX_2 dX_3 = 1 \tag{2.25}$$

the algorithm should provide equivalent results to the full composite likelihood approach. The calculation of the parameter α of equation 2.25 can lead to high computational complexity, though the method can be evaluated in which level provides correct classification.

2.5 Sampling Method

Let $X = (X_1, X_2, ..., X_m)^T$ be a vector of random variables which follows a Multivariate Poisson distribution with parameters $\theta = (\theta_1, \theta_2, ..., \theta_q)$. Assume the transformation: $AX = Y$ where $Y = (Y_1, Y_2, ..., Y_q)$, $q = m + \binom{m}{2} \geq m$ and $Y_r \sim Poisson(\theta_r)$, $r = 1, ..., q$.

$A = [A_1, A_2]$ is a $m \times q$ matrix where A_1 is the identity matrix with dimension m and A_2 is an $m \times (m-1)/2$ matrix where each of its columns contain exactly 2 ones and there are no duplicate columns. The generating function of $X's$ is given by the following expression:

$$g(s) = exp\left\{\sum_{t=1}^{m} \theta_t(s_t - 1) + \sum_{t=1}^{m-1}\sum_{j=t+1}^{m} \theta_{ij}(s_t s_j - 1)\right\},$$

where $s = (s_1, s_2, ...s_m)$ and marginal distributions can be found by setting $si = 1$ to the appropriate index i. The full augmented likelihood that needs to be maximized is:

$$\mathcal{L}(X, \theta) = \prod_{i=1}^{n}\prod_{k=1}^{K} p_k Po_m(\boldsymbol{X}|\boldsymbol{\Theta_k}) \tag{2.26}$$

For multidimensional data where m is large, the 2.26 can be very exhaustive in terms of computational effort. Composite likelihood is a tool that allows less computational effort. Instead as described in previous sections, we focus on maximizing the composite likelihood below:

$$\mathcal{L}_{CL}(X, \theta) = \prod_{i=1}^{n}\prod_{s<t}\sum_{k=1}^{K} p_k Po_2(X_s, X_t|\boldsymbol{\Theta_{stk}}) \tag{2.27}$$

where $\boldsymbol{\Theta_{stk}}$, $s < t \leq m$ are appropriate parameters for each cluster $k = 1, 2, ..., K$ and for each of the bi-variate Poisson p.m.f. of X_s, X_t and $Po_2(X_s, X_t|\boldsymbol{\Theta_{stk}})$ is the probability mass function of the bivariate Poisson distribution for the respective component defined by s & t.

In order to further reduce the computational effort of maximizing the proposed CL likelihood, we introduce a technique that uses systematic or non-systematic sampling to the pairwise bivariate Poisson pmf's as described below with the following example and in the following sections.

Example 2.5:

Let's assume a random vector of size n of a 3-variate Poisson distribution so as $X = (X_1, X_2, X_3) \sim Po_3(X|\theta_1, \theta_2, \theta_3, \theta_{12}, \theta_{13}, \theta_{23})$. The full complete composite mixture likelihood to be optimized is written in the below form:

$$\mathcal{L}(\boldsymbol{X}, \boldsymbol{\Theta}) = \prod_{i=1}^{n}\left\{\sum_{k=1}^{K} p_k Po_2(X_{1i}, X_{2i}|u_{1k})\right\}\left\{\sum_{k=1}^{K} p_k Po_2(X_{1i}, X_{3i}|u_{2k})\right\}\left\{\sum_{k=1}^{K} p_k Po_2(X_{2i}, X_{3i}|u_{3k})\right\}$$

where $k = 1, 2, ..., K$ is the number of clusters and

$$u_{1k} = (\theta_{1k} + \theta_{13k}, \theta_{2k} + \theta_{23k}, \theta_{12k}),\ k = 1, ..., K$$

$$u_{2k} = (\theta_{1k} + \theta_{12k}, \theta_{3k} + \theta_{23k}, \theta_{13k}),\ k = 1, ..., K$$

$$u_{3k} = (\theta_{2k} + \theta_{12k}, \theta_{3k} + \theta_{13k}, \theta_{23k}),\ k = 1, ..., K$$

as defined from the Poisson marginal distributions. The augmented likelihood is now written in the below form:

$$\mathcal{L}(\boldsymbol{X}, \boldsymbol{\Theta}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left\{ p_k Po_2(X_{i1}, X_{2i}|u_{1k}) \right\}^{z_{ik}} \prod_{k=1}^{K} \left\{ p_k Po_2(X_{1i}, X_{3i}|u_{2k}) \right\}^{z_{ik}} \prod_{k=1}^{K} \left\{ p_k Po_2(X_{2i}, X_{3i}|u_{3k}) \right\}^{z_{ik}}$$

where $z_{ik} = 1$, if the observation from the i-th row comes from cluster k, otherwise is $z_{ik} = 0$. The log-likelihood to be optimized via EM algorithm is now written in the form:

$$\begin{aligned} \ell = & \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log p_k + \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log Po_2(X_{1i}, X_{2i}|u_{1k}) \\ & + \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log p_k + \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log Po_2(X_{1i}, X_{3i}|u_{2k}) \\ & + \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log p_k + \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log Po_2(X_{2i}, X_{3i}|u_{3k}) \\ & = \ell_{CL12} + \ell_{CL13} + \ell_{CL23} \end{aligned} \tag{2.28}$$

The equation 2.28 can be optimized separately for each of the components in each row with the use of EM algorithm. For the r^{th} iteration of the algorithm we perform:

E-step: Calculate for $i = 1, ..., n$ and $k = 1, 2, ..., K$

$$w_{ik12}^{(r+1)} = \frac{p_k^{(r)} Po_2(X_{1i}, X_{2i}|u_{1k}^{(r)})}{\sum_{k=1}^{K} p_k^{(r)} Po_2(X_{1i}, X_{2i}|u_{1k}^{(r)}))}$$

$$w_{ik13}^{(r+1)} = \frac{p_k^{(r)} Po_2(X_{1i}, X_{3i}|u_{2k}^{(r)})}{\sum_{k=1}^{K} p_k^{(r)} Po_2(X_{1i}, X_{3i}|u_{2k}^{(r)}))}$$

$$w_{ik23}^{(r+1)} = \frac{p_k^{(r)} Po_2(X_{2i}, X_{3i}|u_{3k}^{(r)})}{\sum_{k=1}^{K} p_k^{(r)} Po_2(X_{2i}, X_{3i}|u_{3k}^{(r)}))}$$

and so

$$w_{ik}^{(r+1)} = \frac{w_{ik12}^{(r+1)} + w_{ik13}^{(r+1)} + w_{ik23}^{(r+1)}}{3}$$

M-step : Update $p_k^{(r+1)} = \sum_{i=1}^{n} w_{ik}^{(r+1)}/n$ for all $k = 1, 2, ..., K$ and then maximize the quantity:

$$\begin{aligned} Q_k &= Q_{k12} + Q_{k13} + Q_{k23} \\ &= \sum_{i=1}^{n} w_{ik12}^{(r+1)} \log Po_2(X_{1i}, X_{2i}|u_{1k}) \\ &+ \sum_{i=1}^{n} w_{ik13}^{(r+1)} \log Po_2(X_{1i}, X_{3i}|u_{2k}) \\ &+ \sum_{i=1}^{n} w_{ik23}^{(r+1)} \log Po_2(X_{2i}, X_{3i}|u_{3k}) \end{aligned} \tag{2.29}$$

to get updated values for u_{1k}, u_{2k}, u_{3k} linked to the $\theta_{1k}, \theta_{2k}, \theta_{3k}, \theta_{12k}, \theta_{13k}, \theta_{23k}$ parameters associated with the k-th cluster $k = 1, 2, ..., K$. The optimization process for the case of count data and Poisson mixtures can be also be optimized with the use of ECM algorithm in a way that the maximization of M-step splits into 4 steps and in each of the steps the estimated parameters are used as input for the next CM step. For the case of the 3-variate Poisson distribution the maximization step of EM provides results directly for the $m + \binom{m}{2} = 6$ θ parameters in one step. For $m = 4$ dimensions M-step should estimate $m + \binom{m}{2} = 4 + \binom{4}{2} = 10$ parameters and so on. The θ_j, $j = 1, 2, ..., m$ participate in $m - 1$ equations out of the $\binom{m}{2}$, while all other θ_{ij}, $i, j = 1, ..., m$ in all $\binom{m}{2}$ equations. So in the general framework of m dimensions the ECM algorithm can be performed into $m + 1$ steps, one for each of the θ_j, $j = 1, 2, ..., m$ and one for the rest θ_{ij}, $i, j = 1, ..., m$. In our example the ECM is as follows:

CM-step 1 : Update $p_k^{(r+1)} = \sum_{i=1}^{n} w_{ik}^{(r+1)}/n$ for all $k = 1, 2, ..., K$ and then maximize the quantity:

$$\begin{aligned} Q_{k1} &= Q_{k12} + Q_{k13} \\ &= \sum_{i=1}^{n} w_{ik12}^{(r+1)} \log Po_2(X_{1i}, X_{2i}|(\theta_{1k} + \theta_{13k}^{(r)}, \theta_{2k}^{(r)} + \theta_{23k}^{(r)}, \theta_{12k}^{(r)})) \\ &+ \sum_{i=1}^{n} w_{ik13}^{(r+1)} \log Po_2(X_{1i}, X_{3i}|(\theta_{1k} + \theta_{12k}^{(r)}, \theta_{3k}^{(r)} + \theta_{23k}^{(r)}, \theta_{13k}^{(r)})) \end{aligned}$$

to get updated values for θ_{1k}, $k = 1, 2, .., K$.

CM-step 2 : Maximize the quantity:

$$\begin{aligned} Q_{k2} &= Q_{k12} + Q_{k23} \\ &= \sum_{i=1}^{n} w_{ik12}^{(r+1)} \log Po_2(X_{1i}, X_{2i}|(\theta_{1k}^{(r+1)} + \theta_{13k}^{(r)}, \theta_{2k} + \theta_{23k}^{(r)}, \theta_{12k}^{(r)})) \\ &+ \sum_{i=1}^{n} w_{ik23}^{(r+1)} \log Po_2(X_{2i}, X_{3i}|(\theta_{2k} + \theta_{12k}^{(r)}, \theta_{3k}^{(r)} + \theta_{13k}^{(r)}, \theta_{23k}^{(r)})) \end{aligned}$$

to get updated values for θ_{2k}, $k = 1, 2, .., K$, with respect to the updated values of $\theta_{1k}^{(r+1)}$.

CM-step 3 : Maximize the quantity:

$$\begin{aligned} Q_{k3} &= Q_{k13} + Q_{k23} \\ &= \sum_{i=1}^{n} w_{ik13}^{(r+1)} \log Po_2(X_{1i}, X_{3i}|(\theta_{1k}^{(r+1)} + \theta_{12k}^{(r)}, \theta_{3k} + \theta_{23k}^{(r)}, \theta_{13k}^{(r)})) \\ &+ \sum_{i=1}^{n} w_{ik23}^{(r+1)} \log Po_2(X_{2i}, X_{3i}|(\theta_{2k}^{(r+1)} + \theta_{12k}^{(r)}, \theta_{3k} + \theta_{13k}^{(r)}, \theta_{23k}^{(r)})) \end{aligned}$$

to get updated values for θ_{3k}, $k = 1, 2, .., K$, with respect to the updated values of $\theta_{1k}^{(r+1)}$,$\theta_{2k}^{(r+1)}$.

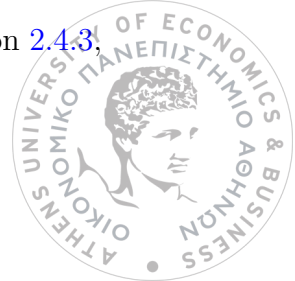
CM-step 4 : Maximize the quantity:

$$\begin{aligned} Q_{k4} &= Q_{k12} + Q_{k13} + Q_{k23} \\ &= \sum_{i=1}^{n} w_{ik12}^{(r+1)} \log Po_2(X_{1i}, X_{2i}|(\theta_{1k}^{(r+1)} + \theta_{13k}, \theta_{2k}^{(r+1)} + \theta_{23k}, \theta_{12k})) \\ &+ \sum_{i=1}^{n} w_{ik13}^{(r+1)} \log Po_2(X_{1i}, X_{3i}|(\theta_{1k}^{(r+1)} + \theta_{12k}, \theta_{3k}^{(r+1)} + \theta_{23k}, \theta_{13k})) \\ &+ \sum_{i=1}^{n} w_{ik23}^{(r+1)} \log Po_2(X_{2i}, X_{3i}|(\theta_{2k}^{(r+1)} + \theta_{12k}, \theta_{3k}^{(r+1)} + \theta_{13k}, \theta_{23k})) \end{aligned}$$

to get updated values for θ_{ijk}, $i, j = 1, 2, 3$, $k = 1, 2, .., K$, with respect to the updated values of $\theta_{1k}^{(r+1)}$,$\theta_{2k}^{(r+1)}$,$\theta_{3k}^{(r+1)}$.

The alternative composite likelihood

For the specific example of Poisson mixtures and for the case of the alternative composite approach, as described for the case of normal mixtures in section 2.4.3,



the alternative composite likelihood and the corresponding log- likelihood are written in the below forms for a dataset of size n:

$$\begin{aligned}\mathcal{L}(X,\boldsymbol{\Theta}) &= \prod_{i=1}^{n}\sum_{k=1}^{K} p_k Po_2(X_{1i},X_{2i}|u_{1k})Po_2(X_{1i},X_{3i}|u_{2k})Po_2(X_{2i},X_{3i}|u_{3k}) \\ &= \prod_{i=1}^{n}\prod_{k=1}^{K}\left\{p_k Po_2(X_{1i},X_{2i}|u_{1k})Po_2(X_{1i},X_{3i}|u_{2k})Po_2(X_{2i},X_{3i}|u_{3k})\right\}^{z_{ik}}\end{aligned} \tag{2.30}$$

$$\begin{aligned}\ell = &\sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log p_k + \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log Po_2(X_{1i},X_{2i}|u_{1k})) \\ &+ \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log Po_2(X_{1i},X_{3i}|u_{2k})) \\ &+ \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log Po_2(X_{2i},X_{3i}|u_{3k}))\end{aligned} \tag{2.31}$$

The expectation maximization algorithm can be performed via the following steps:

With the use of EM algorithm we maximize the augmented log likelihood of the model for the r^{th} iteration as follows:

E-step: Calculate for $i = 1, ..., n$ and $k = 1, 2, ..., K$

$$w_{ik}^{(r+1)} = \frac{p_k^{(r)} Po_2(X_{1i},X_{2i}|u_{1k}^{(r)})Po_2(X_{1i},X_{3i}|u_{2k}^{(r)})Po_2(X_{2i},X_{3i}|u_{3k}^{(r)})}{\sum_{k=1}^{K} p_k^{(r)} Po_2(X_{1i},X_{2i}|u_{1k}^{(r)})Po_2(X_{1i},X_{3i}|u_{2k}^{(r)})Po_2(X_{2i},X_{3i}|u_{3k}^{(r)})}$$

M-step : Update $p_k^{(r+1)} = \sum_{i=1}^{n} w_{ik}^{(r+1)}/n$ for all $k = 1, 2, ..., K$ and with the use of updated values of the quantities w_{ik}, maximize the quantity:

$$\begin{aligned}Q_k = E(\ell(x;z)) = &\sum_{i=1}^{n} w_{ik}^{(r+1)}\log Po_2(X_{1i},X_{2i}|u_{1k}) \\ &+ \sum_{i=1}^{n} w_{ik}^{(r+1)}\log Po_2(X_{1i},X_{3i}|u_{2k}) \\ &+ \sum_{i=1}^{n} w_{ik}^{(r+1)}\log Po_2(X_{2i},X_{3i}|u_{3k})\end{aligned}$$

to get updated values for $\theta_{1k}, \theta_{2k}, \theta_{3k}, \theta_{12k}, \theta_{13k}, \theta_{23k}$ parameters associated through u_{k1}, u_{k2}, u_{k3} and for the k-th component $k = 1, 2, ..., K$.

Sampling methods

Systematic sampling refers to a technique where for a dataset of length n we choose a specific number of the 3 ℓ_{CLst} components of equation 2.29 in a systematic way which ensures that all components are found for almost the same number of observations of the sample X. All n observations of the random vector X participates in the evaluation of the parameters.

Non-systematic approach refers to choosing the components with one probability which means that we choose ℓ_{CLst} components of X_i's and at the same time rows of X_i's. Since all $\theta's$ participate in each of the CL components as in equation 2.29 we expect that the methods works well in estimating the Poisson parameters for the case of multidimensional count data. More details can be found in the following sections.

2.5.1 Systematic Sampling 1 for Poisson mixtures:

Let's assume the previous example 2.5 of n observations of $X = (X_1, X_2, X_3) \sim Po_3(\theta_{1k}, \theta_{2k}, \theta_{3k}, \theta_{12k}, \theta_{13k}, \theta_{23k})$, $k = 1, ..., K$ cluster and the marginal log likelihoods of the bivariate poisson distribution

$$\ell_{CLst} = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} \log p_k + \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} \log Po_2(X_{si}, X_{ti}|u_{stk}), \;\; s < t \leq m, \;\; k = 1, .., K$$

where u_{stk} the corresponding vector of θ's related to the random vector $X = (X_s, X_t)$ and the k^{th} cluster, $k = 1, 2, ..., K$. The log-likelihood that we need to maximize through composite likelihood concept is described in equation 2.28, where $k = 1, 2, ..., K$ is the number of clusters and

$$u_{1k} = (\theta_{1k} + \theta_{13k}, \theta_{2k} + \theta_{23k}, \theta_{12k}), \;\; k = 1, ..., K$$

$$u_{2k} = (\theta_{1k} + \theta_{12k}, \theta_{3k} + \theta_{23k}, \theta_{13k}), \;\; k = 1, ..., K$$

$$u_{3k} = (\theta_{2k} + \theta_{12k}, \theta_{3k} + \theta_{13k}, \theta_{23k}), \;\; k = 1, ..., K.$$

The Systematic Sampling approach 1 proposes to maximize the sampled composite likelihood by choosing 2 out of the 3 pairs of marginal log-likelihoods ℓ_{CLst} described in equation 2.28. The choice of the two pairs correlates to the row of the observation sample data i, $i = 1, ..., n$. For the first observation where mode(i,3)=1 we choose the first 2 components,for the second observation the first and the third

component, while for the third observation the last two components. This is performed for all observations n and the same approach is performed for all clusters $k = 1, 2, ..., K$.

In order to estimate the θ parameters let's define the below indices:

$$I_{1i} = \begin{cases} 1 & \text{if mode}(i,3) = 1 \\ 0 & \text{elsewhere} \end{cases}$$

$$I_{2i} = \begin{cases} 1 & \text{if mode}(i,3) = 2 \\ 0 & \text{elsewhere} \end{cases}$$

$$I_{3i} = \begin{cases} 1 & \text{if mode}(i,3) = 0 \\ 0 & \text{elsewhere} \end{cases}$$

The sampled composite likelihood is now written in the below form:

$$\begin{aligned}
\ell_{Sam1CL} &= 2\sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log p_k \\
&+ \sum_{i=1}^{n}\sum_{k=1}^{K}\Big\{z_{ik}\log Po_2(X_{1i}, X_{2i}|u_{1k})\Big\}I_{1i} + \sum_{i=1}^{n}\sum_{k=1}^{K}\Big\{z_{ik}\log Po_2(X_{1i}, X_{3i}|u_{2k})\Big\}I_{1i} \\
&+ \sum_{i=1}^{n}\sum_{k=1}^{K}\Big\{z_{ik}\log Po_2(X_{1i}, X_{2i}|u_{1k})\Big\}I_{2i} + \sum_{i=1}^{n}\sum_{k=1}^{K}\Big\{z_{ik}\log Po_2(X_{2i}, X_{3i}|u_{3k})\Big\}I_{2i} \\
&+ \sum_{i=1}^{n}\sum_{k=1}^{K}\Big\{z_{ik}\log Po_2(X_{1i}, X_{3i}|u_{2k})\Big\}I_{3i} + \sum_{i=1}^{n}\sum_{k=1}^{K}\Big\{z_{ik}\log Po_2(X_{2i}, X_{3i}|u_{3k})\Big\}I_{3i} \\
&= 2\sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log p_k + \sum_{i=1}^{n}\sum_{k=1}^{K}\Big\{z_{ik}\log Po_2(X_{1i}, X_{2i}|u_{1k})\Big\}(I_{1i} + I_{2i}) \\
&+ \sum_{i=1}^{n}\sum_{k=1}^{K}\Big\{z_{ik}\log Po_2(X_{1i}, X_{3i}|u_{2k})\Big\}(I_{1i} + I_{3i}) \\
&+ \sum_{i=1}^{n}\sum_{k=1}^{K}\Big\{z_{ik}\log Po_2(X_{2i}, X_{3i}|u_{3k})\Big\}(I_{2i} + I_{3i})
\end{aligned} \tag{2.32}$$

2.5.1.1 EM algorithm

Similarly to the full composite likelihood approach, with the use of EM algorithm we maximize the complete-data log likelihood of the model for the r^{th} iteration as

follows:

E-step: Calculate for $i = 1, ..., n$ and $k = 1, 2, ..., K$

if mode$(i, 3) = 1$

$$w_{ik12}^{(r+1)} = \frac{p_k^{(r)} Po_2(X_{1i}, X_{2i}|u_{1k}^{(r)})}{\sum_{k=1}^{K} p_k^{(r)} Po_2(X_{1i}, X_{2i}|u_{1k}^{(r)})}$$

$$w_{ik13}^{(r+1)} = \frac{p_k^{(r)} Po_2(X_{1i}, X_{3i}|u_{2k}^{(r)})}{\sum_{k=1}^{K} p_k^{(r)} Po_2(X_{1i}, X_{3i}|u_{2k}^{(r)})}$$

and

$$w_{ik}^{(r+1)} = \frac{w_{ik12}^{(r+1)} + w_{ik13}^{(r+1)}}{2}$$

if mode$(i, 3) = 2$

$$w_{ik12}^{(r+1)} = \frac{p_k^{(r)} Po_2(X_{1i}, X_{2i}|u_{1k}^{(r)})}{\sum_{k=1}^{K} p_k^{(r)} Po_2(X_{1i}, X_{2i}|u_{1k}^{(r)})}$$

$$w_{ik23}^{(r+1)} = \frac{p_k^{(r)} Po_2(X_{2i}, X_{3i}|u_{3k}^{(r)})}{\sum_{k=1}^{K} p_k^{(r)} Po_2(X_{2i}, X_{3i}|u_{3k}^{(r)})}$$

and

$$w_{ik}^{(r+1)} = \frac{w_{ik12}^{(r+1)} + w_{ik23}^{(r+1)}}{2}$$

if mode$(i, 3) = 0$

$$w_{ik13}^{(r+1)} = \frac{p_k^{(r)} Po_2(X_{1i}, X_{3i}|u_{2k}^{(r)})}{\sum_{k=1}^{K} p_k^{(r)} Po_2(X_{1i}, X_{3i}|u_{2k}^{(r)})}$$

$$w_{ik23}^{(r+1)} = \frac{p_k^{(r)} Po_2(X_{2i}, X_{3i}|u_{3k}^{(r)})}{\sum_{k=1}^{K} p_k^{(r)} Po_2(X_{2i}, X_{3i}|u_{3k}^{(r)})}$$

and

$$w_{ik}^{(r+1)} = \frac{w_{ik13}^{(r+1)} + w_{ik23}^{(r+1)}}{2}$$

M-step: Update $p_k^{(r+1)} = \sum_{i=1}^{n} w_{ik}^{(r+1)}/n$ for all $k = 1, 2, ..., K$ and then maximize the quantity:

$$\begin{aligned} Q_k = &\sum_{i=1}^{n} \Big\{ w_{ik12}^{(r+1)} \log Po_2(X_{1i}, X_{2i}|u_{1k}) \Big\} (I_{1i} + I_{2i}) \\ &+ \sum_{i=1}^{n} \Big\{ w_{ik13}^{(r+1)} \log Po_2(X_{1i}, X_{3i}|u_{2k}) \Big\} (I_{1i} + I_{3i}) \\ &+ \sum_{i=1}^{n} \Big\{ w_{ik23}^{(r+1)} \log Po_2(X_{2i}, X_{3i}|u_{3k}) \Big\} (I_{2i} + I_{3i}) \end{aligned} \tag{2.33}$$

to get updated values for u_{1k}, u_{2k}, u_{3k} parameters associated with the k-th component $k = 1, 2, ..., K$ as defined above. ECM algorithm can be conducted in a similar way in four steps as described in the previous section.

2.5.1.2 Model Selection

The chosen number of clusters results from the lowest value of CLBIC (Composite Likelihood Bayesian Information Criterion) which is typical approach for a family of models running for a range of values of K. The definition of this criterion is:

$$CLBIC = -2\ell(\hat{\theta}) + tr(J(\hat{\theta})H^{-1}(\hat{\theta}))\log(n)$$

where $\hat{\theta}$ is the maximum likelihood estimate of vector θ, $\ell(\hat{\theta})$ is the maximized likelihood, and matrices H & J as defined in section 2.2.3. The difference between full model estimation and the sampling method of these section is that we make use of the estimated w_{igst} mixing weights as defined in section 2.5.1.1.

2.5.2 Systematic Sampling 2 for Poisson mixtures

In the same context of systematic sampling we can further reduce the number of composite components which will be taken into consideration in the estimation of parameters. For the specific example of $m = 3$ dimension of Poisson mixtures we can assume a sampling method depending on the $i-th$ row of the observed dataset and consider only one of the composite components instead of two proposed in the previous section. The adequacy of the number of chosen components depends on the sample size and the number of clusters for evaluation. The larger the number the observations the more accurate the resulted estimators will be.

Let's assume the previous example 2.5 of n observations of $X = (X_1, X_2, X_3) \sim Po_3(\theta_{1k}, \theta_{2k}, \theta_{3k}, \theta_{12k}, \theta_{13k}, \theta_{23k})$, $k = 1, ..., K$ cluster and the marginal log likelihoods of the bivariate poisson distribution

$$\ell_{CLst} = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log p_k + \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log Po_2(X_{si}, X_{ti}|u_{stk}),\ \ s < t \leq m,\ \ k = 1, .., K$$

where u_{stk} the corresponding vector of θ's related to the random vector $X = (X_s, X_t)$ and the k^{th} cluster, $k = 1, 2, ..., K$. The log-likelihood that we need

to maximize through composite likelihood concept is described in equation 2.28, where $k = 1, 2, ..., K$ is the number of clusters and

$$u_{1k} = (\theta_{1k} + \theta_{13k}, \theta_{2k} + \theta_{23k}, \theta_{12k}),\ k = 1, ..., K$$

$$u_{2k} = (\theta_{1k} + \theta_{12k}, \theta_{3k} + \theta_{23k}, \theta_{13k}),\ k = 1, ..., K$$

$$u_{3k} = (\theta_{2k} + \theta_{12k}, \theta_{3k} + \theta_{13k}, \theta_{23k}),\ k = 1, ..., K.$$

.

The Systematic Sampling approach 2 proposes to maximize the sampled composite likelihood by choosing 1 out of the 3 pairs of marginal log-likelihoods ℓ_{CLst} described in equation 2.28. The choice of the component correlates to the row of the observation sample data i, $i = 1, ..., n$. For the first observation where mode(i,3)=1 we choose the first component,for the second observation the second, while for the third observation the last component. This is performed for all observations n and the same approach is performed for all clusters K.

