SEMIPARAMETRIC MULTIVARIATE DENSITY ESTIMATION USING COPULAS AND SHAPE-CONSTRAINTS

SAWITREE BOONPATCHARANON

A DISSERTATION SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN MATHEMATICS AND STATISTICS
YORK UNIVERSITY
TORONTO, ONTARIO

May 2019

Abstract

Maximum likelihood estimation of a log-concave density has certain advantages over other nonparametric approaches, such as kernel density estimation, which requires a bandwidth selection. Furthermore, finding the optimal bandwidth gets more difficult as a dimension increases. On the other hand, the shape-constrained approach is automatic and does not need any tuning parameters. However, for both the kernel and log-concave estimators, the rate of convergence slows down as the dimension d increases. To handle this "curse of dimensionality", we study an intermediate semi-parametric copula approach and we estimate the marginals using the log-concave shape-constrained MLE and use a parametric approach to fit the copula parameters. We prove $\sqrt{n}$ rate of convergence for the parametric estimator and that the joint density converges at a rate of $n^{-2/5}$ regardless of dimension. This is faster than the conjectured rate of $n^{-2/(d+4)}$ for the multivariate log-concave estimators [Cule, 2009]. We examine the performance of our proposed method via simulation studies and real data example.

Acknowledgements

This thesis would not be submitted without support from a number of people. I would like to thanks my supervisor Dr. Hanna Jankowski for her numerous helpful and valuable guidance since I am a Ph.D. student at York University. I also would like to thanks my committee members, Dr. Hélène Massam and Dr. Xin Gao, for their comments both in dissertation subject oral exam and colloquium dissertations. Those comments are significant for improving this thesis. I also want to thanks all the staff members and friends for their helpful in several ways. Moreover, I would like to thanks my family for thier supports and always stand by my side. Finally, I would like to special thanks to my sponsorship, Chulalongkorn University, Thailand for their support throughout the degrees.

Table of Contents

List of Tables

List of Figures

1 Introduction

1.1 Motivation

Multivariate density estimation is a common and much studied problem in statistics. When we have d-dimensional independent and identically distributed (i.i.d.) data, say $X_1, X_2, \ldots, X_n \in \mathbb{R}^d$, the ideal way to estimate a density f is using parametric density estimation in which the true distribution function needs to be known. However, in several cases, we do not have any clue for the distribution of f. The general way to estimate f is to use nonparametric approaches, which require fewer assumptions than parametric approaches and do not need information about the distributions.

When we have no idea about the distribution of data, we can easily consider the distribution function as a step function, which jumps up by $1/n$ at each point of observation. We call this idea an empirical cumulative distribution function. However, it is not a density function, so it is not our preferred approach. The popular

nonparametric density estimation is a kernel density estimation, which requires a good bandwidth selection in order to have a good density estimation. However, a bandwidth parameter is not easy to find, but we can choose this smoothing parameter by using a cross-validation approach or normal reference rule, (see Wasserman [2006] for more details). However, the bandwidth selection will be an issue when the dimension, d, is getting large because, instead of choosing one bandwidth parameter for one-dimensional data, you need to define a bandwidth matrix with its dimension according to the data's dimension. For example, when your data have four dimensions, your bandwidth will be a 4×4 matrix, which has a symmetric and positive definite property. Furthermore, there is a tradeoff between bias and variance, because, if you choose too large a bandwidth, the bias will be huge while the variance small. This event is an oversmoothing problem. On the contrary, if your bandwidth is too small, you will get an undersmoothing density estimation. Further details of kernel density can be found in Section 1.2.2.3.

To overcome the difficulty of bandwidth selection, a shape-restricted density estimation has been introduced and has become more popular in recent years. Instead of dealing with the additional tuning parameters, the shape-constrained density estimation works with some good characteristics of its functions. Such examples include monotonic, unimodal, or log-concavity. Among several shape-constrained models,

our work focuses on the log-concave densities.

One advantage of the log-concave density estimation above the kernel density estimation is that it does not need any tuning parameters. Furthermore, a log-concave maximum likelihood estimator always exists and is unique, and it can be done automatically. However, the multivariate log-concave MLE is computationally intensive, which makes the algorithm not friendly in practice. The univariate log-concave ML estimator converges with rate $n^{-2/5}$, but the conjectured rate of convergence for multivariate log-concave ML estimator is $n^{-2/(d+4)}$, which depends on the dimension d (see Cule [2009]). This makes the estimators from multivariate log-concave distributions have much slower convergence rate than the univariate log-concave, especially when d is large.

Sklar [1959] introduced a dependence modeling called "copula". The copula is a function, which splits joint distribution function to its one-dimensional marginal distributions and dependence part with its parameters, which are called the copula parameters. Suppose we observe n i.i.d. random variables $X = (X_1, \ldots, X_n) \in \mathbb{R}^d$, let $f_j(x_j)$ be the univariate marginal density functions of X in dimension jth with corresponding cumulative distribution functions $F_j(s) = \int_{-\infty}^{s} f_j(r)dr; j = 1, \ldots, d$. The joint density function with the copula density c and the copula parameter θ can

be represented in this form

$$f(x_1, \ldots, x_d) = c(F_1(x_1), \ldots, F_d(x_d); \theta) \prod_{j=1}^{d} f_j(x_j), \tag{1.1}$$

where f_j can be any parametric or nonparametric univariate density functions. Similary, c represents the dependence of d dimensions and its density can be chosen from several well-known copula families (more details in Chapter 3). Copula models have been widely used in recent years, for example, in finance, see Jouanin et al. [2011] and epidemiology, see Chen [2012].

The main objective of this thesis is to improve the rate of convergence and computational time for multivariate log-concave density estimators by using the copula model with univariate log-concave marginals. We propose the semiparametric density estimation for $\mathbb{R}^d$, where the main tools of our study are the copula models together with the univariate log-concave densities. We work on estimating the density in equation (1.1) by applying the method from Joe [2005], which has been widely used in recent years. This method has been called an "inference function for margins" (IFM). It is a two-stage estimation method which estimates marginal densities and copula parameters separately.

Moreover, we also prove that our proposed semiparametric density estimation method converges at a rate of $n^{-2/5}$, and the rate of convergence for the copula estimator is $n^{-1/2}$. This rate improves the conjectured rate of multivariate log-concave

estimators mentioned in Cule [2009], which is $n^{-2/(d+4)}$. However, the convergence rate of the copula estimators is never better than a parametric ML estimators, which converge at a rate of $n^{-1/2}$.

In this chapter, we will discuss each type of density estimation, which belong to the parametric density estimations in Section 1.2.1, to the nonparametric density estimations in Section 1.2.2, and to the semiparametric density estimations in Section 1.2.4. The nonparametric density estimations that we will focus on consist of the empirical cumulative distribution function, the histogram, and the kernel density estimation. The log-concave density estimation will be presented in Section 1.2.3. In addition, the rate of convergence for each method will be summarized in Section 1.3.

1.2 Density estimation

1.2.1 Parametric density estimation

Let $X = \{X_1, X_2, \ldots, X_n\} \sim F$ be the n i.i.d. random variables from a distribution function, F, with a probability density, $f = F'$, the density estimation of f can be represented as $\widehat{f_n}$. A likelihood function of the n observations, $x = \{x_1, \ldots, x_n\}$, for

the density f with distribution's parameter(s) $\theta \in \Theta$ can be expressed as

$$L(x|\theta) = \prod_{i=1}^{n} f_\theta(x_i).$$

Then, we get a log-likelihood function

$$\ell(x|\theta) = \sum_{i=1}^{n} \log f_\theta(x_i). \tag{1.2}$$

To simplify, we denote $\ell(x|\theta)$ as $\ell(\theta)$. Then, we maximize (1.2) to get a maximum likelihood estimator $\widehat{\theta}_n$, which

$$\widehat{\theta}_n = \underset{\theta \in \Theta}{\operatorname{argmax}}\ \ell(\theta).$$

Hence, the density estimator of f can be expressed as $\widehat{f}_n(x) = f_{\widehat{\theta}_n}(x)$. In several distributions, $\widehat{\theta}_n$ has a closed form, for instance, Gaussian, exponential, Poisson, binomial and also other distributions in exponential family. This makes parametric approach easy to use. Moreover, the parametric maximum likelihood estimator under the regularity conditions also satisfies a convergence rate of $n^{-1/2}$, which is the fastest rate among other density estimation methods.

1.2.2 Nonparametric density estimation

Knowing the distribution functions of data is hard in practice. Therefore, the parametric density estimation may not be a good choice. The nonparametric ap-

proaches were introduced because they can model the unknown distribution functions. For example, the upcoming method has the fewest assumptions for estimating distributions by just giving a mass $1/n$ to each observation. For the following nonparametric methods, suppose we observe n i.i.d. random variables, $X = \{X_1, X_2, \ldots, X_n\}$, from an unknown density f.

1.2.2.1 Empirical cumulative distribution function

Definition of the cumulative distribution function (CDF) is $F(t) = P(X \leq t)$. We can estimate $F(t)$ by

$$\widehat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{x_i \leq t}.$$

We call $\widehat{F}_n(t)$ an empirical CDF, which can be found by not assuming any underlying distributions to the data. However, the empirical CDF is a mass function, which is not a density function. It means that this method may not be a desired answer for the density estimation.

1.2.2.2 Histogram

One of the simple nonparametric density estimations is histogram, which is not a smooth function. The density estimator for n observations can be represented as

$$\widehat{f}_n(x) = \sum_{j=1}^{m} \frac{\widehat{p}_j}{h} \mathbb{1}_{B_j}(x),$$

where B_j are the jth bin. $B_1 = \left[0, \frac{1}{m}\right), B_2 = \left[\frac{1}{m}, \frac{2}{m}\right), \ldots, B_m = \left[\frac{m-1}{m}, 1\right]$, where m is the total number of bins. The binwidth h equals $1/m$. Let A_j be the number of observations in B_j, then $\widehat{p}_j = A_j/n$.

Theorem 1.1. *(Wasserman [2006], Theorem 6.9) Consider fixed x and fixed m, let B_j be the bin containing x and let $p_j = \int_{B_j} f(u)du$. Then,*

$$\mathbb{E}\left(\widehat{f}_n(x)\right) = \frac{\mathbb{E}(\widehat{p}_j)}{h} = \frac{p_j}{h} = \frac{\int_{B_j} f(u)du}{h} \approx \frac{f(x)h}{h} = f(x) \quad \textit{and} \quad Var\left(\widehat{f}_n(x)\right) = \frac{p_j(1-p_j)}{nh^2}.$$

The expectation and variance in Theorem 1.1 are calculated with respect to the random quantity $\widehat{f}_n$ which depends on $X_1, \ldots, X_n$. When h approaches zero, $\widehat{f}_n(x)$ will become an unbiased estimator, however its variance will become large. It turns out that choosing a reasonable h is necessary for histogram.

A convergence rate of this estimator is computed via a risk function, which is calculated from an expectation of an integrated square loss function. Let R and L denote a risk function and a loss function, respectively. The risk function is given by

$R = E(L)$, where

$$L = \int \left(\widehat{f}_n(x) - f(x)\right)^2 dx.$$

Theorem 1.2. *(Wasserman [2006], Theorem 6.11) Suppose that f' is absolutely continuous and that $\int (f'(u))^2 du < \infty$. Then*

$$R(\widehat{f}_n, f) = \frac{h^2}{12} \int (f'(u))^2 du + \frac{1}{nh} + o(h^2) + o\left(\frac{1}{n}\right). \tag{1.3}$$

A value h^* that minimizes (1.3) *is*

$$h^* = \frac{1}{n^{1/3}} \left(\frac{6}{\int (f'(u))^2 du} \right)^{1/3}.$$

With this choice of bandwidth,

$$R(\widehat{f}_n, f) \approx \frac{(3/4)^{2/3} \left(\int (f'(u))^2 du\right)^{1/3}}{n^{2/3}}.$$

Hence, when we choose the optimal h^*, the risk decreases to zero at a rate of $n^{-2/3}$. It means the histogram density estimator converges with a rate of $n^{-1/3}$, which is slower than a rate of $n^{-1/2}$ from the parametric density estimation.

1.2.2.3 Kernel density estimation

A kernel density estimation is another simple and well known nonparametric density estimation. A kernel is a smooth function K such that $K(x) \geq 0, \int K(x)dx =$

$1, \int xK(x)dx = 0$, and $\sigma_K^2 = \int x^2K(x)dx > 0$. There are several choices of kernel such as

$$\text{Gaussian kernel: } K(x) = \tfrac{1}{\sqrt{2\pi}}e^{-x^2/2},$$

$$\text{boxcar kernel: } K(x) = \tfrac{1}{2}\mathbb{1}_{[-1,1]}(x),$$

$$\text{Epanechnikov kernel: } K(x) = \tfrac{3}{4}(1-x^2)\mathbb{1}_{[-1,1]}(x).$$

One-dimensional density estimation of n observations with a bandwidth h and a kernel K can be written in this form

$$\widehat{f}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right).$$

An effect of the bandwidth selection can be clearly seen in Figure 1.1. Furthermore, the risk function for a kernel estimator is presented in the following theorem.

Theorem 1.3. *(Wasserman [2006], Theorem 6.28) Let $R = \int E(f(x) - \widehat{f}(x))^2dx$ be the integrated risk. Assume that f'' is absolutely continuous and that $\int (f'''(x))^2dx < \infty$. Then,*

$$R(\widehat{f}_n, f) = \frac{1}{4}\sigma_K^4 h_n^4 \int (f''(x))^2dx + \frac{\int K^2(x)dx}{nh} + O\left(\frac{1}{n}\right) + O(h_n^6), \tag{1.4}$$

where $\sigma_K^2 = \int x^2K(x)dx$. The optimal h^* that minimizes (1.4) is

$$h^* = \left(\frac{\int K(x)^2dx}{n(\int x^2K(x)dx)^2 \int (f''(x))^2dx}\right)^{1/5}.$$

Plug in h^* into (1.4), we get

$$R(\widehat{f}_n, f) = O(n^{-4/5}).$$

Therefore, the kernel density estimator converges with a rate of $n^{-2/5}$, which is faster than $n^{-1/3}$ from the histogram. However, it is still slower than $n^{-1/2}$ from the parametric MLE.

For $d > 1$, suppose we have n i.i.d random variables $X_i = (X_{i1}, \ldots, X_{id}); i = 1, \ldots, n$, then the density estimator with a bandwidth matrix H is given by

$$\widehat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} K_H(x - x_i), \tag{1.5}$$

where H is a $d \times d$ symmetric and positive definite matrix. K is a multivariate kernel function with $K_H(x) = |H|^{-1/2} K(H^{-1/2}x)$. A popular kernel function is a multivariate Gaussian kernel, which can be written in this form

$$K_H(x) = (2\pi)^{-d/2} |H|^{-1/2} \exp\left\{-\frac{1}{2} x^T H^{-1} x\right\},$$

where H plays the role of a covariance matrix. From (1.5), to avoid building the bandwidth matrix H, we can estimate f at $x = (x_1, \ldots, x_d)$ by using the following formula

$$\widehat{f}_n(x) = \frac{1}{n h_1 \cdots h_d} \sum_{i=1}^{n} \left\{ \prod_{j=1}^{d} K\left(\frac{x_j - x_{ij}}{h_j}\right) \right\},$$

where $h_1, \ldots, h_d$ denote the bandwidth for each dimension.

For multivariate kernel estimator, the risk function can be expressed as

$$R(\widehat{f}_n, f) \approx \frac{1}{4} \sigma_K^4 \left[\sum_{j=1}^{d} h_j^4 \int f_{jj}^2(x) dx + \sum_{j \neq k} h_j^2 h_k^2 \int f_{jj} f_{kk} dx \right] + \frac{(\int K^2(x) dx)^d}{n h_1 \cdots h_d},$$

where f_{jj} denote the second partial derivative of f. The optimal bandwidth satisfies $h_i = O(n^{-1/(d+4)})$. Therefore, the risk converges to zero at a rate $O(n^{-4/(d+4)})$, which leads us to the density estimator with a rate of convergence $n^{-2/(d+4)}$.

To summarize, one-dimensional kernel density estimator converges with the rate $n^{-2/5}$, which is slower than $n^{-1/2}$ of the parametric ML estimator. On the other hand, the rate of convergence of multivariate kernel density estimator is $n^{-2/(d+4)}$, which is the slowest density estimation method in our work.

1.2.3 Log-concave density estimation

Suppose we observe n independent random variables $X = \{X_1, X_2, \ldots, X_n\}$ from an unknown density $f : \mathbb{R} \mapsto [0, \infty)$ where $f(x)$ is said to be a univariate log-concave density if $f(x) = \exp \varphi(x)$ for some concave functions $\varphi : \mathbb{R} \mapsto [-\infty, \infty)$. The corresponding cumulative distribution function of f is $F(x) = \int_{-\infty}^{x} f(r)dr$.

A maximum likelihood estimation (MLE) of the log-concave density gets more attention in recent years, because its ML estimator always exists and is unique, (see Dümbgen and Rufibach [2009]). The MLE of log-concave densities has an obvious advantage above the kernel density estimation, because it is fully automatic and it does not need any smoothing parameters. We can clearly see from Figure 1.1 that the log-concave density estimation works better than the kernel density estimation. The

studies of log-concave distributions become more popular - see, for example, the work from Walther [2002]. This is the first paper that discussed the log-concave distributions in a statistical aspect. They worked on one kind of shape restricted maximum likelihood inference, called log-concavity. Walther [2009] showed a good explanation of log-concave distributions and an application in a clustering problem. Moreover, the class of log-concave distributions contains various common parametric distributions such as Gaussian, $\Gamma(\alpha, \beta)$ with $\alpha \geq 1$, $\beta(a, b)$ with both parameters greater than 1, Weibull(α) with $\alpha \geq 1$, Laplace, and exponential distributions. Moreover, Dümbgen and Rufibach [2009] showed the uniform consistency of a nonparametric ML estimator for one-dimensional log-concave densities. Balabdaoui et al. [2009, Theorem 2.1, page 1305] also showed that the log-concave density estimator $\widehat{f_n}$ and the piecewise linear $\widehat{\varphi}_n$ converge pointwise to the true density f and the true φ_0 with the rate $n^{-2/5}$ and characterized the limiting distributions. Furthermore, Doss and Wellner [2016, Theorem 3.2, page 9] proved that the univariate log-concave ML estimator has a global rate of convergence at a rate $n^{-2/5}$ under the Hellinger distance.

For multidimensional log-concave distributions, Cule [2009] studied the limiting behavior of the log-concave ML estimator and conjectured that the rate of convergence with respect to the Hellinger bracketing entropy is $n^{-2/(d+4)}$ for all d. This rate is the same as the convergence rate from multivariate kernel density estimator. Since

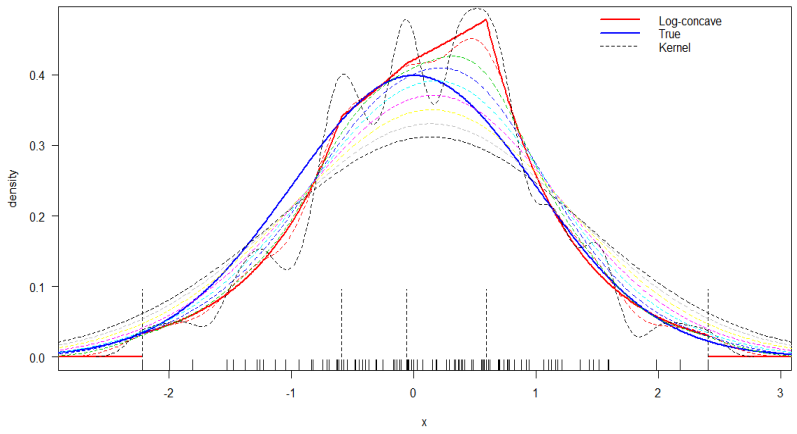


Figure 1.1: Kernel density estimation of standard Gaussian with several bandwidth selections in the same graph of using log-concave density estimation and true density

this rate depends on d, the multidimensional log-concave MLE is computationally intensive especially when d is high. Moreover, Cule et al. [2010] presented some attractive theoretical properties of the multivariate log-concave MLE both when the model is log-concave and when it is a misspecified model.

1.2.4 Semiparametric density estimation

In between parametric and nonparametric models, we have a semiparametric model, which is a statistical model in which some parameters do not belong to a Euclidean space but lie in an infinite dimensional space. A log-concave space, which is a nonparametric space, is the infinite dimensional space that we work on. In order

to estimate the joint density function, we integrate the parametric copula model together with the univariate log-concave densities.