In order to estimate the θ parameters we will make use of the below indices as in the case of Systematic sampling method 1:

$$I_{1i} = \begin{cases} 1 & \text{if mode}(i, 3) = 1 \\ 0 & \text{elsewhere} \end{cases}$$

$$I_{2i} = \begin{cases} 1 & \text{if mode}(i, 3) = 2 \\ 0 & \text{elsewhere} \end{cases}$$

$$I_{3i} = \begin{cases} 1 & \text{if mode}(i, 3) = 0 \\ 0 & \text{elsewhere} \end{cases}$$

The sampled composite likelihood is now written in the below form:

$$
\begin{aligned}
\ell_{Sam2CL} = & \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log p_k \\
& + \sum_{i=1}^{n}\sum_{k=1}^{K}\Big\{z_{ik}\log Po_2(X_{1i},X_{2i}|u_{1k})\Big\}I_{1i} \\
& + \sum_{i=1}^{n}\sum_{k=1}^{K}\Big\{z_{ik}\log Po_2(X_{1i},X_{3i}|u_{2k})\Big\}I_{2i} \\
& + \sum_{i=1}^{n}\sum_{k=1}^{K}\Big\{z_{ik}\log Po_2(X_{2i},X_{3i}|u_{3k})\Big\}I_{3i}
\end{aligned}
\tag{2.34}
$$

2.5.2.1 EM algorithm

Similarly to the full composite likelihood approach, with the use of EM algorithm we maximize the complete-data log likelihood of the model for the r^{th} iteration as follows:

E-step: Calculate for $i = 1, ..., n$ and $k = 1, 2, ..., K$

if mode$(i, 3) = 1$

$$w_{ik12}^{(r+1)} = \frac{p_k^{(r)} Po_2(X_{1i}, X_{2i}|u_{1k}^{(r)})}{\sum_{k=1}^{K} p_k^{(r)} Po_2(X_{1i}, X_{2i}|u_{1k}^{(r)})}$$

if mode$(i, 3) = 2$

$$w_{ik13}^{(r+1)} = \frac{p_k^{(r)} Po_2(X_{1i}, X_{3i}|u_{2k}^{(r)})}{\sum_{k=1}^{K} p_k^{(r)} Po_2(X_{1i}, X_{3i}|u_{2k}^{(r)})}$$

if mode$(i, 3) = 0$

$$w_{ik23}^{(r+1)} = \frac{p_k^{(r)} Po_2(X_{2i}, X_{3i}|u_{3k}^{(r)})}{\sum_{k=1}^{K} p_k^{(r)} Po_2(X_{2i}, X_{3i}|u_{3k}^{(r)})}$$

and

$$w_{ik}^{(r+1)} = w_{ik12}^{(r+1)} + w_{ik13}^{(r+1)} + w_{ik23}^{(r+1)}$$

M-step: Update $p_k^{(r+1)} = \sum_{i=1}^{n} w_{ik}^{(r+1)}/n$ for all $k = 1, 2, ..., K$ and then maximize the quantity:

$$
\begin{aligned}
Q_k = &\sum_{i=1}^{n} \left\{ w_{ik12}^{(r+1)} \log Po_2(X_{1i}, X_{2i}|u_{1k}) \right\} I_{1i} \\
&+ \sum_{i=1}^{n} \left\{ w_{ik13}^{(r+1)} \log Po_2(X_{1i}, X_{3i}|u_{2k}) \right\} I_{2i} \\
&+ \sum_{i=1}^{n} \left\{ w_{ik23}^{(r+1)} \log Po_2(X_{2i}, X_{3i}|u_{3k}) \right\} I_{3i}
\end{aligned} \tag{2.35}
$$

to get updated values for u_{1k}, u_{2k}, u_{3k} parameters associated with the k-th component $k = 1, 2, ..., K$ as defined above. ECM algorithm can be conducted in a similar way in four steps as described in the previous section.

2.5.2.2 Model Selection

The chosen number of clusters results from the lowest value of CLBIC (Composite Likelihood Bayesian Information Criterion). The definition of this criterion is:

$$CLBIC = -2\ell(\hat{\theta}) + tr(J(\hat{\theta})H^{-1}(\hat{\theta}))\log(n)$$

where $\hat{\theta}$ is the maximum likelihood estimate of vector θ, $\ell(\hat{\theta})$ is the maximized likelihood, and matrices H & J as defined in section 2.2.3. The difference between full model estimation and the sampling method of these section is that we make use of the estimated w_{igst} mixing weights as defined in section 2.5.2.1.

2.5.3 Non Systematic Sampling 3 for Poisson mixtures

Let's assume the example 2.5 of n observations of a random vector $X = (X_1, X_2, X_3) \sim Po_3(\theta_{1k}, \theta_{2k}, \theta_{3k}, \theta_{12k}, \theta_{13k}, \theta_{23k})$, $k = 1, ..., K$ cluster and the marginal log likelihoods of the bivariate poisson distribution

$$\ell_{CLst} = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} \log p_k + \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} \log Po_2(X_{si}, X_{ti}|u_{stk}), \;\; s < t \leq m, \;\; k = 1, .., K$$

where u_{stk} the corresponding vector of θ's related to the random vector $X = (X_s, X_t)$ and the k^{th} cluster, $k = 1, 2, ..., K$. The log-likelihood that we need to maximize through composite likelihood concept is described in equation 2.28,

where $k = 1, 2, ..., K$ is the number of clusters and

$$u_{1k} = (\theta_{1k} + \theta_{13k}, \theta_{2k} + \theta_{23k}, \theta_{12k}),\ k = 1, ..., K$$

$$u_{2k} = (\theta_{1k} + \theta_{12k}, \theta_{3k} + \theta_{23k}, \theta_{13k}),\ k = 1, ..., K$$

$$u_{3k} = (\theta_{2k} + \theta_{12k}, \theta_{3k} + \theta_{13k}, \theta_{23k}),\ k = 1, ..., K.$$

.

The Non-Systematic Sampling approach 3 performs maximization of the sampled composite likelihood by choosing a number of the 3 pairs of marginal log-likelihoods l_{CLst} with probability 2/3, this is $I_{ij} \sim Bernoulli(p = 2/3)$. Therefore we produce a matrix I with dimension $n \times 3$ for the specific example of 3-variate Poisson distribution. In general and for the case of m-variate distribution the resulted I matrix is of size $n \times \binom{m}{2}$. The choice of the pairs of components is not related to the row of the observation sample data i, $i = 1, ..., n$, $j = 1, 2, 3$ compared to the Sampling Methods 1 & 2. Due to implementation restrictions we choose the components to be the same for every cluster k, $k = 1, ..., K$ and for every step of the EM algorithm, otherwise we will not achieve convergence. In order to estimate the θ parameters we will make use of the below indices

$$I_{i1} = \begin{cases} 1 & p = 2/3 \\ 0 & \text{elsewhere} \end{cases}$$

$$I_{i2} = \begin{cases} 1 & p = 2/3 \\ 0 & \text{elsewhere} \end{cases}$$

$$I_{i3} = \begin{cases} 1 & p = 2/3 \\ 0 & \text{elsewhere} \end{cases}$$

The sampled composite likelihood is now written in the below form:

$$
\begin{aligned}
\ell_{Sam3CL} = & \sum_{i=1}^{n}\sum_{j=1}^{3} I_{ij} \sum_{k=1}^{K} z_{ik} \log p_k \\
& + \sum_{i=1}^{n}\sum_{k=1}^{K} \Big\{ z_{ik} \log Po_2(X_{1i}, X_{2i}|u_{1k}) \Big\} I_{i1} \\
& + \sum_{i=1}^{n}\sum_{k=1}^{K} \Big\{ z_{ik} \log Po_2(X_{1i}, X_{3i}|u_{2k}) \Big\} I_{i2} \\
& + \sum_{i=1}^{n}\sum_{k=1}^{K} \Big\{ z_{ik} \log Po_2(X_{2i}, X_{3i}|u_{3k}) \Big\} I_{i3}
\end{aligned}
\tag{2.36}
$$

The rows of the observed dataset with all entries $I_{ij} = 0$, $j = 1, 2, 3$ for all components will be eliminated from the calculations. This way we do not only sample down the components but also the size of the data set.

2.5.3.1 EM algorithm

With the use of EM algorithm we maximize the complete-data log likelihood of the model for the r^{th} iteration as follows:

E-step: Calculate for $i = 1, ..., n$ and $k = 1, 2, ..., K$

$$
w_{ik12}^{(r+1)} = \frac{p_k^{(r)} Po_2(X_{1i}, X_{2i}|u_{1k}^{(r)}))I_{i1}}{\sum_{k=1}^{K} p_k^{(r)} Po_2(X_{1i}, X_{2i}|u_{1k}^{(r)}))I_{i1}}
$$

$$
w_{ik13}^{(r+1)} = \frac{p_k^{(r)} Po_2(X_{1i}, X_{3i}|u_{2k}^{(r)})I_{i2}}{\sum_{k=1}^{K} p_k^{(r)} Po_2(X_{1i}, X_{3i}|u_{2k}^{(r)})I_{i2}}
$$

$$
w_{ik23}^{(r+1)} = \frac{p_k^{(r)} Po_2(X_{2i}, X_{3i}|u_{3k}^{(r)})I_{i3}}{\sum_{k=1}^{K} p_k^{(r)} Po_2(X_{2i}, X_{3i}|u_{3k}^{(r)})I_{i3}}
$$

therefore,

$$
w_{ik}^{(r+1)} = \begin{cases} \frac{w_{ik12}^{(r+1)} + w_{ik13}^{(r+1)} + w_{ik23}^{(r+1)}}{\sum_{j=1}^{3} I_{ij}}, & \text{if} \quad \sum_{j=1}^{3} I_{ij} > 0 \\ 0, & \text{if} \quad \sum_{j=1}^{3} I_{ij} = 0 \end{cases}
$$

M-step : Update $p_k^{(r+1)} = \sum_{i=1}^{n} w_{ik}^{(r+1)}/n$ for all $k = 1, 2, ..., K$ and then maximize the quantity

$$
\begin{aligned}
Q_k = &\sum_{i=1}^{n}\sum_{k=1}^{K}\left\{w_{ik12}^{(r+1)}\log Po_2(X_{1i}, X_{2i}|u_{1k})\right\}I_{i1} \\
&+\sum_{i=1}^{n}\sum_{k=1}^{K}\left\{w_{ik13}^{(r+1)}\log Po_2(X_{1i}, X_{3i}|u_{2k})\right\}I_{i2} \\
&+\sum_{i=1}^{n}\sum_{k=1}^{K}\left\{w_{ik23}^{(r+1)}\log Po_2(X_{2i}, X_{3i}|u_{3k})\right\}I_{i3}
\end{aligned} \tag{2.37}
$$

to get updated values for u_{1k}, u_{2k}, u_{3k} parameters associated with the k-th component $k = 1, 2, ..., K$ as defined above.

2.5.3.2 Model Selection

As for previous sampling methods the chosen number of clusters results from the lowest value of CLBIC:

$$CLBIC = -2\ell(\hat{\theta}) + tr(J(\hat{\theta})H^{-1}(\hat{\theta}))\log(n)$$

where $\hat{\theta}$ is the maximum likelihood estimate of vector θ, $\ell(\hat{\theta})$ is the maximized likelihood, and matrices H & J as defined in section 2.2.3. The difference between full model estimation and the sampling method of these section is that we make use of the estimated w_{igst} mixing weights as defined in section 2.5.3.1.

2.6 Simulation Study 1

2.6.1 Data Sample Description

For the particular simulation study we do not assume any mixture model, though the purpose is to compare the above proposed sampling methodologies in terms of efficiency to estimate the parameters and time consumed for each method. Let's assume a 3-variate count dataset of length $n = 200$ from a 3-variate Poisson distribution. In more detail let:

$$X = (X_1, X_2, X_3) \sim Po_3(\theta_1 = 4, \theta_2 = 2, \theta_3 = 3, \theta_{12} = 3, \theta_{13} = 2, \theta_{23} = 1)$$

The marginal distributions of the bi-variate components are written as shown below for $s_3 = 1$:

$$(X_1, X_2) \sim Po_2(u_1 = (\theta_1 + \theta_{13}, \theta_2 + \theta_{23}, \theta_{12})) = Po_2(6, 3, 3)$$

$$(X_1, X_3) \sim Po_2(u_2 = (\theta_1 + \theta_{12}, \theta_3 + \theta_{23}, \theta_{13})) = Po_2(7, 4, 2)$$

$$(X_2, X_3) \sim Po_2(u_3 = (\theta_2 + \theta_{12}, \theta_3 + \theta_{13}, \theta_{23})) = Po_2(5, 5, 1)$$

2.6.2 Models Evaluated

We simulate 1,000 datasets from the $Po_3(\Theta)$ distribution and for each of the produced dataset we estimate the parameters for the different methodologies as described below:

- **Model 1:** Maximize the complete data likelihood of the 3-variate Poisson distribution.
- **Model 2:** Maximize the full composite likelihood without assuming any sampling approach.
- **Model 3:** Maximize the sampled composite likelihood by choosing 2 out of the 3 pairs of marginal log-likelihoods $l_{CL_1 j}$ described in Sampling method 1.
- **Model 4:** Maximize the sampled composite likelihood by choosing 1 out of the 3 pairs of marginal log-likelihoods $l_{CL_2 j}$ described in Sampling method 2.
- **Model 5:** Maximize the sampled composite likelihood by choosing pairs of marginal log-likelihoods via a Bernoulli distribution with probability $p = 2/3$ as described in Sampling method 3.

2.6.3 Results

We performed B=1000 iterations and for each model we constructed $1 - \alpha$= 95% confidence intervals for the estimators of the parameters of each model assumed. The variance of the confidence interval for each parameter was chosen to be the sample (of size B) variance s^2. Then we calculated the number of correct

evaluations, that is how many times the true initial values are included to the intervals.

The table 2.1 demonstrates the results for each model in percentages. Because of the power of the level of statistical significance α we would expect that these percentages are approximately 95%.

Model	θ_1	θ_2	θ_3	θ_{12}	θ_{13}	θ_{23}
Model 1	95.3 %	92.4 %	94.0 %	94.8 %	95.1 %	94.0 %
Model 2	94.5 %	94.5 %	93.5 %	93.7 %	94.6 %	92.2 %
Model 3	95.5 %	93.6 %	93.6 %	95.0 %	94.7 %	94.4 %
Model 4	94.7 %	93.2 %	94.2 %	94.6 %	94.3 %	94.0 %
Model 5	94.5 %	95.3 %	93.7 %	94.8 %	95.2 %	95.0 %

TABLE 2.1: Confidence intervals for the parameters of each model

Table 2.2 demonstrates the average estimated value for each parameter of the models over the 1,000 simulated data sets. We can observe that from the results there is no significance loss of the estimators not for the confidence intervals or for the estimated average values. The sampling methods provide good results, especially for Systematic Sampling method 1, where we choose 2 out of 3 components for every row of the dataset and estimated values are close to the Composite Likelihood method. This observation can be also shown from the boxplots in 2.1. Some deviations are observed in the Sampling method 3, where we do not only eliminate components of the CL algorithm with probability $p = 2/3$ but also rows of the simulated datasets. This is of course depending on the sample size of the data.

Model	θ_1	θ_2	θ_3	θ_{12}	θ_{13}	θ_{23}
Model 1	3.982422	1.893038	2.845538	2.978983	2.031776	1.122784
Model 2	3.937008	1.878251	2.806886	2.995269	2.069563	1.126527
Model 3	3.934525	1.848262	2.791901	3.004987	2.064968	1.146133
Model 4	3.907458	1.850170	2.810340	3.022138	2.065700	1.127180
Model 5	3.891156	1.861447	2.761783	3.003620	2.110253	1.138074

TABLE 2.2: Average estimated values for the parameters of each model

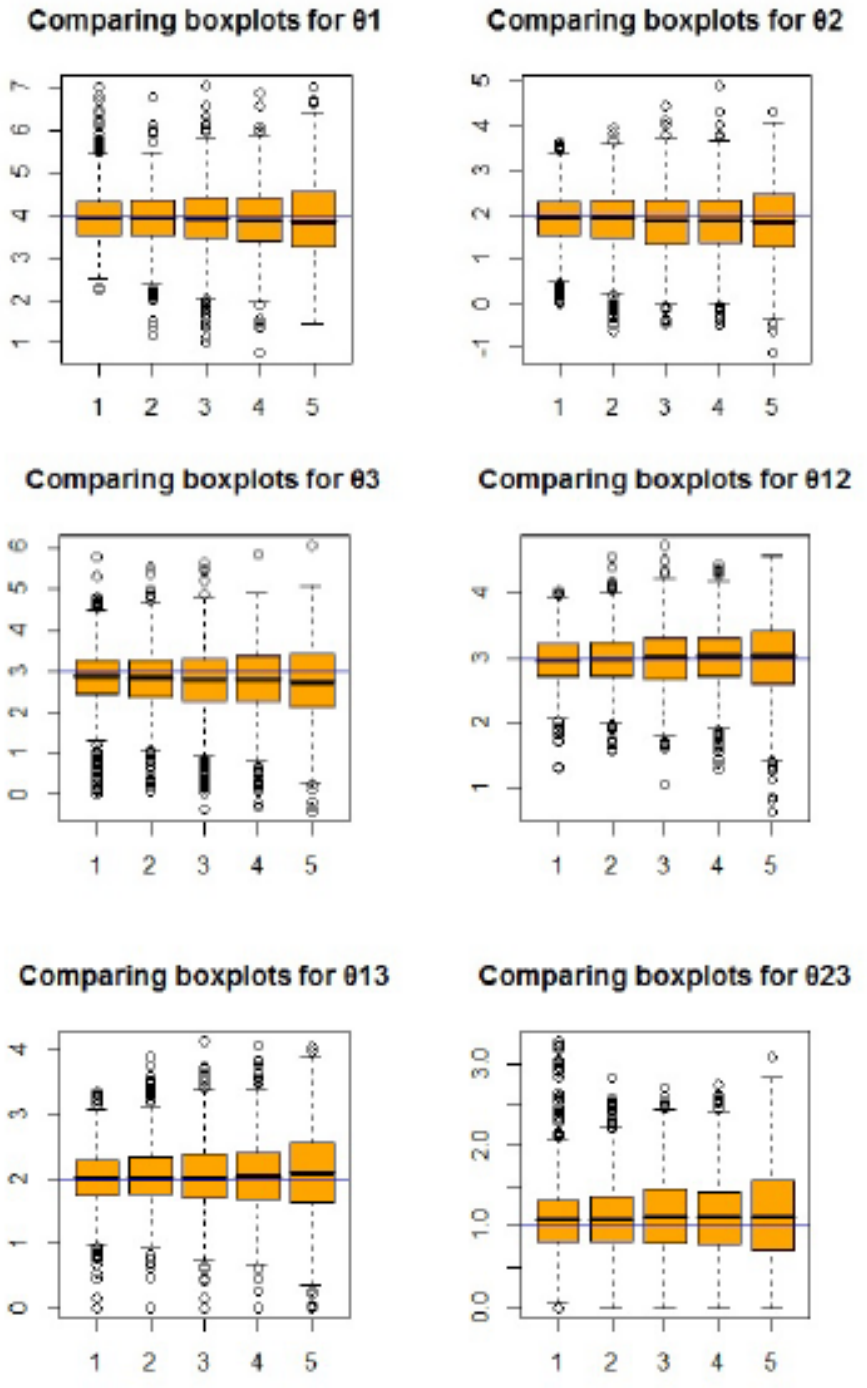


FIGURE 2.1: Boxplots for estimated parameters for all models

The main purpose of the specific simulation study and the introduction of Sampling methods is to evaluate the computational effort needed for each approach. The table 2.3 demonstrates the average time needed by each model to maximize the corresponding likelihood. Maximizing the 3-variate Poisson log-likelihood is time consuming so the equivalent composite likelihood is proposed since it is proven to be efficient in many application in bibliography. The time reduction is significant

for the full Composite Likelihood method, therefore it is prefered, especially for cases where the multivariate Poisson distributions has a large dimension which makes computations and optimization much more inefficient.

Model	Avg time
Model 1	568.57 sec
Model 2	79.38 sec
Model 3	42.61 sec
Model 4	62.99 sec
Model 5	**22.27** sec

TABLE 2.3: Average time needed for each model

Furthermore, among the proposed CL models via Sampling ,we observe that we can further reduce computational effort without significant loss of effciency. In case of Non-Systematic sampling method 3 the average time needed for optimization can be reduced 75% compared to the Composite Likelihood method. We can observe that Systematic Sampling 2 needs more time than the Systematic Sampling 1 even though we choose less components from the bivariate Composite Likelihoods. This is mainly due to slower convergence of the maximization process, to obtain adequate estimations. This observation may be trivial to the specific example of a 3-variate Poisson without mixture but it can be important when it comes to high dimensions and mixtures of Multivariate Poisson, where even the definition of the Multivariate Poisson likelihood is not easy to obtain.

2.7 Simulation Study 2

In this section, we present the results of a simulation study that we conducted to illustrate the effectiveness of our clustering methodology via the composite likelihood method and the Sampling methods described in previous sections. Via simulation,we compare the results of various composite likelihood models with the ones produced of the full mixture model.

2.7.1 Data Sample Description

Let's assume a 3-variate count dataset of length $n = 200$ and $n = 400$ resulting from $K = 2$ components from 3-variate Poisson distributions.

In more detail let's assume for $k = 1$ the 1st cluster:

$$X_1 = (X_{11}, X_{21}, X_{31}) \sim Po_3(\theta_1 = 3, \theta_2 = 3, \theta_3 = 4, \theta_{12} = 0.5, \theta_{13} = 1, \theta_{23} = 1)$$

and for $k = 2$

$$X_2 = (X_{12}, X_{22}, X_{32}) \sim Po_3(\theta_1 = 2, \theta_2 = 4, \theta_3 = 2, \theta_{12} = 1, \theta_{13} = 0, \theta_{23} = 0.5).$$

The marginal distributions of the bi-variate components are written as shown below:
For $k = 1$:

$$(X_{11}, X_{21}) \sim Po_2(u_{11} = (\theta_1 + \theta_{13}, \theta_2 + \theta_{23}, \theta_{12})) = Po_2(4, 4, 0.5)$$

$$(X_{11}, X_{31}) \sim Po_2(u_{21} = (\theta_1 + \theta_{12}, \theta_3 + \theta_{23}, \theta_{13})) = Po_2(3.5, 5, 1)$$

$$(X_{21}, X_{31}) \sim Po_2(u_{31} = (\theta_2 + \theta_{12}, \theta_3 + \theta_{13}, \theta_{23})) = Po_2(3.5, 5, 1)$$

and for $k = 2$:

$$(X_{12}, X_{22}) \sim Po_2(u_{12} = (\theta_1 + \theta_{13}, \theta_2 + \theta_{23}, \theta_{12})) = Po_2(2, 4.5, 1)$$

$$(X_{12}, X_{32}) \sim Po_2(u_{22} = (\theta_1 + \theta_{12}, \theta_3 + \theta_{23}, \theta_{13})) = Po_2(3, 2.5, 0)$$

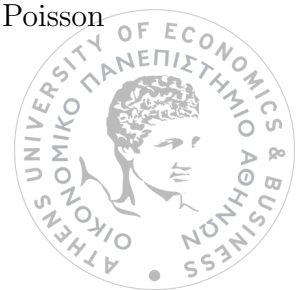
$$(X_{22}, X_{32}) \sim Po_2(u_{32} = (\theta_2 + \theta_{12}, \theta_3 + \theta_{13}, \theta_{23})) = Po_2(5, 2, 0.5)$$

We choose the data from cluster 1 with a probability p=0.7 , while from cluster 2 with a probability p=0.3.

2.7.2 Models Evaluated

We perform 200 iterations of the simulated dataset and for each of the produced datasets we perform Expectation-Maximization algorithm to obtain the estimated parameters of each component for distributions and for different methodologies as described below. Convergence criteria for each EM step has been set to 10^{-10}. Furthermore to the previous simulation study we assess the effectiveness of the alternative composite likelihood method, that is Model 6.

- **Model 1:** Maximize the complete full likelihood of the 3-variate Poisson distributions.



- **Model 2:** Maximize the full composite likelihood.
- **Model 3:** Maximize the sampled composite likelihood by choosing 2 out of the 3 pairs of marginal log-likelihoods l_{CL_j} described in relevant section with Sampling method 1.
- **Model 4:** Maximize the sampled composite likelihood by choosing 1 out of the 3 pairs of marginal log-likelihoods l_{CL_j} described in relevant section with Sampling method 2.
- **Model 5:** Maximize the sampled composite likelihood by choosing pairs of marginal log-likelihoods via a bernoulli distribution with probability $p = 2/3$ with Sampling method 3.
- **Model 6:** Maximize the alternative composite likelihood of equation described in the relevant section.

2.7.3 Model Selection

The optimal number of clusters will be selected with the use of CLBIC defined in section 3.4.2.2 through equation 2.7. For the specific example of the 3-variate Poisson mixture the log likelihood of the mixture composite model that results to the estimators $\hat{\mathbf{\Theta}}_{CL}$ of the parameter space is written in the below form:

$$
\begin{aligned}
\ell &= 3\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\log p_k + \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik12}\log Po_2(X_{i1},X_{i2}|u_{k1}) \\
&+ \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik13}\log Po_2(X_{i1},X_{i3}|u_{k2}) + \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik23}\log Po_2(X_{i2},X_{i3}|u_{k3}) \\
&= 3\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}\log p_k + \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{r=1}^{q} w_{ik12}\log Po(y_{ikr1}|u_{k1}) \\
&+ \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{r=1}^{q} w_{ik13}\log Po(y_{ikr2}|u_{k2}) + \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{r=1}^{q} w_{ik23}\log Po(y_{ikr3}|u_{k3}) \\
&\propto \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik12}\Big\{\log Po(y_{ik11}|\theta_{1k}+\theta_{13k}) + \log Po(y_{ik21}|\theta_{2k}+\theta_{23k}) + \log Po(y_{ik31}|\theta_{12k})\Big\} \\
&+ \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik13}\Big\{\log Po(y_{ik12}|\theta_{1k}+\theta_{12k}) + \log Po(y_{ik22}|\theta_{3k}+\theta_{23k}) + \log Po(y_{ik32}|\theta_{13k})\Big\} \\
&+ \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik23}\Big\{\log Po(y_{ik13}|\theta_{2k}+\theta_{12k}) + \log Po(y_{ik23}|\theta_{3k}+\theta_{13k}) + \log Po(y_{ik33}|\theta_{23k})\Big\} \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik12}\Big\{y_{ik11}\log(\theta_{1k}+\theta_{13k}) + y_{ik21}\log(\theta_{2k}+\theta_{23k}) + y_{ik31}\log(\theta_{12k}) \\
&- (\theta_{1k}+\theta_{2k}+\theta_{12k}+\theta_{13k}+\theta_{23k}) - \log y_{ik11}! - \log y_{ik21}! - \log y_{ik31}!\Big\} \\
&+ \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik13}\Big\{y_{ik12}\log(\theta_{1k}+\theta_{12k}) + y_{ik22}\log(\theta_{3k}+\theta_{23k}) + y_{ik32}\log(\theta_{13k}) \\
&- (\theta_{1k}+\theta_{3k}+\theta_{12k}+\theta_{13k}+\theta_{23k}) - \log y_{ik12}! - \log y_{ik22}! - \log y_{ik32}!\Big\} \\
&+ \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik23}\Big\{y_{ik13}\log(\theta_{2k}+\theta_{12k}) + y_{ik23}\log(\theta_{3k}+\theta_{13k}) + y_{ik33}\log(\theta_{23k}) \\
&- (\theta_{2k}+\theta_{3k}+\theta_{12k}+\theta_{13k}+\theta_{23k}) - \log y_{ik13}! - \log y_{ik23}! - \log y_{ik33}!\Big\}
\end{aligned}
\tag{2.38}
$$

where: $u_{k1} = (\theta_1+\theta_{13}, \theta_2+\theta_{23}, \theta_{12})$, $u_{k2} = (\theta_1+\theta_{12}, \theta_3+\theta_{23}, \theta_{13})$, $u_{k1} = (\theta_2+\theta_{12}, \theta_3+\theta_{13}, \theta_{23})$, $w_{ik} = \frac{w_{ik12}+w_{ik13}+w_{ik23}}{3}$ as defined in 2.5. In order to estimate the matrices $H(\boldsymbol{\theta})$ & $J(\boldsymbol{\theta})$ participating in the calculation of CLBIC of equation 2.7 we have:

$$J(\boldsymbol{\theta}) = n^{-1} \begin{bmatrix} \frac{\partial \ell}{\partial \theta_1} \\ \frac{\partial \ell}{\partial \theta_2} \\ \frac{\partial \ell}{\partial \theta_3} \\ \frac{\partial \ell}{\partial \theta_{12}} \\ \frac{\partial \ell}{\partial \theta_{13}} \\ \frac{\partial \ell}{\partial \theta_{23}} \end{bmatrix} * \begin{bmatrix} \frac{\partial \ell}{\partial \theta_1} & \frac{\partial \ell}{\partial \theta_2} & \frac{\partial \ell}{\partial \theta_3} & \frac{\partial \ell}{\partial \theta_{12}} & \frac{\partial \ell}{\partial \theta_{13}} & \frac{\partial \ell}{\partial \theta_{23}} \end{bmatrix}$$

and

$$H(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial^2 \ell}{\partial \theta_1^2} & \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_2} & \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_3} & \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_{12}} & \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_{13}} & \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_{23}} \\ \frac{\partial^2 \ell}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \ell}{\partial \theta_2^2} & \frac{\partial^2 \ell}{\partial \theta_2 \partial \theta_3} & \frac{\partial^2 \ell}{\partial \theta_2 \partial \theta_{12}} & \frac{\partial^2 \ell}{\partial \theta_2 \partial \theta_{13}} & \frac{\partial^2 \ell}{\partial \theta_2 \partial \theta_{23}} \\ \frac{\partial^2 \ell}{\partial \theta_3 \partial \theta_1} & \frac{\partial^2 \ell}{\partial \theta_3 \partial \theta_2} & \frac{\partial^2 \ell}{\partial \theta_3^2} & \frac{\partial^2 \ell}{\partial \theta_3 \partial \theta_{12}} & \frac{\partial^2 \ell}{\partial \theta_3 \partial \theta_{13}} & \frac{\partial^2 \ell}{\partial \theta_3 \partial \theta_{23}} \\ \frac{\partial^2 \ell}{\partial \theta_{12} \partial \theta_1} & \frac{\partial^2 \ell}{\partial \theta_{12} \partial \theta_2} & \frac{\partial^2 \ell}{\partial \theta_{12} \partial \theta_3} & \frac{\partial^2 \ell}{\partial \theta_{12}^2} & \frac{\partial^2 \ell}{\partial \theta_{12} \partial \theta_{13}} & \frac{\partial^2 \ell}{\partial \theta_{12} \partial \theta_{23}} \\ \frac{\partial^2 \ell}{\partial \theta_{13} \partial \theta_1} & \frac{\partial^2 \ell}{\partial \theta_{13} \partial \theta_2} & \frac{\partial^2 \ell}{\partial \theta_{13} \partial \theta_3} & \frac{\partial^2 \ell}{\partial \theta_{13} \partial \theta_{12}} & \frac{\partial^2 \ell}{\partial \theta_{13}^2} & \frac{\partial^2 \ell}{\partial \theta_{13} \partial \theta_{23}} \\ \frac{\partial^2 \ell}{\partial \theta_{23} \partial \theta_1} & \frac{\partial^2 \ell}{\partial \theta_{23} \partial \theta_2} & \frac{\partial^2 \ell}{\partial \theta_{23} \partial \theta_3} & \frac{\partial^2 \ell}{\partial \theta_{23} \partial \theta_{12}} & \frac{\partial^2 \ell}{\partial \theta_{23} \partial \theta_{13}} & \frac{\partial^2 \ell}{\partial \theta_{23}^2} \end{bmatrix}$$

Indicatively, we can derive the entries of the matrices based on equation 2.38 as follows:

$$\frac{\partial \ell}{\partial \theta_1} = \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik12} \Big\{ \frac{y_{ik11}}{\theta_1 + \theta_{13}} - 1 \Big\} + \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik13} \Big\{ \frac{y_{ik12}}{\theta_1 + \theta_{12}} - 1 \Big\}$$

$$\frac{\partial^2 \ell}{\partial \theta_1^2} = - \sum_{i=1}^{n} \sum_{k=1}^{K} \Big\{ w_{ik12} \frac{y_{ik11}}{(\theta_1 + \theta_{13})^2} + w_{ik13} \frac{y_{ik12}}{(\theta_1 + \theta_{12})^2} \Big\}$$

$$\frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_{12}} = - \sum_{i=1}^{n} \sum_{k=1}^{K} \Big\{ w_{ik12} \frac{y_{ik12}}{(\theta_1 + \theta_{12})^2} \Big\}, \quad \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_{13}} = - \sum_{i=1}^{n} \sum_{k=1}^{K} \Big\{ w_{ik13} \frac{y_{ik11}}{(\theta_1 + \theta_{13})^2} \Big\}$$

$$\frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_2} = 0, \quad \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_3} = 0, \quad \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_{23}} = 0$$

Furthermore, for the case of sampling methods the above calculations are executed similarly, with the only difference that the w_{ikst}, $s < t$ components are defined as in e.g. 2.5.1.1 for Composite Sampling Method 1.