As we discussed in the previous section that finding the multivariate log-concave density estimator is computationally intensive which is because of the curse of dimensionality problem. To overcome this issue, we use the univariate log-concave density estimation and the copula model as our proposed density estimation method. Hence, our proposed semiparametric model is given by

$$f(x_1, \ldots, x_d; F_1, \ldots, F_d, \theta) = c(F_1(x_1), \ldots, F_d(x_d); \theta) \prod_{j=1}^{d} \exp \varphi_j(x_j).$$

Estimating this model will be done in two stages where the first stage is to find the log-concave ML estimators for each marginal density separately. Then, we estimate the copula parameter. This is a reason why the proposed method performs faster than the multivariate log-concave MLE, since it is not depend on the dimension. Note that all marginals do not depend on the copula parameters. A log-likelihood function for estimating the copula parameters is given by

$$\sum_{i=1}^{n} \log c(\widehat{F}_1(x_1), \ldots, \widehat{F}_d(x_d); \theta). \tag{1.6}$$

$\widehat{\theta}$ from maximizing (1.6) is also the θ that would solve

$$n^{-1} \sum_{i=1}^{n} \partial_\theta \log c(\widehat{F}_1(x_1), \ldots, \widehat{F}_d(x_d); \theta) = 0. \tag{1.7}$$

Hence, estimating the copula parameter can be viewed as a Z-estimation, which Z denote the zero from setting the above score function equals to zero. Moreover, the copula estimator from (1.7) can be called as a "Z-estimator".

Furthermore, we prove that the Z-estimator is consistent and converges at rate $n^{-1/2}$. Moreover, the joint density estimator converges at rate $n^{-2/5}$ regardless of dimension. This rate makes our proposed semiparametric density estimator converges faster than the multivariate log-concave MLE and multivariate kernel density estimation. However, it is still slower than the parametric model.

1.3 Rate of convergence

Table 1.1 shows the rates of convergence for all density estimators that have been discussed in the previous sections.

Table 1.1: Convergence rate for density estimators

Estimator from	**Rate of convergence**
parametric model	$n^{-1/2}$
copula model	$n^{-1/2}$
log-concave ($d = 1$)	$n^{-2/5}$
kernel ($d = 1$)	$n^{-2/5}$
histogram	$n^{-1/3}$
log-concave ($d > 1$)*	$n^{-2/(d+4)}$
kernel ($d > 1$)	$n^{-2/(d+4)}$

Note: * is a conjectured rate of convergence for all d.

1.4 Outline

The organization of this thesis is as follows. In Chapter 2, we discuss the log-concave density estimation both in the univariate and multivariate cases. The available R packages, logcondens and LogConcDEAD, that use to calculate the log-concave ML estimators will be mentioned with details of their algorithms. We also show the theoretical parts, which are pointwise limiting distributions and global rates of convergence, for the log-concave ML estimators. Moreover, drawbacks of using

LogConcDEAD package to estimate the multivariate log-concave ML estimators are presented in the form of computational time.

In Chapter 3, we present the copula model to solve the curse of dimensionality problem. This chapter will talk about several copula families such as Gaussian copula and Archimedean copula families. Note that we only focus on parametric copula families. The details of semiparametric Z-estimation will be summarized, and the two-stage estimation method will also be discussed. Moreover, some literature reviews about the asymptotic relative efficiency of the two-stage estimation method and the MLE method will be presented in this chapter too.

Some simulation studies are shown in Chapter 4. The performance on the density estimation of our proposed method and other parametric, nonparametric and semiparametric methods will be presented.

The main theorems and proofs are in Chapter 5. We will show that the copula estimators under the log-concave marginals satisfy the consitency property and has $\sqrt{n}$ convergence rate. Therefore, the joint density estimator by using the copula model with the log-concave marginals converges at a rate $n^{-2/5}$, irrespective of the dimension. Some necessary regularity conditions and assumptions for proving the main theorems also be demonstrated at the end of this chapter.

We also work on a classification problem with real data example which will be

shown in Chapter 7 where the background of finite mixture models and expectation-maximization (EM) algorithm will be presented in Chapter 6. Some further extensions will be discussed in Chapter 8. In this chapter, we focus on vine copulas, which allow us to account for the multiple dependence structures.

Other than the main part of this thesis, we work on another applied project. This project works on the univariate log-concave densities. We do some simulation studies on clustering problems and proposed a new criterion for selecting the number of subpopulations. We call this criterion as a "proposed BIC". This application is presented in Chapter 9.

2 Log-concave density estimation

2.1 Definitions and properties

A function f is said to be concave if

$$f(\lambda x + (1-\lambda)y) \quad \geq \quad \lambda f(x) + (1-\lambda) f(y)$$

for all $x, y \in \mathbb{R}^d$ and $\lambda \in (0,1)$. We also say that a density f will be a log-concave density if $\log f$ is concave. Let $X_1, X_2, \ldots, X_n$ be independent random variables from some unknown densities $f : \mathbb{R} \mapsto [0, \infty)$, the log-concave density function can be expressed as

$$f(x) = \exp \varphi(x). \tag{2.1}$$

From (2.1), we also have $\log f(x) = \varphi(x)$, for some concave functions $\varphi : \mathbb{R} \to [-\infty, \infty)$. The cumulative distibution function (CDF) can be represented as $F(x) = \int_{-\infty}^{x} f(r)dr$. The following are examples of log-concave distributions.

Example 2.1. Gaussian distribution

$$f(x) = \frac{1}{2\pi\sigma^2}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\};\ \mu \in [-\infty, \infty],\ \sigma^2 > 0$$

$$\log f(x) = -\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x-\mu)^2$$

$$\partial_x \log f(x) = -\frac{(x-\mu)}{\sigma^2}$$

$$\partial_x^2 \log f(x) = -\frac{1}{\sigma^2} \qquad \text{concave for all } x$$

Example 2.2. Weibull distribution

$$f(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k};\quad x \geq 0,\ k, \lambda > 0$$

$$\log f(x) = \log\frac{k}{\lambda} + (k-1)\log\frac{x}{\lambda} - \left(\frac{x}{\lambda}\right)^k$$

$$\partial_x \log f(x) = \frac{k-1}{x} - \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}$$

$$\partial_x^2 \log f(x) = -\frac{k-1}{x^2} - \frac{k(k-1)}{\lambda^2}\left(\frac{x}{\lambda}\right)^{k-2} \qquad \text{not concave when } k < 1$$

$$F(x) = 1 - e^{-(x/\lambda)^k};\quad x \geq 0,\ k, \lambda > 0$$

$$\log F(x) = \left(\frac{x}{\lambda}\right)^k$$

$$\partial_x \log F(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}$$

$$\partial_x^2 \log F(x) = \frac{k(k-1)}{\lambda^2}\left(\frac{x}{\lambda}\right)^{k-2} \qquad \text{concave when } k < 1$$

Example 2.3. gamma distribution

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

$$\log f(x) = \alpha \log \beta - \log \Gamma(\alpha) + (\alpha - 1) \log x - \beta x$$

$$\partial_x \log f(x) = \frac{\alpha - 1}{x} - \beta$$

$$\partial_x^2 \log f(x) = \frac{1 - \alpha}{x^2} \qquad \text{concave for } \alpha \geq 1$$

Moreover, the log-concave shape constraint is attractive for various reasons. Some of them are as follows.

1. Most common parametric distributions such as Gaussian, gamma with shape parameter ≥ 1, beta with both parameters ≥ 1, exponential, Laplace, Weibull with shape parameter ≥ 1 are log-concave. In contrast, some distributions are not log-concave for all values of parameters, for instance, Cauchy, log-normal, F, and Student's t-distribution.

2. The cumulative distribution function (CDF) of all log-concave functions are log-concave. Nevertheless, some non log-concave densities have log-concave CDFs such as log-normal, gamma when shape parameter < 1, and Weibull when shape parameter < 1, (see Example 2.2).

3. All marginal and conditional of log-concave densities are again log-concave. The reverse is not necessarily true.

4. All log-concave densities are unimodal, but not all unimodals are log-concave. According to Birgé [1997, Definition 1], a density f of the realines is called unimodal if there exists some number M (not necessary unique) such that f is nondecreasing on $(-\infty, M)$ and nonincreasing on $(M, +\infty)$. Any such M is called a mode of the density. The density f is said to be decreasing if $f = 0$ for $x < M$ and increasing if $f = 0$ for $x > M$.

5. Log-concave is called a strongly unimodal density. (Ibragimov [1956]) A distribution function is called strong unimodal if its composition with any unimodal distribution function is unimodal.

6. The sum of two independent log-concave random variables is log-concave whereas a unimodal class does not satisfy this attractive property.

7. The nonparametric ML estimator of the log-concave density always exists and is unique. The corresponding theorem is shown in Dümbgen and Rufibach [2009, Theorem 2.1] and its proof is in Dümbgen et al. [2011, Section 2]. On the contrary, the nonparametric ML estimator of a unimodal density does not exist, see Birgé [1997].

8. Balabdaoui et al. [2009] proved that the pointwise limiting distribution is $n^{2/5}\big(\widehat{f}_n(x_0) - f_0(x_0)\big)$, where $\widehat{f}_n$ is the univariate log-concave ML estimator.

Furthermore, Doss and Wellner [2016] showed that the univariate log-concave ML estimator has a global rate of convergence at a rate of $n^{-2/5}$. This rate was proved with respect to the Hellinger metric. On the other hand, the nonparametric ML estimator of the unimodal density converges at a slower rate than the log-concave ML estimator. The rate is $n^{-1/3}$.

9. The rate of convergence for univariate log-concave density estimator is better than $n^{-2/(d+4)}$, which is the conjectured rate of multivariate log-concave density estimator and multivariate kernel density estimator.

10. The use of log-concave densities appears in several applications. Chang and Walther [2007] presented clustering with a mixture model. They extend an EM algorithm to work with the univariate log-concave densities and compare the simulation results with the Gaussian mixture model (GMM). It shows that modeling with the log-concave densities has smaller misclassification cases than the GMM especially when the distributions are non-normal.

2.2 One-dimensional log-concave density

2.2.1 Log-concave maximum likelihood estimation

According to the density in (2.1), a log-likelihood function can be expressed as

$$\ell(\varphi) = \frac{1}{n}\sum_{i=1}^{n}\varphi(x_i).$$

Then, we add a Lagrange term to $\ell(\varphi)$ in order to relax a constraint of f being a density. Moreover, the objective function in (2.2) will be set to maximize over all concave functions and will still satisfy the equation of $\int \exp\widehat{\varphi}(t)dt = 1$, see Silverman [1982]. Therefore, the modified log-likelihood function is given by

$$\ell_{mod}(\varphi) = \frac{1}{n}\sum_{i=1}^{n}\varphi(x_i) - \int_{\mathbb{R}}\exp\varphi(t)\ dt. \tag{2.2}$$

Hence, the nonparametric ML estimator of φ is the maximizer of the function (2.2) over all concave functions, which can be represented as

$$\widehat{\varphi}_n = \underset{\varphi \text{ concave}}{\operatorname{argmax}}\ \ell_{mod}(\varphi).$$

We also show a comparison between the estimated density from the log-concave ML estimator and the true density in Figure 2.1. Moreover, Dümbgen and Rufibach [2009] showed that there exists a unique concave function $\widehat{\varphi}_n$ that maximizes the $\ell_{mod}(\varphi)$ function. In the next section, we will present some properties of $\widehat{\varphi}_n$.

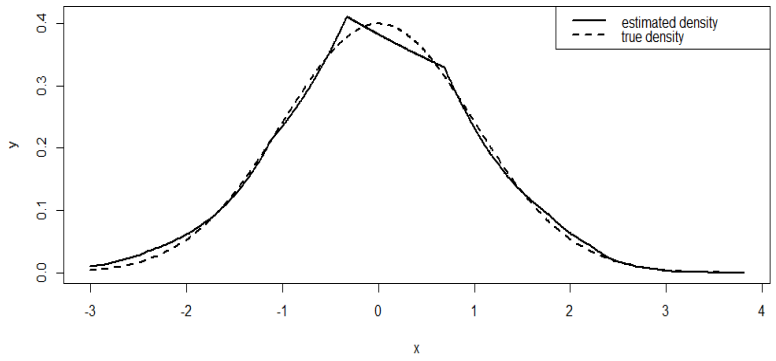


Figure 2.1: Estimated density from log-concave MLE with a true density of standard Gaussian distribution

2.2.2 Log-concave density estimator $\widehat{\varphi}_n$

We denote $\mathcal{S}_n$ as a set of all knots from some continuous piecewise linear functions $g_n : [X_{(1)}, X_{(n)}] \mapsto \mathbb{R}$, where $X_{(1)} < \cdots < X_{(n)}$ denote an order statistics of $X_1, \ldots, X_n$. The set of knots can be represented as

$$\mathcal{S}_n(g_n) := \{u \in (X_{(1)}, X_{(n)}) : g_n'(u-) > g_n'(u+)\} \cup \{X_{(1)}, X_{(n)}\}. \tag{2.3}$$

As we can see, knots occur when the function changes slope. The minimum and maximum observations always are knots. The density estimation is of the form $\widehat{f}_n(x) = \exp \widehat{\varphi}_n(x)$. Figure 2.2 shows the estimated logarithm function of standard Gaussian distribution where its knots represent at the vertical dashed line. More-

over, $\widehat{f}_n = 0$ outside the data range since $\widehat{\varphi}_n = -\infty$. The followings are some other characterizations of knots.

- $\widehat{\varphi}_n$ occur at some points of data in $[X_{(1)}, X_{(n)}]$. This is different from k-monotone density for $k > 1$ where the knots always lie between observations.

- According to Dümbgen and Rufibach [2009], for $x \geq X_{(1)}$, let

$$\widehat{F}_n(x) := \int_{X_{(1)}}^{x} \exp \widehat{\varphi}_n(u) du,$$
$$\widehat{G}_n(x) := \int_{X_{(1)}}^{x} \widehat{F}_n(u) du,$$
$$\mathbb{G}_n(x) := \int_{X_{(1)}}^{x} \mathbb{F}_n(u) du = \int_{-\infty}^{x} \mathbb{F}_n(u) du.$$

  Then, the concave function $\widehat{\varphi}_n$ is the ML estimator of the log-density φ_0 if and only if

$$\widehat{G}_n(x) \begin{cases} \leq \mathbb{G}_n(x) & \forall x \geq X_{(1)}, \\ = \mathbb{G}_n(x) & \text{if } x \in \mathcal{S}_n(\widehat{\varphi}_n). \end{cases}$$

- A consequence from the previous characterization of $\widehat{\varphi}_n$ is that the estimator of the distribution function $\widehat{F}_n$ is close to the empirical distribution function $\mathbb{F}_n$ on $\mathcal{S}_n(\widehat{\varphi}_n)$.

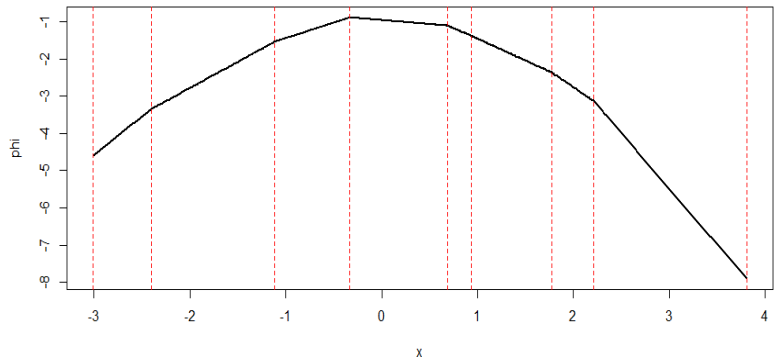


Figure 2.2: Logarithm density of standard Gaussian distribution with the vertical dotted lines represent the locations of knots

2.2.3 A computational aspect of the univariate log-concave MLE

Theorem 2.4. *[Dümbgen and Rufibach, 2009] The nonparametric ML estimator* $\widehat{\varphi}_n$ *exists and is unique. It is linear on all intervals* $[X_i, X_{i+1}]$, $1 \le i < n$. *Moreover,* $\widehat{\varphi}_n = -\infty$ *on* $\mathbb{R}\backslash[X_{(1)}, X_{(n)}]$.

From Theorem 2.4, $\exp\varphi(t)$ from (2.2) can be written as a linear function for each interval of $[X_i, X_{i+1}]$. We define S_{i+1} as a slope of $x \in [X_i, X_{i+1}]$. Hence,

$$S_{i+1} = \frac{\varphi_{i+1} - \varphi_i}{x_{i+1} - x_i}.$$

Then, the second term of $\ell_{mod}(\varphi)$ in (2.2) can be expressed as

$$\begin{aligned}\int_{\mathbb{R}} \exp\varphi(t)\ dt &= \sum_{i=1}^{n-1} \int_{x_i}^{x_{i+1}} e^{\varphi_i + (t-x_i)S_{i+1}}\ dt \\ &= \sum_{i=1}^{n-1} (e^{\varphi_{i+1}} - e^{\varphi_i}) \left(\frac{x_{i+1} - x_i}{\varphi_{i+1} - \varphi_i} \right).\end{aligned}$$

Now we can write (2.2) in an explicit form, which is

$$\ell^*(\varphi) = \frac{1}{n} \sum_{i=1}^{n} \varphi(x_i) - \sum_{i=1}^{n-1} (e^{\varphi_{i+1}} - e^{\varphi_i}) \left(\frac{x_{i+1} - x_i}{\varphi_{i+1} - \varphi_i} \right).$$

Finding one-dimensional log-concave ML estimator is quite convenient because there is an available package in R called logcondens. This package is built by Dümbgen and Rufibach [2011] and can be accessible from CRAN at `http://CRAN.R-project.org/package=logcondens`. According to their work, they presented two algorithms for calculating the univariate log-concave ML estimator, which are iterative convex minorant algorithm (ICMA) and active set algorithm (ASA). According to Walther [2009], ASA appears to be an efficient algorithm to calculate the MLE nowadays. Thus, we decide to use ASA, which is implemented by Dümbgen et al. [2011], in our thesis. The ASA is a useful tool from optimization theory, see Fletcher [1987]. The main idea of this algorithm is that it solves a finite number of unconstrained optimization problems, see Dümbgen et al. [2011, Section 3]. A function for finding the log-concave ML estimator in logcondens package with ASA is activeSetLogCon. An example code can be found in Appendix A.3.1.

2.2.4 Auxiliary results for $d = 1$

Gradient and Hessian matrices of f are also important to be studied. Since the expressions of these two matrices are complicated, we introduce a new auxiliary function J, which can rewrite the partial derivatives of (2.2) in terms of the J functions. The following J functions will be discussed again in Chapter 9 when we calculate a criterion for choosing the number of subpopulations in the clustering problem. We will show the expressions of these two matrices only for one-dimensional data. First, the modified log-likelihood function can be represented in the term of J function, which is given by

$$\ell^*(\varphi) = \frac{1}{n}\sum_{i=1}^{n}\varphi(x_i) - \sum_{i=1}^{n-1} J(\varphi_i, \varphi_{i+1})(x_{i+1} - x_i).$$

The J function can be expressed as

$$J(\varphi_j, \varphi_k) = J(\varphi_k, \varphi_j) = \begin{cases} \frac{\exp(\varphi_k) - \exp(\varphi_j)}{\varphi_k - \varphi_j} & \text{if } \varphi_j \neq \varphi_k, \\ \exp(\varphi_j) & \text{if } \varphi_j = \varphi_k, \end{cases} \tag{2.4}$$

with the fact that $J(\varphi_j, \varphi_k) = \exp(\varphi_j)J(0, \varphi_k - \varphi_j)$. In addition, $J(0,0) = 1$ and $J(0,r) = \frac{\exp(r)-1}{r}$. Letting $J_{pq}(\varphi_j, \varphi_k) = \partial^p_{\varphi_j}\partial^q_{\varphi_k} J(\varphi_j, \varphi_k)$ and $\Delta_j = x_{j+1} - x_j$, then the

gradient and Hessian matrices of $\ell^*(\varphi)$ when we have m knots are given by

$$\partial_{\varphi_j}\ell^*(\varphi) = \begin{cases} \frac{1}{n} - \Delta_1 J_{10}(\varphi_1, \varphi_2) & \text{for } j = 1, \\ \frac{1}{n} - \Delta_j J_{10}(\varphi_j, \varphi_{j+1}) - \Delta_{j-1} J_{01}(\varphi_{j-1}, \varphi_j) & \text{for } 2 \le j < m, \\ \frac{1}{n} - \Delta_{n-1} J_{01}(\varphi_{m-1}, \varphi_m) & \text{for } j = m. \end{cases} \tag{2.5}$$

$$-\partial_{\varphi_j}\partial_{\varphi_k}\ell^*(\varphi) = \begin{cases} \Delta_1 J_{20}(\varphi_1, \varphi_2) & \text{for } j = k = 1, \\ \Delta_j J_{20}(\varphi_j, \varphi_{j+1}) - \Delta_{j-1} J_{02}(\varphi_{j-1}, \varphi_j) & \text{for } 2 \le j = k < m, \\ \Delta_{n-1} J_{02}(\varphi_{n-1}, \varphi_m) & \text{for } j = k = m, \\ \Delta_j J_{11}(\varphi_k, \varphi_j) & \text{for } 1 < j = k+1 \le m, \\ 0 & \text{for } |j-k| > 1. \end{cases} \tag{2.6}$$

More details of J functions are in Appendix A.2.

2.2.5 Pointwise limiting distributions of the log-concave ML estimator

Balabdaoui et al. [2009] derived the pointwise limiting distributions of $n^{k/(2k+1)}\left(\widehat{f}_n(x_0) - f_0(x_0)\right)$, $n^{k/(2k+1)}\left(\widehat{\varphi}_n(x_0) - \varphi_0(x_0)\right)$ and also $n^{k/(2k+1)}\left(\widehat{f}'_n(x_0) - f'_0(x_0)\right)$, $n^{k/(2k+1)}\left(\widehat{\varphi}'_n(x_0) - \varphi'_0(x_0)\right)$, where k is the smallest kth derivative of φ which $\varphi^{(k)} \neq 0$. They showed that these limiting distributions depend on the lower invelope of an integrated Brownian motion process starting at 0 minus a drift term t^{k+2}, which depends on the value of k. Fix

$x_0 \in \mathbb{R}$, let $W(t)$ be a standard Brownian motion starting from zero and define

$$Y_k(t) = \begin{cases} \int_0^t W(s)\ ds - t^{k+2}, & \text{if } t \geq 0, \\ \int_t^0 W(s)\ ds - t^{k+2}, & \text{if } t < 0. \end{cases} \tag{2.7}$$

Let $f_0 = \exp \varphi_0$ denote the true density and satisfies the following assumptions:

(a1) The density function $f_0 \in$ the class of log-concave densities $\mathcal{F}_{lcd}$.

(a2) $f_0(x_0) > 0$.

(a3) The function is at least twice continuously differentiable in a neighborhood of x_0.

(a4) If $\varphi_0''(x_0) \neq 0$, then $k = 2$, see Groeneboom et al. [2001a] and Groeneboom et al. [2001b].

(a5) The random continuous function H in Theorem 2.5 satisfying $H_k(t) \leq Y_k(t)$ for all $t \in \mathbb{R}$. Thus, the function H is everywhere below Y.

(a6) H_k has a second derivative in which H_k'' is concave. On top of that, $H_k(t) = Y_k(t)$, if the slope of H_2'' is strictly decreasing at t.

(a7) With probability 1, H is three times differentiable at $t = 0$ and $\int_{\mathbb{R}} \{Y(t) - H(t)\}\, dH'''(t) = 0$.

According to (a4), we have $\varphi_0''(x_0) \neq 0$, so $k = 2$. Therefore, we have the process of Y as

$$Y_2(t) = \begin{cases} \int_0^t W(s)ds - t^4, & \text{if } t \geq 0, \\ \int_t^0 W(s)ds - t^4, & \text{if } t < 0. \end{cases}$$

Moreover, (a1) to (a7) are also true for $k = 2$. Hence, we get Theorem 2.5.

Theorem 2.5. *[Balabdaoui et al., 2009, Corollary 2.2, page 1306] Suppose that assumptions (a1) - (a7) hold. Then,*

$$n^{2/5}\left(\widehat{f}_n(x_0) - f_0(x_0)\right) \xrightarrow{d} c_2(x_0, \varphi_0)H_2''(0)$$

and

$$n^{2/5}\left(\widehat{\varphi}_n(x_0) - \varphi_0(x_0)\right) \xrightarrow{d} C_2(x_0, \varphi_0)H_2''(0),$$

where $H_2''(0)$ is the second derivative at 0 of the invelope H of Y. The constant c_2 and C_2 are given by

$$c_2(x_0, \varphi_0) = \left(\frac{\{f_0(x_0)\}^3 |\varphi_0''(x_0)|}{24}\right)^{1/5}$$

and

$$C_2(x_0, \varphi_0) = \left(\frac{|\varphi_0''(x_0)|}{24\{f_0(x_0)\}^2}\right)^{1/5},$$

where $\varphi_0''(x_0)$ denote the second derivative of $\varphi_0(x_0)$.

2.2.6 Global rates of convergence for the log-concave ML estimator

Doss and Wellner [2016] studied the global rate of convergence for one-dimensional log-concave ML estimator. They proved that the log-concave ML estimator converges with a rate of $n^{-2/5}$ with respect to the Hellinger distance. Let Ω be any measurable space, and if $\widehat{f}_n,\ f_0$ are the estimated and true densities of the measures P, respectively. The Hellinger distance, d_H is given by

$$d_H(\widehat{f}_n, f_0) = \left[\int_\Omega \left(\sqrt{\widehat{f}_n} - \sqrt{f_0}\right)^2 dP\right]^{1/2}. \tag{2.8}$$

Let $\log N_{[]}(\varepsilon, \mathcal{F}_{M,lcd}, d_H)$ denote a bracketing entropy of an appropriate subclass $\mathcal{F}_{M,lcd}$ of log-concave densities $\mathcal{F}_{lcd}$ with respect to the Hellinger distance d_H, where $\mathcal{F}_{M,lcd} = \{f \in \mathcal{F}_{lcd} : \sup_{s\in\mathbb{R}} f(s) \le M \text{ and } 1/M \le f(s) \text{ if } s \in [-1,1]\}$. More details of the bracketing entropy can be found in Definition A.12. Doss and Wellner [2016, Theorem 3.1, page 8] showed that this bracketing entropy obtains a bound of the form

$$\log N_{[]}(\varepsilon, \mathcal{F}_{M,lcd}, d_H) \ \le \ A_M \varepsilon^{-1/2}, \tag{2.9}$$

where the constant A_M depends on M and $\varepsilon > 0$. The equation (2.9) is the main result to obtain the rate of convergence for the log-concave ML estimator. Under this bound, we get Theorem 2.6, which is similar to Doss and Wellner [2016, Theorem 3.2].

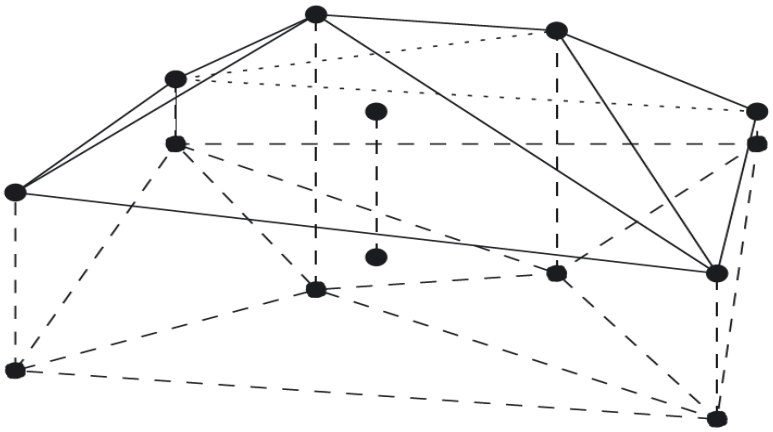


Figure 2.3: Tent-like structure for the logarithm of MLE when $d = 2$ (Figure from Cule et al. [2010])

Theorem 2.6. *Suppose that $\widehat{f}_n$ is the univariate log-concave ML density estimator of f_0, then*

$$d_H(\widehat{f}_n, f_0) = O_p(n^{-2/5}). \tag{2.10}$$

2.3 Multi-dimensional log-concave density

2.3.1 Log-concave maximum likelihood estimation

In multivariate cases, the log-concave MLE was studied by Cule et al. [2010] and Cule [2009]. Cule et al. [2010] showed that with probability one, the log-concave ML estimator $\widehat{f}_n$ of f_0 again exists and is unique. Computation of the multivariate log-

concave ML estimator is different from univariate log-concave ML estimator because we cannot write an objective function in the terms of slopes. Figure 2.3 shows an example of a two-dimensional log-concave ML estimator in a log-scale. We can view this ML estimator as pulling a tent or a sheet in the vertical way, where the heights of the tent poles represent the values of $\log \widehat{f_n}$, which is built from bivariate data that are the black dots in the Figure. However, it will be harder to visualize when we deal with data more than two dimensions, because the illustrations are not obvious. Let $X_1, \ldots, X_n$ be random samples from f_0 on $\mathbb{R}^d$ and denote $C_n = \text{conv}(X_1, \ldots, X_n)$ as a convex hull of data. According to Cule et al. [2010], an objective function for finding the ML estimator is

$$\sigma(y) = -\frac{1}{n}\sum_{i=1}^{n} y_i + \int_{C_n} \exp\left(\bar{h}_y(x)\right) dx. \tag{2.11}$$

Theorem 2.7. *[Cule et al., 2010, Theorem 3] The function σ is a convex function. It has a unique minimum at $y^* \in \mathbb{R}^n$, say, and $\log \widehat{f_n} = \bar{h}_{y^*}$.*

According to Cule et al. [2010], they called $\log \widehat{f_n}$ as a 'tent function', which is a function $\bar{h}_y : \mathbb{R}^d \to \mathbb{R}$ for a fixed vector $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$. $\bar{h}_y$ is a least concave function where

$$\bar{h}_y(x) = \inf\{h(x) : h \text{ is concave, } h(X_i) \geq y_i, i = 1, \ldots, n\}.$$

2.3.2 A computational aspect of the multivariate log-concave ML estimator

The idea is to use Shor's r-algorithm, which presented in Cule et al. [2010]. It is built for solving a convex and non-differentiable problems. This algorithm is to generate a sequence y^t, which $\sigma(y^t) \to \min_{y \in \mathbb{R}^n} \sigma(y)$ as $t \to \infty$. $\sigma(y^t)$ and $\partial\sigma(y^t)$ will be required at each iteration where $\partial\sigma(y^t)$ represents the direction moving from y^t to y^{t+1}.

Maximizing the multivariate log-concave objective function can be viewed as the infinite dimensional optimization problem. It can be reduced to the problem of maximizing function $\bar{h}_y$ for some suitable vector y. In other words, we can imagine that the function $\bar{h}_y$ is when we place the pole height y_i at X_i and pull the sheet over the top of the pole. Thus, a key for finding the log-concave ML estimator for multidimensional data is to find an appropriate vector $y^* \in \mathbb{R}^n$, where y^* comes from minimizing $\sigma(y)$ in (2.11). From this minimization problem, we will get a unique $y^* = (y_1^*, \ldots, y_n^*) \in \mathbb{R}^n$.

In order to calculate $\sigma(y)$, we need to evaluate $\int_{C_n} \exp\left(\bar{h}_y(x)\right) dx$. We can write the closed form of this integral by triangulating the convex hull of data, C_n. An example of the triangulations for $d = 2$ can be found in Figure 2.3. Each simplex represents an affine function of $\log \widehat{f}_n$. This step uses much computational time to

find a proper y^* which makes $\bar{h}_{y^*}$ as a tent function that all tent poles touch the tent.

It can also be noticed that there is an available package in R that builts from Chen et al. [2015] for finding the multidimensional log-concave ML estimator. The package is called LogConcDEAD and the useful function is "mlelcd". In this function, the stopping criteria after the $(r+1)$th iteration are given by

$$\begin{aligned}
&\left|y_i^{r+1} - y_i^r\right| \le \delta\left|y_i^r\right| \quad \text{for} \quad i = 1, \ldots, n, \\
&\left|\sigma(y^{r+1}) - \sigma(y^r)\right| \le \epsilon\left|\sigma(y^r)\right|, \\
&\Big|\int_{C_n} \exp \bar{h}_{y^r}(x)dx - 1\Big| \le \eta,
\end{aligned}$$

for some small values of δ, ϵ, and η.

In the algorithm, these tolerances has been set to $\delta = 10^{-4}, \epsilon = 10^{-8}$, and $\eta = 10^{-4}$. However, these stopping criteria can be set by a user in the mlelcd function with parameters ytol, sigmatol, and integraltol, respectively. An example code for finding the multivariate log-concave ML estimator by using the LogConcDEAD package is in Appendix A.3.2.

2.3.3 Rate of convergence

Cule [2009] showed that the multivariate ML estimator $\widehat{f_n}$ is a consistent estimator of the true density f_0. Moreover, they conjectured that the optimal rate of

convergence with respect to the Hellinger distance is $n^{-2/(d+4)}$.

Theorem 2.8. *[Cule, 2009, Theorem 5.11] Let f_0 be a log-concave density and let $\widehat{f_n}$ denote the log-concave maximum likelihood estimator. Then, with probability 1, $d_H(\widehat{f_n}, f_0) \to 0$ as $n \to \infty$.*

Moreover, Cule [2009, page 97] conjectured that $d_H(\widehat{f_n}, f_0) = O_p(\delta_n)$ where

$$\delta_n = \begin{cases} n^{-2/(d+4)} & \text{when } d < 4, \\ n^{-1/4}(\log n)^{1/2} & \text{when } d = 4, \\ n^{-1/d} & \text{when } d > 4. \end{cases}$$

Then, they use the results from Cule [2009, Section 5.2.6] and conjectured that $d_H(\widehat{f_n}, f_0) = O_p(n^{-2/(d+4)})$ for all d.

Furthermore, Kim and Samworth [2016, Theorem 5, page 2762] proved that the actual rates of convergence for the log-concave ML estimator with respect to the Hellinger distance converges up to the logarithmic factors. However, they stated the results only for $d \leq 3$.

Theorem 2.9. *[Kim and Samworth, 2016, Theorem 5, page 2762] Let $X_1, \ldots, X_n$ be i.i.d. random vectors with density $f_0 \in \mathcal{F}_d$, and let $\widehat{f_n}$ denote the corresponding*

log-concave ML estimator. Then,

$$d_H(\widehat{f}_n, \mathcal{F}_d) = \begin{cases} O(n^{-2/5}) & \text{if } d = 1, \\ O(n^{-1/3}\sqrt{\log n}) & \text{if } d = 2, \\ O(n^{-1/4}\sqrt{\log n}) & \text{if } d = 3 \end{cases}$$

where $\mathcal{F}_d$ denote the set of upper semi-continuous, log-concave densities on $\mathbb{R}^d$.

2.3.4 Computational time

As we mentioned before, the conjectured rate of convergence of the multivariate log-concave ML estimator is $n^{-2/(d+4)}$, which is computationally intensive. The running time for four-dimensional data with sample size 1,000 is 18 minutes for a 1.60GHz/8GB RAM desktop PC. Unlike, for one-dimensional data, finding the ML estimator with the ASA in the logcondens package takes under one second. Because it is a time-consuming algorithm, we propose a new method that works well with multivariate density estimation and is also applicable in practice. This method will combine the knowledge of one-dimensional log-concave MLE with a copula model, which will be presented in the next Chapter.

3 Dependence modeling with copulas

3.1 Introduction

Because of the computationally intensive problem when we find the multivariate log-concave ML estimator with Shor's r-algorithm, we propose another useful method that works with the "copula model". Copula can use to model the dependencies between variables and allows us to form a multivariate model in which its margins are modeled separately from the dependence structure. We can find the estimators of each marginal density separately. Since our marginal densities are univariate log-concave densities, their ML estimators give us a better convergence rate than the multivariate log-concave ML estimators. This is how the convergence rate can be improved.

Copula model has been widely used in several fields such as economics and finance, see Patton [2012]. He applied the copula model with time series of the stock index returns and also presented the goodness of fit test for choosing an appropriate copula

family. Rémillard et al. [2012] presented the copula model with Archimedean copulas to work with the multivariate time series on the Canadian/US exchange rate and the values of oil in the future ten-year period.

In this chapter, we show how to find the estimators under the copula model. The estimation can be done in two steps. First, we estimate the univariate log-concave marginals. Then, we estimate the copula parameters. This two-stage estimation is called inference function for margins (IFM). As we mentioned before, univariate log-concave ML estimator gives the better rate of convergence than multivariate log-concave ML estimator. Therefore, modeling under the copula model improves the performance of the density estimation in terms of the convergence rate. Moreover, it gaurantees that the convergence rate of our proposed method is much faster than $n^{-2/(d+4)}$, which is from the conjectured rate of multivariate log-concave ML estimator. However, our proposed rate is never better than the convergence rate of parametric estimator which is $n^{-1/2}$.

3.1.1 Definitions and properties

Copula is a multivariate function with uniform marginal distribution functions. Moreover, it can be called as a uniform representation or a dependence function, which describes the dependencies between each margin. Sklar [1959] introduced a

concept of copula to work with a d-dimensional distribution function F. We can split F into two parts, the marginal distribution functions F_j and the copula distribution function C with its parameters $\theta \in \mathbb{R}^k$.

Definition 3.1. The joint distribution function F is a function with its domain in $\mathbb{R}^d$ which

- F is nondecreasing.
- $F_1, \ldots, F_d$ are distribution functions.
- F has margins $F_1, \ldots, F_d$ such that $F_j(x) = F(\infty, \ldots, x_j, \ldots, \infty)$ for $j = 1, \ldots, d$.
- $F(x_1, \ldots, -\infty, \ldots, x_d) = 0$ especially for $d = 2$. $F(x, -\infty) = F(-\infty, y) = 0$ and $F(\infty, \ldots, \infty) = 1$.

Theorem 3.2. *(Sklar's theorem) Let $X = (X_1, \ldots, X_d)$ be a random vector with distribution function F and $F \in \mathcal{F}(F_1, \ldots, F_d)$ be a d-dimensional distribution function with margins $F_1, \ldots, F_d$. Then there exists a d-copula C with uniform margins such that*

$$F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)). \tag{3.1}$$

Theorem 3.2 tells us that every joint distribution function F has at least one copula function. Moreover, if C is a copula, then it is the distribution of a multivariate

uniform random vector [Joe, 1997]. Note that a copula, C, is defined as a cumulative distribution function with support in $[0,1]^d$. Moreover, a copula density function of the copula distribution function C is given by

$$c(F_1,\ldots,F_d) = \frac{\partial^d C(F_1,\ldots,F_d)}{\partial F_1 \cdots \partial F_d}. \tag{3.2}$$

Furthermore, some properties of C are as follows.

1. The copula function is always unique if all marginal functions are continuous. Conversely, if C is a d-copula with distribution function $F_1,\ldots,F_d$, then F from (3.1) is a d-dimensional distribution function with margins $F_1,\ldots,F_d$.

2. For $F_j \in [0,1]; j = 1,\ldots,d$, when $\partial^d C(F_1,\ldots,F_d)/(\partial F_1 \cdots \partial F_d)$ exists, C is absolutely continuous.

3. Every copula C is continuous and satisfies the following inequality $\left|C(F_1,\ldots,F_d) - C(G_1,\ldots,G_d)\right| \leq \sum_{j=1}^{d} \left|F_j - G_j\right|$ when $\forall 1 \leq j \leq d$, and $0 \leq F_j, G_j \leq 1$.

4. For all $1 \leq j \leq d,\ 0 \leq F_j \leq 1$, we have $C_j(F_j) = C(1,\ldots,1,F_j,\ldots,1,\ldots,1) = F_j$ and $C_j(F_j) = C(F_1,\ldots,0,\ldots,F_j,\ldots,F_d) = 0$.

5. If $g_1,\ldots,g_d$ are monotone, nondecreasing mappings of $\mathbb{R}$ in itself, any copula function of $(X_1,\ldots,X_d)$ is also a copula function of $(g_1(X_1),\ldots,g_d(X_d))$.

6. For a bivariate copula, there are some interesting properties as follows.

   - $C(1,u) = C(u,1) = u$ and $C(0,u) = C(u,0) = 0$ for all $u \in [0,1]$.
   - C is nondecreasing in each variable.
   - For every $s,t,u,v \in [0,1]$, such that $s \le t$ and $u \le v$, then

$$C(t,v) - C(t,u) - C(s,v) + C(s,u) \ge 0.$$

   - For every $s,t,u,v \in [0,1]$, C satisfies the following Lipschitz condition

$$|C(t,v) - C(s,u)| \le |t-s| + |v-u|.$$

Moreover, copulas have their universal bound called "Fréchet-Hoeffding bounds inequality" as given in Theorem 3.3.

Theorem 3.3. *Let C be any d-copula with $F_1, \ldots, F_d$ be marginal distribution functions with support in $[0,1]^d$. Then,*

$$W(F_1(x_1), \ldots, F_d(x_d)) \;\le\; C(F_1(x_1), \ldots, F_d(x_d)) \;\le\; M(F_1(x_1), \ldots, F_d(x_d))$$

where $W(F_1(x_1), \ldots, F_d(x_d)) = \max(F_1(x_1) + \cdots + F_d(x_d) - d + 1, 0)$ and $M(F_1(x_1), \ldots, F_d(x_d)) = \min(F_1(x_1), \ldots, F_d(x_d))$. Moreover, an independence copula can be expressed as $\Pi(F_1(x_1), \ldots, F_d(x_d)) = F_1(x_1) \cdots F_d(x_d)$.

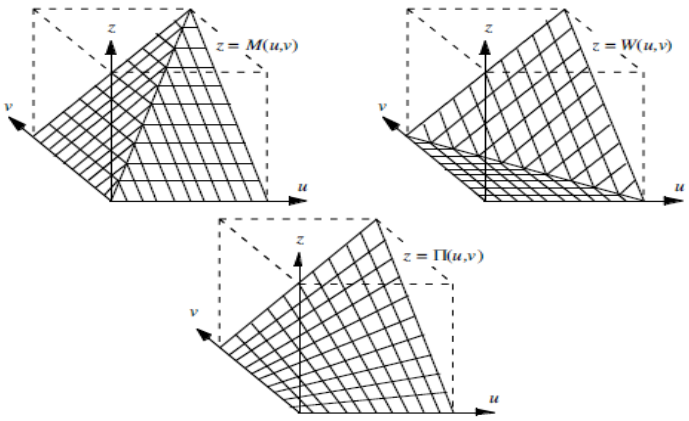


Figure 3.1: Graphics of M, W, and Π (figure from Nelsen [2006])

W and M are called lower and upper Fréchet-Hoeffding bounds. The copula $C(F_1(x_1), \ldots, F_d(x_d)) = M(F_1(x_1), \ldots, F_d(x_d))$ represents the most positive dependence. The functions M and Π are d-copulas for all $d \geq 2$. On the other hand, W is a copula only when $d \leq 2$. When $C(F_1(x_1), \ldots, F_d(x_d)) = \max(F_1(x_1) + \cdots + F_d(x_d) - d + 1, 0)$, it represents the most negative dependence. Furthermore when $X_1, \ldots, X_d$ are independent, $C(F_1(x_1), \ldots, F_d(x_d)) = \Pi(F_1(x_1), \ldots, F_d(x_d)) = F_1(x_1) \cdots F_d(x_d)$.