2.7.4 Results

For each iteration of the simulation we calculate the log-likelihood of each model which of-course is an alteration of the full model's likelihood as described in above sections. Obtained likelihoods are not comparable for the models, though it can be used for optimal number of clusters to be chosen.

Tables 2.4 & 2.5 provides the average estimated values for each model for both mixing probabilities and the true values for which the dataset was constructed.

Model	θ_1	θ_2	θ_3	θ_{12}	θ_{13}	θ_{23}
Model 1	2.596719	2.574240	3.519074	0.671335	1.171275	1.235230
Model 2	2.394133	2.815689	3.545578	0.649722	1.244587	1.080638
Model 3	2.170110	2.677306	3.293150	0.664401	1.388533	1.174902
Model 4	1.668290	2.768176	2.597681	0.690340	1.562008	1.199620
Model 5	2.125479	2.555044	3.183359	0.815358	1.268933	1.212717
Model 6	3.344200	3.080976	4.244008	0.523754	0.619867	1.121181
True Value	3	3	4	0.5	1	1

TABLE 2.4: Average estimated values for the parameters of each model, $\pi = 0.3$

Model	θ_1	θ_2	θ_3	θ_{12}	θ_{13}	θ_{23}
Model 1	1.882426	3.931840	1.846500	0.993276	0.189492	0.557063
Model 2	1.946750	4.044773	1.752004	0.902916	0.253646	0.539348
Model 3	1.955200	4.030634	1.738097	0.883362	0.303975	0.569315
Model 4	1.868037	4.010956	1.580126	0.772625	0.638154	0.632163
Model 5	1.914751	4.016434	1.773638	0.866491	0.389403	0.589284
Model 6	1.898210	3.846895	1.616314	0.977497	0.068181	0.621895
True Value	2	4	2	1	0	0.5

TABLE 2.5: Average estimated values for the parameters of each model,$\pi = 0.7$

The table 2.6 demonstrates the selected number of clusters through CLBIC for the 200 simulated datasets of both n=200 and n=400. We observe that for the Alternative Composite method (Model 6) the number of clusters selected for all iterations is K=3, while for all other methods the minimum value of the criterion is achieved for K=2 which is the true value the dataset was constructed. For a data sample of the same structure we observe that the chosen number of clusters is further improved. With the use of the estimated parameters for each model's

	$n = 200$		$n = 400$	
Model	**K=2**	**K=3**	**K=2**	**K=3**
Model 1	200	0	200	0
Model 2	199	1	199	1
Model 3	195	5	197	3
Model 4	191	9	195	5
Model 5	188	12	193	7
Model 6	199	1	197	3

TABLE 2.6: Number of clusters selected over 200 iterations.

components we re-calculate the log-likelihood of the 3-variate Poisson model. This will allow the models to be comparable in terms of Bayesian Information Criterion. The selection of the proper number of clusters is now performed with the estimated BIC values. Table 2.7 demonstrates the updated number of selected clusters. We observe that even though the alternative composite models fails to choose the correct number of clusters if we use the estimated likelihood from the model, it succeeds if we use the re-calculated formula. Similar results are derived for all the sampling methods. A higher deviation is achieved for the Systematic Sampling method 3 (Model 4), where we estimate the parameters by choosing one out of the three Composite Likelihood components.

	$n = 200$		$n = 400$	
Model	**K=2**	**K=3**	**K=2**	**K=3**
Model 1	200	0	199	1
Model 2	198	2	197	3
Model 3	197	3	196	4
Model 4	188	12	179	21
Model 5	194	6	185	15
Model 6	200	0	200	0

TABLE 2.7: Results from fitting the different models-Estimated number of clusters selected over 200 iterations.

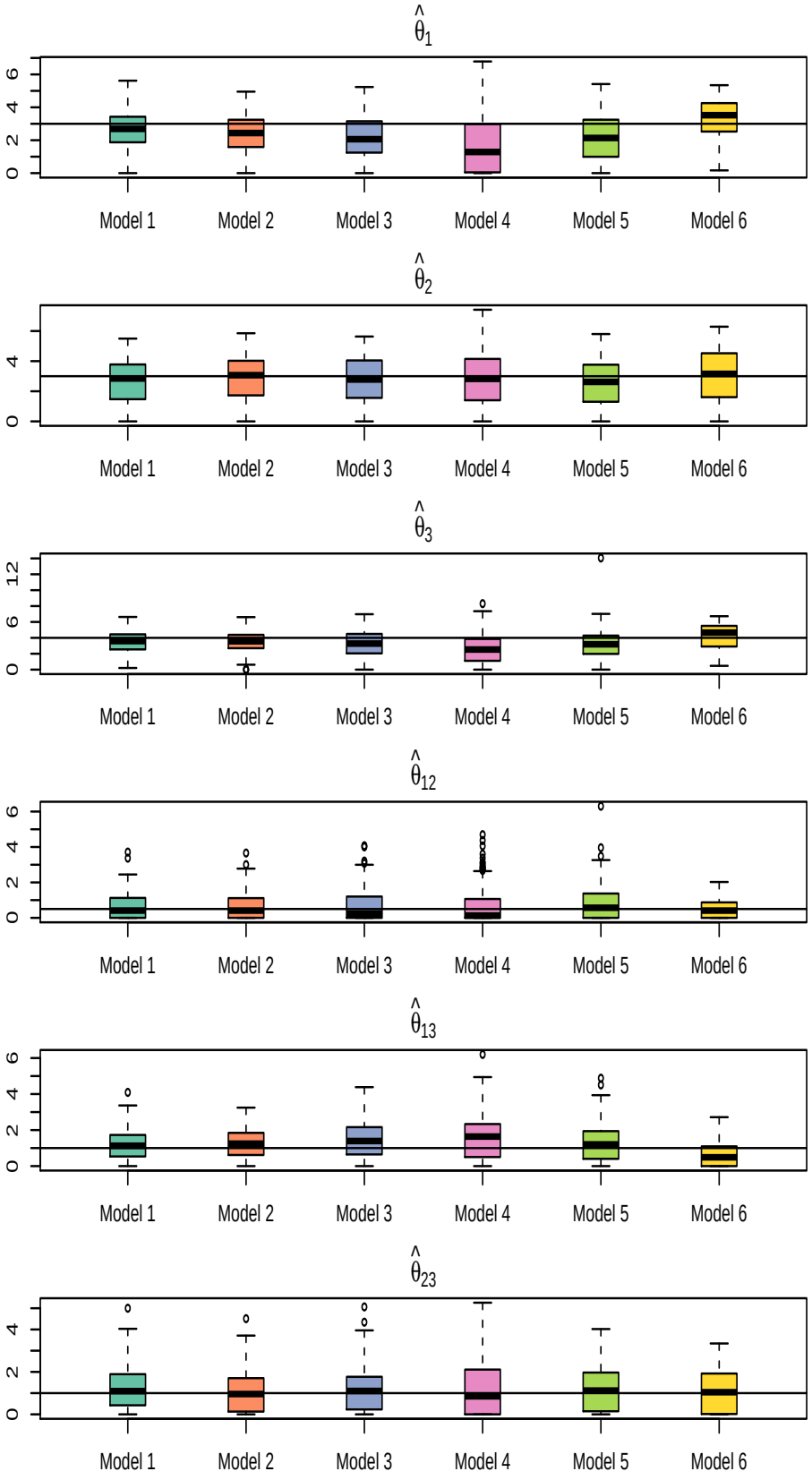


FIGURE 2.2: Estimated Parameters' Boxplots for k=1 cluster

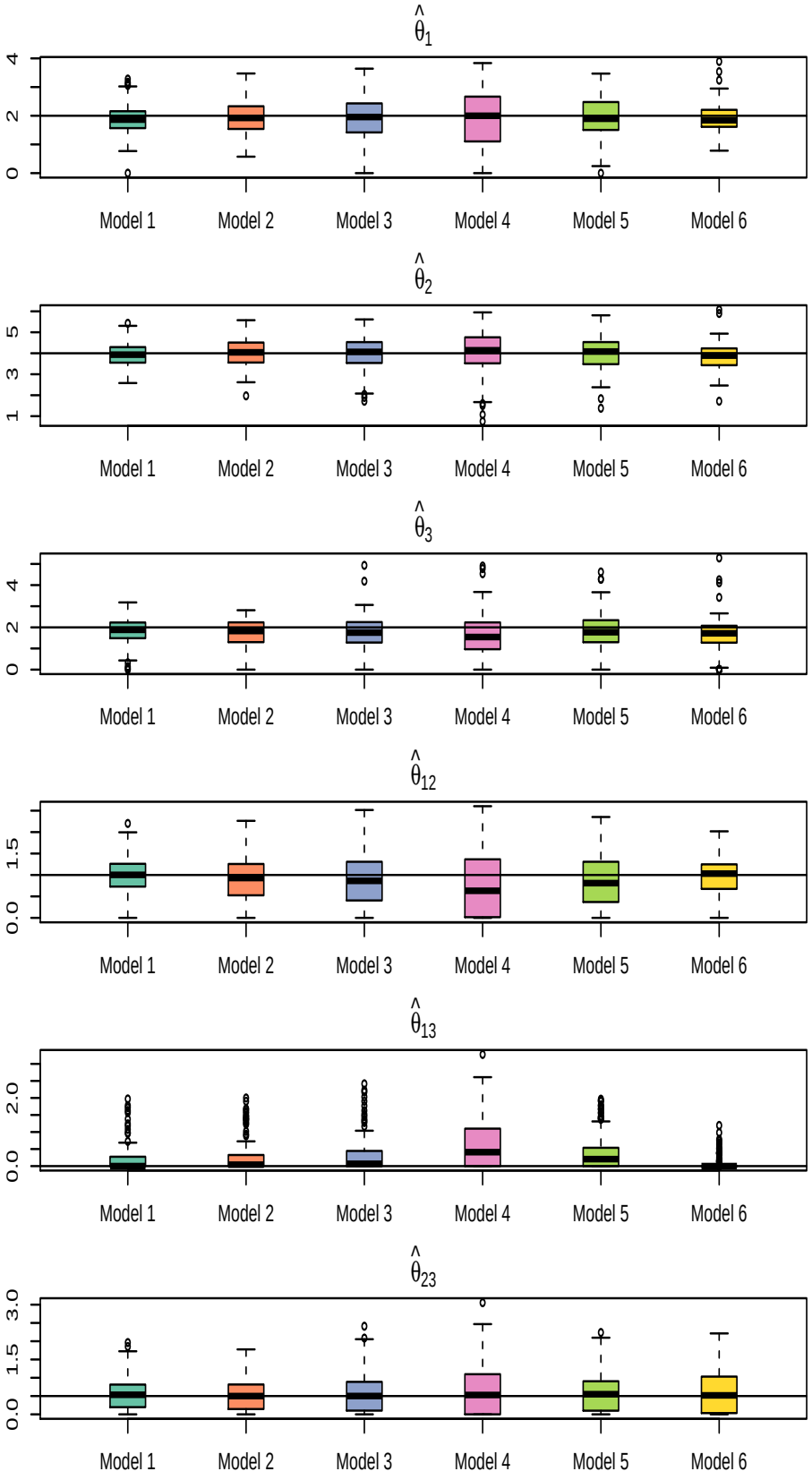


FIGURE 2.3: Estimated Parameters' Boxplots for k=2 cluster

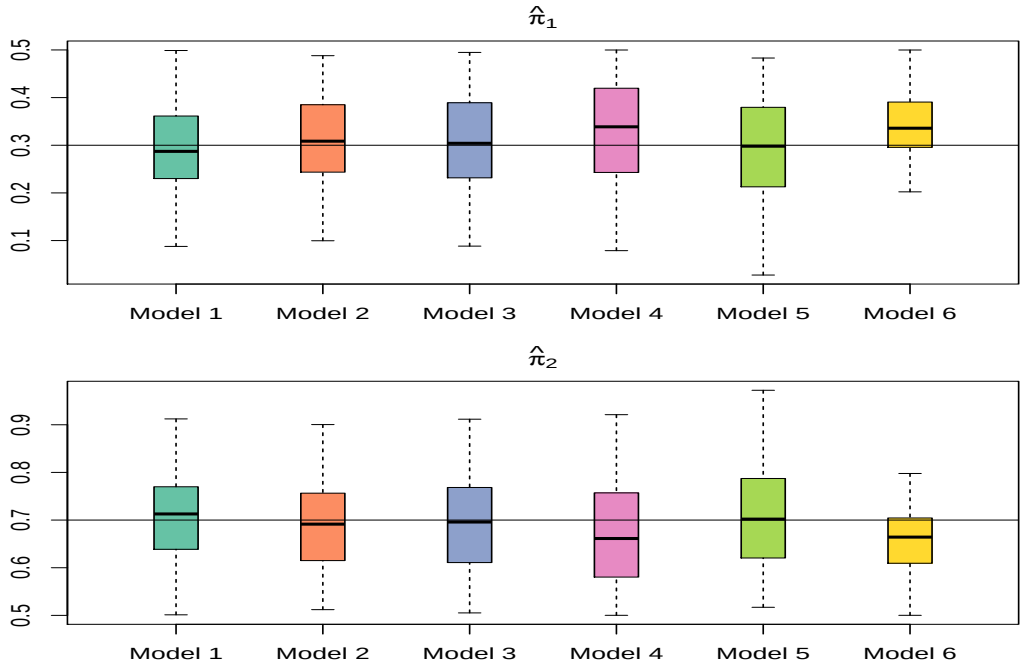


FIGURE 2.4: Estimated Mixing Probabilities' Boxplots

Figures 2.2, 2.3 & 2.4 provides the results for the fitted values for the parameters of the Poisson distributions of each cluster as well as for the mixing probabilities $\pi_j, j = 1, 2$ for all the models examined. In general models methods provide adequate results compared to the 3-variate Poisson distribution estimations. Less effective seems to be the Systematic Sampling 2 especially for the component with mixing probability $\hat{\pi} = 0.3$.

The measurement of effieciency of the classification will be performed with the use of Rand Index (RI) and Adjusted Rand Index (ARI).

$X \backslash Y$	Y_1	Y_2	$\cdots$	Y_s	sums
X_1	n_{11}	n_{12}	$\cdots$	n_{1s}	a_1
X_2	n_{21}	n_{22}	$\cdots$	n_{2s}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_r	n_{r1}	n_{r2}	$\cdots$	n_{rs}	a_r
sums	b_1	b_2	$\cdots$	b_s	n

TABLE 2.8: Results from fitting the different models-Estimated number of clusters selected over 200 iterations.

For a contingency table as shown in the table 2.8 the Rand Index is calculated as:

$$RI = \frac{\sum\sum_{i=j} n_{ij}}{n},$$

while

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] \Big/ \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] \Big/ \binom{n}{2}}$$

where n_{ij}, a_i, b_j, n are values from the contingency table.

For our simulation study table 2.10 summarizes for every model the results of the classification compared to the true values of clusters of the simulated data.

		Assigned Component		
Model	**Real Component**	**K=1**	**K=2**	**Total**
Model 1	**K=1**	**8,356**	3,609	11,965
(3-Variate	**K=2**	2,643	**25,392**	28,035
Poisson)	**Total**	10,999	29,001	40,000
Model 2	**K=1**	**7,709**	4,256	11,965
(Full	**K=2**	2,276	**25,759**	28, 035
Composite)	**Total**	9,985	30,015	40,000
Model 3	**K=1**	**7,388**	4,577	11,965
(Sampling	**K=2**	2,405	**25,630**	28,035
1)	**Total**	9,793	30,207	40,000
Model 4	**K=1**	**6,992**	4,973	11,965
(Sampling	**K=2**	5,141	**22,894**	28,035
2)	**Total**	12,133	27,867	40,000
Model 5	**K=1**	**7,118**	4,370	11,488
(Sampling	**K=2**	3,605	**23,362**	26,967
3)	**Total**	10,723	27,732	38,455
Model 6	**K=1**	**9,465**	2,500	11,965
(Alter.	**K=2**	3,961	**24,074**	28,035
Composite	**Total**	13,426	26,574	40,000

TABLE 2.9: Results from fitting the different models-Contingency table for every row of the sample datasets of n=200

For every simulated dataset there are 40,000 combinations examined resulting from the combinations of 200 rows × 200 iterations. For the case of Model 5 - Non

Systematic Sampling, since we eliminate rows of the simulated datasets, the number of combinations examined for similarity is 38,455. The sampling algorithm is random therefore different rows are excluded in each iteration.

Similarly to the structure of the sample data of n=200, we perform a simulation study for a data set of n=400 rows. For this case there are 80,000 combinations examined resulting from the combinations of 400 rows × 200 iterations. For the case of Model 5 - Non Systematic Sampling the number of combinations examined is 77,028.

		Assigned Component		
Model	**Real Component**	**K=1**	**K=2**	**Total**
Model 1	**K=1**	**16,640**	7,330	23,970
(3-Variate	**K=2**	3,792	**52,238**	56,030
Poisson)	**Total**	20,432	59,568	80,000
Model 2	**K=1**	**15,076**	8,894	23,970
(Full	**K=2**	2,841	**53,189**	56,030
Composite)	**Total**	17,917	62,083	80,000
Model 3	**K=1**	**14,801**	9,169	23,970
(Sampling	**K=2**	3,611	**52,419**	56,030
1)	**Total**	18,412	61,588	80,000
Model 4	**K=1**	**13,332**	10,638	23,970
(Sampling	**K=2**	6,602	**49,428**	56,030
2)	**Total**	19,934	60,066	80,000
Model 5	**K=1**	**14,540**	8,524	23,064
(Sampling	**K=2**	5,574	**48,390**	53,964
3)	**Total**	20,114	56,914	77,028
Model 6	**K=1**	**18,923**	5,047	23,970
(Alter.	**K=2**	7,262	**48,768**	56,030
Composite	**Total**	26,185	53,815	80,000

TABLE 2.10: Results from fitting the different models-Contingency table for every row of the sample datasets of n=400

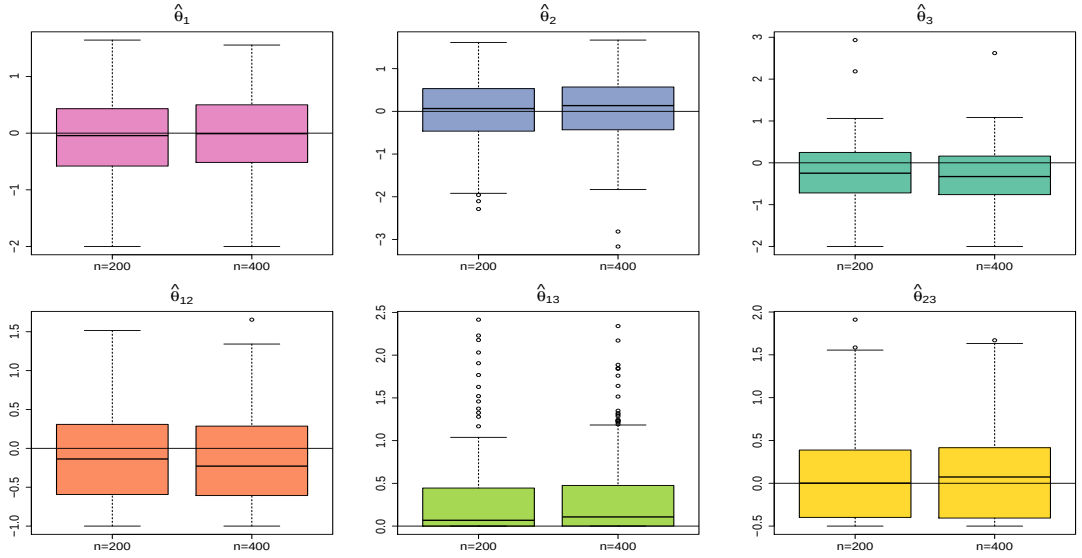


FIGURE 2.5: Residuals of the parameters of $\pi = 0.7$ for n=200 & n=400

	$n = 200$		$n = 400$	
Model	**RI**	**ARI**	**RI**	**ARI**
Model 1	0.8437	0.4546	0.8609	0.5020
Model 2	0.8367	0.4304	0.8533	0.4735
Model 3	0.8255	0.3984	0.8403	0.4367
Model 4	0.7472	0.2250	0.7845	0.2952
Model 5	0.7926	0.3210	0.8169	0.3790
Model 6	0.8385	0.4486	0.8461	0.4690

TABLE 2.11: RI & ARI Results of classification

Results of the RI and ARI values for each model are shown in table 2.11. From the resulted values of RI and ARI we can observe that the Composite Likelihood approach and the Alternative Composite Likelihood approach provides similar results. There is no significant loss for the case of Systematic Sampling approach 1 though the Systematic Sampling approach 2 provides more precarious results.

Figure 2.6 provides the resulted bi-variate contours for the density of the fitted values of each model assumed. Methods provide similar results with an exception of model 4 (choose one out of the three composite likelihood components), where we observe a smoother density failing to capture the mixture schema.

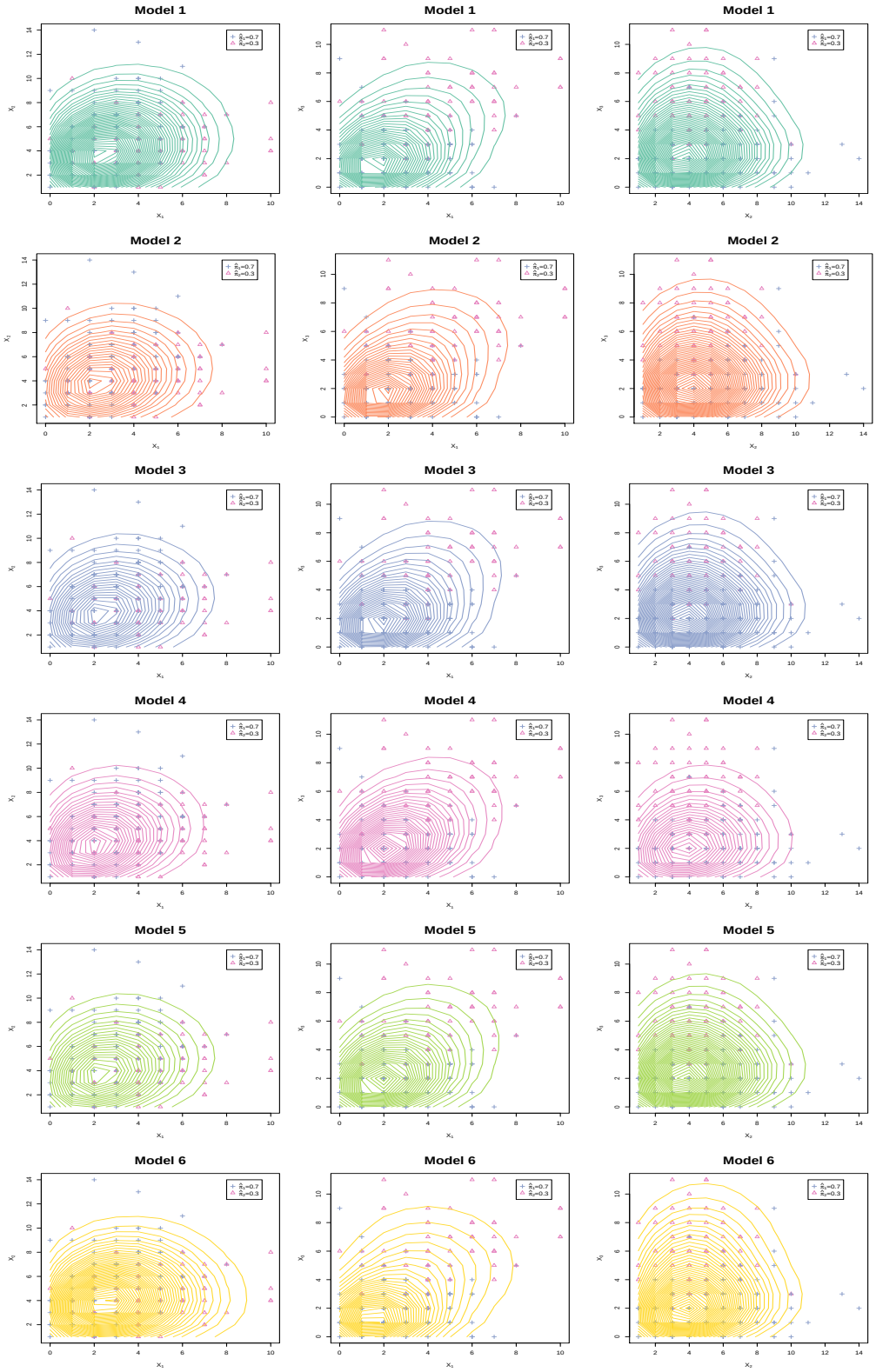


FIGURE 2.6: Contours for the mixture bi-variate density of the simulated data for all models

2.8 Concluding Remarks

Sampling methods have been introduced to further reduce computational effort of the Composite Likelihood approach. For our Simulation studies purposes the dimension reduction was not extended nevertheless the time consumed reduction was important compared even compared to the full composite likelihood method. For cases of high dimensional count data, where the full Multivariate Poisson density function is not easy to be defined the composite likelihood concept offers flexibility in calculations. We can further reduce the complexity of calculations via the sampling methods.

Among the Sampling methods the efficiency is much dependable on the sample data size, the highest the row size the more components or rows we can eliminate in calculations. For the specific simulation studies with $n = 200$ rows the Sampling method where we choose only one of the composite likelihood components provides poorer results, though for highest data points the method can further provide adequate results.

Alternative composite likelihood method which is less complex than the full traditional composite likelihood method can be also provide good classification without significant loss of miss-classified data points. This method can provide adequate values for the parameters and which can also be used as starting values of the full multivariate model estimation or the composite likelihood estimation. This method can be further investigated in order to provide adequate results.