3.1.2 Dependence

For a set of distributions $\mathcal{F}(F_1, \ldots, F_d)$, there are several statistics for measuring the level of dependences between random variables. In our work, we first discuss concordance. Then, we present two famous measures, which are Spearman's rho and

Kendall's tau.

3.1.2.1 Concordance

The random variables with distribution function F and G are said to be concordant if the large values of F are associated with the large values of G and also the small values of F and G are being small together.

Definition 3.4. (Concordance) Let F and G be distribution functions in $\mathcal{F}(F_1, \ldots, F_d)$ where $X \sim F, Y \sim G$ and X, Y are continuous random variables such that (x_i, y_i) and (x_j, y_j) are the two observations from a vector (X, Y). Then, (x_i, y_i) and (x_j, y_j) are concordant if $x_i > x_j$ and $y_i > y_j$ or $x_i < x_j$ and $y_i < y_j$. Conversely, we say that they are discordant when $x_i > x_j$ but $y_i < y_j$ or if $x_i < x_j$ and $y_i > y_j$. The formula can be represented as $(x_i - x_j)(y_i - y_j) > 0$ for concordance and $(x_i - x_j)(y_i - y_j) < 0$ for discordance.

Theorem 3.5. *[Nelsen, 2006, Theorem 5.1.1] Let (X_1, Y_1) and (X_2, Y_2) be independent vectors of continuous random variables with joint distribution functions H_1 and H_2, respectively, with common margins F (of X_1 and X_2) and G (of Y_1 and Y_2). Let $u = F(x), v = G(y)$, and C_1 and C_2 denote the copulas of (X_1, Y_1) and (X_2, Y_2), respectively, so that $H_1(x, y) = C_1(F(x), G(y))$ and $H_2(x, y) = C_2(F(x), G(y))$. Let Q denote the difference between the probabilities of concordance and discordance of*

(X_1, Y_1) *and* (X_2, Y_2)*, i.e., let*

$$Q = P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0].$$

Then,

$$Q = Q(C_1, C_2) = 4 \int_0^1 \int_0^1 C_2(u, v) dC_1(u, v) - 1.$$

3.1.2.2 Spearman's rho

Spearman's rho correlation is based on both concordance and discordance. We will show details of this correlation via examples of three independent random vectors. Let $(X_1, Y_1), (X_2, Y_2), (X_3, Y_3)$ be three independent random vectors from a joint distribution function H where F and G are the marginal distribution functions of X and Y, respectively. The Spearman's rho for (X_1, Y_1) and (X_3, Y_2) is given by

$$\rho_C = 3\{P[(X_1 - X_3)(Y_1 - Y_2) > 0] - P[(X_1 - X_3)(Y_1 - Y_2) < 0]\}. \tag{3.3}$$

The equation (3.3) represents a probability of concordance minus a probability of discordance times a normalizing constant. Note that we can also use (X_2, Y_3) instead of (X_3, Y_2). The idea is that one vector has the joint distribution function H, which is (X_1, Y_1), and another vector (X_3, Y_2) is independent. Thus, the joint distribution function of (X_3, Y_2) is $F(x)G(y)$. The copula of (X_1, Y_1) is C and because X_3 and

Y_2 are independent, the copula of (X_3, Y_2) is Π. Then from Theorem 3.5, we get

$$Q(C, \Pi) = 4 \int_0^1 \int_0^1 uv \; dC(u, v) - 1. \tag{3.4}$$

The Spearman's rho can also be viewed as the measurement of how far from independent of the variables. To study the range of $Q(C, \Pi)$, we work on the boundaries of (3.4). The lower and upper bounds of $Q(C, \Pi)$ are given by

$$Q(W, \Pi) = 4 \int_0^1 \int_0^1 uv \; dW(u, v) - 1, \quad \text{and} \quad Q(M, \Pi) = 4 \int_0^1 \int_0^1 uv \; dM(u, v) - 1.$$

Because the support of W is the second diagonal $G(y) = 1 - F(x)$, therefore

$$\int_0^1 \int_0^1 h(u, v) \; dW(u, v) = \int_0^1 h(u, 1 - u) \; du \tag{3.5}$$

for all integrable function h, which domain is in $[0, 1]^2$. Likewise, the support of M is the main diagonal $G(y) = F(x)$ in $[0, 1]^2$. Because M has a uniform margin, then

$$\int_0^1 \int_0^1 h(u, v) \; dM(u, v) = \int_0^1 h(u, u) \; du. \tag{3.6}$$

Therefore, we have

$$Q(W, \Pi) = 4 \int_0^1 u(1 - u) \; du - 1 = -\frac{1}{3} \quad \text{and}$$
$$Q(M, \Pi) = 4 \int_0^1 u^2 \; du - 1 = \frac{1}{3}.$$

Consequently, for any copula C, $Q(C, \Pi) \in [-1/3, 1/3]$. A multiplier 3 in (3.3) is added to make $Q(C, \Pi)$ covers the whole possibility range of dependence.

Theorem 3.6. *[Nelsen, 2006, Theorem 5.1.6] Let X and Y be continuous random variables whose copula is C. Then the population version of Spearman's rho for X and Y is given by*

$$\begin{aligned}\rho_C &= 3Q(C,\Pi)\\ &= 12\int_0^1\int_0^1 uv\ dC(u,v) - 3\\ &= 12\int_0^1\int_0^1 C(u,v)\ dudv - 3\\ &= 12\int_0^1\int_0^1 \{C(u,v) - uv\}\, dudv.\end{aligned}$$

Example 3.7. Farlie-Gumbel -Morgenstern (FGM) copula

$$C(u,v) = uv + \theta uv(1-u)(1-v); \quad \theta \in [-1,1], \tag{3.7}$$

then

$$\begin{aligned}\rho_C &= 12\int_0^1\int_0^1 \{uv + \theta uv(1-u)(1-v) - uv\}\, dudv\\ &= 12\left(\frac{1}{6}\right)\int_0^1 \theta v(1-v)\ dv\\ &= \frac{\theta}{3}.\end{aligned}$$

Hence, $\rho_C \in [-1/3, 1/3]$.

3.1.2.3 Kendall's tau

A population version of Kendall's tau is also related to the concordance and discordance of random variables. Let $(X_1, Y_1), (X_2, Y_2)$ be independent and identically

distributed random vectors from the same joint distribution function H. Therefore, the population version of Kendall's tau is in the form of

$$\tau_C = P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0]. \tag{3.8}$$

Theorem 3.8. *[Nelsen, 2006, Theorem 5.1.3] Let X and Y be continuous random variables whose copula is C. Then, the population version of Kendall's tau for X and Y is given by*

$$\begin{aligned}
\tau_C &= Q(C,C) \\
&= 4\int_0^1\int_0^1 C(u,v)\ dC(u,v) - 1 \\
&= 4\int_0^1\int_0^1 C(u,v)c(u,v)\ dudv - 1.
\end{aligned}$$

The lower and upper bounds of $Q(C,C)$ can be calculated from $Q(W,W)$ and $Q(M,M)$, respectively. We use the calculations in (3.5) and (3.6), hence we get

$$\begin{aligned}
Q(W,W) &= 4\int_0^1\int_0^1 \max(u+v-1,0)\ dW(u,v) - 1 \\
&= 4\int_0^1 0\ du - 1 = -1, \\
Q(M,M) &= 4\int_0^1\int_0^1 \min(u,v)\ dM(u,v) - 1 \\
&= 4\int_0^1 u\ du - 1 = 1.
\end{aligned}$$

Therefore, $Q(C,C) \in [-1,1]$.

Example 3.9. Farlie-Gumbel -Morgenstern (FGM) copula

Refer to the FGM copula distribution function in (3.7), we get

$$\partial_u C(u,v) = v + \theta v(1-v)(1-2u)$$

$$\partial_u \partial_v C(u,v) = 1 + \theta(1-2u)(1-2v) = c(u,v).$$

Then,

$$\begin{aligned}\tau_C &= 4\int_0^1\int_0^1 \{uv + \theta uv(1-u)(1-v)\}\{1+\theta(1-2v)(1-2u)\}\,dudv - 1\\ &= 4\left(\frac{1}{4} + \frac{\theta}{18}\right) = \frac{2\theta}{9}.\end{aligned}$$

Because of $\theta \in [-1,1]$, therefore $\tau_C \in [-2/9, 2/9]$. According to Example 3.7 and 3.9, FGM has restricted usefulness because ρ_C and τ_C have the limited ranges of dependence.

Although, both Spearman's rho (ρ) and Kendall's tau (τ) are the measurements of dependence. There are some differences. First, the range of dependence that ρ_C and τ_C can cover are different as shown in Example 3.7 and 3.9. Second, Nelsen [2006] showed universal inequality for these two measures.

Theorem 3.10. *[Nelsen, 2006, Theorem 5.1.10] Let X and Y be continuous random variables, and let ρ and τ denote Spearman's rho and Kendall's tau, defined by* (3.3) *and* (3.8)*, respectively. Then,*

$$-1 \;\le\; 3\tau - 2\rho \;\le\; 1.$$

3.1.3 Copula families

In this thesis, we focus on parametric copula families. However, there are also nonparametric copulas such as empirical copula and kernel copula. Some parametric copula families contain one parameter. Some contain more than one parameter, see Durrleman et al. [2000], Nelsen [2006] and Yan [2007]. We show some examples of bivariate Gaussian and Archimedean copula families. Gaussian copula has one parameter but Archimedean copulas contain both one parameter and two parameters families. In each example, we present the copula distribution function, the copula density that derives from the representation in (3.2), and also the explicit form of Spearman's rho and Kendall's tau if they can be shown explicitly. For examples below, let u, v be the uniform representations of $F(x)$ and $G(y)$.

3.1.3.1 Gaussian copula

$$C(u,v) = N_\theta\left(\Phi^{-1}(u), \Phi^{-1}(v)\right);\ u,v \in (0,1)$$

The bivariate Gaussian copula density function can be represented as

$$c(u,v) = \frac{1}{\sqrt{\det R}} \exp\left(-\frac{1}{2}\begin{bmatrix}\Phi^{-1}(u)\\ \\ \Phi^{-1}(v)\end{bmatrix}^T \left(R^{-1} - I\right)\begin{bmatrix}\Phi^{-1}(u)\\ \\ \Phi^{-1}(v)\end{bmatrix}\right).$$

Let consider the correlation matrix $R = \begin{bmatrix} 1 & \theta \\ \theta & 1 \end{bmatrix}$ where $\det(R) = 1 - \theta^2$, then

$$\begin{aligned} R^{-1} - I &= \frac{1}{1-\theta^2}\begin{bmatrix} 1 & -\theta \\ -\theta & 1 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\ &= \frac{\theta}{1-\theta^2}\begin{bmatrix} \theta & -1 \\ -1 & \theta \end{bmatrix}. \end{aligned}$$

Therefore,

$$\begin{aligned} c(u,v) &= \frac{1}{\sqrt{1-\theta^2}} \exp\left(-\frac{1}{2}\left(\frac{\theta}{1-\theta^2}\right) \begin{bmatrix} \theta\Phi^{-1}(u) - \Phi^{-1}(v) \\ -\Phi^{-1}(u) + \theta\Phi^{-1}(v) \end{bmatrix}^T \begin{bmatrix} \Phi^{-1}(u) \\ \Phi^{-1}(v) \end{bmatrix} \right) \\ &= \frac{1}{\sqrt{1-\theta^2}} \exp\left(-\frac{\theta}{2(1-\theta^2)} \left[\theta(\Phi^{-1}(u))^2 - 2\Phi^{-1}(u)\Phi^{-1}(v) + \theta(\Phi^{-1}(v))^2 \right] \right). \end{aligned}$$

Kendall's tau is given by

$$\tau = \frac{2}{\pi}\arcsin(\theta), \qquad \text{where } \theta \text{ is the Spearman's rho.}$$

3.1.3.2 The t copula

Let $x = (x_1, x_2)^T$,

$$C(u,v) = \int_{-\infty}^{t_\delta^{-1}(u)} \int_{-\infty}^{t_\delta^{-1}(v)} \frac{\Gamma(\frac{\delta+2}{2})}{\Gamma(\frac{\delta}{2})\pi\delta\sqrt{1-\theta^2}} \left(1 + \frac{x^T R^{-1} x}{\delta}\right)^{-(\delta+2)/2} dx,$$

where t_δ^{-1} denote the quantile function of a standard univariate t_δ distribution with δ degrees of freedom, and R is the correlation matrix with off-diagonal elements equal to θ. The density of t copula has a form

$$c(u,v) = \frac{f_{\delta,\theta}\left(t_\delta^{-1}(u), t_\delta^{-1}(v)\right)}{f_\delta\left(t_\delta^{-1}(u)\right) f_\delta\left(t_\delta^{-1}(v)\right)}; u, v \in (0,1),$$

where $f_{\delta,\theta}$ is the joint density of bivariate standard t-distributed random vectors and f_δ is the standard t density function with degrees of freedom δ. The t copula has the same Spearman's rho and Kendall's tau as the Gaussian copula.

3.1.3.3 Archimedean copulas

Let $\phi^{[-1]}$ denote a pseudo-inverse of ϕ and considers ϕ as a continuously strictly decreasing convex function from $[0,1]$ to $[0,\infty]$. The function ϕ is called a generator of the copula. Thus, we have

$$\phi^{[-1]}(t) = \begin{cases} \phi^{-1}(t) & \text{for } t \in [0, \phi(0)], \\ 0 & \text{for } t \geq \phi(0). \end{cases}$$

Note that $\phi^{[-1]}$ is continuous and nonincreasing on $[0,\infty]$ but strictly decreasing on $[0,\phi(0)]$. The distribution function of Archimedean copula is given by

$$C(u,v) = \phi^{[-1]}(\phi(u) + \phi(v)). \tag{3.9}$$

We say that C is strict, when $\phi(0) = \infty$ and $C(u,v) > 0$ for all $(u,v) \in [0,1]^2$. On the contrary, C is non-strict when $\phi(0) < \infty$. When the copula C is strict, $\phi^{[-1]} = \phi^{-1}$. The copula in (3.9) is said to be a strict Archimedean copula, which equals to $\phi^{-1}(\phi(u) + \phi(v))$. Kendall's tau of the Archimedean copula has an explicit expression, which relates to their generator. An expression is given by

$$\tau = 1 + 4 \int_0^1 \frac{\phi(t)}{\phi'(t)} dt, \tag{3.10}$$

where $\phi'(t) = d\phi(t)/dt$.

We show some examples of one-parameter Archimedean copula as follows.

Example 3.11. Clayton copula

The generator is

$$\phi(t) = \frac{t^{-\theta} - 1}{\theta}, \qquad \theta \in [-1, \infty) \backslash \{0\}.$$

Then, the copula distribution function can be expressed as

$$C(u,v) = \left\{ \max(u^{-\theta} + v^{-\theta} - 1, 0) \right\}^{-1/\theta}.$$

The copula density is given by

$$c(u,v) = \begin{cases} (1+\theta)(uv)^{-\theta-1}(u^{-\theta} + v^{-\theta} - 1)^{(-1/\theta)-2} & \text{when } u^{-\theta} + v^{-\theta} > 1, \\ 0, & \text{otherwise.} \end{cases}$$

Kendall's tau derived by (3.10) can be represented as

$$\begin{aligned}\tau &= 1 + 4\int_0^1 \frac{t^{\theta+1} - t}{\theta} dt \quad \text{when } \theta \neq 0 \\ &= \frac{\theta}{\theta + 2}.\end{aligned}$$

Therefore, $\tau \in [-1, 1]$. In contrast, the closed form of ρ is complicated to find.

Example 3.12. Gumbel-Hougaard

The generator is

$$\phi(t) = (-\ln t)^\theta, \qquad \theta \in [1, \infty).$$

Thus, the copula function is

$$C(u, v) = \exp\left[-\left((-\ln u)^\theta + (-\ln v)^\theta\right)^{1/\theta}\right].$$

Note that the copula density function has a complex form, which will not be stated here. There is also no closed form for Spearman's rho whereas Kendall's tau has a simple form, which is

$$\tau = 1 - \frac{1}{\theta}.$$

However, there is a restriction that τ does not cover the negative correlations because when $\theta \in [1, \infty), \tau \in [0, 1]$.

Example 3.13. Frank copula

The generator is

$$\phi(t) = -\ln\left[\frac{e^{-\theta t}-1}{e^{-\theta}-1}\right], \qquad \theta \in (-\infty,\infty)\backslash\{0\}.$$

Hence, the copula function is

$$C(u,v) = -\frac{1}{\theta}\ln\left(1+\frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\right).$$

Then, the copula density function is given by

$$c(u,v) = \frac{\theta e^{-\theta u-\theta v}(1-e^{-\theta})}{\left[(e^{-\theta}-1)+(e^{-\theta u}-1)(e^{-\theta v}-1)\right]^2}.$$

Kendall's tau and Spearman's rho can be calculated; however, they depend on the Debye function $D_k(\theta)$, which is defined for any positive integer k by

$$D_k(\theta) = \frac{k}{\theta^k}\int_0^\theta \frac{t^k}{e^t-1}\,dt.$$

Spearman's rho can be expressed as

$$\rho \;=\; 1-\frac{12}{\theta}\left[D_1(\theta)-D_2(\theta)\right]$$

and Kendall's tau is given by

$$\tau \;=\; 1-\frac{4}{\theta}\left[1-D_1(\theta)\right].$$

Example 3.14. Joe copula

The generator is

$$\phi(t) = -\log(1-(1-t)^{\theta}), \qquad \theta \in [1,\infty).$$

Hence, the copula function is

$$C(u,v) = 1 - \left\{(1-u)^{\theta} + (1-v)^{\theta} - (1-u)^{\theta}(1-v)^{\theta}\right\}^{1/\theta}.$$

The copula density function has a complicated form. Similarly, there are no explicit formula for Kendall's tau and Spearman's rho.

3.1.3.4 Empirical copulas

Sometimes choosing one copula from the parametric copula families is not easy and may cause a misspecification problem. The nonparametric copulas are another choice that we can consider. First, we focus on empirical copulas, which first introduced by Deheuvels [1979].

Definition 3.15. [Nelsen, 2006, Definition 5.6.1] Let $\{(x_k, y_k)\}_{k=1}^{n}$ denote a sample of size n from a continuous bivariate distribution. The empirical copula C_n is given by

$$C_n\left(\frac{i}{n}, \frac{j}{n}\right) = \frac{\text{number of pairs } (x,y) \text{ in the sample with } x \le x_{(i)}, y \le y_{(j)}}{n}$$

where $x_{(i)}$ and $y_{(j)}, 1 \le i, j \le n$, denote order statistics from the sample. The empirical copula frequency c_n is given by

$$c_n\left(\frac{i}{n}, \frac{j}{n}\right) = \begin{cases} \frac{1}{n} & \text{if } (x_{(i)}, y_{(j)}) \text{ is an element of the sample,} \\ 0 & \text{otherwise.} \end{cases}$$

Theorem 3.16. *[Nelsen, 2006, Theorem 5.6.2] Let C_n and c_n be the empirical copula and the empirical copula frequency function for sample $\{(x_k, y_k)\}_{k=1}^n$, respectively. If r and t denote respectively the sample version of Spearman's rho and Kendall's tau, then*

$$r = \frac{12}{n^2 - 1} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[C_n\left(\frac{i}{n}, \frac{j}{n}\right) - \left(\frac{i}{n}\right)\left(\frac{j}{n}\right) \right],$$

$$t = \frac{2n}{n-1} \sum_{i=2}^{n} \sum_{j=2}^{n} \sum_{p=1}^{i-1} \sum_{q=1}^{j-1} \left[c_n\left(\frac{i}{n}, \frac{j}{n}\right) c_n\left(\frac{p}{n}, \frac{q}{n}\right) - c_n\left(\frac{i}{n}, \frac{q}{n}\right) c_n\left(\frac{p}{n}, \frac{j}{n}\right) \right].$$

3.1.3.5 Kernel copulas

Nagler [2017] presented a bivariate kernel copula for n i.i.d. observations, see Charpentier et al. [2007]. We denote (U_i, V_i) as the i.i.d. observations from the copula C where $i = 1, \ldots, n$. U_i and V_i are the empirical functions where

$$U_i = \frac{1}{n+1} \sum_{i=1}^{n} \mathbb{1}(X_i \le x)$$

and

$$V_i = \frac{1}{n+1} \sum_{i=1}^{n} \mathbb{1}(Y_i \le y).$$

An idea of $n+1$ is to avoid the boundary problems. Hence, the kernel copula is given by

$$c(u,v) = \frac{1}{nh^2}\sum_{i=1}^{n} K\left(\frac{u-U_i}{h}\right) K\left(\frac{v-V_i}{h}\right), \quad \text{where } (u,v) \in [0,1]^2,$$

with the kernel function K and the bandwidth parameter h as described in 1.2.2.3.

3.2 Copula Selection

Among several copulas, we need some good criteria for the copula selection. We study Akaike information criterion (AIC), Bayesian information criterion (BIC), and distance method. AIC and BIC are model selection criteria which penalized the log-likelihood function by depending on the size of parameters. For n observations with a copula C and uniform marginal distribution functions u and v, let n observations be u_{ij} where $i = 1,\ldots,n$, and $j = 1,\ldots,d$ with a log-likelihood function $\sum_{i=1}^{n} \log[c(u_{i1},\ldots,u_{id}|\theta)]$, the AIC of a d-copula density c with parameter θ is given by

$$AIC := -2\sum_{i=1}^{n} \log[c(u_{i1},\ldots,u_{id}|\theta)] + 2k,$$

where k is the size of copula parameters. Similarly, the BIC can be expressed as

$$BIC := -2\sum_{i=1}^{n} \log[c(u_{i1},\ldots,u_{id}|\theta)] + k\log(n).$$

Copula which has the minimum AIC or BIC values will be chosen. Usually, BIC has a stronger penalty term than AIC.

The distance method [Durrleman et al., 2000] is a criterion that measure distances between finite M interested copulas, $\{C_m\}_{1\leq m\leq M}$, and empirical copula, $\widehat{C}_{(T)}$. Let us consider the discrete L^2 norm, the distance formula can be represented as

$$\begin{aligned} d(\widehat{C}_{(T)}, C_m) &= \|\hat{C}_{(T)} - C_m\|_{L^2} \\ &= \left[\sum_{t_1=1}^{T}\cdots\sum_{t_b=1}^{T}\cdots\sum_{t_d=1}^{T}\left(\widehat{C}_{(T)}\left(\frac{t_1}{T},\ldots,\frac{t_b}{T},\ldots,\frac{t_d}{T}\right) - C_m\left(\frac{t_1}{T},\ldots,\frac{t_b}{T},\ldots,\frac{t_d}{T}\right)\right)^2\right]^{1/2}, \end{aligned}$$

where $1 \leq b \leq d$. The best copula is the copula that has a minimum distance.

3.3 Density estimation

3.3.1 Estimator

Suppose that we observe $X_1, \ldots, X_n$ random variables from an unknown density $f : \mathbb{R}^d \mapsto [0, \infty)$ with cumulative distribution function $F : \mathbb{R}^d \mapsto [0, 1]$. Let c represents the copula density function of a copula distribution function C. For each j where $j = 1, \ldots, d$, the density f_j are modeled as log-concave densities, which has a form $f_j(x_j) = \exp\varphi_j(x_j)$ where $\varphi_j : \mathbb{R} \mapsto [-\infty, \infty)$ is a concave function for all $j = 1, \ldots, d$, and $F_j(s) = \int_{-\infty}^{s} \exp\varphi_j(r)dr : \mathbb{R} \mapsto [0, 1]$. For simplicity, let $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$,

then the joint density function is given by

$$f(x,\theta) = c(F_1(x_1),\ldots,F_d(x_d);\theta)\prod_{j=1}^{d}\exp\varphi_j(x_j). \tag{3.11}$$

$c(F_1(x_1),\ldots,F_d(x_d);\theta)$ is a copula density function with uniform margins and parameters $\theta \in \Theta \subset \mathbb{R}^k$. To simplify notation, we write $c(F_1(x_1),\ldots,F_d(x_d);\theta) = c(F(x);\theta)$, likewise functions $F_1,\ldots,F_d$ denote $F_1(x_1),\ldots,F_d(x_d)$. Hence, the log-likelihood function of (3.11) for n observations can be defined as

$$\begin{aligned}\ell(F_1,\ldots,F_d,\theta) &= \sum_{i=1}^{n}\log c(F(x);\theta) + \sum_{i=1}^{n}\sum_{j=1}^{d}\varphi_j(x_j)\\ &= \ell_c(F_1,\ldots,F_d;\theta) + \sum_{j=1}^{d}\ell_j(F_j).\end{aligned} \tag{3.12}$$

We can present the ℓ_j as a function of F_j instead of φ_j which is easy to follow because these F_j are represented as uniform margins for copula density. (3.12) can be estimated by using the maximum likelihood estimation (MLE). However, this method is time consuming when the dimension is large. Joe [2005] introduced a useful method called inference function for margins (IFM), which estimates (3.12) in two stages. We will discuss the MLE first, after that we will discuss the IFM method.

3.3.2 One-stage estimation: maximum likelihood estimation (MLE)

From the log-likelihood in (3.12), let $\phi = (F_1, \ldots, F_d, \theta)$ denote a set of parameters and $\tilde{\phi} = (\tilde{F}_1, \ldots, \tilde{F}_d, \tilde{\theta})$ denote a set of corresponding ML estimator. Therefore,

$$\tilde{\phi} = \text{argmax } \ell(F_1, \ldots, F_d, \theta).$$

3.3.3 Two-stage estimation: inference function for margins (IFM)

IFM method is a two-stage estimation, which estimates each margin and copula density separately. Steps of IFM can be described as follows.

Step 1: We find the log-concave estimators for each margin by maximizing each $\ell_j(F_j)$ separately. Then, we get

$$\widehat{\varphi}_j := \underset{\varphi \text{ concave}}{\text{argmax}} \sum_{i=1}^{n} \varphi_j(x_{ij}).$$

We also get the corresponding $\widehat{F}_j$. After that, we plug-in $\widehat{F}_1, \ldots, \widehat{F}_d$ in (3.12). Therefore, we get

$$\ell(\widehat{F}_1, \ldots, \widehat{F}_d, \theta) = \ell_c(\widehat{F}_1, \ldots, \widehat{F}_d; \theta) + \sum_{j=1}^{d} \ell_j(\widehat{F}_j). \tag{3.13}$$

Step 2: We can clearly see that $\widehat{\theta}$ from maximizing (3.13) is the same as maximizing only $\ell_c(\widehat{F}_1, \ldots, \widehat{F}_d; \theta)$ because all marginal densities do not depend on the copula parameter θ. Hence,

$$\widehat{\theta} := \underset{\theta \in \Theta}{\text{argmax}}\ \ell_c(\widehat{F}_1, \ldots, \widehat{F}_d; \theta).$$

Generally, IFM procedure is simpler than MLE. However, IFM and MLE are the same in Gaussian copula with correlation matrix R and F_j corresponds to $N(\mu_j, \sigma_j^2)$.

3.3.4 Asymptotic relative efficiency of MLE and IFM

MLE is a well-known optimization method with good properties under regularity conditions. However, IFM is more flexible and consumes less computational time. In some situations the performance of MLE and IFM are really close to each other. For example, Kim et al. [2007] presented the simulation study of bivariate data to compare the performance between MLE and IFM via six one-parameter copulas, which are Ali–Mikhail–Haq (AMH), Clayton, Frank, Gumbel, Joe, and Plackett copulas. The marginal distributions that were used are assumed to be normal. Moreover, they studied the misspecification of marginal distribution by considering t and other skewed distributions, which are skew t and chi-square distributions. Dependences of the model are from Kendall's tau and sample sizes to be used are 40, 100, and 500. The performance is measured with an efficiency of estimated mean square error (MSE), which can be given by an estimated MSE of IFM/an estimated MSE of MLE. They showed that in a normal case, IFM and MLE are equally good. Similarly, for misspecification cases, MLE and IFM still give small values of bias.

Joe [2005] also studied the asymptotic efficiency of IFM. Let $\eta = (\alpha_1, \ldots, \alpha_d, \theta)$ be a set of parameters in the model where $\alpha_1, \ldots, \alpha_d$ represent marginal parameters and θ is a copula parameter. We can study the performance of IFM compare to MLE from asymptotic relative efficiency (ARE) of the IFM estimators $\tilde{\eta} = (\tilde{\alpha}_1, \ldots, \tilde{\alpha}_d, \tilde{\theta})$ and the ML estimators $\widehat{\eta} = (\widehat{\alpha}_1, \ldots, \widehat{\alpha}_d, \widehat{\theta})$ via covariance matrices of IFM (M) and MLE (I^{-1}), where M and I are given below. Let subscript $1, \ldots, d$ represent each dimension of the marginal distributions and subscript m is for the copula density function,

$$I = \begin{bmatrix} I_{11} & \ldots & I_{1d} & I_{1m} \\ \vdots & \ddots & \vdots & \vdots \\ I_{d1} & \ldots & I_{dd} & I_{dm} \\ I_{m1} & \ldots & I_{md} & I_{mm} \end{bmatrix},$$

where $I_{jk} = -E[\partial^2 \ell / \partial \alpha_j \alpha_k^T]$ for $1 \leq j, k \leq d$, $I_{jm} = -E[\partial^2 \ell / \partial \alpha_j \partial \theta^T]$ and $I_{mj} = I_{jm}^T$ for $j = 1, \ldots, d$. Unlike I, matrix M is more complicated.

Let $V_{jk} = E[(\partial \ell_j / \partial \alpha_j)(\partial \ell_k^T / \partial \alpha_k^T)]; 1 \leq j, k \leq d$, so that V_{jj} equal to the information

matrix of the jth marginal log-likelihood. We get

$$-A = \begin{bmatrix} V_{11} & \dots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & V_{dd} & 0 \\ I_{m1} & \dots & I_{md} & I_{mm} \end{bmatrix}.$$

Let $D = \mathrm{cov}(\partial \ell(\alpha_1, \dots, \alpha_d, \theta)/\partial \eta)$, so we get

$$D = \begin{bmatrix} V_{11} & \dots & V_{1d} & 0 \\ \vdots & \ddots & \vdots & \vdots \\ V_{d1} & \dots & V_{dd} & 0 \\ 0 & \dots & 0 & I_{mm} \end{bmatrix}.$$

Therefore, the covariance matrix of IFM is

$$M = \left(-A^{-1}\right) D \left(-A^{-1}\right)^T.$$

For an analytic purpose, an estimated covariance matrix for $\tilde{\eta}$ and $\widehat{\eta}$ are $n^{-1}\tilde{M}$ and $(n-1)^{-1}\widehat{I}^{-1}$, respectively, where $\tilde{M}$ is the consistent estimator of M and $\widehat{I}$ is the observed Fisher information matrix of $\widehat{\eta}$. Hence, the ARE of $\tilde{\eta}$ and $\widehat{\eta}$ is given by

$$ARE(\tilde{\eta}, \widehat{\eta}) = \frac{(n-1)\tilde{M}}{n\hat{I}^{-1}}.$$

Then, the conclusions are

$$ARE(\tilde{\eta}, \widehat{\eta}) \begin{cases} < 1, & \tilde{\eta} \text{ is more efficient than } \widehat{\eta}. \\ = 1, & \text{the efficiency of } \tilde{\eta} \text{ and } \widehat{\eta} \text{ are the same.} \\ > 1, & \tilde{\eta} \text{ is less efficient than } \widehat{\eta}. \end{cases}$$

Furthermore, covariance matrices of IFM and MLE are the same when independence copula is approached. Let θ_I denote the parameter values of independence copula. Then, Theorem 3.17 is applied.

Theorem 3.17. *[Joe, 2005] As $\theta \to \theta_I$, under the usual regularity conditions for maximum likelihood, $M - I^{-1} \to 0$. That is, the covariance matrix for the IFM estimator becomes the same as the covariance matrix of the MLE when the independence copula is approached.*

3.4 Our work

An objective is to study the performance of multivariate density estimation by using the copula model and univariate log-concave marginals. We do some simulation studies, which are in Chapter 4, by using the two-stage IFM as our density estimation method. As mentioned before, the joint density function is in the form of (3.11) where its marginal density fucntions are estimated with univariate log-concave distributions

and the copula density is from one of the six well-known copulas, which are Gaussian, t, Clayton, Gumbel, Frank, and Joe. Details of all copulas are already stated in Section 3.1.3.

We also work on the asymptotic behavior of the copula estimator $\widehat{\theta}$. Because the marginal density function is log-concave, which is a nonparametric density, and we work on the parametric copula. Hence, our problem turns out to be the semi-parametric model. We prove that the copula estimator $\widehat{\theta}$ is consistent and converges at rate $n^{-1/2}$. Moreover, we also prove that our proposed joint density estimator is consistent and converges with rate $n^{-2/5}$. We will show in Chapter 5 for the details of our proofs and also some necessary assumptions and regularity conditions.

4 Simulation study

4.1 Density estimation

A goal of this simulation study is to show the performance of multivariate density estimation by using parametric copulas with log-concave marginals. Recalling that our proposed model is given in the form of

$$f(x,\theta) = c(F_1(x_1),\ldots,F_d(x_d);\theta)\prod_{j=1}^{d}\exp\varphi_j(x_j),$$

where $c(F_1(x_1),\ldots,F_d(x_d);\theta)$ is a copula selected with the BIC as described in Section 3.2. We use inference functions for margins (IFM) as our density estimation method, which its details are described in Section 3.3.3. The copula in each simulation is selected from six well-known one parameter copulas, which can be summarized in Table 4.1. In the simulation studies, we use a Gaussian copula as our true copula.

We compare our proposed method with other density estimation methods. They consist of parametric, semiparametric, and nonparametric density estimations, which

Table 4.1: Copulas in the simulation study

Copula	$C(u,v) =$	$\theta \in$
Gaussian	$N_\theta\left(\Phi^{-1}(u), \Phi^{-1}(v)\right)$	[-1,1]
t	$\int_{-\infty}^{t_\delta^{-1}(u)} \int_{-\infty}^{t_\delta^{-1}(v)} \frac{\Gamma(\frac{\delta+2}{2})}{\Gamma(\frac{\delta}{2})\Pi\delta\sqrt{1-\theta^2}} \left(1+\frac{x^T R^{-1} x}{\delta}\right)^{-(\delta+2)/2} dx$	[-1,1]
Clayton	$\{\max(u^{-\theta}+v^{-\theta}-1,0)\}^{-1/\theta}$	$[-1,\infty)\backslash\{0\}$
Gumbel-Hougaard	$\exp\left[-\left((-\ln u)^\theta+(-\ln v)^\theta\right)^{1/\theta}\right]$	$[1,\infty)$
Frank	$-\frac{1}{\theta}\ln\left(1+\frac{(e^{-\theta u}-1)(e^{-\theta v}-1)}{e^{-\theta}-1}\right)$	$(-\infty,\infty)\backslash\{0\}$
Joe	$1-\{(1-u)^\theta+(1-v)^\theta-(1-u)^\theta(1-v)^\theta\}^{1/\theta}$	$[1,\infty)$

details of each method are presented below. To simplify, we create the short terms for each method, which are given in the brackets with italic text after their details.

1. Parametric methods

   - We use parametric MLE with parametric marginal densities. *(parametric MLE)*
   - We do the same as previous method but the IFM method is used instead of the MLE. *(parametric IFM)*

2. Semiparametric methods

- (**proposed method**) We use IFM method with univariate log-concave marginals. *(log-concave IFM)*
- We use IFM method with univariate kernel marginals. We use Gaussian kernel with least-squares cross-validation bandwidth (h). The copula densities are still selected with BIC. *(kernel IFM).*
- We use IFM method with univariate kernel marginals. We use Gaussian kernel with bandwidth from Goldenshluger-Lepski (G-L) method, (see [Chagny, 2016, Section 4.4, page 115]). *(kernel IFM with G-L)*

3. Nonparametric methods

- We do multivariate log-concave density estimation. *(multivariate log-concave)*
- We do multivariate kernel density estimation with Gaussian kernel and least-squares cross-validation bandwidth matrix (H). *(multivariate kernel)*

The concept of G-L method is to select a reasonable bandwidth $\widehat{h}$ among $(\widehat{f}_h)_{h\in\mathcal{H}_n}$. The objective is to choose $\widehat{h}$ that satisfies the following criterion;

$$\widehat{h} = \underset{h\in\mathcal{H}_n}{\operatorname{argmin}}\ Crit_{GL}(h),$$

with $Crit_{GL}(h) = A(h) + V(h)$, where

$$A(h) = \max_{h' \in \mathcal{H}_n} \left(\|\widehat{f}_{h'} - \widehat{f}_{h,h'}\|^2 - V(h') \right)_+ ,$$
$$V(h) = \frac{(\kappa' \|K\|_1^2 \|K\|^2)}{nh}.$$

We choose $\mathcal{H}_n = \{2^{-k}; k = 1, 2, \ldots, \log_2(n)\}$, which satisfies the assumptions in [Chagny, 2016, Theorem 4.5 page 116]. Moreover, there is no systematic way for choosing a constant κ', so we use $\kappa' = 1$. The definition of $x_+, \widehat{f}_{h,h'}$, $\|K\|_1^2$, and $\|K\|^2$ are as follows.

$$x_+ = \max(x, 0), \qquad \widehat{f}_{h,h'} = \int_{\mathbb{R}} K_h(x - x') \widehat{f}_{h'}(x') dx'$$
$$\|K\|_1^2 = \int_{\mathbb{R}} |K(x)| dx, \qquad \|K\|^2 = \int_{\mathbb{R}} K(x)^2 dx.$$

We perform simulation studies for $d = 2, 4, 5$, and 6 with various sample sizes $n = 100, 200, 500, 1000$ and different levels of dependence $\tau = 0, 0.2, 0.6$. We consider both symmetric and skew marginal densities. The details of these simulation studies are in Table 4.2.

However, in the kernel IFM with G-L, we consider only for $d = 2$ cases. Each element in $\mathcal{H}_n$ is considered as a candiate for a bandwidth selection. The results are shown in Figure 4.1 and 4.2.

Furthermore, we also highlight the performance of our proposed method under misspecification. We choose t distribution because it is not a log-concave distribution

Table 4.2: Details of specification cases

Dimensions	Distributions	τ
2	N(0,1), N(0,1)	$0, 0.2, 0.6$
	N(0,1), Γ(2,1)	
	Γ(2,1), Γ(2,1)	
4	N(0,1), Γ(2,1), N(0,1), $\Gamma(2,1)$	$0.2, 0.6$
	Γ(2,1), Γ(2,1), Γ(2,1), $\Gamma(2,1)$	0.2
	Γ(2,1), Γ(2,1), β(2,5), β(2,5)	0.6
5*	N(0,1), Γ(2,1), Γ(2,1), Γ(2,1), N(0,1)	$0.2, 0.6$
	Γ(2,1), Γ(2,1), β(2,5), β(2,5), Γ(2,1)	
6*	N(0,1), Γ(2,1), Γ(2,1), Γ(2,1), N(0,1), Γ(2,1)	$0.2, 0.6$
	Γ(2,1), Γ(2,1), β(2,5), β(2,5), Γ(2,1), Γ(2,1)	

* means there are no multivariate log-concave results because of the computationally intensive problem.

for any degrees of freedom. Details of simulation studies under misspecification are in Table 4.3.

Moreover, we also perform some simulations when the copula densities are from two-parameter copula families. In each set of simulation, we choose one copula from these four copulas, which are BB1 (Clayton-Gumbel), BB6 (Joe-Gumbel), BB7 (Joe-Clayton), and BB8 (Joe-Frank), (see Brechmann and Schepsmeier [2013]). The simulation results can be found in Figure 4.7.

Because finding the MLE for multivariate log-concave distribution is time-consuming, so we do 100 sets of simulation. In our simulation, we use some available packages in

Table 4.3: Details of misspecification cases

Dimensions	Distributions	τ
2	t_2, t_2	0
4	N(0,1), $\Gamma(2,1), t_2, t_2$	0.2
5*	N(0,1), Γ(2,1), Γ(2,1), β(2,5), t_2	0.6
6*	N(0,1), Γ(2,1), Γ(2,1), β(2,5), β(2,5), t_2	0.6

* means there are no multivariate log-concave results because of the computationally intensive problem.

R. For kernel density estimation, we use "kedd" [Guidoum, 2015] and "ks" [Duong, 2017] packages to find bandwidth estimators for univariate and multivariate kernel density estimations, respectively. Package "copula" [Hofert et al., 2017] is used for estimating copula densities and finding copula estimators. For copula selection, we use "BiCopSelect" in VineCopula package [Schepsmeier et al., 2018], which can be accessible from CRAN `https://github.com/tnagler/VineCopula`. For finding the log-concave MLE, the packages are already stated in Chapter 2.