Chapter 3

Copulas

3.1 Introduction

Model based clustering (MBC) has found a large number of applications in recent years as opposed to distance based and partition clustering. Most of the existing MBC literature is based on multivariate Gaussian mixtures and their variants, like multivariate t- mixtures. Both approaches assume that each cluster has an elliptical shape which is very restrictive for real data. To correct on this, one may consider skewed multivariate distributions like mixtures of multivariate skew-normal or skew-t distributions. For restricted domain, e.g. for data only on the positive axes, the literature is less developed, a common approach transforms the data to the real line to apply the models mentioned above, or apply a conditional independence assumption, i..e within each cluster the variables are independent, for such an example with multivariate beta mixtures see Sahu et al. (2016).

For non-continuous data models are less developed. For multivariate count data there are attempts with multinomial distribution (see, Jorgensen (2004)) and multivariate Poisson models (see, Karlis and Meligkotsidou (2007)). Also conditional independent models are described in Wedel and Kamakura (2000) while a model with block conditional assumption in Tom Brijs (2004). For mixed mode data there are works based on latent models and/or conditionally independence assumption (see, e.g. , McParland and Gormley (2016), Browne (2012) and Marbac et al. (2017)) The main reason is that it is not easy to create multivariate models that are simple enough for clustering purposes. For a recent review on clustering mixed mode data see Foss (2016), Foss et al. (2018) and Hennig and Liao (2013).

Recently Kosmidis and Karlis (2016) proposed the use of copulas to define the joint distributions that characterize the distribution of each component. This was a very generic approach that actually contains all the models as special cases, since one can define the marginal properties for each component and the dependence structure in a flexible way.

Copulas are well known as flexible models which allow creating multivariate distributions with given marginals. Hence, they can create a wealth of multivariate models including models with different marginal distributions. The purpose of the present thesis is mainly to expand the derived so far results of using copula based models for MBC applications.

We think that there are several advantages of using copulas in model based applications. These are:

- Copulas can be used to create a variety of multivariate models.
- using copulas we may define multivariate distributions with different marginal distributions and hence expand a lot our tank of potential models. For example, we can create a bivariate model with one normal and one t distribution. Such flexibility is not offered with recent models that restrict the models to have the same marginal distribution properties.
- The dependence structure as captured by the copula can have several different shapes, beyond the elliptical as we usually have with multivariate normal and their extensions. Therefore we are able to define appropriate multivariate models to allow for possible flexible shapes but stay in the context of model based clustering.
- It is easy to create multivariate models for several kinds of data (e.g. discrete) or even mixed model data. Hence one can create realistic models for data that have for example continuous, discrete and ordinal variables at the same time. Recent models typically assume conditional independence in order to derive joint distribution, i.e. they assume that conditional on the cluster we have independence.
- In some circumstances while the marginal distributions are the same, the dependence structure changes. Such a behavior occurs for example in finance (regarding the behavior of a portfolio when news come), sports (scoring behavior depends on the current score), marketing (purchase frequency patterns depends on household decomposition) etc. Changes in dependence

structure can be captured by different copulas and hence mixtures of copulas can be used to cluster data with respect to their dependence behavior.

As mentioned in Kosmidis and Karlis (2016) the approach has some important advantages, the more important being that the appropriate choice of copulas provides the ability to obtain a range of exotic shapes for the clusters, which is not easy with standard models and the explicit choice of marginal distributions for the clusters allows the modelling of multivariate data of various modes (discrete, continuous, both discrete and continuous) in a natural way.

In the present thesis we exploit the use of copulas based MBC for mixed mode data. A similar attempt has been made in Marbac et al. (2017) using a Bayesian approach to estimate parameters. Our approach uses full likelihood methods using also dimension reduction approaches to reduce the number of parameters we have to estimate. We propose two approaches for dimension reduction, the first one is based on factor analyzers McLachlan et al. (2003) where the correlation matrix of the Gaussian copula is expressed via the standard factor decomposition. The second approach makes use of a parsimonious representation of the correlation matrix due to Tsay and Pourahmadi (2017) that allows to put together variables and assume a conditional independence between group of variables leading to reduced parameter space. Note that while the chapter focuses on the mixed mode data clustering problems, the dimension reduction methods apply to any model based through copula case.

3.2 Background

3.2.1 Finite Mixture models

The use of finite mixture models in clustering is finding a large number of applications, mainly because it allows standard statistical modelling tools to be used in order to assess and evaluate the clustering. The density or probability mass function of a finite mixture model is defined as

$$h(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\pi}) = \sum_{j=1}^{k} \pi_j f_j(\boldsymbol{x};\boldsymbol{\theta}_j) \quad (\boldsymbol{x} \in \Re^p)\,, \tag{3.1}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_k^T)^T \in \Theta_1 \times \ldots \times \Theta_k$, and $\pi_j \in (0,1)$ with $\sum_{j=1}^{k} \pi_j = 1$. Appropriate choices of $f_j(\boldsymbol{x};\boldsymbol{\theta}_j)$ can result in flexible models of small complexity.

Banfield and Raftery (1993) and the book of McLachlan and Peel (2000) provide a detailed treatment of the framework of finite mixture modelling for clustering.

For continuous data, a common choice for the component densities $f_j(\boldsymbol{x};\boldsymbol{\theta}_j)$ $(j = 1,\ldots,k)$ is the density of the multivariate Gaussian distribution. This is mainly because of the convenience it offers in estimation (closed-form maximization steps in the EM algorithm) and interpretation (easy marginalization for visualising fitted components and the mixture density). The resultant clusters, though, are limited to be elliptical in shape, and as is demonstrated in Hennig (2010), one may need more than one multivariate Gaussian components in order to fit a single non-elliptical cluster.

Such restrictions of multivariate Gaussian finite mixtures have resulted in an expanding literature where other special component distributions are considered. Prominent examples of alternative component densities include multivariate t distributions (see, Andrews and McNicholas, 2011), multivariate skew-Gaussian and skew-t distribution (see, for example, Frühwirth-Schnatter and Pyne, 2010; Lee and McLachlan, 2014), multivatiate skew student-t-Gaussian distributions (Lin et al., 2014), multivariate Gaussian inverse Gaussian distributions (Karlis and Santourian, 2009). Other attempts can be found in Forbes and Wraith (2014) for finite mixtures of multivariate scaled Gaussian distributions and (Morris and McNicholas, 2013) for mixtures of shifted asymmetric Laplace distributions. The results of such studies indicate that the introduction of heavy tails and/or skewness allows the construction of more parsimonious models than multivariate Gaussian mixtures, that can also bridge the gap between the number of clusters present in the data and the number of components used in the mixture.

For non-continuous data, one needs to specify $f_j(\boldsymbol{x};\boldsymbol{\theta}_j)$ $(j = 1,\ldots,k)$ in (3.1) through probability mass functions. While there is a wealth of choices for univariate non-continuous distributions, the use of multivariate non-continuous distributions for the definition of mixture models is limited due to the difficulty in constructing easy to work with models that allow practical flexibility on the dependence structure. Some successful, but limited in application examples, are finite mixtures of multivariate Poisson distributions (Karlis and Meligkotsidou, 2007), finite mixtures of multinomial distributions (Jorgensen, 2004) and models based on conditionally independent Poisson distributions (see, for example Alfo et al., 2011). Mixture models with latent structures have been considered in Browne and McNicholas (2012), but these can have limitations because of assumptions like conditional independence.

3.2.2 Mixture models through copulas

A copula $C(u_1, \ldots, u_p)$ is a distribution function with uniform marginals. The importance of copulas in statistical modelling stems from Sklar's theorem (see, Nelsen, 2006, §2.3), which shows that every multivariate distribution can be represented via the choice of an appropriate copula and, more importantly, it provides a general mechanism to construct new multivariate models in a straightforward manner.

The copula-based mixture model is defined as in (3.1) but now $\boldsymbol{\theta}_j$ is partitioned as $(\boldsymbol{\gamma}_j^\top, \boldsymbol{\psi}_j^\top)^\top$ and $f_j(\boldsymbol{x}; \boldsymbol{\theta}_j)$ is the density (or probability mass function) corresponding to a distribution function

$$F_j(\boldsymbol{x}; \boldsymbol{\psi}_j, \boldsymbol{\gamma}_j) = C_j(G_1(x_1, \boldsymbol{\gamma}_{j1}), \ldots, G_p(x_p; \boldsymbol{\gamma}_{jp}); \boldsymbol{\psi}_j) \quad (j = 1, \ldots, k)\,, \tag{3.2}$$

where $G_1, \ldots, G_p$ are univariate marginal cumulative distribution functions. As far as the model parameters are concerned, $\boldsymbol{\gamma}_j$ contains the parameter vectors $\boldsymbol{\gamma}_{jt}$ for all marginals for jth component $(t = 1, \ldots, p)$ and $\boldsymbol{\psi}_j$ contains the parameters of the copula used for the j-th component.

3.2.3 Construction of mixture models for any type of data

The definition of the component density F_j through the choice of a copula C_j and the choice of marginal distributions $G_1, \ldots, G_p$ leads to a flexible framework for model-based clustering that according to Sklar's theorem necessarily encompasses all known mixture models and allows the convenient construction of new mixture models that can handle any of continuous, discrete data.

Temporarily omitting the component index and suppressing the dependence on the parameters, assume that the density of the copula $C(u_1, \ldots, u_p)$ exists and is $c(u_1, \ldots, u_p) = \partial^p C(u_1, \ldots, u_p)/\partial u_1 \ldots \partial u_p$. Then the component density for continuous marginals is

$$f(\mathbf{x}) = c(G_1(x_1), \ldots, G_p(x_p)) \prod_{t=1}^{p} g_t(x_t)\,.$$

For discrete data, the probability mass function is given in Panagiotelis et al. (2012, expression (1.2)), and results from finite differences of the distribution function as

$$P(\mathbf{x}) = \sum_{d} \operatorname{sgn}(\boldsymbol{d}) C(G_1(d_1), \ldots, G_p(d_p))\,, \tag{3.3}$$

with $\boldsymbol{d} = (d_1, \ldots, d_p)$ vertices, where each d_t is equal to either x_t or $x_t - 1$ ($t = 1, \ldots, p$), and

$$\operatorname{sgn}(\boldsymbol{d}) = \begin{cases} 1\,, & \text{if } d_t = x_t - 1 \text{ for an even number of } t\text{'s} \\ -1\,, & \text{if } d_t = x_t - 1 \text{ for an odd number of } t\text{'s} \end{cases}.$$

The model defined from (3.1) and (3.2) being a finite mixture allows for inferential procedures based on the standard theory of finite mixtures, like use of the EM algorithm for maximum likelihood estimation and the use of model selection criteria.

3.2.4 Full Expectation Maximization

Following Kosmidis and Karlis (2016), suppose that a sample of n p-vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ is available, which are assumed to be realizations of independent random variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ each with distribution with density or probability mass function as defined by (3.1) and (3.2). The maximization of the likelihood function based on that sample can be performed using the EM algorithm. At the ℓth iteration of the algorithm ($\ell = 2, 3, \ldots$),

- *E-step:* Calculate

$$w_{ij}^{(\ell+1)} = \frac{\pi_j^{(\ell)} f_j(\boldsymbol{x}_i; \boldsymbol{\theta}_j^{(\ell)})}{\sum_{j=1}^{k} \pi_j^{(\ell)} f_j(\boldsymbol{x}_i; \boldsymbol{\theta}_j{}^{(\ell)})} \quad (i = 1, \ldots, n; j = 1, \ldots, k)\,.$$

- *M-step 1:* Set $\pi_j^{(\ell+1)} = \sum_{i=1}^{n} w_{ij}^{(\ell+1)}/n$ ($j = 1, \ldots, k$).

- *M-step 2:* Maximize

$$\sum_{j=1}^{k} \sum_{i=1}^{n} w_{ij}^{(\ell+1)} \log\left\{f_j(\boldsymbol{x}_i; \boldsymbol{\theta}_j)\right\}\,,$$

  with respect to $\boldsymbol{\theta}$ to obtain an updated value $\boldsymbol{\theta}^{(\ell+1)}$ for the copula and marginal parameters.

The algorithm iterates between the E-step and the M-step until some convergence criterion is satisfied. In all the examples in the current work the terminating criterion that is used is that the relative increase $\{l(\theta^{(\ell+1)}, \pi^{(\ell+1)}) - l(\theta^{(\ell)}, \pi^{(\ell)})\}/l(\theta^{(\ell)}, \pi^{(\ell)})$ of the log-likelihood $l(\theta, \pi)$ in two successive iterations is less than $\epsilon = 10^{-8}$.

For calculating the starting values for $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ the following procedure is proposed which takes into account both the copula and the marginal specification of each component in the mixture model. The procedure is an application of the Inference Functions from Margins (IFM) method (Joe, 1997, Chapter 10) for each component, and relies on an initial partitioning of the observation indices $A = \{1, \ldots, n\}$ into exclusive subsets $S_1, \ldots, S_k$, with $\cup_{j=1}^{k} S_j = A$, of cardinality $N_1, \ldots, N_k$, respectively. More specifically, the procedure for obtaining starting values consists of the following steps:

S1 Set the starting values for $\boldsymbol{\pi}_j$ using $\pi_j^* = N_j/n$ $(j = 1, \ldots, k)$.

S2 Use maximum likelihood to fit the marginal g_t on data x_{it} for $i \in S_j$ in order to obtain starting values $\boldsymbol{\gamma}_{jt}^*$ for $\boldsymbol{\gamma}_{jt}$ $(t = 1, \ldots, p)$.

S3 Use maximum likelihood to fit the copula $C_j(u_1, \ldots, u_p; \boldsymbol{\psi}_j)$ on observations $u_{it} = G_t(x_{it}, \boldsymbol{\gamma}_{jt}^*)$ $(i \in S_j; t = 1, \ldots, p)$, in order to get starting values $\boldsymbol{\psi}_j^*$ for the copula parameters $\boldsymbol{\psi}_j$.

The initial classification vector that is required for this procedure can be obtained either using a hard-partitioning distance-based algorithm (like k-means for continuous data or k-medoids more generally) or by randomly sampling k observations and using the minimum distance of each those from all other observations in order to form $S_1, \ldots, S_k$.

For the analysis of continuous data, *M-step 2* takes the form

- *M-step 2:* Maximize the log-likelihood

$$\sum_{j=1}^{k} \sum_{i=1}^{n} w_{ij}^{(\ell+1)} \left[\log c_j(G_1(x_{i1}; \boldsymbol{\gamma}_{j1}), \ldots, G_p(x_{ip}; \boldsymbol{\gamma}_{jp}); \boldsymbol{\psi}_j) + \sum_{t=1}^{p} \log g_t(x_{it}; \boldsymbol{\gamma}_{jt}) \right], \tag{3.4}$$

with respect to $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_k, \boldsymbol{\gamma}_{11}, \ldots, \boldsymbol{\gamma}_{1p}, \boldsymbol{\gamma}_{k1}, \ldots, \boldsymbol{\gamma}_{kp}$, where $\boldsymbol{\gamma}_{jt}$ is the vector of parameters of the tth marginal distribution for the jth component of the mixture $(t = 1, \ldots, p; j = 1, \ldots, k)$.

As is apparent from (3.4) the only necessary ingredients for implementing the EM algorithm for mixtures of copulas for continuous data are the specification of the copula densities $c_1, \ldots, c_k$ and the specification of the marginal density and distribution functions $g_1, \ldots, g_p$ and $G_1, \ldots, G_p$, respectively.

The particular form of the complete data log-likelihood for continuous data allows here the use of the Expectation/Conditional Maximization (ECM) algorithm

of Meng and Rubin (1993), where the full maximization of the complete data log-likelihood is relaxed to maximization in blocks; first with respect to the marginal parameters given the current value of the copula parameter and then with respect to the copula parameter given the updated values for the marginal parameters. In mathematical notation, *M-step 2* is replaced by the steps

- *CM-step 1:* Maximize

$$\sum_{j=1}^{k}\sum_{i=1}^{n} w_{ij}^{(\ell+1)} \left[\log c_j(G_1(x_{i1};\boldsymbol{\gamma}_{j1}),\ldots,G_p(x_{ip};\boldsymbol{\gamma}_{jp});\boldsymbol{\psi}_j^{(\ell)}) + \sum_{t=1}^{p}\log g_t(x_{it};\boldsymbol{\gamma}_{jt})\right], \tag{3.5}$$

  with respect to $\boldsymbol{\gamma}_{11}$, …, $\boldsymbol{\gamma}_{1p}$, $\boldsymbol{\gamma}_{k1}$, …, $\boldsymbol{\gamma}_{kp}$ to obtain updated values $\boldsymbol{\gamma}_{11}^{(\ell+1)}$, …, $\boldsymbol{\gamma}_{1p}^{(\ell+1)}$, $\boldsymbol{\gamma}_{k1}^{(\ell+1)}$, …, $\boldsymbol{\gamma}_{kp}^{(\ell+1)}$ for the marginal parameters.

- *CM-step 2:* Maximize

$$\sum_{j=1}^{k}\sum_{i=1}^{n} w_{ij}^{(\ell+1)} \left[\log c_j(G_1(x_{i1};\boldsymbol{\gamma}_{j1}^{(\ell+1)}),\ldots,G_p(x_{ip};\boldsymbol{\gamma}_{jp}^{(\ell+1)});\boldsymbol{\psi}_j)\right], \tag{3.6}$$

  with respect to $\boldsymbol{\psi}_1,\ldots,\boldsymbol{\psi}_k$ to obtain updated values $\boldsymbol{\psi}_1^{(\ell+1)},\ldots,\boldsymbol{\psi}_k^{(\ell+1)}$ for the copula parameters.

According to the definitions and results in Meng and Rubin (1993), the ECM algorithm that results by replacing *M-step 2* with the pair *CM-step 1* and *CM-step 2* shares all the convergence properties of the full EM algorithm, and, in this particular case, is more computationally efficient and stable, because *CM-step 2* consists of a simple maximization with respect to the copula parameters. Furthermore, *CM-step 1* and *CM-step 2* can each be broken down into parallel optimizations across components, as in the case of the full EM, which significantly reduces computation time in multicore systems.

For the pair of *CM-step 1* and *CM-step 2* their difference lies in *CM-step 1* where instead of maximizing the weighted sum of marginal log-likelihoods, a valid ECM algorithm requires the maximization of a penalized version of it where the penalty depends on the log copula density at the current value for the copula parameter.

3.3 Gaussian copula for mixed mode data

Multivariate Gaussian copula is defined as (see, e.g. Joe (2014))

$$C^N(u_1, u_2, ..., u_p; R) = \mathbf{\Phi}_p(\mathbf{\Psi}(u_1), ..., \mathbf{\Psi}(u_p); R), \tag{3.7}$$

where $\mathbf{\Phi}_p$ is the distribution function of a standard p-variate Gaussian distribution with correlation matrix R, $\mathbf{\Psi}(\cdot) = \mathbf{\Phi}^{-1}(\cdot)$ is the inverse distribution function of a standard univariate Gaussian distribution and u_i, $i = 1, ..., p$ are the marginal probability distributions $F_i(x_i; \boldsymbol{\gamma_i})$ for variables $X_1, ..., X_p$. The correlation matrix R has the form:

$$R = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{12} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1p} & \rho_{2p} & \cdots & 1 \end{bmatrix}, \tag{3.8}$$

where ρ_{kt} is the correlation between the k-th and t-th distribution of variables $k, t = 1, ..., p$, $k \neq t$. In order to represent the joint probability mass function of a set of variables through a Gaussian copula one has to estimate the parameters θ_j associated with the marginals and the correlation matrix R associated with the copula. For the continuous case the density of a set of variables results from derivative of the copula function for every marginal and for discrete case at the same way from finite differences (see e.g. Panagiotelis et al. (2012)). For mixed mode data where some of the marginals are continuous and some discrete the procedure of modelling the joint probability is described below.

Example 3.1

Let's assume X is a continuous random variable and Y a discrete one, with marginal c.d.f.s $u_1 = F(x)$ and $u_2 = G(y)$ then:

$$\begin{aligned} f(x, y) &= \frac{\partial F(x, y)}{\partial x} - \frac{\partial F(x, y-1)}{\partial x} \\ &= \left(\frac{\partial C^N(u_1, G(y))}{\partial u_1} - \frac{\partial C^N(u_1, G(y-1))}{\partial u_1} \right) f(x) \\ &= \Big(P\big(T_2 \leq \Psi(u_2) | T_1 = \Psi(u_1)\big) - P(T_2 \leq \Psi(G(y-1)) | T_1 = \Psi(u_1)\Big) \\ &\times P(T_1 = \Phi^{-1}(u_1)) \ \frac{1}{\phi(\Phi^{-1}(u_1))} f(x) \end{aligned}$$

where $T_2|T_1 \sim N\big(\rho_{12}\Phi^{-1}(u_1),\, 1-\rho_{12}^2\big)$ and $T_1 \sim N(0, \rho_{12})$

Let $X_1, X_2, ..., X_k$ be a set of continuous random variables with distribution functions $F_i(X_i;\gamma_i)$, $i = 1, ..., k$ and $X_{k+1}, X_{k+2}, ..., X_{k+l}$, a set of discrete random variables with p.d.f. $F_j(X_j;\gamma_j)$, $j = k+1, ..., k+l$, respectively.

The joint p.d.f. of $X_1, ..., X_{k+l}$ results from 2^{l-1} finite differences of the k-th partial derivative of the distribution function $F(x_1, x_2, ..., x_p)$. That is:

$$f(x_1, x_2, ..., x_{k+l}) = \sum_d sgn(\boldsymbol{d}) \frac{\partial^{(k)} F(x_1, x_2, ..., x_k, d_1, d_2, ..., d_l)}{\partial x_1 \partial x_2 ... \partial x_k} \tag{3.9}$$

where d_i is either x_i or $x_i - 1$, $i = 1, ..., l$. For an even number of $x_i - 1$, $sgn(\boldsymbol{d})$ is positive, and negative otherwise (see, e.g. Czado et al. (2012)).

The k-th partial derivative, for the chosen Gaussian copula, can be written to a simpler form by Leibnitz rule and 3.9 as:

$$C^N(u_1, u_2, ..., u_k, u_{k+1}, ..., u_{k+l}; R) = \boldsymbol{\Phi}_p(\boldsymbol{\Psi}(u_1), ..., \boldsymbol{\Psi}(u_{k+l}); R)$$

$$\propto \int_{-\infty}^{\Psi(u_1)} ... \int_{-\infty}^{\Psi(u_k)} \int_{-\infty}^{\Psi(u_{k+1})} ... \int_{-\infty}^{\Psi(u_{k+l})} \exp\left\{ -\frac{1}{2} \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_{k+l} \end{bmatrix}' R^{-1} \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_{k+l} \end{bmatrix} \right\} dt_1 ... dt_{k+l}$$

$$\frac{\partial^{(k)} F(x_1, ..., x_k, d_1, d_2, ..., d_l)}{\partial x_1 ... \partial x_k} = \frac{\partial^{(k)} C^N(u_1, u_2, ..., u_k, u_{k+1}, ..., u_{k+\ell}; R)}{\partial u_1 \partial u_2 \cdots \partial u_k} \times \prod_{i=1}^{k} f(x_i)$$

$$= P\Big(t_{k+1} \le \Psi(u_{k+1}), \cdots, t_{k+\ell} \le \Psi(u_{k+\ell}) | t_1 = \Psi(u_1), \cdots, t_k = \Psi(u_k)\Big)$$

$$\times P\Big(t_1 = \Psi(u_1), \cdots, t_k = \Psi(u_k)) \times \prod_{i=1}^{k} \frac{1}{\phi(\Phi^{-1}(u_i))}\Big) \times \prod_{i=1}^{k} f(x_i)$$

$$\frac{\partial^{(1)} C^N(u_1, u_2, ..., u_k, u_{k+1}, ..., u_{k+l}; R)}{\partial u_1} \propto \frac{1}{\phi(\Phi^{-1}(u_1))} \times$$

$$\int_{-\infty}^{\Psi(u_2)} ... \int_{-\infty}^{\Psi(u_k)} ... \int_{-\infty}^{\Psi(u_{k+l})} \exp\left\{ -\frac{1}{2} \begin{bmatrix} \Psi(u_1) \\ t_2 \\ \vdots \\ t_{k+l} \end{bmatrix}' R^{-1} \begin{bmatrix} \Psi(u_1) \\ t_2 \\ \vdots \\ t_{k+l} \end{bmatrix} \right\} dt_2 ... dt_{k+l}$$

By induction:

$$\frac{\partial^{(k)} C^N(u_1, u_2, ..., u_k, u_{k+1}, ..., u_{k+l}; R)}{\partial u_1 \partial u_2 \cdots \partial u_k} = \prod_{i=1}^{k} \frac{1}{\phi(\Phi^{-1}(u_i))} \times$$

$$(2\pi)^{-\frac{k+l}{2}} R^{-\frac{1}{2}} \int_{-\infty}^{\Psi(u_{k+1})} ... \int_{-\infty}^{\Psi(u_{k+l})} \exp\left\{ -\frac{1}{2} \begin{bmatrix} \Psi(u_1) \\ \vdots \\ \Psi(u_k) \\ t_{k+1} \\ \vdots \\ t_{k+l} \end{bmatrix}' R^{-1} \begin{bmatrix} \Psi(u_1) \\ \vdots \\ \Psi(u_k) \\ t_{k+1} \\ \vdots \\ t_{k+l} \end{bmatrix} \right\} dt_{k+1} ... dt_{k+l}$$

$$= P(t_{k+1} \leq \Psi(u_{k+1}), t_{k+2} \leq \Psi(u_{k+2}), \cdots, t_{k+l} \leq \Psi(u_{k+l}), t_1 = \Psi(u_1), \cdots, t_k = \Psi(u_k))$$

$$\times \prod_{i=1}^{k} \frac{1}{\phi(\Phi^{-1}(u_i))}$$

$$= P(t_{k+1} \leq \Psi(u_{k+1}), \cdots, t_{k+l} \leq \Psi(u_{k+l}) | t_1 = \Psi(u_1), \cdots, t_k = \Psi(u_k))$$

$$\times P(t_1 = \Psi(u_1), \cdots, t_k = \Psi(u_k)) \prod_{i=1}^{k} \frac{1}{\phi(\Phi^{-1}(u_i))}$$

(3.10)

Lemma 1. Let $\underline{\boldsymbol{x}} = \begin{bmatrix} \underline{\boldsymbol{x}}_1 \\ \underline{\boldsymbol{x}}_2 \end{bmatrix}$ be a p-dimensional vector with $\underline{\boldsymbol{x}} \sim MN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu_1} \\ \boldsymbol{\mu_2} \end{bmatrix}$ and $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma_{11}} & \boldsymbol{\Sigma_{12}} \\ \boldsymbol{\Sigma_{21}} & \boldsymbol{\Sigma_{22}} \end{bmatrix}$. The conditional distribution of $\underline{\boldsymbol{x}}_2 | \underline{\boldsymbol{x}}_1 = \boldsymbol{a}$ is also

multivariate Gaussian distribution with parameters $\boldsymbol{\mu}'$ and $\boldsymbol{\Sigma}'$ where

$$\boldsymbol{\mu}' = \boldsymbol{\mu_2} + \boldsymbol{\Sigma_{21}}\boldsymbol{\Sigma_{11}^{-1}}(\boldsymbol{a} - \boldsymbol{\mu_1})$$

and

$$\boldsymbol{\Sigma}' = \boldsymbol{\Sigma_{22}} - \boldsymbol{\Sigma_{21}}\boldsymbol{\Sigma_{11}^{-1}}\boldsymbol{\Sigma_{12}}$$

The correlation matrix R defined by 3.8 can be split into blocks where R_{11} is the correlation matrix between the first k continuous variables, R_{22} the correlation matrix for the ℓ discrete ones and R_{12} includes the correlations between both of them.

$$R = \left[\begin{array}{ccc:ccc} 1 & \cdots & \rho_{1k} & \rho_{1(k+1)} & \cdots & \rho_{1(k+l)} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \rho_{1k} & \cdots & 1 & \rho_{1(k+1)} & \cdots & \rho_{k(k+l)} \\ \hdashline \rho_{1(k+1)} & \cdots & \rho_{k(k+1)} & 1 & \cdots & \rho_{(k+1)(k+l)} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \rho_{1(k+l)} & \cdots & 1 & \rho_{(k+1)(k+l)} & \cdots & 1 \end{array}\right] = \left[\begin{array}{c:c} R_{11} & R_{12} \\ \hdashline R_{21} & R_{22} \end{array}\right]$$

Based on the above Lemma 1 the conditional probability in equation 3.10 $P\Big(t_{k+1} \leq \Psi(u_{k+1}), \cdots, t_{k+l} \leq \Psi(u_{k+l}) | t_1 = \Psi(u_1), \cdots, t_k = \Psi(u_k)\Big)$ can be estimated as the probability mass function of a multivariate normal distribution of size l, with mean $\boldsymbol{\mu} = R_{21}R_{11}^{-1}\Big[\Psi(u_1), \cdots, \Psi(u_k)\Big]'$ and covariance matrix $\boldsymbol{\Sigma} = R_{22} - R_{21}R_{11}^{-1}R_{12}$. Moreover, the density $P\Big(t_1 = \Psi(u_1), \cdots, t_k = \Psi(u_k)\Big)$ can also be estimated as the probability density function of a multivariate normal distribution of size k, with zero mean and covariance matrix R_{11}.