Steps for these simulation studies can be summarized as follows:

1. We choose the best copula among six well-known copulas with BIC. Copula which has the smallest BIC will be chosen.

2. We estimate the density by using methods represented above.

3. We calculate the mean integrated square error (MISE), which is our criterion

for the performance measurement. Let $\widehat{f_n}(x)$ and $f_0(x)$ be the estimated and the true density functions of $x \in \mathbb{R}^d$. The estimated MISE is given by

$$\widehat{MISE} = \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}^d}\left(\widehat{f_n}(x) - f_0(x)\right)^2 dx. \tag{4.1}$$

Moreover, (4.1) can be estimated by Riemann sum. Let $\Delta x = \Delta x_1 \cdots \Delta x_d$, x_{ij} be the ith observation of dimension jth, and $\Delta x_j = x_{ij} - x_{i-1,j}; i = 1, \ldots, n, j = 1, \ldots, d$, then the estimated MISE can be expressed as

$$\widehat{MISE} = \frac{1}{n}\sum_{i=1}^{n}\left(\widehat{f_n}(x_i) - f_0(x_i)\right)^2 \Delta x.$$

The simulation results from Figure 4.1, 4.2, 4.3, 4.4, and 4.5 can be summarized in the following bullet points.

- When sample size increases, $\widehat{MISE}$ decreases regardless of the level of dependences and marginal density functions. This makes sense because when the sample size is large, $\widehat{f_n}$ of each density estimation method converges to the true density function f_0. This follows by the law of large number (LLN).

- When we consider the results by type of methods, we can conclude that parametric models, which are parametric MLE and parametric IFM, perform almost the same and they give the best performance. Log-concave IFM performs better than kernel IFM while nonparametric models, which are multivariate log-concave and multivariate kernel, are the worst.

- The best to the worst results are parametric MLE, parametric IFM, log-concave IFM, and multivariate kernel, regardless of the marginal distributions, levels of dependence, and sample sizes. The performance of kernel IFM and multivariate log-concave are close to each other and usually lie between the log-concave IFM and the multivariate kernel.

- Among the semiparametric and nonparametric models, log-concave IFM provides the best results. However, it is worse than parametric MLE and parametric IFM.

- The kernel IFM with G-L method performs well when sample sizes are moderate ($n = 100, 200$), but gives higher MISE when sample sizes are large. However, it is still better than multivariate kernel.

- In general, higher dependence gives higher $\widehat{MISE}$.

For misspecification cases, results are shown in Figure 4.6 which can be summarized as follows:

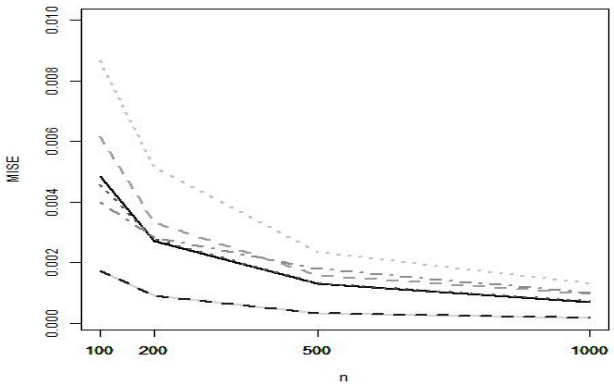
- It is obvious that the proposed method does not perform as good as the nonparametric methods which makes sense because the marginal densities are not log-concave densities. However, it is still better than nonparametric density estimation for moderate sample sizes. This similar result has been presented

Table 4.4: % of how often each copula has been selected with BIC

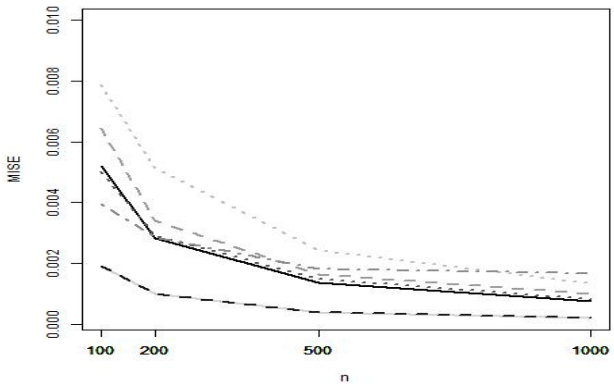
Figure	Gaussian	t	Clayton	G-H	Frank	Joe
4.1 (a)	40.25	0.75	2.25	3	**41.75**	12
4.1 (b)	**54.75**	0.25	1.25	13.75	23.5	6.5
4.1 (c)	**85**	0.75	3.25	7	3.75	0.25
4.1 (d)	40.25	1.25	1.75	2.75	**42.5**	11.5
4.1 (e)	**54.75**	0.25	0.75	13.5	24	6.75
4.1 (f)	**85**	0.75	3.25	7	3.75	0.25
4.2 (g)	40.5	1.25	1	2.75	**42.5**	12
4.2 (h)	**55.25**	0.25	0.25	13.25	24	7
4.2 (i)	**82.25**	1	4	7.25	5.5	0
4.3 (j)	**95.75**	0	1.5	1.75	1	0
4.3 (k)	**97.5**	2.25	0	0	0.25	0
4.3 (l)	**82**	0.25	0.25	3	11.5	3
4.3 (m)	**97.5**	2.25	0	0	0.25	0
4.4 (n)	**98.5**	0	0	1.25	0.25	0
4.4 (o)	**93.5**	6.25	0	0	0.25	0
4.4 (p)	**90**	4.75	0	0	0.25	0
4.4 (q)	**98.5**	0	0	1.25	0.25	0
4.5 (r)	**98.75**	0	0	0.5	0.75	0
4.5 (s)	**97.75**	2.25	0	0	0	0
4.5 (t)	**98.75**	0	0	0.5	0.75	0
4.5 (u)	**96.25**	3.75	0	0	0	0
4.6 (v)	**58.5**	0.25	0.75	1.75	31.5	7.25
4.6 (w)	**96**	0	1.5	1.5	1	0
4.6 (x)	**94.5**	5.25	0	0	0.25	0
4.6 (y)	**97.25**	2.75	0	0	0	0

before in Cule et al. [2010, Figure 3, 4]. Moreover, this result also stated the robustness property in Cule et al. [2010, Theorem 4].

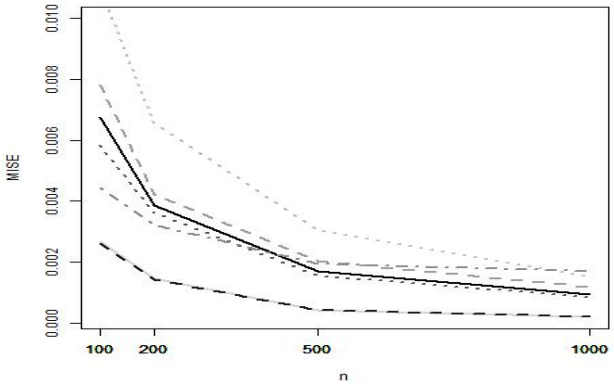
Furthermore, Table 4.4 shows percentages of how often each copula has been used in each Figure. Copulas that have been selected with the highest percentage will be presented as the bold numbers.



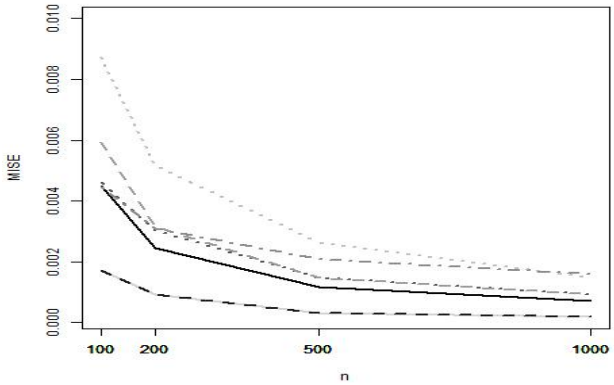
(a) $\tau = 0$, N(0,1), N(0,1)



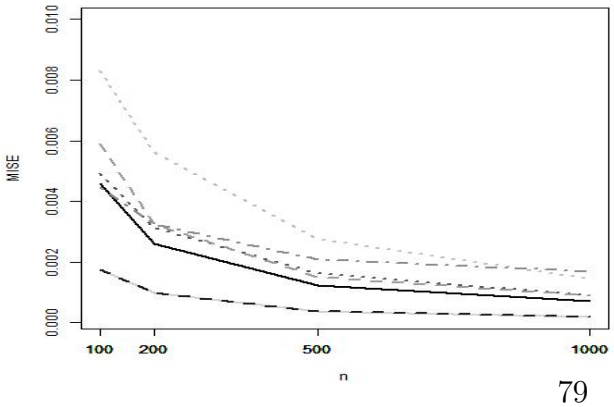
(b) $\tau = 0.2$, N(0,1), N(0,1)



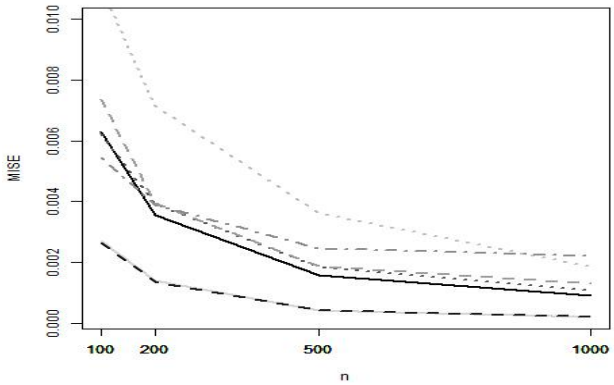
(c) $\tau = 0.6$, N(0,1), N(0,1)



(d) $\tau = 0$, N(0,1), $\Gamma(2,1)$

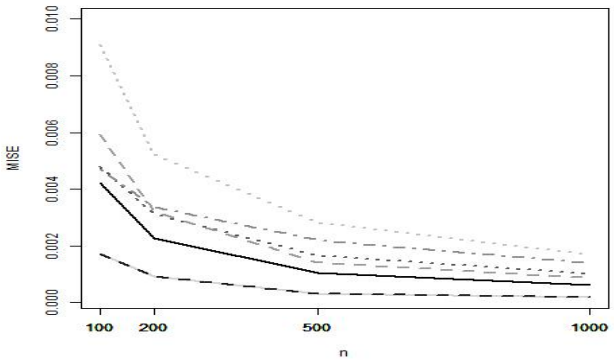


(e) $\tau = 0.2$, N(0,1), $\Gamma(2,1)$

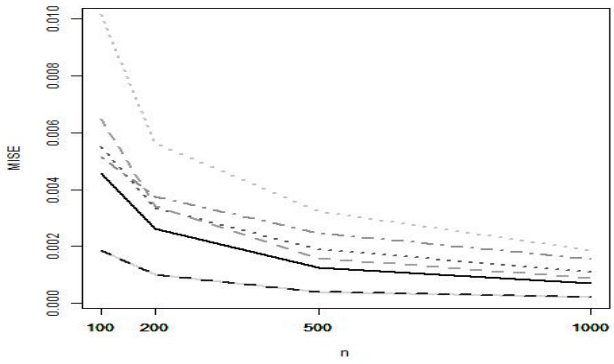


(f) $\tau = 0.6$, N(0,1), $\Gamma(2,1)$

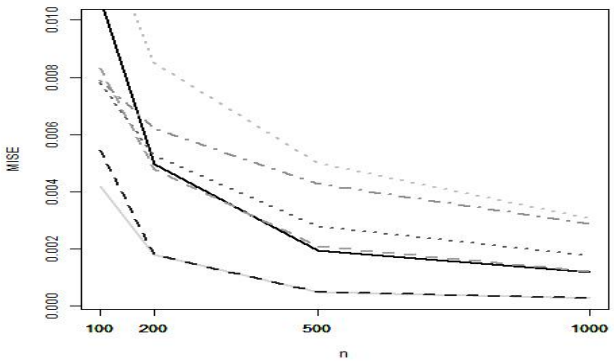
Figure 4.1: MISE for $d = 2$ —— MLE, - - - parametric IFM, —— log-concave IFM, ····· kernel IFM, -·-· kernel IFM with G-L, - - - multivariate log-concave, and ····· multivariate kernel



(g) $\tau = 0$, $\Gamma(2,1)$, $\Gamma(2,1)$

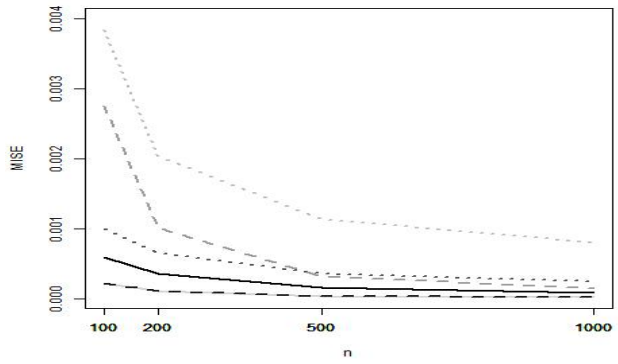


(h) $\tau = 0.2$, $\Gamma(2,1)$, $\Gamma(2,1)$

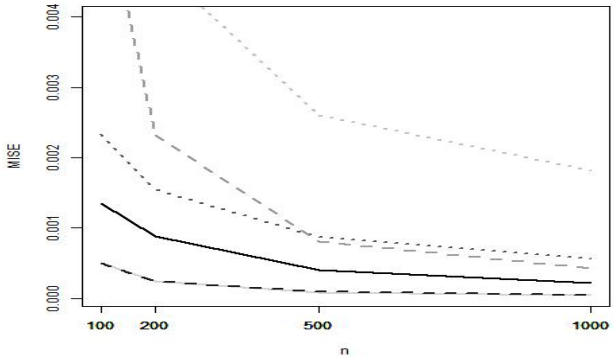


(i) $\tau = 0.6$, $\Gamma(2,1)$, $\Gamma(2,1)$

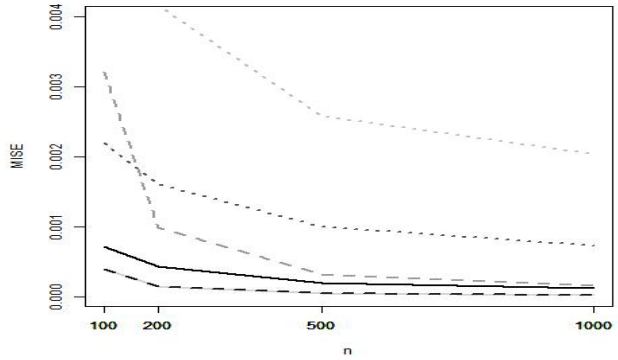
Figure 4.2: MISE for $d = 2$ —— MLE, - - - parametric IFM, —— log-concave IFM, ····· kernel IFM, -·-· kernel IFM with G-L, - - - multivariate log-concave, and ····· multivariate kernel



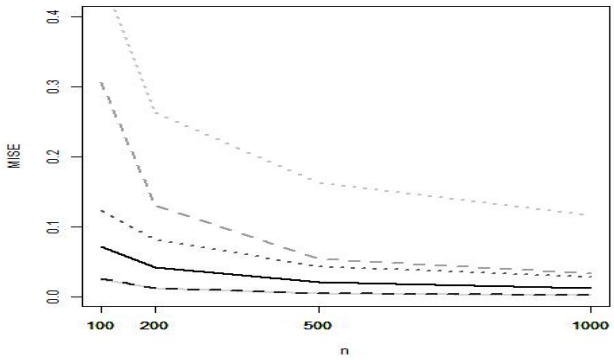
(j) τ=0.2,N(0,1),Γ(2,1),N(0,1),$\Gamma(2,1)$



(k) τ=0.6,N(0,1), Γ(2,1),N(0,1),$\Gamma(2,1)$

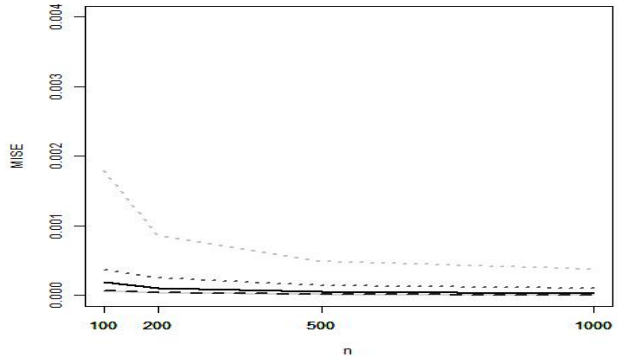


(l) τ=0.2,Γ(2,1), Γ(2,1),Γ(2,1),$\Gamma(2,1)$

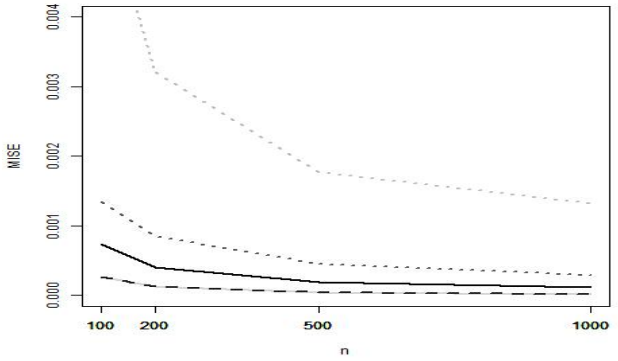


(m) τ=0.6,Γ(2,1),Γ(2,1),β(2,5),β(2,5)

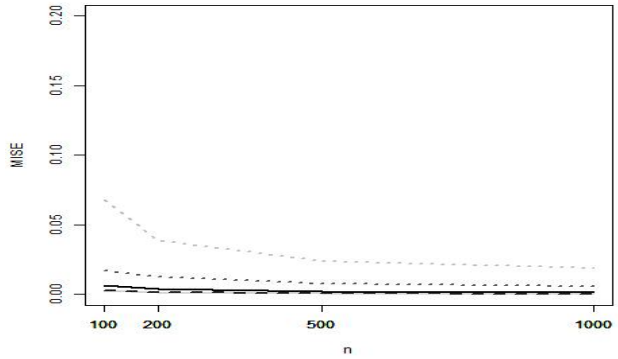
Figure 4.3: MISE for d = 4 —— MLE, - - - parametric IFM, ═══ log-concave IFM, ····· kernel IFM, - - - multivariate log-concave, and ····· multivariate kernel.



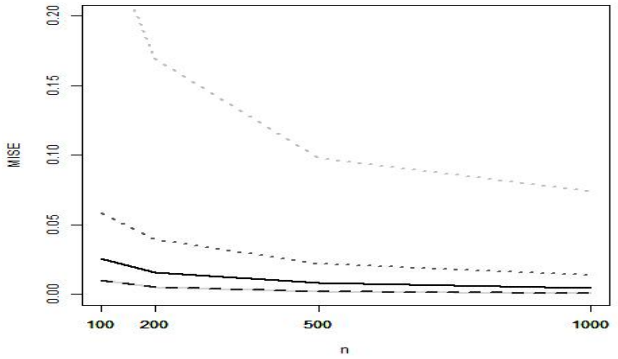
(n) τ=0.2,N(0,1),Γ(2,1),Γ(2,1),Γ(2,1),N(0,1)



(o) τ=0.6,N(0,1),Γ(2,1),Γ(2,1),Γ(2,1),N(0,1)

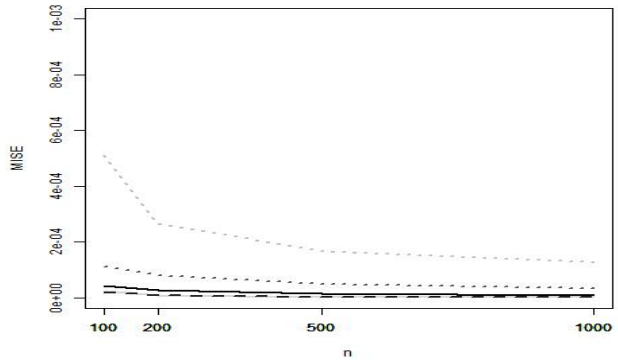


(p) τ=0.2,Γ(2,1),Γ(2,1),β(2,5),β(2,5),Γ(2,1)

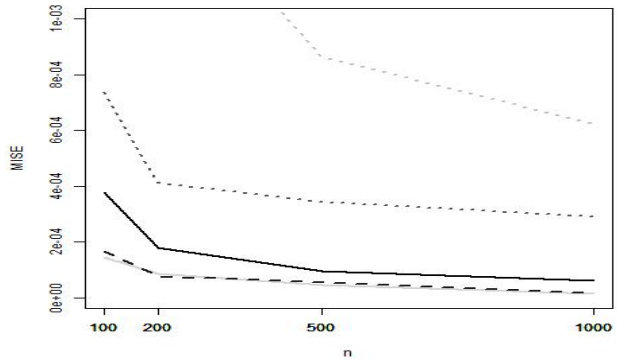


(q) τ=0.6,Γ(2,1),Γ(2,1),β(2,5),β(2,5),Γ(2,1)

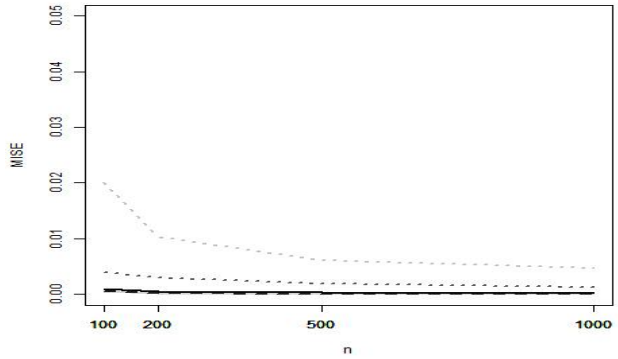
Figure 4.4: MISE for d = 5 —— MLE, - - - parametric IFM, ——— log-concave IFM, ····· kernel IFM, and ····· multivariate kernel.



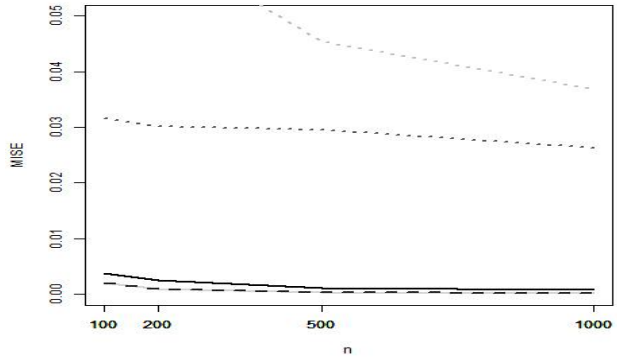
(r) τ=0.2,N(0,1),Γ(2,1),Γ(2,1),Γ(2,1),N(0,1),Γ(2,1)



(s) τ=0.6,N(0,1),Γ(2,1),Γ(2,1),Γ(2,1),N(0,1),Γ(2,1)

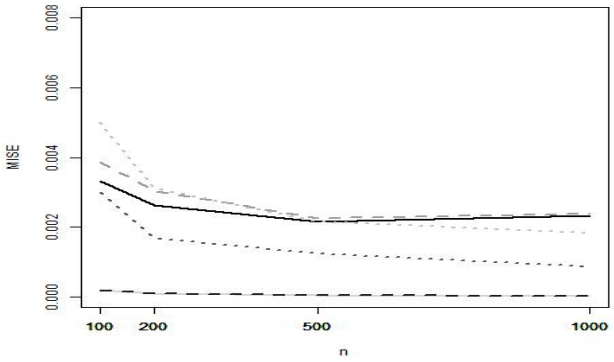


(t) τ=0.2,Γ(2,1),Γ(2,1),β(2,5),β(2,5),Γ(2,1),Γ(2,1)

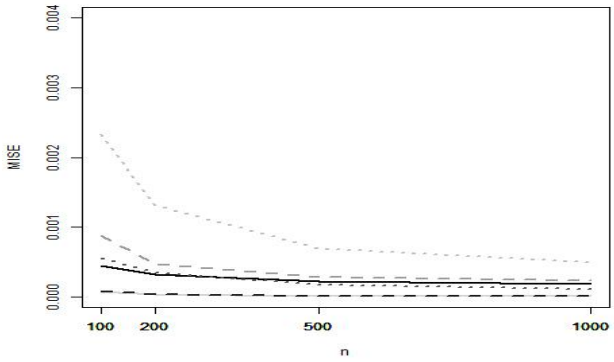


(u) τ=0.6,Γ(2,1),Γ(2,1),β(2,5),β(2,5),Γ(2,1),Γ(2,1)

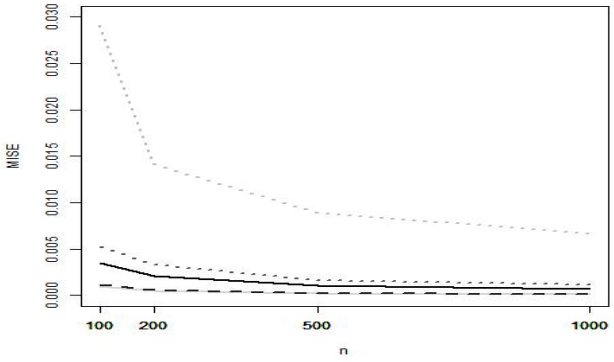
Figure 4.5: MISE for d = 6 —— MLE, - - - parametric IFM, ——— log-concave IFM, ····· kernel IFM, and ····· multivariate kernel.



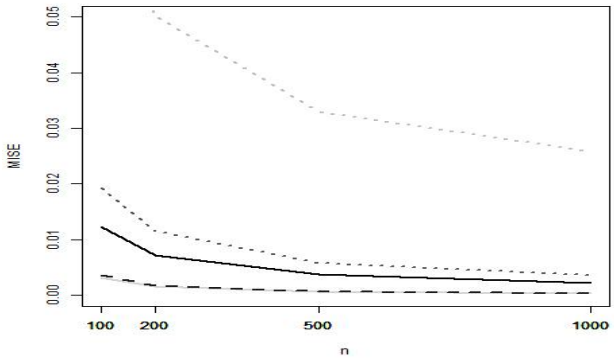
(v) $\tau = 0, t_2, t_2$



(w) $\tau = 0.2, N(0,1), \Gamma(2,1), t_2, t_2$

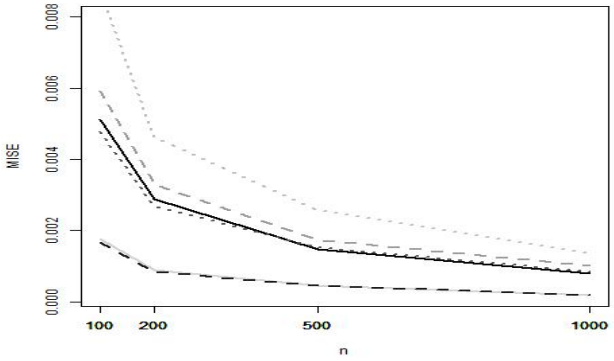


(x) τ=0.6,N(0,1),Γ(2,1),Γ(2,1),β(2,5),t_2

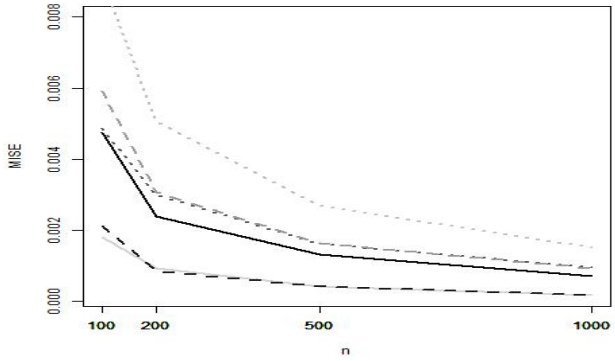


(y) τ=0.6,N(0,1),Γ(2,1),Γ(2,1),β(2,5),β(2,5),t_2

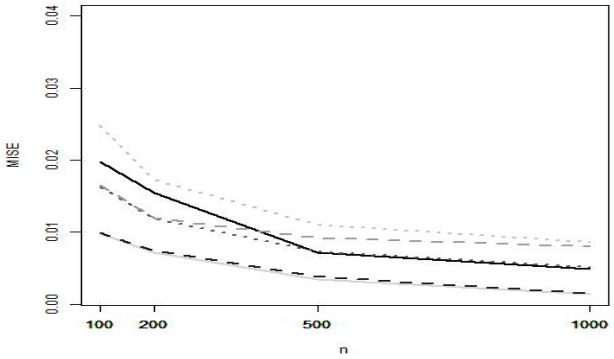
Figure 4.6: MISE from top left to bottom right: $d = 2$, $d = 4$, $d = 5$, and $d = 6$ —— MLE, - - - parametric IFM, —— log-concave IFM, ····· kernel IFM, - - - multivariate log-concave, and ····· multivariate kernel.



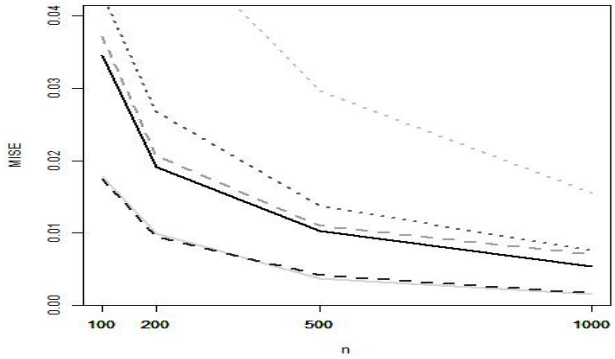
$\tau = 0.2$, N(0,1), N(0,1)



$\tau = 0.6$, N(0,1), Γ(2,1)



$\tau = 0.6$, Γ(2,1), Γ(2,1)



(q) $\tau = 0.6$, Γ(2,1), β(2,5)

Figure 4.7: MISE for two-parameter copula when $d = 2$ — MLE, - - - parametric IFM, — log-concave IFM, ····· kernel IFM, and ····· multivariate kernel.

5 Main theorem and proof

5.1 Define estimators

Suppose that we observe $X_1, \ldots, X_n$ random variables from an unknown density $f : \mathbb{R}^d \mapsto [0, \infty)$ with cumulative distribution function $F : \mathbb{R}^d \mapsto [0, 1]$. Let c represents the copula density function of copula distribution function C with parameters $\theta \in \Theta \subset \mathbb{R}^k$. For each j where $j = 1, \ldots, d$, the density f_j are modeled as log-concave densities, which has the form $f_j(x) = \exp \varphi_j(x_j)$ where $\varphi_j : \mathbb{R} \mapsto [-\infty, \infty)$ is a concave function for all $j = 1, \ldots, d$, and $F_j(s) = \int_{-\infty}^{s} \exp \varphi_j(r) dr : \mathbb{R} \mapsto [0, 1]$. For simplicity, let $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$. We write $c(F_1(x_1), \ldots, F_d(x_d); \theta) = c(F(x); \theta)$, likewise functions $F = (F_1, \ldots, F_d)$ denote $F(x) = (F_1(x_1), \ldots, F_d(x_d))$. Then, the log-likelihood function is given by

$$\begin{aligned} \ell(F_1, \ldots, F_d, \theta) &= \sum_{i=1}^{n} \log c(F_1(x_{i1}), \ldots, F_d(x_{id}); \theta) + \sum_{i=1}^{n} \sum_{j=1}^{d} \log f_j(x_{ij}) \\ &= \ell_c(F_1, \ldots, F_d; \theta) + \sum_{j=1}^{d} \ell_j(F_j) \end{aligned}$$

where

$$\ell_c(F_1, \ldots, F_d; \theta) = \sum_{i=1}^{n} \log c(F_1(x_{i1}), \ldots, F_d(x_{id}); \theta),$$
$$\ell_j(F_j) = \sum_{i=1}^{n} \log f_j(x_{ij}).$$

Let $\mathcal{F}_{lcd}$ denote the class of log-concave densities $f : \mathbb{R} \mapsto [0, \infty)$. We also let $\overline{\mathcal{F}}_{lcd}$ denote the class of corresponding CDF where

$$\overline{\mathcal{F}}_{lcd} = \left\{ F : F(s) = \int_{-\infty}^{s} f(r)dr, f \in \mathcal{F}_{lcd} \right\}$$

where $\overline{\mathcal{F}}_{lcd}$ is an infinite-dimensional space with supremum norm, that is $\|F\|_\infty = \sup_{s \in \mathbb{R}} |F(s)|$. We define the univariate log-concave density estimators for each $j = 1, \ldots, d$ as

$$\widehat{f}_{nj} = \underset{f_j \in \mathcal{F}_{lcd}}{\operatorname{argmax}} \ \ell_j(F_j)$$

with $\widehat{F}_{nj}(s) = \int_{-\infty}^{s} \widehat{f}_{nj}(r)dr$ where $s \in \mathbb{R}$. Next, we estimate the copula parameters θ. Finding the copula estimators from maximizing $\ell_c(\widehat{F}_{n1}, \ldots, \widehat{F}_{nd}; \theta)$ is also a solution of a score function. That is,

$$\Psi_n(\theta, \widehat{F}) = n^{-1} \partial_\theta \ell_c(\widehat{F}_{n1}, \ldots, \widehat{F}_{nd}; \theta) = 0. \tag{5.1}$$

We call the estimators $\widehat{\theta}_n$ from (5.1) as Z-estimator.

Hence, the joint density estimator with copula estimator $\widehat{\theta}_n$, univariate log-concave density estimators $\widehat{f}_{nj}$, and corresponding $\widehat{F}_{nj}$ can be represented as

$$\widehat{f}_n(x) = c(\widehat{F}_{n1}(x_1), \ldots, \widehat{F}_{nd}(x_d); \widehat{\theta}_n) \prod_{j=1}^{d} \widehat{f}_{nj}(x_j).$$

5.2 Main theoretical results

Let $f_0(x)$ denote the true joint density function with true copula parameters θ_0 and true univariate log-concave marginal f_{0j}. Hence, the true joint density function is given by

$$f_0(x) = c(F_{01}(x_1), \ldots, F_{0d}(x_d); \theta_0) \prod_{j=1}^{d} f_{0j}(x_j),$$

where $f_{0j} \in \mathcal{F}_{lcd}$ and consequently $F_{0j} \in \overline{\mathcal{F}}_{lcd}$.

For asymptotic results, let $\Psi(\theta, F) = E_{f_0}[\partial_\theta \log c(F(x); \theta)]$ denote the asymptotic version of $\Psi_n(\theta, F)$. We first prove consistency and rate of convergence for $\widehat{\theta}_n$. Then, we prove rate of convergence for joint density estimator. To prove these theorems, we use empirical processes theory and also the covering numbers and bracketing numbers. The proofs are done under some regularity conditions which have been stated in Section 5.2.4. From the definition in van der Vaart and Wellner [1996], let $(\Omega, \mathcal{A}, P)$ denote an arbitrary probability space. We use p^* to denote the "in outer probability" where the outer probability P^* of an arbitrary subset B of Ω is

$P^*(B) = \inf\{P(A) : A \supset B, A \in \mathcal{A}\}$. Our results are in outer probability, as is typical of empirical process theory results.

First of all, we need to define the notations of derivatives that will be used in the following theorems. Let $\partial_F \Psi_n(\theta, F)[h]$ denote the $d-$dimensional vector with jth element given by

$$[\partial_F \Psi_n(\theta, F)[h]]_j = n^{-1} \sum_{i=1}^{n} \partial_{u_j} \partial_\theta \log c(F(x_i); \theta) h_j(x_{ij}).$$

The supremum norm $\|F\|_\infty$ can also be written as

$$\|F\|_\infty = \max_{j=1,\ldots,d} \sup_{s \in \mathbb{R}} |F_j(s)|.$$

Also the $\|\theta\|_\infty = \max_{l=1,\ldots,k} |\theta_l|$.

5.2.1 Consistency

Theorem 5.1. *The estimators $\widehat{f}_n, \widehat{f}_{n1}, \ldots, \widehat{f}_{nd}$ and $\widehat{\theta}_n$ are consistent:*

1. *[Dümbgen and Rufibach, 2009, Theorem 4.2] The log-concave estimators satisfy*

$$\int \left|\widehat{f}_{nj}(s) - f_{0j}(s)\right| ds \to 0 \quad \text{and} \quad \sup_{s \in \mathbb{R}} \left|\widehat{F}_{nj}(s) - F_{0j}(s)\right| \to 0,$$

   in probability, for each $j = 1, \ldots, d$.

2. *Assume that (U), (B) and that the regularity conditions (R1), (R2), (R3) hold. Then $\widehat{\theta}_n \to \theta_0$ in outer probability.*

3. Assume that (U), (B) and that the regularity conditions (R1), (R2), (R3), (R8), (R9) hold. Then,

$$\int \left|\widehat{f}_n(x) - f_0(x)\right| dx \to 0$$

in probability.

Proof of Theorem 5.1. Consistency of the log-concave density estimators follows well-known results from Dümbgen and Rufibach [2009, Corollary 4.2, page 48]. Next, we prove the consistency result for copula estimator $\widehat{\theta}_n$. In order to establish the result, we follow Nan and Wellner [2013, Lemma 1, page 1157]. We need to check three conditions from this Lemma.

(i) For the first condition, we assume that θ_0 is the unique solution to $\Psi(\theta, F_0) = 0$ which is the assumption (U).

(ii) Secondly, $\|\widehat{F}_n - F_0\|_\infty = o_{p^*}(1)$

Thus, we need to show that $\|\widehat{F}_n - F_0\|_\infty = \max\limits_{j=1,\ldots,d} \sup\limits_{s\in\mathbb{R}} \left|\widehat{F}_{nj}(s) - F_{0j}(s)\right| \to 0$. This has already been proved in Dümbgen and Rufibach [2009, Corollary 4.2, page 48].

(iii) Finally, we prove that

$$\sup_{\theta\in\Theta, \|F-F_0\|\leq\delta_n} \frac{|\Psi_n(\theta, F) - \Psi(\theta, F_0)|}{1 + |\Psi_n(\theta, F)| + |\Psi(\theta, F_0)|} = o_{p^*}(1)$$

for every sequence $\{\delta_n\} \downarrow 0$. We will show a stronger version of this condition by showing that

$$\sup_{\theta \epsilon \Theta, \|F - F_0\| \leq \delta_n} |\Psi_n(\theta, F) - \Psi(\theta, F_0)| = o_{p^*}(1).$$

First, we show that

$$\begin{aligned}
|\Psi_n(\theta, F) - \Psi(\theta, F_0)| &= |\Psi_n(\theta, F) - \Psi_n(\theta, F_0) + \Psi_n(\theta, F_0) - \Psi(\theta, F_0)| \\
&= \left| n^{-1} \sum_{i=1}^{n} \partial_\theta \log c(F(x_i); \theta) - n^{-1} \sum_{i=1}^{n} \partial_\theta \log c(F_0(x_i); \theta) \right. \\
&\qquad \left. + n^{-1} \sum_{i=1}^{n} \partial_\theta \log c(F_0(x_i); \theta) - E_{f_0}[\partial_\theta \log c(F_0(X); \theta)] \right| \\
&\leq \left| n^{-1} \sum_{i=1}^{n} \partial_\theta \log c(F(x_i); \theta) - n^{-1} \sum_{i=1}^{n} \partial_\theta \log c(F_0(x_i); \theta) \right| \\
&\qquad + \left| n^{-1} \sum_{i=1}^{n} \partial_\theta \log c(F_0(x_i); \theta) - E_{f_0}[\partial_\theta \log c(F_0(X); \theta)] \right|.
\end{aligned}$$

We will prove these two terms in the last inequality separately. For the first

term, we use Taylor's expansion. Then, we get

$$
\begin{aligned}
|\Psi_n(\theta, F) - \Psi_n(\theta, F_0)| &\leq n^{-1} \sum_{i=1}^{n} |\partial_\theta \log c(F(x_i); \theta) - \partial_\theta \log c(F_0(x_i); \theta)| \\
&= n^{-1} \sum_{i=1}^{n} \Bigg| \sum_{j=1}^{d} \partial_{u_j} \partial_\theta \log c(F_0(x_i); \theta)(F_j - F_{0j})(x_{ij}) \\
&\qquad + 2^{-1} \sum_{j,l=1}^{d} \partial_{u_j} \partial_{u_l} \partial_\theta \log c(F^*(x_i); \theta)(F_j - F_{0j})(x_{ij})(F_l - F_{0l})(x_{il}) \Bigg| \\
&\leq n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{d} \left| \partial_{u_j} \partial_\theta \log c(F_0(x_i); \theta)(F_j - F_{0j})(x_{ij}) \right| \\
&\qquad + (2n)^{-1} \sum_{i=1}^{n} \sum_{j,l=1}^{d} \left| \partial_{u_j} \partial_{u_l} \partial_\theta \log c(F^*(x_i); \theta)(F_j - F_{0j})(x_{ij})(F_l - F_{0l})(x_{il}) \right| \\
&\leq n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{d} \left| \partial_{u_j} \partial_\theta \log c(F_0(x_i); \theta) \right| \|F_j - F_{0j}\|_\infty \\
&\qquad + (2n)^{-1} \sum_{i=1}^{n} \sum_{j,l=1}^{d} \left| \partial_{u_j} \partial_{u_l} \partial_\theta \log c(F^*(x_i); \theta) \right| \|F_j - F_{0j}\|_\infty \|F_l - F_{0l}\|_\infty \\
&\leq n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{d} \left| \partial_{u_j} \partial_\theta \log c(F_0(x_i); \theta) \right| \|F - F_0\|_\infty \\
&\qquad + (2n)^{-1} \sum_{i=1}^{n} \sum_{j,l=1}^{d} \left| \partial_{u_j} \partial_{u_l} \partial_\theta \log c(F^*(x_i); \theta) \right| \|F - F_0\|_\infty^2.
\end{aligned}
$$

By assumptions (R1), (R2), and applying law of large numbers, we can show

that

$$\begin{aligned}
\sup_{\theta\in\Theta}|\Psi_n(\theta,F)-\Psi_n(\theta,F_0)| &\leq \sup_{\theta\in\Theta}\left\{n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{d}\left|\partial_{u_j}\partial_\theta \log c(F_0(x_i);\theta)\right|\right\}\|F-F_0\|_\infty + \\
&\qquad \sup_{\theta\in\Theta}\left\{(2n)^{-1}\sum_{i=1}^{n}\sum_{j,l=1}^{d}\left|\partial_{u_j}\partial_{u_l}\partial_\theta \log c(F^*(x_i);\theta)\right|\right\}\|F-F_0\|_\infty^2 \\
&\leq \sum_{j=1}^{d}\max_{j=1,\ldots,d}\sup_{\theta\in\Theta}\left\{n^{-1}\sum_{i=1}^{n}\left|\partial_{u_j}\partial_\theta \log c(F_0(x_i);\theta)\right|\right\}\|F-F_0\|_\infty + \\
&\qquad \sum_{j,l=1}^{d}\max_{j=1,\ldots,d}\sup_{\theta\in\Theta}\left\{(2n)^{-1}\sum_{i=1}^{n}\left|\partial_{u_j}\partial_{u_l}\partial_\theta \log c(F^*(x_i);\theta)\right|\right\}\|F-F_0\|_\infty^2 \\
&\leq 2dM_1\|F-F_0\|_\infty + d^2M_2\|F-F_0\|_\infty^2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\sup_{\theta\in\Theta,\|F-F_0\|\leq\delta_n}|\Psi_n(\theta,F)-\Psi_n(\theta,F_0)| &\leq 2dM_1\delta_n + d^2M_2\delta_n^2 \\
&\leq B\delta_n
\end{aligned}$$

for constant $B = 2dM_1 + d^2M_2\delta_n$. Hence, when $\delta_n \downarrow 0$,

$$\sup_{\theta\in\Theta,\|F-F_0\|\leq\delta_n}|\Psi_n(\theta,F)-\Psi_n(\theta,F_0)| = o_{p^*}(1).$$

For the second term, we use law of large numbers. For a fixed value of θ,

$$|\Psi_n(\theta,F_0)-\Psi(\theta,F_0)| = \left|n^{-1}\sum_{i=1}^{n}\partial_\theta \log c(F_0(x_i);\theta) - E_{f_0}[\partial_\theta \log c(F_0(X);\theta)]\right| = o_{p^*}(1).$$

Next, we will prove that the convergence is uniformly in θ. We need to show that the class of $\{\partial_\theta \log c(F_0(x);\theta)\}_{\theta\in\Theta}$ is Glivenko-Cantelli where $\theta \in \Theta \subset \mathbb{R}^k$,

see Lemma A.4. Hence, we need to show that $N_{[]}(\varepsilon, \Theta, L_1(P)) < \infty$. From the well-known result in van der Vaart [1998, Example 19.7, page 271],

$$N_{[]}(\varepsilon, \Theta, L_r(P)) \quad \leq \quad \left(\frac{2\,\mathrm{diam}(\Theta)}{\varepsilon}\right)^k < \infty,$$

where k is the dimension of Θ. To prove that this bracketing number is finite, we need assumption (B) and (R3). Therefore,

$$\sup_{\theta \in \Theta} |\Psi_n(\theta, F_0) - \Psi(\theta, F_0)| = o_{p^*}(1).$$

Then,

$$\sup_{\theta \in \Theta, \|F - F_0\| \leq \delta_n} |\Psi_n(\theta, F) - \Psi(\theta, F_0)| = o_{p^*}(1).$$

This complete all three conditions. Hence, $\widehat{\theta}_n \to \theta_0$ in outer probability.