In conclusion the log likelihood for a multi-dimensional vector $\boldsymbol{x} = (x_1, ..., x_{k+\ell})$ with respect of the parameters of the marginals γ and the copula's ψ, is now

written as:

$$\ell = \log\left\{f(\boldsymbol{x};\boldsymbol{\gamma},\boldsymbol{\Psi})\right\} = \sum_{t=1}^{k}\log(f_t(\boldsymbol{x};\gamma_t)) - \sum_{t=1}^{k}\log\left\{\phi(\Phi^{-1}(u_t))\right\} + \log\left\{\phi_k\Big(\Psi(u_1),\cdots,\Psi(u_k)\,;\mathbf{0},R_{11}\Big)\right\} + \log\left\{\sum_{\boldsymbol{d}} sgn(\boldsymbol{d})\,\boldsymbol{\Phi}_\ell\Big(\Psi(u_{k+1}),\cdots,\Psi(u_{k+l})\,;\,\boldsymbol{\mu},\boldsymbol{\Sigma}\Big)\right\} \tag{3.11}$$

where $\boldsymbol{d}$ as defined above, $u_t = F(x_t;\gamma_t)$ $t = 1,...,k$ the marginal distribution functions, and $u_1,...,u_t$ quantities that follow the standard uniform distribution and $\Psi(u_i)$ quantities that follow $MVN_{k+\ell}(\mathbf{0},\boldsymbol{R})$.

3.3.1 CEM for mixtures of copulas

When managing with mixtures of copulas for mixed mode data, the unconditional density function for an observation $\boldsymbol{x} = (x_1,...,x_k,x_{k+1},...,x_{k+\ell})$ with k continuous variables and ℓ discrete, is now written in the following form:

$$f(\boldsymbol{x}|\boldsymbol{\Theta}) = \sum_{g=1}^{G}\pi_g\Big(c\big(F_{1g}(x_1),...,F_{(k+l)g}(x_{k+l})\big)\prod_{t=1}^{k}f_t(x_t)\Big)$$

and $\boldsymbol{\Theta}$ includes all parameters for the marginals, the Gaussian copula and π_g. In order to evaluate the appropriate model, someone has to estimate not only the parametric space $\boldsymbol{\Theta}$ but also the number of components G. We can produce $\boldsymbol{\Theta}$ with the use of Expectation Conditional Maximization algorithm customized for mixed mode data. We are interest in maximizing the complete-data log likelihood of the model

$$\begin{aligned}\ell &= \log\Big\{\prod_{i=1}^{n}\sum_{g=1}^{G}\pi_g f_g(\boldsymbol{x_i};\boldsymbol{\gamma_j})\Big\} = \log\Big\{\prod_{i=1}^{n}\prod_{g=1}^{G}f_g(\boldsymbol{x_i};\boldsymbol{\gamma_j})^{w_{ig}}\Big\}\\ &= \sum_{i=1}^{n}\sum_{g=1}^{G}\Big\{w_{ig}\log f_g(\boldsymbol{x_i};\boldsymbol{\gamma_g})\Big\}\Big\} = \sum_{i=1}^{n}\sum_{g=1}^{G}\Big\{w_{ig}\ell_{ig}\Big\}\end{aligned} \tag{3.12}$$

where ℓ_{ig} as described in equation 3.11 for i, $i = 1,...,n$ sample of $(k+\ell)$-dimension vector $\boldsymbol{x}$ and g, $g = 1,...,G$ component. At the r^{th}-iteration of ECM

algorithm we have:

E-step: Calculate for $i = 1, ..., n$ and $g = 1, ..., G$

$$w_{ig}^{(r+1)} = \frac{\pi_g^{(r)} f_g(x_i; \gamma_g^{(r)}, \psi_g^{(r)})}{\sum_{g=1}^{G} \pi_g^{(r)} f_g(x_i; \gamma_g^{(r)}, \psi_g^{(r)})}$$

CM-step 1: Update $\pi_g^{(r+1)} = \sum_{i=1}^{n} w_{ig}^{(r+1)}/n$ for all $g = 1, ..., G$ and then maximize

$$\sum_{i=1}^{n} \Big\{ w_{ig} \ell_{ig1} \Big\}$$

with respect to the γ_{gt} to get updated values for $\gamma_{gt}^{(r+1)}$ parameters associated with the t-th marginal distribution, $t = 1, ..., k+\ell$, and the g-th component $g = 1, ..., G$. As in equation 3.11 for the case of mixed mode data with k continuous and ℓ discrete marginals ℓ_{ig1} is written in the following form

$$\begin{aligned}
\ell_{ig1}^{(r+1)} = & \sum_{t=1}^{k} \log(f_{tg}(x_{it}; \boldsymbol{\gamma}_{gt})) - \sum_{t=1}^{k} \log \Big\{ \phi(\Phi^{-1}(F(x_{it}\,; \gamma_{gt})) \Big\} \; + \\
& \log \Big\{ \phi_k \Big(\Phi^{-1}(F(x_{i1}\,; \gamma_{g1})), \cdots, \Phi^{-1}(F(x_{ik}\,; \gamma_{gk})); \mathbf{0}, R_{11g}^{(r)} \Big) \Big\} \; + \\
& \log \Big\{ \sum_{\boldsymbol{d}} sgn(\boldsymbol{d}) \; \Phi_\ell \Big(\Phi^{-1}(F(x_{i(k+1)}\,; \gamma_{g(k+1)})), \cdots, \Phi^{-1}(F(x_{i(k+\ell)}\,; \gamma_{g(k+\ell)}))\,; \boldsymbol{\mu_g}, \boldsymbol{\Sigma_g^{(r)}} \Big) \Big\}
\end{aligned} \tag{3.13}$$

where

$\boldsymbol{\mu_g} = R_{21g}^{(r)} (R_{11g}^{(r)})^{-1} \Big[\Phi^{-1}(F(\boldsymbol{x}_1\,; \gamma_{g1})), \cdots, \Phi^{-1}(F(\boldsymbol{x}_k\,; \gamma_{gk})) \Big]'$ and $\boldsymbol{\Sigma_g^{(r)}} = R_{22g}^{(r)} - R_{21g}^{(r)} (R_{11g}^{(r)})^{-1} R_{12g}^{(r)}$.

CM-step 2: Maximize for all $g = 1, ..., G$

$$\sum_{i=1}^{n} \Big\{ w_{ig} \ell_{ig2} \Big\}$$

with respect to the matrix R_{11g}, given the updated values $\boldsymbol{\gamma^{(r+1)}}$, where

$$\ell_{ig2}^{(r+1)} = \log\left\{\phi_k\Big(\Phi^{-1}(F(x_{i1}\,;\gamma_{g1}^{(r+1)})),\cdots,\Phi^{-1}(F(x_{ik}\,;\gamma_{gk}^{(r+1)}));\mathbf{0},R_{11g}\Big)\right\} + \log\left\{\sum_{\boldsymbol{d}} sgn(\boldsymbol{d})\;\boldsymbol{\Phi_\ell}\Big(\Phi^{-1}(F(x_{i(k+1)}\,;\gamma_{g(k+1)}^{(r+1)})),\cdots,\Phi^{-1}(F(x_{i(k+\ell)}\,;\gamma_{g(k+\ell)}^{(r+1)}))\,;\boldsymbol{\mu_g},\boldsymbol{\Sigma_g}\Big)\right\}, \tag{3.14}$$

where

$\boldsymbol{\mu_g} = R_{21g}^{(r)}R_{11g}^{-1}\left[\Phi^{-1}(F(\boldsymbol{x}_1\,;\gamma_{g1}^{(r+1)})),\cdots,\Phi^{-1}(F(\boldsymbol{x}_k\,;\gamma_{gk}^{(r+1)}))\right]'$ and $\boldsymbol{\Sigma_g} = R_{22g}^{(r)} - R_{21g}^{(r)}R_{11g}^{-1}R_{12g}^{(r)}$.

CM-step 3: Maximize for all $g = 1, ..., G$

$$\sum_{i=1}^{n}\Big\{w_{ig}\ell_{ig3}\Big\}$$

with respect to the matrix $\boldsymbol{R_{12g}}$ and it's transpose $\boldsymbol{R_{21g}}$, given the updated values $\boldsymbol{\gamma^{(r+1)}}$ and $\boldsymbol{R_{11g}^{(r+1)}}$ from previous steps, where

$$\ell_{ig3}^{(r+1)} = \log\left\{\phi_k\Big(\Phi^{-1}(F(x_{i1}\,;\gamma_{g1}^{(r+1)})),\cdots,\Phi^{-1}(F(x_{ik}\,;\gamma_{gk}^{(r+1)}));\mathbf{0},\boldsymbol{R}_{11g}^{(r+1)}\Big)\right\} + \log\left\{\sum_{\boldsymbol{d}} sgn(\boldsymbol{d})\;\Phi_\ell\Big(\Phi^{-1}(F(x_{i(k+1)}\,;\gamma_{g(k+1)}^{(r+1)})),\cdots,\Phi^{-1}(F(x_{i(k+\ell)}\,;\gamma_{g(k+\ell)}^{(r+1)}))\,;\boldsymbol{\mu_g},\boldsymbol{\Sigma_g}\Big)\right\} \tag{3.15}$$

where

$\boldsymbol{\mu_g} = R_{21g}(R_{11g}^{(r+1)})^{-1}\left[\Phi^{-1}(F(\boldsymbol{x}_1\,;\gamma_{g1}^{(r+1)})),\cdots,\Phi^{-1}(F(\boldsymbol{x}_k\,;\gamma_{gk}^{(r+1)}))\right]'$ and $\boldsymbol{\Sigma_g} = R_{22g}^{(r)} - R_{21g}(R_{11g}^{(r+1)})^{-1}R_{12g}$.

CM-step 4: Maximize

$$\sum_{i=1}^{n}\Big\{w_{ig}\ell_{ig4}\Big\}$$

with respect to the matrix $\boldsymbol{R_{22}}$, given the updated values $\boldsymbol{\gamma^{(r+1)}}$ and $\boldsymbol{R_{11g}^{(r+1)}}$,$\boldsymbol{R_{12g}^{(r+1)}}$ from previous steps, where

$$\ell_{ig4}^{(r+1)} = \log\left\{ \sum_{\boldsymbol{d}} sgn(\boldsymbol{d})\ \Phi_\ell\Big(\Phi^{-1}(F(x_{i(k+1)}\,;\gamma_{g(k+1)}^{(r+1)})), \cdots, \Phi^{-1}(F(x_{i(k+\ell)}\,;\gamma_{g(k+\ell)}^{(r+1)}))\,;\boldsymbol{\mu_g^{(r+1)}}, \boldsymbol{\Sigma_g}\Big)\right\} \tag{3.16}$$

where

$\boldsymbol{\mu_g^{(r+1)}}$ as described in previous steps and $\boldsymbol{\Sigma_g} = R_{22g} - R_{21g}^{(r+1)}(R_{11g}^{(r+1)})^{-1}R_{12g}^{(r+1)}$.

That means that the maximization step is executed in 4 steps to evaluate the parameters γ_{gt} associated with the t-th, $(t = 1, ..., k)$ marginal distribution of continuous variable, for g,$(g = 1, ..., G)$ cluster, and the correlation matrices R_g split into blocks for all components $g = 1, ..., G$. The maximization steps for the G components are independent to each other so these can be performed in parallel in order to reduce computational time. The algorithm continues to iterate until a stopping criterion is met. In our case that is the difference in log-likelihood between the (r) and (r+1) iteration should be less than 10^{-12}.

3.3.2 Starting values

One approach to get starting values of the γ_{gt}, $t = 1, ..., (k + \ell)$ parameters for the g-clustering component for all marginals distributions, for the $(k+\ell)\times(k+l-1)/2$ correlations between variables and the $G-1$ in total π_g probabilities of clusters, is to use the independence framework. This means that the initial values of all $\rho_{ij} = 0$. This of-course is biased because we started with the assumption that we want to model the dependencies between distributions and EM is doomed to converge slow. Another approach is to use an alternative EM algorithm which estimates all γ's for the marginals and the correlation matrix for each component is estimated from the sampling correlation matrix of quantities $\Psi(u_{tg}) = \Phi^{-1}(F_t(x_{tg}; \gamma_g))$ which by the definition of Gaussian copula, follow a multivariate normal distribution with $\boldsymbol{\mu} = \mathbf{0}$ and correlation matrix $\boldsymbol{R_g}$. Those estimators are closer to the real values and so we can get adequate starting values.

3.3.3 Model Selection

Once we have decided the extracted from Expectation-Maximization algorithm parameters associated with the copula, arises the problem of choosing the proper number of components G. The best model is the one with the lowest value of BIC (Bayesian Information Criterion) which is typical approach for a family of models running for a range of values of G. The definition of this criterion is:

$$BIC = -2\ell(\hat{\theta}) + \rho \log(n)$$

where $\hat{\theta}$ is the maximum likelihood estimate of vector θ,$\ell(\hat{\theta})$ is the maximized likelihood, ρ is the number of free parameters in the model and n the sample size. As the component number grows, the number of free parameters also grows and BIC gets higher values because it penalizes these extra parameters.

3.4 Towards Parsimonious models

Model-based clustering is becoming increasing popular tool for clustering purposes mainly due to its probabilistic foundations and its flexibility. However, in the big data era, high-dimensional data are nowadays more and more frequent and, unfortunately, classical model-based clustering techniques needs to improve to cope with high-dimensional spaces. This is mainly due to the fact that model-based clustering methods are over-parametrized in this case. However, high-dimensional spaces have specific characteristics which are useful for clustering and recent techniques exploit those characteristics. Parsimonious alternatives have been proposed based on dimension reduction approaches like factor analysis (see, e.g. McLachlan et al. (2003)) or clever representations of the correlation matrices (see, e.g. McNicholas and Murphy (2008)). For a broad review see Bouveyron and Brunet-Saumard (2014). In this section we propose two such approaches for reducing the number of parameters of the correlation matrix in the Gaussian copula to model the correlation structure.

3.4.1 Correlation matrix decompositions

The Gaussian copula for modelling the joint probability function is an appropriate choice for mixed mode data because it offers flexibility on the choice of marginals but also takes account all the internal dependencies which are included to the copula with the correlation matrix R. For the fully parametrized mixture model,

described in previous sections, it is obvious that the added components in model based clustering and the added variables lead to computational complexity. This drawback is intense because of the correlation matrix. For example for a set of 10 variables and $G = 4$ clusters we need to estimate $(10 \times 9/2) \times 4 = 45 \times 4 = 180$ along with $G - 1 = 3$ probabilities of clusters and the parameters associated with the marginals. To avoid this in literature we meet a proposed model where all components may have the same unconstrained corelation matrix which is called the homogeneous model. This of course can add limitations to the chosen distributions and the shape of clusters. Here we propose 2 correlation matrices decompositions in order to achieve parsimony in computations. The first approach involves factor analysis as data reduction technique and the second structural correlation matrices customized for mixed data.

3.4.1.1 Factor analysis decomposer

Factor analysis is a data reduction technique that replaces the observed variables by latent factors with smaller dimension. This method works well when the latent factors explain a satisfactory amount of the variability of the observed variables. This approach can offer along with dimension reduction, interpretable results. Consider n independent $k + \ell$-dimensional random variables $\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_n$. The factor analysis model is written in the form

$$\boldsymbol{X}_i = \boldsymbol{\mu} + \boldsymbol{\Lambda}\, \boldsymbol{U}_i + \boldsymbol{\epsilon}_i$$

for $i = 1, 2, ..., n$, where Λ is a $(k + \ell) \times q$, $q < k + \ell$ matrix of factor loadings, the latent factor $\boldsymbol{U}_i \sim N(\boldsymbol{0}, \boldsymbol{I})$, and $\boldsymbol{\epsilon}_i \sim N(\boldsymbol{0}, \boldsymbol{\Psi})$,where $\boldsymbol{\Psi}$ diagonal matrix with the called communalities. From definition the marginal distribution of X_i's under the factor analysis model $\sim N(\boldsymbol{\mu}, \boldsymbol{\Lambda\Lambda'} + \boldsymbol{\Psi})$. Under this definition the factor analysis is appropriate for decomposing the correlation matrix of the fully parametrized mixture model of Gaussian copulas for mixed mode data since it includes high dependencies between the variables and the correlation matrix R_g for each component $g = 1, ..., G$ can be written in the form $R = \boldsymbol{\Lambda\Lambda'} + \boldsymbol{\Psi}$. So there is the need of finding a $(k + \ell) \times q$ matrix Λ with λ_{iu}, $i = 1, ..., k + \ell$, $u = 1, ..., q$ elements so as

$$\rho_{ij} = \sum_{u=1}^{q} \lambda_{iu}\lambda_{ju}, \quad i \neq j$$

and $\rho_{ii} = 1, \quad i = j$. With this approach the problem of estimating $(k+\ell) \times (k+\ell-1)/2$ correlations reduces in estimating $(k+\ell) \times q$ elements λ_{iu}, $q << (k+\ell)$. There is no need of estimating the diagonal matrix of communalities Ψ since the diagonal elements of R are always units. The choice of the number of factors is an important consideration in factor analysis. One approach is to choose the number of factors that captures a certain proportion of the variation of the data. In our case we run the factor analysis model for a variety of number of factors. The best model is revealed from the lowest value of BIC.

The optimization process is performed with the use of CEM algorithm similarly to the fully parametrized model. Here, we can make use of the fact that for a number of p factors the correlation matrix can be written in the form $R = \Lambda_1\Lambda_1^T + \Psi_1 + \cdots + \Lambda_p\Lambda_p^T + \Psi_p$ and so the logarithmic likelihood can be maximized separately for the parameters of the marginal distributions and for each factor. The correlation matrix R_g for every component g $g = 1, ..., G$ can also split into blocks $R_{11g}, R_{12g}, R_{22g}$ as for full model.

E-step: Calculate for $i = 1, ..., n$ and $g = 1, ..., G$

$$w_{ig}^{(r+1)} = \frac{\pi_g^{(r)} f_g(x_i; \gamma_g^{(r)}, \psi_g^{(r)})}{\sum_{g=1}^{G} \pi_g^{(r)} f_g(x_i; \gamma_g^{(r)}, \psi_g^{(r)})}$$

CM-step 1: Update $\pi_g^{(r+1)} = \sum_{i=1}^{n} w_{ig}^{(r+1)}/n$ for all $g = 1, ..., G$ and then maximize

$$\sum_{i=1}^{n} \left\{ w_{ig} \ell_{ig1} \right\}$$

with respect to the γ_{gt} to get updated values for $\gamma_{gt}^{(r+1)}$ parameters associated with the t-th marginal distribution, $t = 1, ..., k+\ell$, and the g-th component $g = 1, ..., G$. As in equation 3.11 for the case of mixed mode data with k continuous and ℓ discrete marginals ℓ_{ig1} is written in the following form

$$\ell_{ig1}^{(r+1)} = \sum_{t=1}^{k} \log(f_{tg}(x_{it};\boldsymbol{\gamma}_{gt})) - \sum_{t=1}^{k} \log\Big\{\ \phi(\Phi^{-1}(F(x_{it}\,;\gamma_{gt}))\Big\}\ + \\ \log\Big\{\phi_k\Big(\Phi^{-1}(F(x_{i1}\,;\gamma_{g1})),\cdots,\Phi^{-1}(F(x_{ik}\,;\gamma_{gk}));\mathbf{0},\boldsymbol{R}_{11g}^{(r)}\Big)\Big\}\ + \\ \log\left\{\sum_{\boldsymbol{d}} sgn(\boldsymbol{d})\ \Phi_\ell\Big(\Phi^{-1}(F(x_{i(k+1)}\,;\gamma_{g(k+1)})),\cdots,\Phi^{-1}(F(x_{i(k+\ell)}\,;\gamma_{g(k+\ell)}))\,;\boldsymbol{\mu_g},\boldsymbol{\Sigma_g^{(r)}}\Big)\right\} \tag{3.17}$$

where

$\boldsymbol{\mu_g} = R_{21g}^{(r)}(R_{11g}^{(r)})^{-1}\left[\Phi^{-1}(F(\boldsymbol{x}_1\,;\gamma_{g1})),\cdots,\Phi^{-1}(F(\boldsymbol{x}_k\,;\gamma_{gk}))\right]'$ and $\boldsymbol{\Sigma_g^{(r)}} = R_{22g}^{(r)} - R_{21g}^{(r)}(R_{11g}^{(r)})^{-1}R_{12g}^{(r)}$.

CM-step 2 j: Maximize the log-likelihood for the j-th factor $j = 1, ..., p$ and for every component g, $g = 1, ..., G$

$$\sum_{i=1}^{n}\Big\{w_{ig}\ell_{igj}\Big\}$$

with respect to the matrix $\boldsymbol{R_{gj}}$, given the updated values $\boldsymbol{\gamma^{(r+1)}}$ and $\boldsymbol{R_{gt}}$ from previous steps, so as

$$R_g = R_{gj} + \sum_{t\neq j} R_{gt}$$

where

$$\ell_{igj}^{(r+1)} = \log\Big\{\phi_k\Big(\Phi^{-1}(F(x_{i1}\,;\gamma_{g1}^{(r+1)})),\cdots,\Phi^{-1}(F(x_{ik}\,;\gamma_{gk}^{(r+1)}));\mathbf{0},\boldsymbol{R}_{11gj}\Big)\Big\}\ + \\ \log\left\{\sum_{\boldsymbol{d}} sgn(\boldsymbol{d})\ \Phi_\ell\Big(\Phi^{-1}(F(x_{i(k+1)}\,;\gamma_{g(k+1)}^{(r+1)})),\cdots,\Phi^{-1}(F(x_{i(k+\ell)}\,;\gamma_{g(k+\ell)}^{(r+1)}))\,;\boldsymbol{\mu_{gj}},\boldsymbol{\Sigma_{gj}}\Big)\right\} \tag{3.18}$$

and

$\boldsymbol{\mu_{gj}} = R_{21gj}(R_{11gj})^{-1}\left[\Phi^{-1}(F(\boldsymbol{x}_1\,;\gamma_{g1}^{(r+1)})),\cdots,\Phi^{-1}(F(\boldsymbol{x}_k\,;\gamma_{gk}^{(r+1)}))\right]'$ and $\boldsymbol{\Sigma_{gj}} = R_{22gj} - R_{21gj}(R_{11gj})^{-1}R_{12gj}$.

3.4.1.2 Structured correlation matrices

Let R denote a symmetric positive definite $(k+\ell)$ correlation matrix corresponding to a random vector X. The Choleski factor L is an upper triangular matrix such

as $R = LL'$ and with respect to the angles reparametrization matrix Θ we have an one to one relation between the correlation matrix R and Θ through L such:

$$L_{11} = 1, \quad L_{ii} = \prod_{u=1}^{i-1} \sin\theta_{ui}, \quad L_{ij} = \cos\theta_{ij} \prod_{u=1}^{i-1} \sin\theta_{uj}, \quad (i, j = 1, ..., k+\ell, i < j) \tag{3.19}$$

where the angles are measured in radians. We require $\theta_{ij} \in (0, \pi]$ so that the Choleski factor is unique. Generally the Choleski factor for a given covariance matrix is not unique, but it is unique for the case of correlation matrices where the diagonal entries are units. The entries of the correlation matrix are related to the angles:

$$\rho_{ij} = \sum_{u=1}^{i} L_{ui}L_{uj}, \qquad \theta_{ij} = cos^{-1}\left\{\frac{L_{ij}}{\prod_{u=1}^{i-1}\sin\theta_{uj}}\right\}, \qquad i < j, \qquad \theta_{ij} = 0, \qquad i \geq j \tag{3.20}$$

Because of the angles reparametrization we can assume structures of correlation matrix where we produce Θ in a way that the R matrix retains the properties of symmetry and positive definition. Here, for the case of mixed mode data we can assume at east 2 blocks of correlations; one for the continuous marginal and one for the discrete, and the maximum number of blocks is equal to the different entries of the unstructured correlation matrix. For the data set described in section 3 we rank in descend order the variables based on the observed correlation matrix $\hat{R}$ separately for continuous and discrete variables so as to achieve the optimal groups of variables with equal correlations.

The optimization process is performed with the use of CEM algorithm similarly to the fully parametrized model and the factor analyzers. The correlation matrix R_g for every component g $g = 1, ..., G$ can also split into blocks $R_{11g}, R_{12g}, R_{22g}$ as for full model, though every ρ_{ij} element of the correlation matrix can be written through the angles re-parametrization as a result of equations 3.19 and 3.20.

E-step: Calculate for $i = 1, ..., n$ and $g = 1, ..., G$

$$w_{ig}^{(r+1)} = \frac{\pi_g^{(r)} f_g(x_i; \gamma_g^{(r)}, \psi_g^{(r)})}{\sum_{g=1}^{G} \pi_g^{(r)} f_g(x_i; \gamma_g^{(r)}, \psi_g^{(r)})}$$

CM-step 1: Update $\pi_g^{(r+1)} = \sum_{i=1}^{n} w_{ig}^{(r+1)}/n$ for all $g = 1, ..., G$ and then maximize

$$\sum_{i=1}^{n} \Big\{ w_{ig} \ell_{ig1} \Big\}$$

with respect to the γ_{gt} to get updated values for $\gamma_{gt}^{(r+1)}$ parameters associated with the t-th marginal distribution, $t = 1, ..., k+\ell$, and the g-th component $g = 1, ..., G$. Similar to the case of mixed mode data with k continuous and ℓ discrete marginals ℓ_{ig1} is written in the following form

$$\begin{aligned}
\ell_{ig1}^{(r+1)} = & \sum_{t=1}^{k} \log(f_{tg}(x_{it}; \boldsymbol{\gamma}_{gt})) - \sum_{t=1}^{k} \log \Big\{ \phi(\Phi^{-1}(F(x_{it}\,; \gamma_{gt})) \Big\} \; + \\
& \log \Big\{ \phi_k \Big(\Phi^{-1}(F(x_{i1}\,; \gamma_{g1})), \cdots, \Phi^{-1}(F(x_{ik}\,; \gamma_{gk})); \mathbf{0}, \boldsymbol{R}_{11g}^{(r)} \Big) \Big\} \; + \\
& \log \Bigg\{ \sum_{\boldsymbol{d}} sgn(\boldsymbol{d}) \; \Phi_\ell \Big(\Phi^{-1}(F(x_{i(k+1)}\,; \gamma_{g(k+1)})), \cdots, \Phi^{-1}(F(x_{i(k+\ell)}\,; \gamma_{g(k+\ell)})) \,; \boldsymbol{\mu_g}, \boldsymbol{\Sigma_g^{(r)}} \Big) \Bigg\}
\end{aligned} \tag{3.21}$$

where

$\boldsymbol{\mu_g} = R_{21g}^{(r)} (R_{11g}^{(r)})^{-1} \Big[\Phi^{-1}(F(\boldsymbol{x}_1\,; \gamma_{g1})), \cdots, \Phi^{-1}(F(\boldsymbol{x}_k\,; \gamma_{gk})) \Big]'$ and $\boldsymbol{\Sigma_g^{(r)}} = R_{22g}^{(r)} - R_{21g}^{(r)} (R_{11g}^{(r)})^{-1} R_{12g}^{(r)}$ and $\rho_{ij}, \;\; i, j = 1, ..., k+\ell$ as a result of equations 3.19 and 3.20.

CM-step 2 : Maximize the log-likelihood for every component g, $g = 1, ..., G$

$$\sum_{i=1}^{n} \Big\{ w_{ig} \ell_{ig2} \Big\}$$

with respect to the matrix $\boldsymbol{R_g}$ to get updated values for $\boldsymbol{\theta}$, given the updated values $\boldsymbol{\gamma^{(r+1)}}$

$$\begin{aligned}
\ell_{ig2}^{(r+1)} = & \log \Big\{ \phi_k \Big(\Phi^{-1}(F(x_{i1}\,; \gamma_{g1}^{(r+1)})), \cdots, \Phi^{-1}(F(x_{ik}\,; \gamma_{gk}^{(r+1)})); \mathbf{0}, \boldsymbol{R}_{11g} \Big) \Big\} \; + \\
& \log \Bigg\{ \sum_{\boldsymbol{d}} sgn(\boldsymbol{d}) \; \Phi_\ell \Big(\Phi^{-1}(F(x_{i(k+1)}\,; \gamma_{g(k+1)}^{(r+1)})), \cdots, \Phi^{-1}(F(x_{i(k+\ell)}\,; \gamma_{g(k+\ell)}^{(r+1)})) \,; \boldsymbol{\mu_g}, \boldsymbol{\Sigma_g} \Big) \Bigg\},
\end{aligned}$$

$\boldsymbol{\mu_g} = R_{21g}(R_{11g})^{-1}\left[\Phi^{-1}(F(\boldsymbol{x}_1\,;\gamma_{g1}^{(r+1)})),\cdots,\Phi^{-1}(F(\boldsymbol{x}_k\,;\gamma_{gk}^{(r+1)}))\right]'$ and $\boldsymbol{\Sigma_g} = R_{22g} - R_{21g}(R_{11g})^{-1}R_{12g}$,

$$\rho_{ij} = \sum_{u=1}^{i} L_{ui}L_{uj}$$

where for $i, j = 1, ..., k + \ell$.

$$L_{11} = 1, \quad L_{ii} = \prod_{u=1}^{i-1} \sin\theta_{ui}$$

$$L_{ij} = \cos\theta_{ij} \prod_{u=1}^{i-1} \sin\theta_{uj}, \quad i < j.$$

3.4.1.3 Structured correlation matrices and relation with factor decomposer

The above structures of correlation matrices imply analogous results to factor loadings. From the definition of the relation between ρ_{ij} and λ_{iu}, if $\rho_{ij} = \rho_{ir}$ then $\lambda_{ju} = \lambda_{ru}$, for every factor $u = 1, ..., q$. For the structure where we consider different correlations only for continuous and discrete variables, every factor j, $j = 1, ..., p$ element is obtained by letting $\lambda_{jt} = \lambda_{jk}$, $t = 1, ..., k$, and $\lambda_{jt} = \lambda_{j\ell}$, $t = 1, ..., \ell$. This can be done due to uniqueness of factor loadings and Choleski decomosition. The matrix of factors loadings of size $(k + \ell) \times p$ has the form:

$$\Lambda_1 = \left(\begin{array}{cccc} \lambda_{k1} & \lambda_{k2} & \cdots & \lambda_{kp} \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_{k1} & \lambda_{k2} & \cdots & \lambda_{kp} \\ \hdashline \lambda_{\ell 1} & \lambda_{\ell 2} & \cdots & \lambda_{\ell p} \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_{\ell 1} & \lambda_{\ell 2} & \cdots & \lambda_{\ell p} \end{array}\right)$$

and the parameters that we need to estimate are of size $2 \times p$.

Following the same rationale, for any structure of correlation matrix , every group of correlations creates a block to the matrix of factor loadings. Here for group sizes $N_d = \{n_1, n_2, ..., n_d\}$, $n_i \geq 1$ is

$$\Lambda_2 = \begin{pmatrix} \lambda_{n_{11}} & \lambda_{n_{12}} & \cdots & \lambda_{n_{1q}} \\ \vdots & \vdots & \vdots & \vdots \\ \hline \lambda_{n_{1i}} & \lambda_{n_{i2}} & \cdots & \lambda_{n_{iq}} \\ \vdots & \vdots & \vdots & \vdots \\ \hline \lambda_{n_{d1}} & \lambda_{n_{d2}} & \cdots & \lambda_{n_{dq}} \\ \vdots & \vdots & \vdots & \vdots \\ \lambda_{n_{d1}} & \lambda_{n_{d2}} & \cdots & \lambda_{n_{dq}} \end{pmatrix}$$

where $\lambda_{n_{ji}}$, $j = 1, ..., p$, $i = 1, ..., d$ refers to the entry for j-th factor and for i-block of parameters. Here we have to estimate $d \times p$ parameters in total.

3.4.2 Penalized Mixtures of Copulas

Let R denote a symmetric positive definite $(k+\ell)$ correlation matrix corresponding to a random vector X. The Choleski factor L as described in section 3.4.1.2 is an upper triangular matrix such as $R = LL'$ and with respect to the angles reparametrization matrix Θ we have an one to one relation between the correlation matrix R and Θ through L such:

$$L_{11} = 1, \quad L_{ii} = \prod_{u=1}^{i-1} \sin\theta_{ui}, \quad L_{ij} = \cos\theta_{ij} \prod_{u=1}^{i-1} \sin\theta_{uj}, \quad (i, j = 1, ..., k+\ell, i < j) \tag{3.22}$$

where the angles are measured in radians. We require $\theta_{ij} \in (0, \pi]$ so that the Choleski factor is unique. The entries of the correlation matrix are related to the angles:

$$\rho_{ij} = \sum_{u=1}^{i} L_{ui} L_{uj}, \; \theta_{ij} = cos^{-1}\left\{ \frac{L_{ij}}{\prod_{u=1}^{i-1} \sin\theta_{uj}} \right\}, \; i < j, \; \theta_{ij} = 0, \; i \geq j \tag{3.23}$$

Because of the angles reparametrization we can assume structures of correlation matrix where we produce Θ in a way that the R matrix retains the properties of symmetry and positive definition. Here, for the case of mixed mode data we can assume at east 2 blocks of correlations; one for the continuous marginal and one for the discrete, and the maximum number of blocks is equal to the different entries of the unstructured correlation matrix. For any data set we change the order of the variables so as for continuous variables to appear first. No other ordering is required for the case of penalized algorithm.

Following the methodology of Structured Correlation Matrices and based on the fact that unstructured model is equivalent to the independent model where all $\theta_i = \pi/2$ we propose to add a penalty to the unconditional log-likelihood such as:

$$\begin{aligned}\ell = &\sum_{i=1}^{n}\sum_{g=1}^{G}\log\left\{f(\boldsymbol{x_i};\boldsymbol{\gamma_g},\boldsymbol{\Psi_g})\right\} = \sum_{i=1}^{n}\sum_{g=1}^{G}\sum_{t=1}^{k}\log(f_t(\boldsymbol{x_i};\gamma_{tg}))\\ &-\sum_{i=1}^{n}\sum_{g=1}^{G}\sum_{t=1}^{k}\log\left\{\phi(\Phi^{-1}(u_{tg}))\right\} \;+\sum_{i=1}^{n}\sum_{g=1}^{G}\log\left\{\phi_k\Big(\Psi(u_{1g}),\cdots,\Psi(u_{kg})\,;\mathbf{0},\boldsymbol{R_{11g}}\Big)\right\}\\ &+\sum_{i=1}^{n}\sum_{g=1}^{G}\log\left\{\sum_{\boldsymbol{d}} sgn(\boldsymbol{d})\,\Phi_\ell\Big(\Psi(u_{(k+1)g}),\cdots,\Psi(u_{(k+l)g})\,;\,\boldsymbol{\mu_g},\boldsymbol{\Sigma_g}\Big)\right\}\\ &+\lambda\sum_{g=1}^{G}\sum_{j=1}^{p(p-1)/2}\sin^2\theta_{gj}\end{aligned} \tag{3.24}$$

where all correlation matrices R are written through the correlation matrix re-parametrization matrix $\boldsymbol{\Theta}$ as defined from equations 3.22 and 3.23, and λ takes values into a grid ($\lambda > 0$). For $\lambda = 0$ the approach is equivalent to the full model evaluation without any constrain to the correlation matrix, while λ inf the log-likelihood shrinks to the independent model where all θ's are equal to $pi/2$.

3.4.2.1 ECM algorithm

The optimization process is performed with the use of CEM algorithm similarly to the fully parametrized model and the factor analyzers. The correlation matrix R_g for every component g $g = 1, ..., G$ can also split into blocks $R_{11g}, R_{12g}, R_{22g}$ as for full model, though every ρ_{ij} element of the correlation matrix can be written through the angles re-parametrization as a result of equations 3.19 and 3.20.

E-step: Calculate for $i = 1, ..., n$ and $g = 1, ..., G$

$$w_{ig}^{(r+1)} = \frac{\pi_g^{(r)} f_g(x_i;\gamma_g^{(r)},\psi_g^{(r)})}{\sum_{g=1}^{G}\pi_g^{(r)} f_g(x_i;\gamma_g^{(r)},\psi_g^{(r)})}$$

CM-step 1: Update $\pi_g^{(r+1)} = \sum_{i=1}^{n} w_{ig}^{(r+1)}/n$ for all $g = 1, ..., G$ and then maximize

$$\sum_{i=1}^{n}\Big\{w_{ig}\ell_{ig1}\Big\}$$

with respect to the γ_{gt} to get updated values for $\gamma_{gt}^{(r+1)}$ parameters associated with the t-th marginal distribution, $t = 1, ..., k+\ell$, and the g-th component $g = 1, ..., G$. Similar to the case of mixed mode data with k continuous and ℓ discrete marginals ℓ_{ig1} is written in the following form

$$\begin{aligned}\ell_{ig1}^{(r+1)} = &\sum_{t=1}^{k}\log(f_{tg}(x_{it};\boldsymbol{\gamma}_{gt})) - \sum_{t=1}^{k}\log\Big\{\,\phi(\Phi^{-1}(F(x_{it}\,;\gamma_{gt}))\Big\} \;+\\ &\log\Big\{\phi_k\Big(\Phi^{-1}(F(x_{i1}\,;\gamma_{g1})),\cdots,\Phi^{-1}(F(x_{ik}\,;\gamma_{gk}));\mathbf{0},\boldsymbol{R}_{11g}^{(r)}\Big)\Big\} \;+\\ &\log\left\{\sum_{\boldsymbol{d}} sgn(\boldsymbol{d})\;\Phi_\ell\Big(\Phi^{-1}(F(x_{i(k+1)}\,;\gamma_{g(k+1)})),\cdots,\Phi^{-1}(F(x_{i(k+\ell)}\,;\gamma_{g(k+\ell)}))\,;\boldsymbol{\mu_g},\boldsymbol{\Sigma_g^{(r)}}\Big)\right\}\end{aligned} \tag{3.25}$$

where

$\boldsymbol{\mu_g} = R_{21g}^{(r)}(R_{11g}^{(r)})^{-1}\left[\Phi^{-1}(F(\boldsymbol{x}_1\,;\gamma_{g1})),\cdots,\Phi^{-1}(F(\boldsymbol{x}_k\,;\gamma_{gk}))\right]'$ and $\boldsymbol{\Sigma_g^{(r)}} = R_{22g}^{(r)} - R_{21g}^{(r)}(R_{11g}^{(r)})^{-1}R_{12g}^{(r)}$ and $\rho_{ij},\ \ i,j = 1, ..., k+\ell$ as a result of equations 3.19 and 3.20.

CM-step 2 : Maximize the log-likelihood for every component g, $g = 1, ..., G$

$$\sum_{i=1}^{n}\Big\{w_{ig}\ell_{ig2}\Big\}$$

with respect to the matrix $\boldsymbol{R_g}$ to get updated values for $\boldsymbol{\theta}$, given the updated values $\boldsymbol{\gamma^{(r+1)}}$

$$\begin{aligned}\ell_{ig2}^{(r+1)} = &\log\Big\{\phi_k\Big(\Phi^{-1}(F(x_{i1}\,;\gamma_{g1}^{(r+1)})),\cdots,\Phi^{-1}(F(x_{ik}\,;\gamma_{gk}^{(r+1)}));\mathbf{0},\boldsymbol{R}_{11g}\Big)\Big\}\\ &+\log\left\{\sum_{\boldsymbol{d}} sgn(\boldsymbol{d})\;\Phi_\ell\Big(\Phi^{-1}(F(x_{i(k+1)}\,;\gamma_{g(k+1)}^{(r+1)})),\cdots,\Phi^{-1}(F(x_{i(k+\ell)}\,;\gamma_{g(k+\ell)}^{(r+1)}))\,;\boldsymbol{\mu_g},\boldsymbol{\Sigma_g}\Big)\right\}\\ &+\lambda\sum_{j=1}^{p(p-1)/2}\sin^2\theta_{gj},\end{aligned}$$

$\boldsymbol{\mu_g} = R_{21g}(R_{11g})^{-1}\left[\Phi^{-1}(F(\boldsymbol{x}_1\,;\gamma_{g1}^{(r+1)})),\cdots,\Phi^{-1}(F(\boldsymbol{x}_k\,;\gamma_{gk}^{(r+1)}))\right]'$ and $\boldsymbol{\Sigma_g} = R_{22g} - R_{21g}(R_{11g})^{-1}R_{12g}$,

$$\rho_{ij} = \sum_{u=1}^{i} L_{ui}L_{uj}$$

where for $i, j = 1, ..., k + \ell$.

$$L_{11} = 1, \quad L_{ii} = \prod_{u=1}^{i-1} \sin\theta_{ui}$$

$$L_{ij} = \cos\theta_{ij} \prod_{u=1}^{i-1} \sin\theta_{uj}, \quad i < j.$$

The λ factor is chosen in a grid $(0, \text{inf})$ and the algorithm is performed for all values of λ. For $\lambda = 0$ we simply estimate the parameters for the full model with unstructured correlation matrix. The estimation for the next value of λ is performed using as initial values the ones resulted from the previous step. In this way the computational effort is reduced.

3.4.2.2 Model Selection

The ECM algorithm produces the penalized log likelihood for every λ value into the chosen grid. Once we have decided the extracted from Expectation-Maximization algorithm parameters associated with the copula, arises the problem of choosing the proper number of components G and the proper value of λ. The best model is the one with the lowest value of BIC (Bayesian Information Criterion) which is typical approach for a family of models running for a range of values of G. The definition of this criterion is:

$$BIC = -2\ell(\hat{\theta}) + \rho(\lambda)\log(n) \tag{3.26}$$

where $\hat{\theta}$ is the maximum likelihood estimator of vector θ,$\ell(\hat{\theta})$ is the maximized log likelihood, ρ is the number of free parameters in the model related to the number of components and n the sample size.

Among models of the same λ value, as the component number grows, the number of free parameters also grows and BIC gets higher values because it penalizes these extra parameters.

Among models for different values of λ value in the chosen grid the number of free parameters also changes. Following Tingjin Chu and Wang (2011), as $\lambda \to \infty$ since the correlation matrix becomes more sparse and allows entries to shrink towards $\pi/2$. Therefore, the number of correlations such as $|\rho_{ij}| < 5 \cdot 10^{-3}$ or $|\theta_{ij} - \pi/2| < 5 \cdot 10^{-3}$ are omitted from calculation of BIC. The resulted λ is revealing the optimal structured correlation matrix which yields better fit to the data.

3.5 Simulation Study

In this section, we present the results of a simulation study that we conducted to illustrate the effectiveness of our clustering methodology. Via simulation, we compared the performance of the Penalized Gaussian Copulas method for mixed mode data for various λ penalties.

3.5.1 Data Sample Description

Let's assume a 4-variate mixed mode dataset of length $n = 200$ resulting from $G = 2$ components, where consists of 2 continuous variables and 2 discrete variables. In more detail let's assume for $G = 1$ the:

$$X_{11} \sim Normal(\mu_{11} = 10, \sigma_{11} = 2)$$

$$X_{12} \sim Gamma(\alpha_{12} = 10, \gamma_{12} = 2)$$

$$X_{13} \sim Poisson(\lambda_{13} = 7)$$

$$X_{14} \sim Bernoulli(p_{14} = 0.8)$$

which are correlated through a correlation matrix R_1 such as shown below:

$$\boldsymbol{R_1} = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{12} & 1 & \rho_{23} & \rho_{24} \\ \rho_{13} & \rho_{23} & 1 & \rho_{34} \\ \rho_{14} & \rho_{24} & \rho_{34} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.6 & 0.6 & 0.05 \\ 0.6 & 1 & 0.05 & 0.4 \\ 0.6 & 0.05 & 1 & 0.3 \\ 0.05 & 0.4 & 0.3 & 1 \end{bmatrix}$$

and for $G = 2$

$$X_{21} \sim Normal(\mu_{21} = 1, \sigma_{21} = 2)$$

$$X_{22} \sim Gamma(\alpha_{22} = 10, \gamma_{22} = 2)$$

$$X_{23} \sim Poisson(\lambda_{23} = 10)$$

$$X_{24} \sim Bernoulli(p_{24} = 0.5)$$

which are correlated through a correlation matrix R_2 such as shown below:

$$\boldsymbol{R_2} = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{12} & 1 & \rho_{23} & \rho_{24} \\ \rho_{13} & \rho_{23} & 1 & \rho_{34} \\ \rho_{14} & \rho_{24} & \rho_{34} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.4 & 0.4 & 0.05 \\ 0.4 & 1 & 0.05 & 0.5 \\ 0.5 & 0.05 & 1 & 0.1 \\ 0.05 & 0.4 & 0.1 & 1 \end{bmatrix}$$

The sample data X have been produced through the R package Copula and through the Gaussian Copula with the respective correlations and parameters of each of the components. We choose the data from component 1 with a probability $\pi = 0.7$, while from component 2 with a probability $\pi = 0.3$.

We perform 60 iterations of the simulated dataset and for each of the produced datasets we perform Conditional Expectation-Maximization algorithm to obtain the estimated parameters of each component for distributions and correlation matrices with respect to the λ penalty, as described in section 3.3.2. We choose λ in a sequence of values in $[0, 5000)$ with a step of 10. This means that for every λ and for the same dataset and component we produce the estimated values which are then used as starting values for the sequential value of λ. Convergence criteria for each ECM step has been set to 10^{-10}.

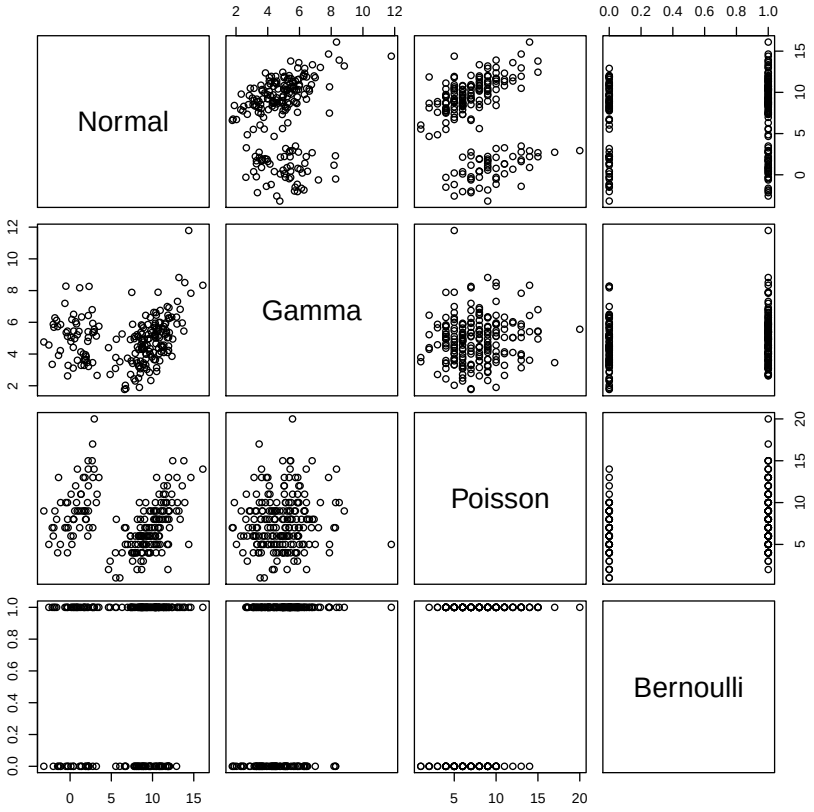


FIGURE 3.1: Simulated Data Illustration

3.5.2 Results

For the 60 iterations of the simulation study we calculate the respective BIC values as defined in section 3.4.2.2 through equation 3.26 with the constraints arizen from the different values of λ. For the simulated data and for the various λ's, table 3.1 provides a summary of the average BIC values of all ierations, while table 3.2 provides a summary of the chosen number of components through BIC. We can observe that for growing values of the λ in the grid the minimum value of Bayesian Information criterion selects higher number of clusters, since the number of free parameters has been decreased and the correlation matrix becomes sparse.

Model	K=2	K=3
$\lambda = 0$	**2856.46**	2868.34
$\lambda = 10$	**2859.94**	2877.58
$\lambda = 50$	**2911.58**	2924.45
$\lambda = 100$	**2976.66**	2983.74
$\lambda = 200$	**3025.49**	3025.95
$\lambda = 500$	3052.28	**3040.95**
$\lambda = 1000$	3054.61	**3039.28**
$\lambda = 5000$	3043.43	**3028.10**

TABLE 3.1: Average BIC value for various values of λ.

The lowest value of BIC for any number of clusters and any value of λ, is achieved for the full model evaluation, where $\lambda = 0$, which means that all data variables are correlated. This is expected due to the structure of the simulated dataset, which assumes a highly correlated sample which can be seen from correlation matrices R_1 & R_2 in section 3.5.1.

Model	K=2	K=3
$\lambda = 0$	55	5
$\lambda = 10$	55	5
$\lambda = 50$	54	6
$\lambda = 100$	43	17
$\lambda = 200$	33	27
$\lambda = 500$	9	51
$\lambda = 1000$	6	54
$\lambda = 5000$	5	55

TABLE 3.2: Results from fitting the different models-Estimated number of clusters selected over 60 iterations.

From the above tables, a change point in the chosen number of clusters is observed at around $\lambda = 500$, where at this point we start to observe some convergence to the estimated BIC values.

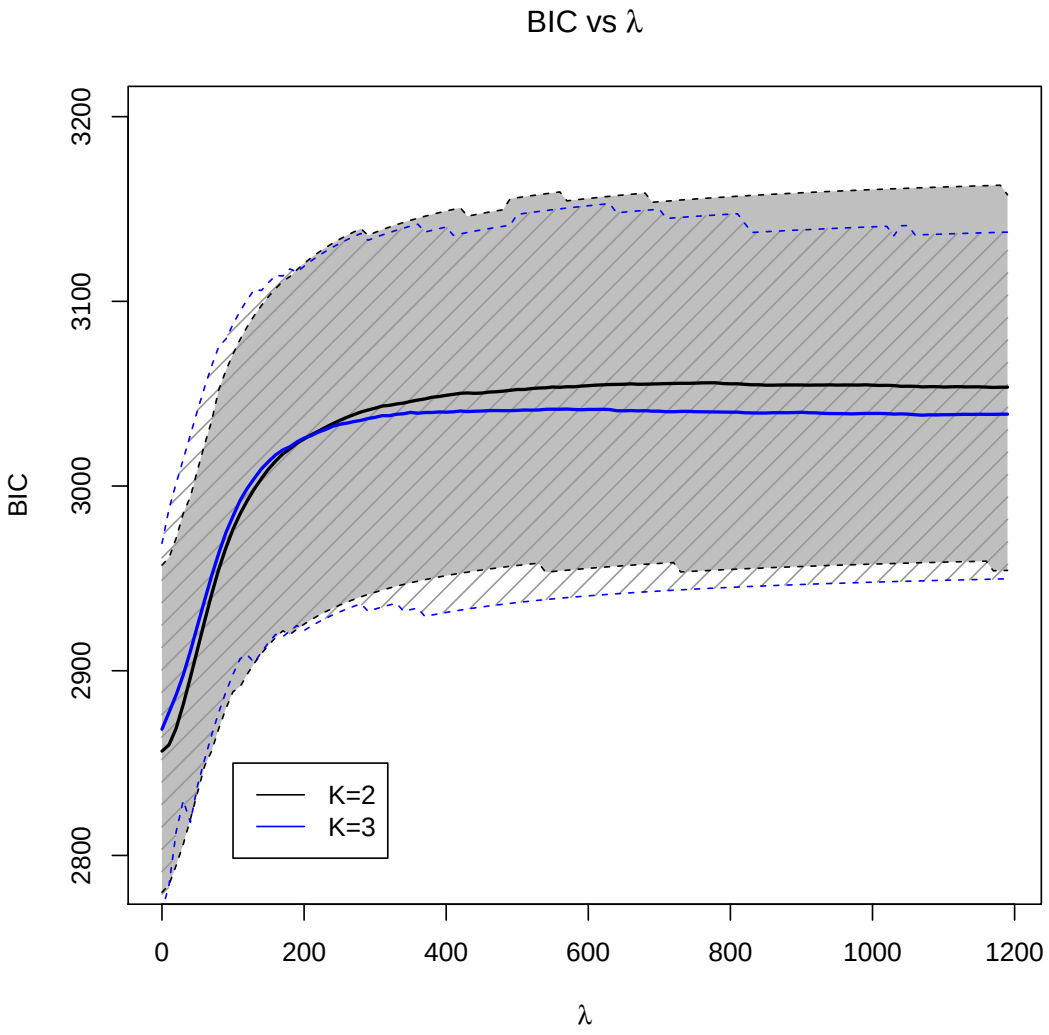


FIGURE 3.2: BIC values for the 60 iterations and for the various values of λ penalty

Figure 3.2 illustrates the confidence intervals and the average values of estimated BIC for the different number of components. For small values, and more specifically $\lambda \leq 210$, the clusters selected are $G = 2$, while this number increases as the λ increases. This change point is achieved for a large value of the penalized parameter compared to other penalized models, since the penalized factor $\sin^2 \theta$ takes values in the interval (0,1), therefore, low penalty values apply slight impact to the penalized likelihood.

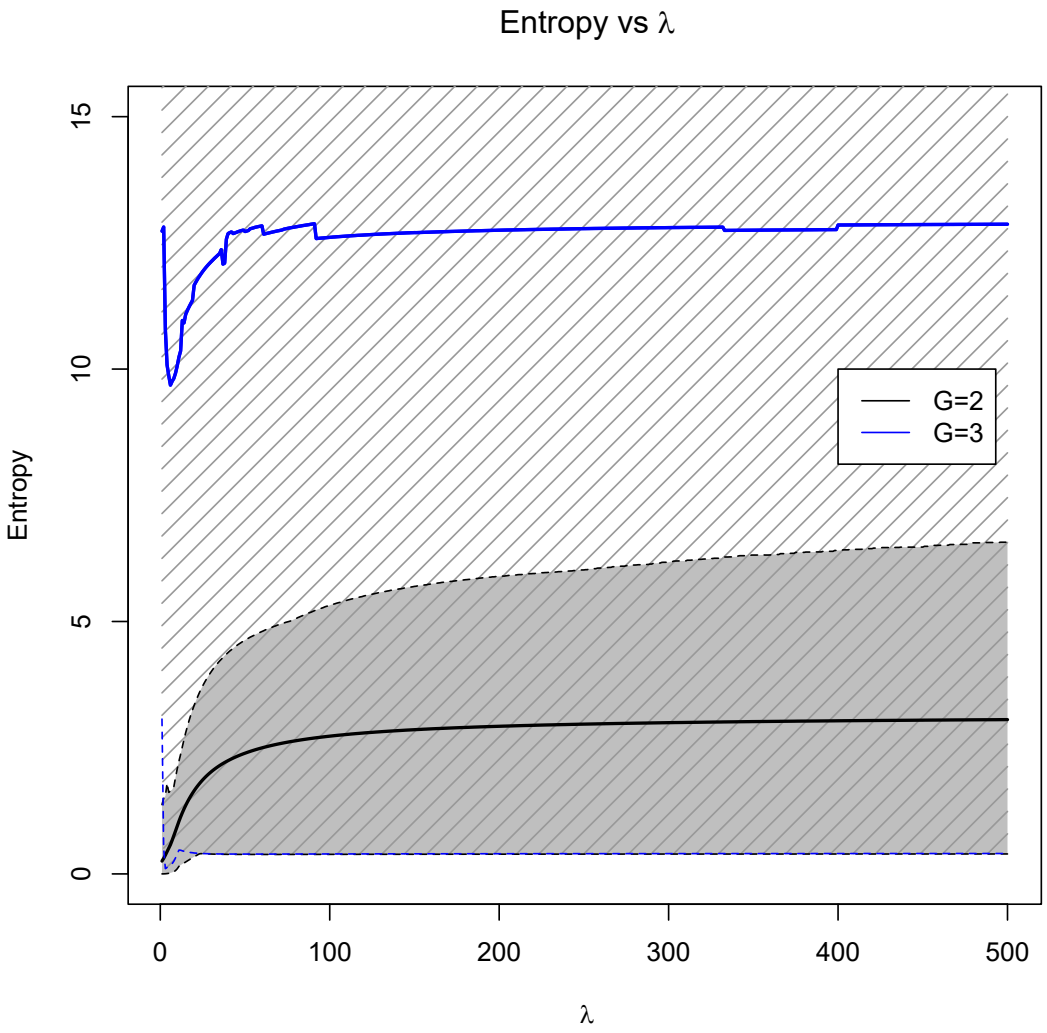


FIGURE 3.3: Entropy values for the 60 iterations and for the various values of λ penalty

For the estimated values of $G = 2$ clusters and $G = 3$ clusters we calculate an estimation of the fitted entropy of the mixture model for various values of the penalty factor λ. The estimation is performed through the below equation:

$$E(\lambda) = -\sum_{i=1}^{n}\sum_{g=1}^{G} w_{ig}(\lambda)\log w_{ig}(\lambda) \tag{3.27}$$

where w_{ig} are the mixing weights of the mixture model for each iteration of λ . From figure 3.3 we observe similar results such as in the case of BIC in terms of convergence for the various penalty values. For $G = 2$ we achieve a lower value of entropy which is expected because of the lower number of clusters. As penalty

values increase we do not have significant loss of clustering efficiency since the entropy converges to some value.

Table 3.3 illustrates the structure of the correlation matrix for various values of λ for a selected number of $G = 2$ components. We can observe that for large values of the penalty the correlation matrix is close to the identity matrix.

$$R_0 = \begin{bmatrix} 1 & 0.588 & 0.600 & 0.049 \\ & 1 & 0.042 & 0.407 \\ & & 1 & 0.307 \\ & & & 1 \end{bmatrix} \qquad R_{100} = \begin{bmatrix} 1 & 0.317 & 0.312 & -0.006 \\ & 1 & -0.039 & 0.132 \\ & & 1 & 0.091 \\ & & & 1 \end{bmatrix}$$

$\lambda = 0, \;\; \hat{\pi} = 0.7$ $\qquad$ $\lambda = 100, \;\; \hat{\pi} = 0.7$

$$R_{1000} = \begin{bmatrix} 1 & 0.039 & 0.037 & 0.001 \\ & 1 & 0.001 & 0.016 \\ & & 1 & 0.011 \\ & & & 1 \end{bmatrix} \qquad R_{5000} = \begin{bmatrix} 1 & 0.008 & 0.007 & 0.000 \\ & 1 & 0.000 & 0.003 \\ & & 1 & 0.002 \\ & & & 1 \end{bmatrix}$$

$\lambda = 1000, \;\; \hat{\pi} = 0.7$ $\qquad$ $\lambda = 5000, \;\; \hat{\pi} = 0.7$

$$R_0 = \begin{bmatrix} 1 & -0.022 & 0.408 & 0.035 \\ & 1 & -0.002 & -0.001 \\ & & 1 & 0.088 \\ & & & 1 \end{bmatrix} \qquad R_{100} = \begin{bmatrix} 1 & -0.006 & 0.090 & 0.005 \\ & 1 & 0.001 & -0.001 \\ & & 1 & 0.013 \\ & & & 1 \end{bmatrix}$$

$\lambda = 0, \;\; \hat{\pi} = 0.3$ $\qquad$ $\lambda = 100, \;\; \hat{\pi} = 0.3$

$$R_{1000} = \begin{bmatrix} 1 & -0.001 & 0.010 & 0.000 \\ & 1 & 0.000 & 0.000 \\ & & 1 & 0.001 \\ & & & 1 \end{bmatrix} \qquad R_{5000} = \begin{bmatrix} 1 & 0.000 & 0.002 & 0.000 \\ & 1 & 0.000 & 0.000 \\ & & 1 & 0.000 \\ & & & 1 \end{bmatrix}$$

$\lambda = 1000, \;\; \hat{\pi} = 0.3$ $\qquad$ $\lambda = 5000, \;\; \hat{\pi} = 0.3$

TABLE 3.3: Correlation matrix for various values of λ and for both components

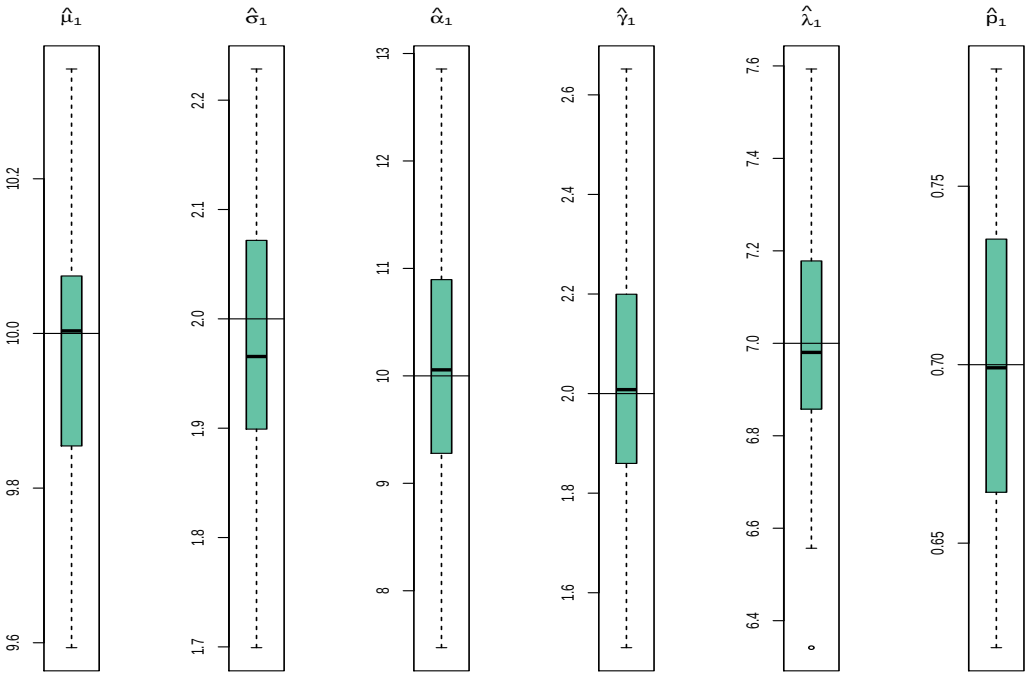


FIGURE 3.4: Boxplots for the estimated parameters of G=1 for the 60 iterations for the selected full model

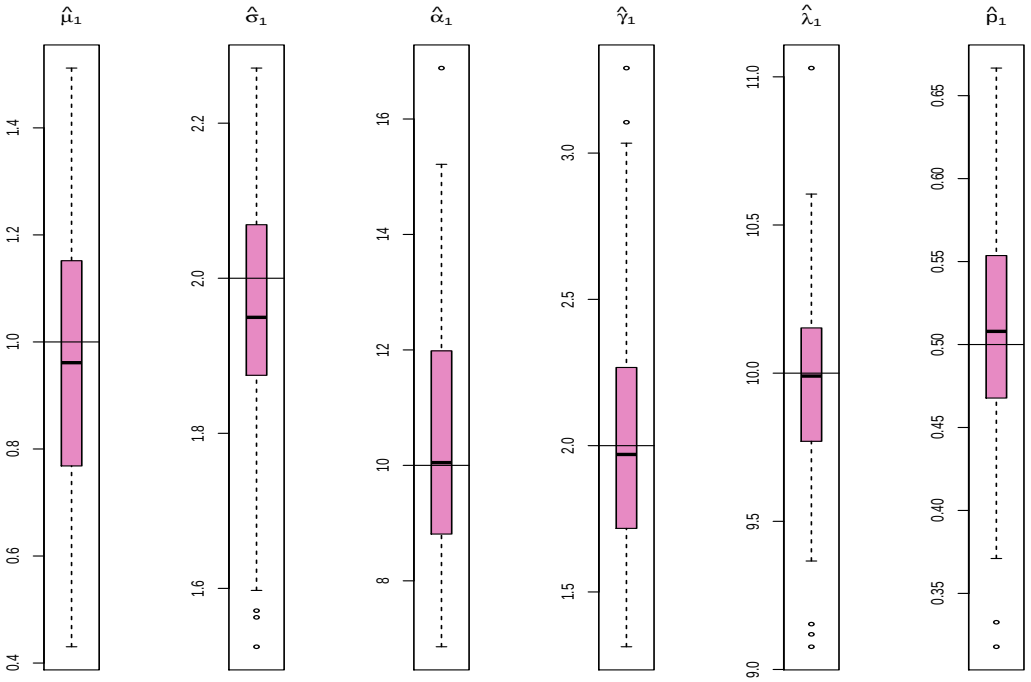


FIGURE 3.5: Boxplots for the estimated parameters of G=2 for the 60 iterations for the selected full model

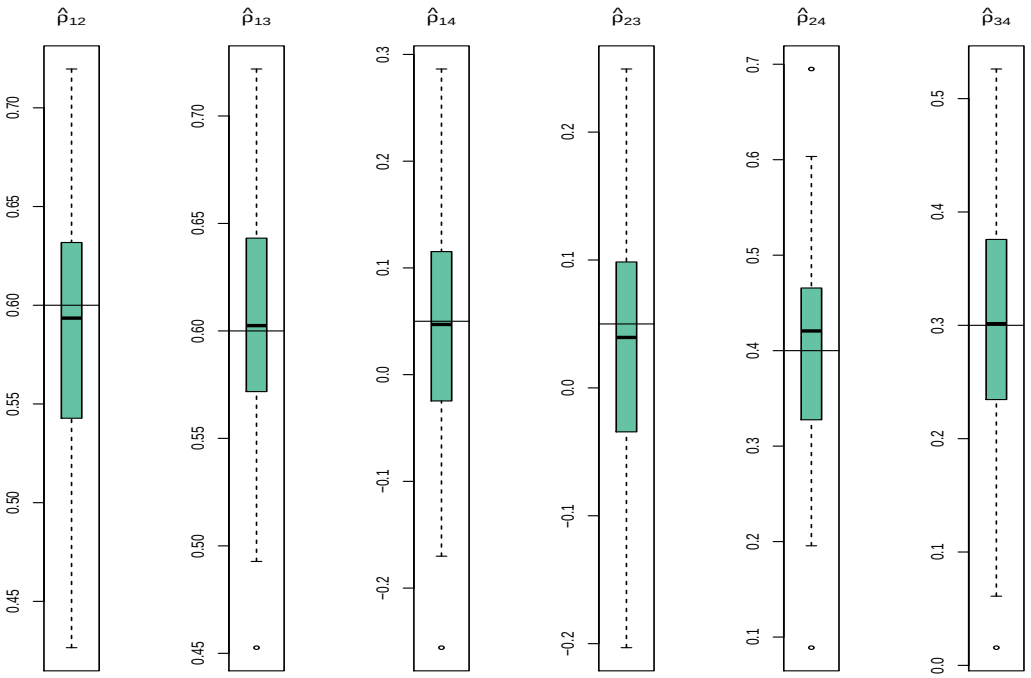


FIGURE 3.6: Boxplots for the estimated correlations of G=1 for the 60 iterations for the selected full model

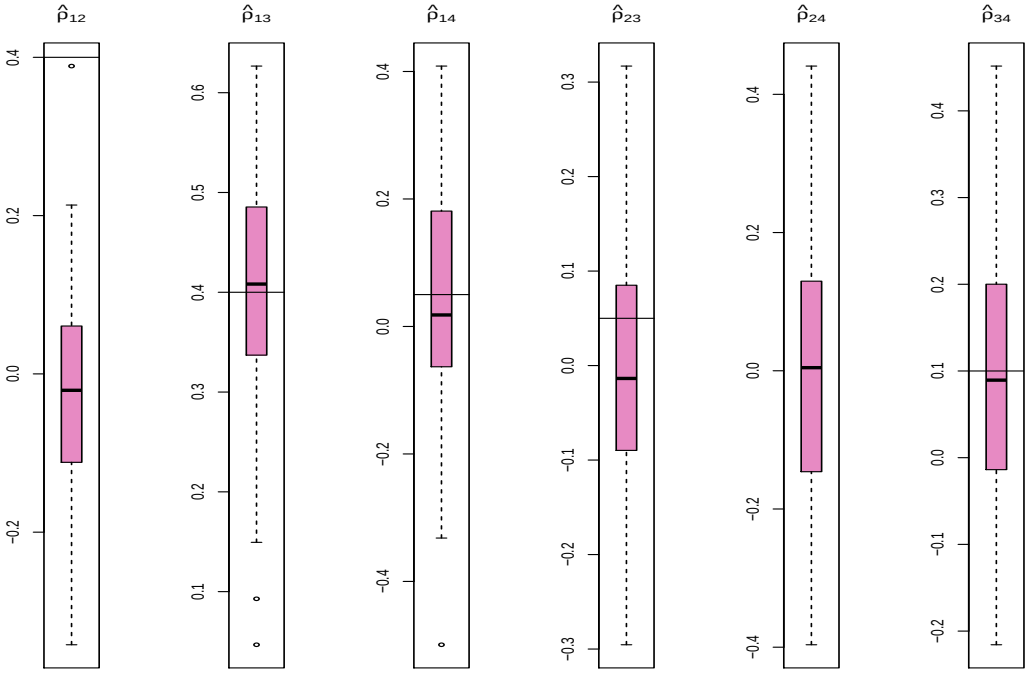


FIGURE 3.7: Boxplots for the estimated correlations of G=2 for the 60 iterations for the selected full model

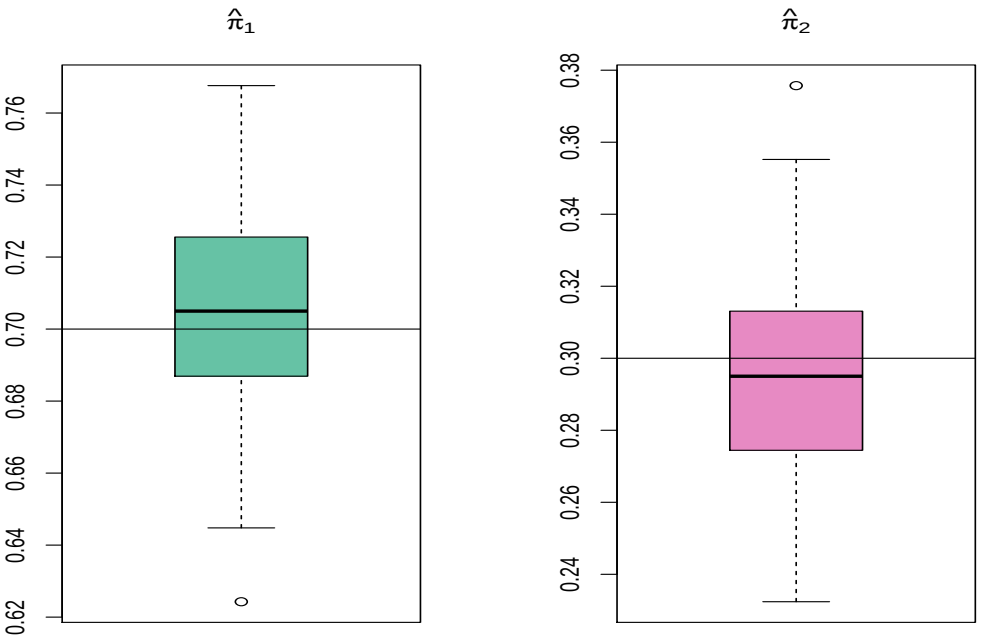


FIGURE 3.8: Boxplots for the estimated correlations of G=2 for the 60 iterations for the selected full model

For the selected model of G=2 clusters and $\lambda = 0$ figures 3.4 -3.8 provides boxplots of the estimated parameters for all γ_g related to the marginals of each cluster and for all ρ_g related to the correlations of the marginal distributions, over the 60 iterations. We observe more precarious results for the correlations of the second component, which is mainly due to the fact that the mixing probability is 0.3. For a larger dataset than n=200 we would expect better estimations. Another reason could be that the simulated datasets comes from the same distribution for both clusters when it comes to the Gamma marginal distribution. This fact might have forced the estimated correlations ρ_{12} and ρ_{24} of the second component to be close to 0.

Figure 3.9 provides some of the bi-variate contours of the mixture model for the combinations of continuous-continuous variables and continuous-discrete case, for several values of the penalty factor. We observe the change in the shape of the mixture probabilities as $\lambda \to \infty$, the shape becomes smoother as well as the density of the mixed clusters.

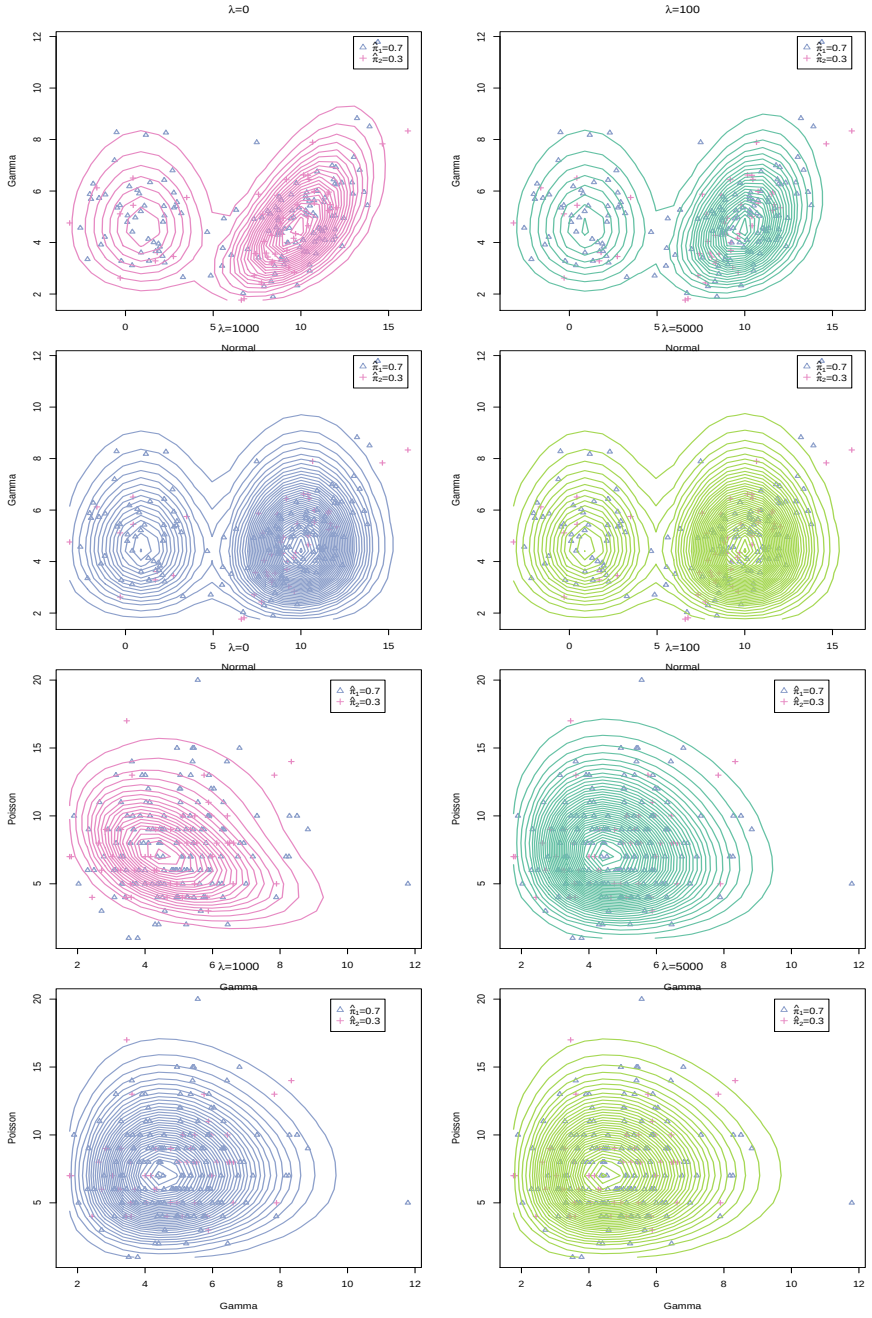


FIGURE 3.9: Contours of indicative bi-variate distributions for various values of λ.

3.6 Application Study

3.6.1 Data description

Data are composed of 517 forest fires recorded in north-east Portugal. Cortez and Morais (2007). Fires are described by 10 meteorological variables based on the Canadian Forest Fire Weather Index (FWI) System. Some of them are continuous while some others are binary. We will apply a model based clustering for mixed data based on the Gaussian copula based model described previously. To achieve parsimony the models described will be applied. Note that, to show the flexibility of the copula approach we assume different marginals for the continuous variables. The available variables are:

- **Fine Fuel Moisture Code (FFMC)**: The Fine Fuel Moisture Code (FFMC) is a numeric rating of the moisture content of litter and other cured fine fuels. This code is an indicator of the relative ease of ignition and the flammability of fine fuel.It is a continuous variable, which takes positive values. The distribution of FFMC based on recordings is usually right skewed so we assume a Weibull distribution to describe the data.

- **Duff Moisture Code (DMC)**: The Duff Moisture Code (DMC) is a numeric rating of the average moisture content of loosely compacted organic layers of moderate depth. This code gives an indication of fuel consumption in moderate duff layers and medium-size woody material. Also a Weibull distribution will be fitted for description.

- **Relative Humidity (RH)**: Relative humidity in %. Also the family of Weibull distributions is used.

- **Wind**: Wind speed in km/h. Continuous variable which takes positive values. Here a Weibull distribution is used fit the data.

- **Drought code (DC)**: The Drought Code (DC) is a numeric rating of the average moisture content of deep, compact organic layers. This code is a useful indicator of seasonal drought effects on forest fuels and the amount of smoldering in deep duff layers and large logs. We assume a normal distribution.

- **Initial Spread Index (ISI)**: The Initial Spread Index (ISI) is a numeric rating of the expected rate of fire spread. It combines the effects of wind

and the FFMC on rate of spread without the influence of variable quantities of fuel. We assume a normal distribution.

- **Temperature (temp)**: Temperature in Celsius degrees. We assume a normal distribution.
- **Summer index (season)**: Binary variable which takes value equal to 1 if the season was the summer. Here, we choose a Bernoulli distribution to fit to the data values.
- **Rain index (rain)**: Binary variable which takes value 1 if there was outside rainfall in the last 24 hours. It can be modelled as Bernoulli trials.
- **Weekend index (day)**: Binary variable which takes value 1 if the fire burst into flames at the weekend. Rationally, we assume a Bernoulli distribution.

The same data set has been used in Marbac et al. (2017) but assuming only normal distributions for the continuous variables. We have selected Weibull marginal distributions for some variables to better describe their shape, this is rather simple for copula defined models. We will apply our parsimonious models and a full likelihood approach for estimating the parameters.

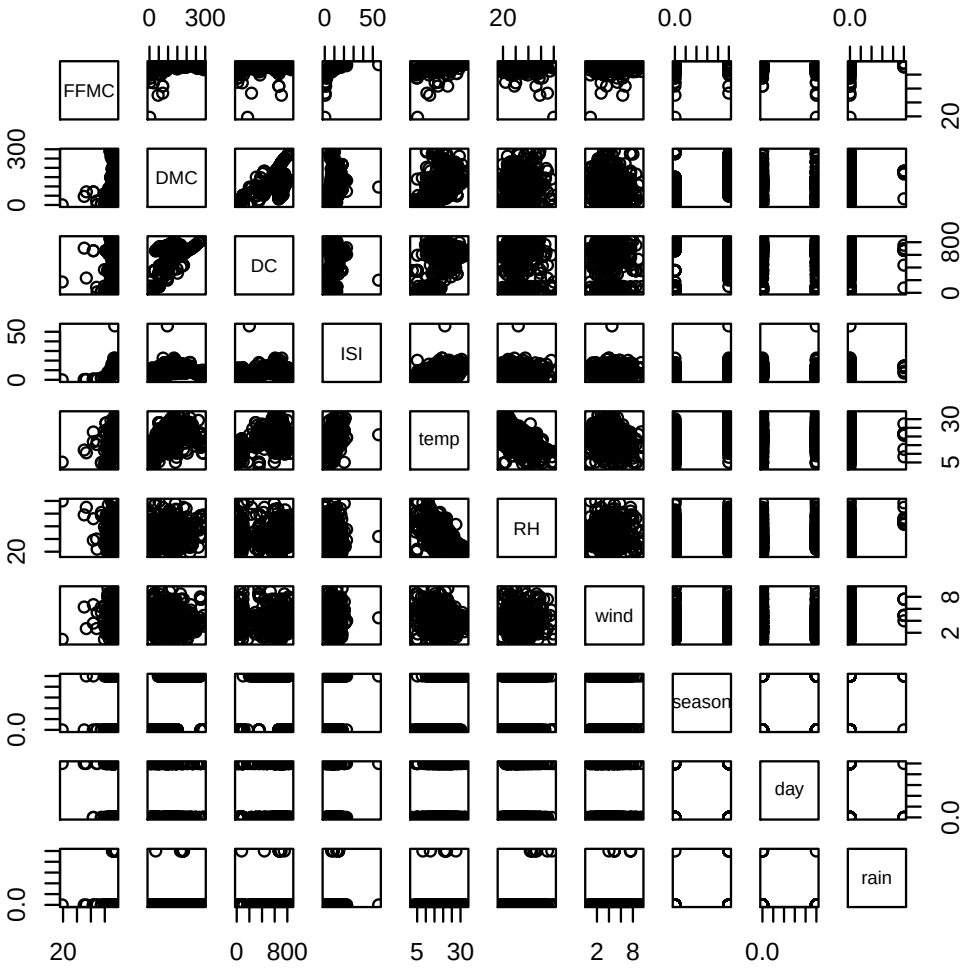


FIGURE 3.10: Data visualization

3.6.2 Results

We have fitted a series of different models to the data to achieve parsimony. Table 3.6 presents results from several different models. The models fitted to the data are: a) the independent model where the clusters are considered as conditionally independent, this is a typical model used so far as it is very simple and does not need to consider complicated multivariate models b) the model with all clusters having a full correlation matrix which is different for each cluster, this implies the larger structure and the larger number of parameters to estimate, c-e)factor models with 1 up to 3 factors, and f-h) the structured correlation models with different structure. The first structure just teats differently the continuous to the binary variables, hence blocks the continuous and the binary variables separately.

The second one creates some blocks within each category, while the 3rd structure creates a more refined structure at the cost of a large number of parameters. One can see that the model have a certain amount of parsimony. We have fitted up to 6 clusters. For each model we report the log-likelihood and the BIC, the BIC can be also seen in Figure 3.11. The number of parameters for each model can be also seen. All models are fitted using the algorithms described in the previous sections.

To start with the model that assumes conditional independence fails for all numbers of clusters. This implies that we need to model the correlation structure inside the clusters. The structured models while capture part of the correlation lead to an increasing number of parameters. The structure correlation models implying an increasing structure do not improve a lot and we see based on BIC that the 2-factors model has the best BIC.

Table 3.4 outlines the number of parameters to be estimated for the chosen set of marginal distributions and for the chosen structure of correlation matrix. As the number of factor components increases the number of free parameters also increases though the likelihood is close to the one of the fully parametrized model.

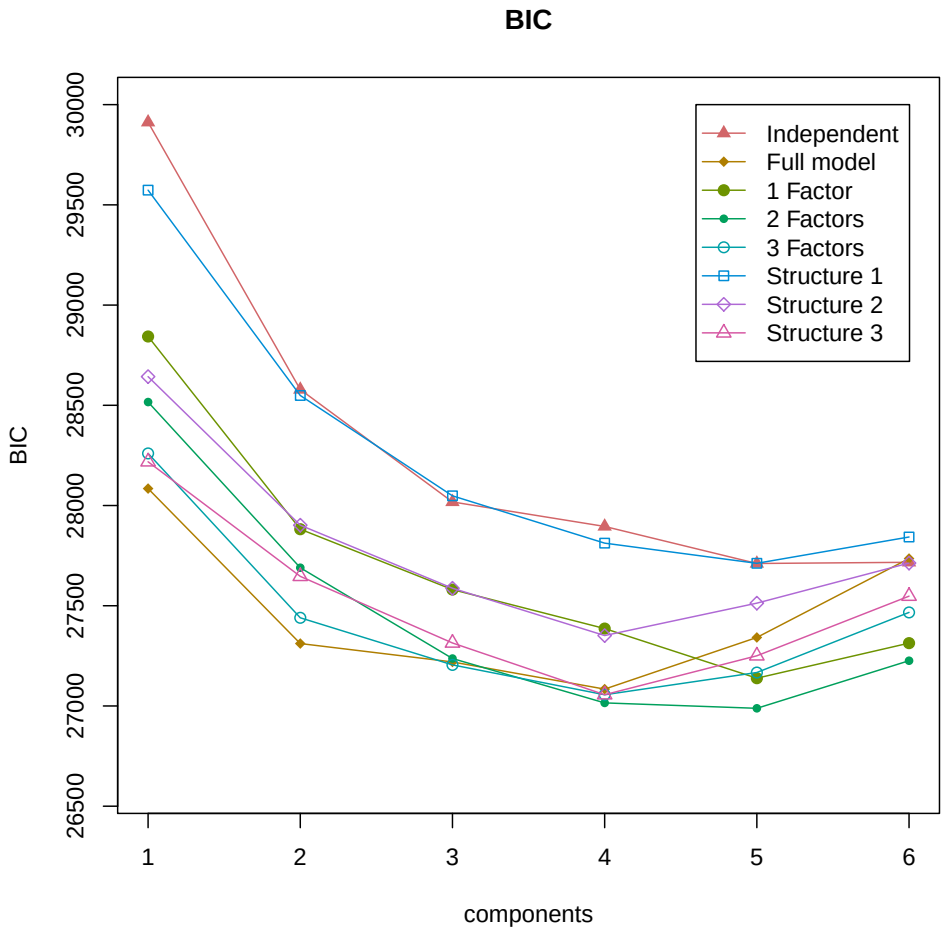


FIGURE 3.11: The BIC for the different models for increasing number of clusters.

Model	1	2	3	4	5	6
Independent (a)	17	35	53	71	89	107
Fully Parametrized Model(b)	62	125	188	251	314	377
1 Factor (c)	27	55	83	111	139	167
2 Factors (d)	37	75	113	151	189	227
3 Factors (e)	47	95	143	191	239	287
1^{st} Structure Ng= (7,3) (f)	20	41	62	83	104	125
2^{nd} Structure Ng=(3,2,2,1,2) (g)	31	63	95	127	159	191
3^{rd} Structure Ng= (1,1,1,2,1,1,1,2)(h)	47	95	143	191	239	287

TABLE 3.4: Results from fitting several different models-Number of estimated parameters.

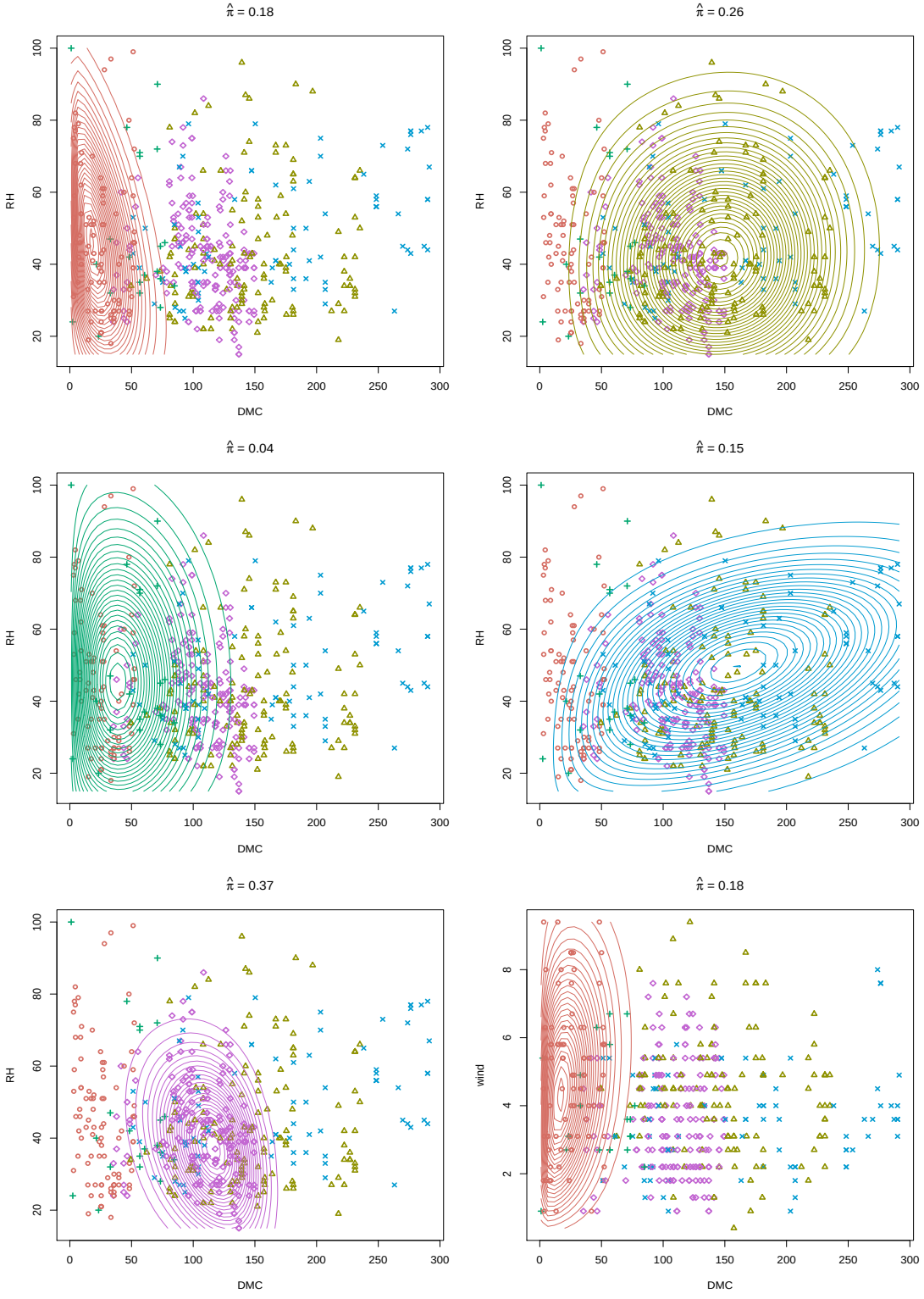
Table 3.5 provides the estimated log-likelihood values, while 3.6 the estimated BIC values. We observe that the fully parametrized model chooses as best $G = 4$ number of clusters, while as we allow the correlation matrix to be more structured as well as for the independent model, the number of chosen components increases to $G = 5$.

Model	**1**	**2**	**3**	**4**	**5**	**6**
a	-14902.9	-14180.1	-13843.4	-13726.0	-13596.1	-13596.1
b	-13848.5	-13265.2	-13022.4	-12757.9	-12689.9	-12689.9
c	-14337.3	-13769.1	-13531.0	-13346.3	-13135.0	-13135.0
d	-14142.5	-13610.5	-13265.3	-13035.2	-12903.7	-12903.7
e	-13983.2	-13423.2	-13155.7	-13931.3	-12836.8	-12836.8
f	-14724.2	-14146.4	-13830.5	-13646.8	-13530.9	-13530.9
g	-14224.7	-13754.1	-13496.9	-13278.9	-13259.7	-13259.7
h	-13962.3	-13526.5	-13210.3	-12931.3	-12878.55	-12877.3

TABLE 3.5: Results from fitting several different models-Log likelihood values.

Model	**1**	**2**	**3**	**4**	**5**	**6**
a	29912.1	28578.7	28017.9	27895.6	**27748.3**	27860.6
b	28084.5	27311.4	27219.6	**27084.0**	27341.8	27735.4
c	28843.3	27881.8	27580.6	27386.2	**27138.6**	27313.5
d	28516.3	27689.6	27236.6	27015.3	**26988.3**	27225.8
e	28260.0	27439.9	27204.9	**27055.9**	27166.8	27466.7
f	29573.4	28549.0	28048.4	27812.3	**27711.7**	27842.9
g	28643.0	27901.9	27587.3	**27351.4**	27512.9	27712.9
h	28218.2	27646.5	27314.2	**27056.1**	27520.3	27547.8

TABLE 3.6: Results from fitting several different models BIC values.



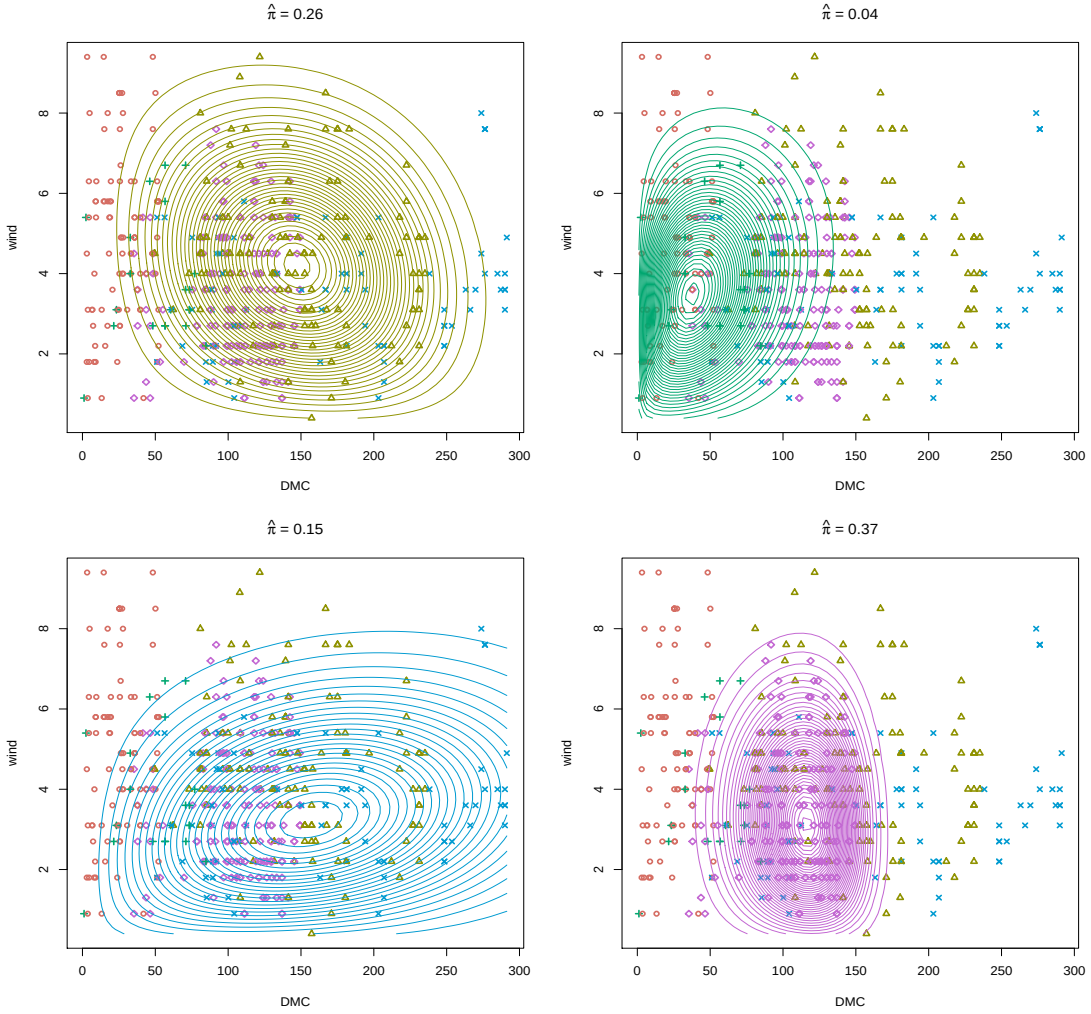


FIGURE 3.12: Contours of bi-variate distributions for the selected model with 2 Factors.

Figure 3.12 provides the bi-variate contours of the selected model of 2 factors for the chosen number of clusters $G = 5$ and for the continuous variables wind, DMC and RH.

3.7 Concluding Remarks

Sampling methods have been introduced to further reduce computational effort of the Composite Likelihood approach. For our Simulation studies purposes the dimension reduction was not extended nevertheless the time consumed reduction was important compared even compared to the full composite likelihood method. For cases of high dimensional count data, where the full Multivariate Poisson density function is not easy to be defined the composite likelihood concept offers

flexibility in calculations. We can further reduce the complexity of calculations via the sampling methods.

Among the Sampling methods the efficiency is much dependable on the sample data size, the highest the row size the more components or rows we can eliminate in calculations. For the specific simulation studies with $n = 200$ rows the Sampling method where we choose only one of the composite likelihood components provides poorer results, though for highest data points the method can further provide adequate results.

Alternative composite likelihood method which is less complex than the full traditional composite likelihood method can be also provide good classification without significant loss of miss-classified data points. This method can provide adequate values for the parameters and which can also be used as starting values of the full multivariate model estimation or the composite likelihood estimation. This method can be further investigated in order to provide adequate results.

In the present work, we focused also on the mixed data problem. The models and the derivations of parsimonious representations of the correlation matrix of the Gaussian copula are applicable to all models of model based clustering through copulas, like only continuous or only discrete random variables. The problem of parsimony in the model base clustering literature is an important one. Other kind of parsimonious representations like the representation in Celeux et al. (1995) can be an alternative approach. Since in copula based models we wok with the correlation rather the covariance matrix the approach needs to adjust.

Moreover, the model used here is related to the Gaussian copula factor model as described in Murray JS (2013). Another representation of factor models in the copula setting has been proposed in Krupskii and Joe (2013) which is perhaps more general. In the present work we have used the former representation and not the latter.

The approach based on the representation of Tsay and Pourahmadi (2017) leads to some interesting model selection problems in the sense that we would like to identify the best structure of the correlation matrices with some automatic approach. This problem is described and addressed through the simulation study and the penalized log likelihood.

Similar results can be applied to other copulas expressed through a correlation matrix like the t-copulas or some of the elliptical copulas.

Chapter 4

Concluding Remarks

The present thesis contributes towards the analysis of finite mixture models for model based clustering for count data and mixed data. In chapter 2, in order to overcome problems related to the problem of defining a model for count data in high dimensions, we opted to the use of composite likelihood methodology. Such an approach overcomes the problem of fully specifying the model in high dimensions but requires to define the marginal models in a lower dimension. We used pairwise approach in this thesis, namely by specifying only the bivariate marginals. This allows to work in higher dimensions. It is known that composite likelihood gains computational efficiently sacrificing statistical efficiency. In the context of model based clustering where the efficiency is not the main issue but rather the ability to identify clusters we examined the performance of the methodology. We made use of different approaches, an heuristic one that while used an alternative surrogate function it is still able to recognize the clustering.

Sampling methods have been introduced to further reduce computational effort of the Composite Likelihood approach. Such methods aims at reducing the computational pattern since the surrogate function used is not consisted of all possible pairs but fewer either by randomly selecting them or via some stratified sampling approach. Note the recent work on this randomized approach in Mazo et al. (2021). For our simulation studies purposes the dimension reduction was not extended nevertheless the time consumed reduction was important compared even compared to the full composite likelihood method. For cases of high dimensional count data, where the full Multivariate Poisson density function is not easy to be defined the composite likelihood concept offers flexibility in calculations. We can further reduce the complexity of calculations via the sampling methods.

Among the Sampling methods the efficiency is much dependable on the sample data size, the highest the row size the more components or rows we can eliminate in calculations. For the specific simulation studies with $n = 200$ rows the Sampling method where we choose only one of the composite likelihood components provides poorer results, though for highest data points the method can further provide adequate results.

Alternative composite likelihood method which is less complex than the full traditional composite likelihood method can be also provide good classification without significant loss of miss-classified data points. This method can provide adequate values for the parameters and which can also be used as starting values of the full multivariate model estimation or the composite likelihood estimation.

It is still an open problem to find proper randomized algorithm to deal with high dimensional count data so as to achieve parsimony. For applying composite likelihood an EM type algorithm was developed but the underlying structure generates some interesting questions since we can get a classifier based on different approaches, and we have not pursued this further on.

A related problem that has attracted much less work in the literature is the model based clustering of mixed mode data. For such models, applying the finite mixture approach has certain limitations since it is not easy to develop multivariate models for such data. In this thesis we followed an approach developed in Kosmidis and Karlis (2016) via copulas. In particular one may define an appropriate multivariate distribution to describe jointly variables of different kind via copulas. This offers full flexibility and in effect contains several other models as special cases. For example one may define/select the marginal distributions and couple them via a copula in order to define the multivariate models. A Gaussian copula allowing for full structure can provide such a tool. However some issues arise, especially with respect the parsimony of such models as dimension increases. Here we worked two such approaches for a parsimonious representation of the Gaussian copula correlation matrix, one based on a factor decomposition of the correlation matrix and the other based on the Tsay and Pourahmadi (2017) representation with angles.

We emphasize that while we applied such parsimonious copula based finite mixtures for mixed mode data, the models and the derivations of parsimonious representations of the correlation matrix of the Gaussian copula are applicable to all models of model based clustering through copulas, like only continuous or only discrete random variables. The problem of parsimony in the model base clustering

literature is an important one. Other kind of parsimonious representations like the representation in Celeux et al. (1995) can be an alternative approach. Since in copula based models we work with the correlation rather the covariance matrix the approach needs to adjust.

Moreover, the model used here is related to the Gaussian copula factor model as described in Murray JS (2013). Another representation of factor models in the copula setting has been proposed in Krupskii and Joe (2013) which is perhaps more general. In the present work we have used the former representation and not the latter.

The approach based on the representation of Tsay and Pourahmadi (2017) leads to some interesting model selection problems in the sense that we would like to identify the best structure of the correlation matrices with some automatic approach.

Finally note that we applied a penalized version of the Tsay and Pourahmadi (2017) representation that selects the structure in an automatic way by penalizing more complicated ones. This problem is described and addressed through the simulation study including penalized log likelihood approach.

Similar results can be applied to other copulas expressed through a correlation matrix like the t-copulas or some of the elliptical copulas.

4.1 Future Work

Further work that can be exploit in the future consider:

- Computational issues, how can we improve the computing time, like using parallel computing and tricks in the maximization steps, including the penalization approach where some more clever search for the optimum lambda may be used.
- Apply composite likelihood for the copula based model based clustering for mixed mode data, no need to specify complicated models
- Complement the composite likelihood approach with the penalized approach in order to achieve parsimony and reduce the need to estimate a huge number of parameters.
- Examine in more depth the effect of the CL to the clustering problem, can we get back the clustering (or at least help on that) by a simplified CL approach?

- Expand and check the properties of the randomized version of composite likelihood in the context of model based clustering through copulas. Create proper model selection approaches for such an extension.
- Apply the concepts (both the CL and the parsimony) to related copula based models.

Bibliography

Ahmad, A. and Khan (2019). Survey of state of the art mixed data clustering algorithms. *Advances in Statistical Analysis* (7), 31883–31902.

Aitchinson, J., H. C. (1989). The multivariate poisson-log normal distribution. *Biometrika 75*, 621–629.

Al-Hussaini, E.K., A. K. (1981). On the identifiability of finite mixtures of distributions. *IEEE Trans. Inf. Theory 27*, 664–668.

Alfo, M., A. Maruotti, and G. Trovato (2011). A finite mixture model for multivariate counts under endogenous selectivity. *Statistics and Computing 21*(2), 185–202.

Andrews, J. L. and P. D. McNicholas (2011). Mixtures of modified t-factor analyzers for model-based clustering, classification, and discriminant analysis. *Journal of Statistical Planning and Inference 141*, 1479–1486.

Bandeen-roche, K., M. D. L. Z. S. L. . R. P. J. (1997). Latent variable regression for multiple discrete outcomes. *J. Am. Stat. Assoc. 92*(440), 1375–1386.

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

Berkhout, P. and E. Plug (2004). A bivariate poisson count data model using conditional probabilities. *Statistica Neerlandica 58*(3), 349–364.

Bohning, D. (2000). Computer assisted analysis of mixtures and applications in meta-analysis. *Disease Mapping and Others. CRC Press, New York.*.

Boudreault, M. and A. Charpentier (2011). Multivariate integer-valued autoregressive models applied to earthquake counts. *http://arxiv.org/abs/1112.0929*.

Bouveyron, C. and C. Brunet-Saumard (2014). Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis 71*, 52–78.

Browne, R., . M. P. (2012). Model-based clustering, classification, and discriminant analysis of data with mixed type. *Journal of Statistical Planning and Inference 11*(142), 2976–2984.

Browne, R. and P. McNicholas (2012). Model-based clustering, classification, and discriminant analysis of data with mixed type. *Journal of Statistical Planning and Inference 142*(11), 2976–2984.

Celeux, G., G. Govaert, and G. Govaert (1995). Gaussian parsimonious clustering models. *Pattern Recognition 28*, 781–793.

Chib, S. and R. Winkelmann (2001). Markov Chain Monte Carlo Analysis of Correlated Count Data. *Journal of Business & Economic Statistics 19*(4), 428–435.

Chib, S., W. R. (2001). Markov chain monte carlo analysis of correlated count data. *J. Bus. Econ. Stat. 19*, 428—-435.

Cortez, P. and A. d. J. R. Morais (2007). A data mining approach to predict forest fires using meteorological data.

Czado, C., R. Kastenmeier, E. C. Brechmann, and A. Min (2012). A mixed copula model for insurance claims and claim sizes. *Scandinavian Actuarial Journal 2012*(4), 278–305.

D., L. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*.

Davis, R. A. and C. Y. Yau (2011). Comments on pairwise likelihood in time series models. *Statistica Sinica*, 255–277.

DB., H. (2000). Zero-inflated poisson and binomial regression with random effects:a case study. *Biometrics*.

Dillon, J. V. and G. Lebanon (2010). Stochastic composite likelihood. *Journal of Machine Learning Research 11*(Oct), 2597–2633.

Everitt, B. (1988). A finite mixture model for the clustering of mixed-mode data. *Statistics and Probability Letters 6*(5), 305–309.

Fieuws, S. and G. Verbeke (2006). Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics 62*(2), 424–431.

Forbes, F. and D. Wraith (2014). A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweight: application to robust clustering. *Statistics and Computing 24*(6), 971–984.

Foss, A. H., M. Markatou, and B. Ray (2018). Distance metrics and clustering methods for mixed-type data. *International Statistical Review*.

Foss, A., M. M. . R. A. H. (2016). A semiparametric method for clustering mixed data. *Mach. Learn. 105*(3), 419–458.

Fraley, C. and A. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association 97*(458), 611–631.

Fraley, C., A. E. Raftery, T. B. Murphy, and L. Scrucca (2012). mclust version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation. Technical Report 597, Department of Statistics, University of Washington.

Frühwirth-Schnatter, S. and S. Pyne (2010). Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. *Biostatistics 11*(2), 317–336.

Genest, C. and J. Nešlehová (2007). A primer on copulas for count data. *Astin Bulletin 37*(2), 475–515.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Advances in Statistical Analysis 27*(4), 857–871.

Hausman JA, Hall BH, G. Z. (1984). Econometric models for count data with an application to the patents-r&d relationship. *Cambridge: National Bureau Of Economic Research*.

Hennig, C. (2010). Methods for merging gaussian mixture components. *Advances in Data Analysis and Classification 4*(1), 3–34.

Hennig, C. and T. F. Liao (2013). How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 62*(3), 309–369.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery 2*(3), 283–304.

Hui, F. K., S. Müller, and A. Welsh (2018). Sparse pairwise likelihood estimation for multivariate longitudinal mixed models. *Journal of the American Statistical Association 113*(524), 1759–1769.

Hunt, L. and M. Jorgensen (2011). Clustering mixed data. *WIREs Data Mining and Knowledge Discovery 1*, 352–361.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*. Chapman & Hall Ltd.

Joe, H. (2014). *Dependence Modeling with Copulas*. Chapman and Hall/CRC.

Joe, H. and Y. Lee (2009). On weighting of bivariate margins in pairwise likelihood. *Journal of Multivariate Analysis 100*(4), 670–685.

Johnson, N., S. Kotz, and N. Balakrishnan (1997). *Multivariate Discrete Distributions*.

Jorgensen, M. (2004). Using multinomial mixture models to cluster internet traffic. *Australian and New Zealand Journal of Statistics 46*(2), 205–218.

Kano, K., K. K. (1991). On recurrence relations for the probability function of multivariate generalized poisson distribution. *Commun. Stat. Theory Methods 20*, 165–178.

Karazsia BT, V. D. M. (2008). Regression models for count data: illustrations using longitudinal predictors of childhood injury. *J Pediatr Psychol.*.

Karlis, D. (2003). An em algorithm for multivariate poisson distribution and related models. *J. Appl. Stat. 30*, 63–77.

Karlis, D. and L. Meligkotsidou (2005). Multivariate Poisson regression with covariance structure. *Statistics and Computing 15*(4), 255–265.

Karlis, D. and L. Meligkotsidou (2007). Finite multivariate Poisson mixtures with applications. *Journal of Statistical Planning and Inference 137*, 1942–1960.

Karlis, D. and A. Santourian (2009). Model-based clustering with non-elliptically contoured distributions. *Statistics and Computing 19*(1), 73–83.

Karlis, D., X. E. (2005). Mixed poisson distributions. *Int. Stat. Rev. 73*, 35–58.

Kaufman, L., . R. P. (1990). Finding groups in data. *New York: Wiley*.

Kocherlakota, S. and K. Kocherlakota (1992). *Bivariate Discrete Distributions, Statistics: textbooks and monographs*, Volume 132. New York: Markel Dekker.

Kosmidis, I. and D. Karlis (2016). Model-based clustering using copulas with applications. *Statistics and computing 26*(5), 1079–1099.

Krupskii, P. and H. Joe (2013). Factor copula models for multivariate data. *Journal of Multivariate Analysis 120*, 85–101.

Krzanowski, W. (1993). The location model for mixtures of categorical and continuous variables. *Journal of Classification 10*(1), 25–49.

Kuk, A. Y. (2007). A hybrid pairwise likelihood method. *Biometrika 94*(4), 939–952.

Kuk, A. Y. and D. J. Nott (2000). A pairwise likelihood approach to analyzing correlated binary data. *Statistics and Probability Letters 47*(4), 329–335.

Lawrence, C. and W. Krzanowski (1996). Mixture separation for mixed-mode data. *Statistics and Computing 6*(1), 85–92.

Lee, S. and G. McLachlan (2014). Finite mixtures of multivariate skew t-distributions: some recent and new results. *Statistics and Computing 24*, 181–202.

Lin, T.-I., H. Ho, and C.-R. Lee (2014). Flexible mixture modelling using the multivariate skew-t-normal distribution. *Statistics and Computing 24*(4), 531–546.

Lindsay, B. (1995). Mixture models: theory, geometry and applications. in: Regional conference series in probability and statistics. *Institute of Mathematical Statistics and American Statistical Association. 5*.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary mathematics 80*(1), 221–239.

Long SJ, Long JS, F. J. (2006). Regression models for categorical dependent variables using stata. *Texas: Stata Press*.

Marbac, M., C. Biernacki, and V. Vandewalle (2017). Model-based clustering of gaussian copulas for mixed data. *Communications in Statistics-Theory and Methods 46*(23), 11635–11656.

Marbac, M. & Sedki, M. (2018). Varsellcm:an r/c++ package for variable selection in model-based clustering of mixed-data with missing values. *Bioinformatics 35*(7), 1255–1257.

Mazo, G., D. Karlis, and A. Rau (2021, 01). A randomized pairwise likelihood method for complex statistical inferences.

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

McLachlan, G. J., D. Peel, and R. Bean (2003). Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics & Data Analysis 41*(3-4), 379–388.

McNicholas, P. D. (2016). *Mixture model-based classification*. Chapman and Hall/CRC.

McNicholas, P. D. and T. B. Murphy (2008). Parsimonious Gaussian mixture models. *Statistics and Computing 18*(3), 285–296.

McParland, D. and I. C. Gormley (2016). Model based clustering for mixed data: clustmd. *Advances in Data Analysis and Classification 10*(2), 155–169.

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

Morris, K. and P. McNicholas (2013). Dimension reduction for model-based clustering via mixtures of shifted asymmetric laplace distributions. *Statistics and Probability Letters 83*(9), 2088–2093.

Murray JS, Dunson DB, C. L. L. J. (2013). Bayesian gaussian copula factor models for mixed data. *J Am Stat Assoc. 108*(502), 656–665.

Nelsen, R. (2006). *An introduction to copulas*. Springer series in statistics. Springer.

Nikoloulopoulos, A. and D. Karlis (2009). Finite normal mixture copulas for multivariate discrete data modeling. *Journal of Statistical Planning and Inference 139*, 3878–3890.

Nott, D. J. and T. Rydén (1999). Pairwise likelihood methods for inference in image models. *Biometrika 86*(3), 661–676.

Panagiotelis, A., C. Czado., and M. Joe (2012). Pair copula constructions for multivariate discrete data. *Journal of the American Statistical Association 107*(499), 1063–1072.

Papageorgiou, I. and I. Moustaki (2018). Sampling of pairs in pairwise likelihood estimation for latent variable models with categorical observed variables. *Statistics and Computing*, 1–15.

Preud, G. e. a. (2021). Head-to-head comparison of clustering methods for heterogeneous data: a simulation-driven benchmark. *Scientific Reports*.

R., W. (2016). Models for count outcomes. *Notre Dame: University of Notre Dame*.

Ranalli, M. and R. Rocci (2017). Mixture models for mixed-type data through a composite likelihood approach. *Computational Statistics & Data Analysis 110*, 87–102.

Sahu, T. P., N. K. Nagwani, and S. Verma (2016). Multivariate beta mixture model for automatic identification of topical authoritative users in community question answering sites. *IEEE Access 4*, 5343–5355.

Sklar, M. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris 8*, 229–231.

Tingjin Chu, a. J. Z. and H. Wang (2011). Penalized maximum likelihood estimation and variable selection in geostatistics. *The Annals of Statistics 39*(5), 2607–2625.

Tom Brijs, Dimitris Karlis, G. S. K. V. G. W. P. M. (2004). A multivariate poisson mixture model for marketing applications. *Statistica Neerlandica 58*(3), 322–348.

Tsay, R. S. and M. Pourahmadi (2017). Modelling structured correlation matrices. *Biometrika 104*(1), 237–242.

Tsiamyrtzis, P., K. D. (2004). Strategies for efficient computation of multivariate probabilities. *Commun. Stat. Simulation Comput. 33*, 271–293.

Varin, C., G. Høst, and Ø. Skare (2005). Pairwise likelihood inference in spatial generalized linear mixed models. *Computational Statistics and Data Analysis 49*(4), 1173–1191.

Varin, C., N. Reid, and D. Firth (2011). An overview of composite likelihood methods. *Statistica Sinica 21*(1), 5–42.

Varin, C. and P. Vidoni (2005). A note on composite likelihood inference and model selection. *Biometrika 92*(3), 519–528.

Vasdekis, V. G., D. Rizopoulos, and I. Moustaki (2014). Weighted pairwise likelihood estimation for a general class of random effects models. *Biostatistics 15*(4), 677–689.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc. 58*(301), 236–244.

Wedel, M. and W. Kamakura (2000, 01). *Market Segmentation: Conceptual and Methodological Foundations*, Volume 8.