Next, we prove consistency for the joint density estimator.

$$\begin{aligned}
&\int \left|\widehat{f}_n(x) - f_0(x)\right| dx \\
&= \int \left| c(\widehat{F}_n(x); \widehat{\theta}_n) \prod_{j=1}^d \widehat{f}_{nj}(x_j) - c(F_0(x); \theta_0) \prod_{j=1}^d f_{0j}(x_j) \right| dx \\
&= \int \left| c(\widehat{F}_n(x); \widehat{\theta}_n) \prod_{j=1}^d f_{0j}(x_j) - c(F_0(x); \theta_0) \prod_{j=1}^d f_{0j}(x_j) - c(\widehat{F}_n(x); \widehat{\theta}_n) \prod_{j=1}^d f_{0j}(x_j) + c(\widehat{F}_n(x); \widehat{\theta}_n) \prod_{j=1}^d \widehat{f}_{nj}(x_j) \right| dx \\
&\leq \int \left| c(\widehat{F}_n(x); \widehat{\theta}_n) - c(F_0(x); \theta_0) \right| \prod_{j=1}^d f_{0j}(x_j) dx + \int \left| \prod_{j=1}^d \widehat{f}_{nj}(x_j) - \prod_{j=1}^d f_{0j}(x_j) \right| c(\widehat{F}_n(x); \widehat{\theta}_n)\, dx \\
&= \int \left| c(\widehat{F}_n(x); \widehat{\theta}_n) - c(F_0(x); \theta_0) \right| \prod_{j=1}^d f_{0j}(x_j) dx + \sum_{j=1}^d \int \left| \widehat{f}_{nj}(x_j) - f_{0j}(x_j) \right| \left\{ \prod_{i<j} \widehat{f}_{ni}(x_i) \right\} \left\{ \prod_{i>j} f_{0i}(x_i) \right\} c(\widehat{F}_n(x); \widehat{\theta}_n)\, dx.
\end{aligned}$$

Now, each $\widehat{f}_{nj}$ is consistent and each f_{0j} is bounded since it is log-concave, and hence for large enough n we assume that for some $B < \infty, \left\{ \prod_{i<j} \widehat{f}_{ni}(x_i) \right\} \left\{ \prod_{i>j} f_{0i}(x_i) \right\} \leq$

B^{d-1}. Therefore,

$$\begin{aligned}
&\int \left|\widehat{f}_{nj}(x_j) - f_{0j}(x_j)\right|\Big\{\prod_{i<j}\widehat{f}_{ni}(x_i)\Big\}\Big\{\prod_{i>j} f_{0i}(x_i)\Big\}c(\widehat{F}_n(x);\widehat{\theta}_n)dx \\
&\quad\leq\quad B^{d-1}\int \left|\widehat{f}_{nj}(x_j) - f_{0j}(x_j)\right|c(\widehat{F}_n(x);\widehat{\theta}_n)dx \\
&\quad\leq\quad B^{d-1}\left\{\int \left|\widehat{f}_{nj}(x_j) - f_{0j}(x_j)\right|^2 dx\right\}^{1/2}\left\{\int \left(c(\widehat{F}_n(x);\widehat{\theta}_n)\right)^2 dx\right\}^{1/2} \\
&\quad\leq\quad B^{d-1}\left\{\int \left|\widehat{f}_{nj}(x_j) - f_{0j}(x_j)\right|\left|\widehat{f}_{nj}(x_j) - f_{0j}(x_j)\right|dx\right\}^{1/2}\left\{\int \left(c(\widehat{F}_n(x);\widehat{\theta}_n)\right)^2 dx\right\}^{1/2} \\
&\quad\leq\quad B^{d-1}\left\{2B\int \left|\widehat{f}_{nj}(x_j) - f_{0j}(x_j)\right|dx\right\}^{1/2}\left\{\int \left(c(\widehat{F}_n(x);\widehat{\theta}_n)\right)^2 dx\right\}^{1/2} \\
&\quad\leq\quad B^{d-1}\left\{2B\int \left|\widehat{f}_{nj}(x_j) - f_{0j}(x_j)\right|dx\right\}^{1/2}\left\{\sup_{(F,\theta)\in\mathcal{N}}\int \left(c(F(x);\theta)\right)^2 dx\right\}^{1/2} \\
&\quad\leq\quad B_2^{1/2}B^{d-1}\left\{2B\int \left|\widehat{f}_{nj}(x_j) - f_{0j}(x_j)\right|dx\right\}^{1/2}
\end{aligned}$$

with the consistency of $\widehat{F}_n$, $\widehat{\theta}_n$, and (R8), $\sup_{(F,\theta)\in\mathcal{N}}\int \left(c(F(x);\theta)\right)^2 dx$ is bounded with B_2. Then, we combine this result with the consistency of $\widehat{f}_{nj}(x_j)$ for each $j = 1,\ldots,d$. Therefore,

$$\int \left|\widehat{f}_{nj}(x_j) - f_{0j}(x_j)\right|\Big\{\prod_{i<j}\widehat{f}_{ni}(x_i)\Big\}\Big\{\prod_{i>j} f_{0i}(x_i)\Big\}c(\widehat{F}_n(x);\widehat{\theta}_n)dx \quad = \quad o_{p^*}(1).$$

For the first part of the last inequality of $\int \left|\widehat{f}_n(x) - f_0(x)\right|dx$, we use Taylor's

expansion and the fact that f_{0j} is bounded. Hence,

$$\begin{aligned}
&\int \left|c(\widehat{F}_n(x);\widehat{\theta}_n) - c(F_0(x);\theta_0)\right| \prod_{j=1}^{d} f_{0j}(x_j)dx \\
&\leq \int \left|\sum_{j=1}^{d} \partial_{u_j} c(F^*(x);\theta^*)(\widehat{F}_{nj} - F_{0j})(x_j) + \sum_{l=1}^{k} \partial_{\theta_l} c(F^*(x);\theta^*)(\widehat{\theta}_{n,l} - \theta_{0l})\right| \prod_{j=1}^{d} f_{0j}(x_j)dx \\
&\leq B^d \left\{\sum_{j=1}^{d} \int \left|\partial_{u_j} c(F^*(x);\theta^*)(\widehat{F}_{nj} - F_{0j})(x_j)\right|dx + \sum_{l=1}^{k} \int \left|\partial_{\theta_l} c(F^*(x);\theta^*)(\widehat{\theta}_{nl} - \theta_{0l})\right|dx\right\} \\
&\leq B^d \sum_{j=1}^{d} \left\{\int \left|\partial_{u_j} c(F^*(x);\theta^*)\right|dx\right\} \|\widehat{F}_{nj} - F_{0j}\|_\infty + B^d \sum_{l=1}^{k} \left\{\int \left|\partial_{\theta_l} c(F^*(x);\theta^*)\right|dx\right\} |\widehat{\theta}_{nl} - \theta_{0l}| \\
&\leq B^d \sum_{j=1}^{d} \left\{\int \left|\partial_{u_j} c(F^*(x);\theta^*)\right|dx\right\} \|\widehat{F}_n - F_0\|_\infty + B^d \sum_{l=1}^{k} \left\{\int \left|\partial_{\theta_l} c(F^*(x);\theta^*)\right|dx\right\} \|\widehat{\theta}_n - \theta_0\|_\infty \\
&\leq B_3 \left\{\|\widehat{F}_n - F_0\|_\infty + \|\widehat{\theta}_n - \theta_0\|_\infty\right\}, \qquad (5.2)
\end{aligned}$$

with a new constant $B_3 = 2dB^dD_1 + 2kB^dD_2$. The integral terms are bounded with (R9). This result combines with the consistency of $\widehat{F}_n$ and $\widehat{\theta}_n$. Then,

$$\int \left|c(\widehat{F}_n(x);\widehat{\theta}_n) - c(F_0(x);\theta_0)\right| \prod_{j=1}^{d} f_{0j}(x_j)dx \quad = \quad o_{p^*}(1).$$

Hence, we can conclude that

$$\int \left|\widehat{f}_n(x) - f_0(x)\right|dx = o_{p^*}(1).$$

□

5.2.2 Rate of convergence

Theorem 5.2. *Assume that consistency, (U), (B), (LC), and that the regularity conditions (R1), (R2), (R3), (R4), (R5), (R6), (R7) hold. Then, $\widehat{\theta}_n$ is $\sqrt{n}-$consistent*

and

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = \sqrt{n}\left\{-\dot{\Psi}_\theta(\theta_0, F_0)\right\}^{-1}\left\{(\Psi_n - \Psi)(\theta_0, F_0) + \dot{\Psi}_F(\theta_0, F_0)[\widehat{F}_n - F_0]\right\} + o_{p^*}(1).$$

Note that from the simulation study in Section 5.2.3.3, Theorem 5.2 still holds without (LC).

Theorem 5.3. *Assume that the conditions of Theorem 5.1 and 5.2 hold. Then,*

$$\int \left|\widehat{f}_n(x) - f_0(x)\right| dx \quad = \quad O_{p^*}(n^{-2/5}).$$

Proving Theorem 5.2, we need some lemmas as follows:

Lemma 5.4.

$$\|\widehat{F}_n - F_0\|_\infty = O_{p^*}(n^{-2/5})$$

Proof of Lemma 5.4.

$$\|\widehat{F}_n - F_0\|_\infty \quad = \quad \max_{j=1,\ldots,d} \sup_{s\in\mathbb{R}} \left|\widehat{F}_{nj}(s) - F_{0j}(s)\right|$$

It is enough to show that $\sup_{s\in\mathbb{R}} \left|\widehat{F}_{nj}(s) - F_{0j}(s)\right| = O_{p^*}(n^{-2/5})$ for any $j = 1, \ldots, d$. From [Gibbs and Su, 2002], we have the relationship of $d_K(F, G) \leq d_H(f, g)$ where $d_K(F, G)$ and $d_H(f, g)$ are defined in Definition A.6 and (2.8), respectively. Moreover, from Theorem 2.6, we have $d_H(\widehat{f}_{nj}, f_{0j}) = O_p(n^{-2/5})$ for each $j = 1, \ldots, d$. Hence,

$$d_K(\widehat{F}_{nj}, F_{0j}) = \sup_{s\in\mathbb{R}} \left|\widehat{F}_{nj}(s) - F_{0j}(s)\right| \leq d_H(\widehat{f}_{nj}, f_{0j}) = O_p(n^{-2/5}).$$

Therefore, we can conclude that $\|\widehat{F}_n - F_0\|_\infty = O_{p^*}(n^{-2/5})$ □

Lemma 5.5. *Let $\mathcal{F}_{M,c}$ denote the class of functions*

$$\mathcal{F}_{M,c} \quad = \quad \{\partial_\theta \log c(F(x);\theta) : \theta \in \Theta, F \in \overline{\mathcal{F}}_{M,lcd}\},$$

where

$$\mathcal{F}_{M,lcd} \quad = \quad \{f \in \mathcal{F}_{lcd} : \sup_{s\in\mathbb{R}} f(s) \le M \text{ and } 1/M \le f(s) \text{ if } s \in [-1,1]\}.$$

and $\overline{\mathcal{F}}_{M,lcd} = \{F(x) = \int_{-\infty}^{x} f(s)ds : f \in \mathcal{F}_{M,lcd}\}$. Assume that (B), (R6), and (R7) hold. Then there exists a constant A such that

$$\log N_{[]}(\varepsilon, \mathcal{F}_{M,c}, L_2(P)) \quad \le \quad A\varepsilon^{-1/2}.$$

Proof of Lemma 5.5. We provide the details when $k, d = 2$, as the proof generalizes easily. Let $\psi(\theta, F(x)) = \partial_\theta \log c(F(x);\theta)$. By Taylor's expansion, we have that

$$\psi(\theta, F(x)) = \psi(\theta_0, F_0(x)) + \sum_{l=1}^{2} \partial_{\theta_l}\psi(\theta^*, F^*(x))(\theta_l - \theta_{0l}) + \sum_{j=1}^{2} \partial_{u_j}\psi(\theta^*, F^*(x))(F_j - F_{0j})(x_j).$$

Then,

$$\psi(\theta_2, F_2(x)) - \psi(\theta_1, F_1(x)) \quad = \quad \sum_{l=1}^{2} \partial_{\theta_l}\psi(\theta^*, F^*(x))(\theta_{2l} - \theta_{1l}) + \sum_{j=1}^{2} \partial_{u_j}\psi(\theta^*, F^*(x))(F_{2j} - F_{1j})(x_j).$$

Next, we need to find the bracketing number for the class $\mathcal{F}_{M,c}$. Let $N(\varepsilon_\theta, \Theta, \|\cdot\|_\infty)$, $N(\varepsilon, \overline{\mathcal{F}}_{M,lcd}, \|\cdot\|_\infty)$ denote covering numbers of the sets Θ and $\overline{\mathcal{F}}_{M,lcd}$. Since, we consider the case where $k = 2$, we cover the two spaces of Θ. Let θ_1^i, θ_2^j denote centers of the balls of the two coverings of Θ with $\varepsilon_\theta = \varepsilon_{\theta_1}, \varepsilon_{\theta_2}$. Also we consider the case where $d = 2$. We cover two spaces $\overline{\mathcal{F}}_{M,lcd}$. Therefore, let F_1^k, F_2^l denote the

centers of the balls of the two coverings of $\overline{\mathcal{F}}_{M,lcd}$ with $\varepsilon = \varepsilon_1, \varepsilon_2$, respectively. This differentiation is not technically necessary, but it makes the exposition below slightly easier to follow. Then,

$$\begin{aligned}
&\psi(\theta, F(x)) \\
&\quad = \quad \psi(\theta_1, \theta_2, F_1(x), F_2(x)) \\
&\quad = \quad \psi(\theta_1, \theta_2, F_1(x), F_2(x)) - \psi(\theta_1^i, \theta_2^j, F_1^k(x), F_2^l(x)) + \psi(\theta_1^i, \theta_2^j, F_1^k(x), F_2^l(x)) \\
&\quad = \quad \psi(\theta_1^i, \theta_2^j, F_1^k(x), F_2^l(x)) + \partial_{\theta_1}\psi(\theta^*, F^*(x))(\theta_1 - \theta_1^i) + \partial_{\theta_2}\psi(\theta^*, F^*(x))(\theta_2 - \theta_2^j) \\
&\qquad\qquad + \partial_{u_1}\psi(\theta^*, F^*(x))(F_1 - F_1^k)(x_1) + \partial_{u_2}\psi(\theta^*, F^*(x))(F_2 - F_2^l)(x_2).
\end{aligned}$$

Let $\theta_1, \theta_2, F_1, F_2$ be such that each lies in the covering ball with center $\theta_1^i, \theta_2^j, F_1^k, F_2^l$. From assumption (R6) it therefore follows that

$$\begin{aligned}
\psi(\theta, F(x)) \quad &\geq \quad \psi(\theta_1^i, \theta_2^j, F_1^k(x), F_2^l(x)) - \xi_{\theta_1}(x)\varepsilon_{\theta_1} - \xi_{\theta_2}(x)\varepsilon_{\theta_2} - \xi_1(x)\varepsilon_1 - \xi_2(x)\varepsilon_2, \\
\psi(\theta, F(x)) \quad &\leq \quad \psi(\theta_1^i, \theta_2^j, F_1^k(x), F_2^l(x)) + \xi_{\theta_1}(x)\varepsilon_{\theta_1} + \xi_{\theta_2}(x)\varepsilon_{\theta_2} + \xi_1(x)\varepsilon_1 + \xi_2(x)\varepsilon_2.
\end{aligned}$$

Hence, $\psi(\theta, F) = \psi(\theta_1, \theta_2, F_1, F_2)$ is inside the bracket $[l_{ijkl}, u_{ijkl}]$ where

$$\begin{aligned}
l_{ijkl}(x) \quad &= \quad \psi(\theta_1^i, \theta_2^j, F_1^k(x), F_2^l(x)) - \xi_{\theta_1}(x)\varepsilon_{\theta_1} - \xi_{\theta_2}(x)\varepsilon_{\theta_2} - \xi_1(x)\varepsilon_1 - \xi_2(x)\varepsilon_2, \\
u_{ijkl}(x) \quad &= \quad \psi(\theta_1^i, \theta_2^j, F_1^k(x), F_2^l(x)) + \xi_{\theta_1}(x)\varepsilon_{\theta_1} + \xi_{\theta_2}(x)\varepsilon_{\theta_2} + \xi_1(x)\varepsilon_1 + \xi_2(x)\varepsilon_2.
\end{aligned}$$

The size of the bracket is

$$\begin{aligned}
&\int (u_{ijk}(x) - l_{ijk}(x))^2 f_0(x)dx \\
&= \int (2\xi_{\theta_1}(x)\varepsilon_{\theta_1} + 2\xi_{\theta_2}(x)\varepsilon_{\theta_2} + 2\xi_1(x)\varepsilon_1 + 2\xi_2(x)\varepsilon_2)^2 f_0(x)dx \\
&\leq 16\varepsilon_{\theta_1}^2 \int \xi_{\theta_1}^2(x) f_0(x)dx + 16\varepsilon_{\theta_2}^2 \int \xi_{\theta_2}^2(x) f_0(x)dx + 16\varepsilon_1^2 \int \xi_1^2(x) f_0(x)dx + 16\varepsilon_2^2 \int \xi_2^2(x) f_0(x)dx \\
&\leq \varepsilon^2
\end{aligned}$$

if $\varepsilon_{\theta_1}^2 = \varepsilon^2/(64\int \xi_{\theta_1}^2(x) f_0(x)dx)$, $\varepsilon_{\theta_2}^2 = \varepsilon^2/(64\int \xi_{\theta_2}^2(x) f_0(x)dx)$, $\varepsilon_1^2 = \varepsilon^2/(64\int \xi_1^2(x) f_0(x)dx)$, and $\varepsilon_2^2 = \varepsilon^2/(64\int \xi_2^2(x) f_0(x)dx)$. The bracketing number for the space $\mathcal{F}_{M,c}$ is then

$$\begin{aligned}
&N_{[]}(\varepsilon, \mathcal{F}_{M,c}, \|\cdot\|_{2,f_0}) \\
&\leq N(\varepsilon_{\theta_1}, \Theta, \|\cdot\|_\infty) \times N(\varepsilon_{\theta_2}, \Theta, \|\cdot\|_\infty) \times N(\varepsilon_1, \overline{\mathcal{F}}_{M,lcd}, \|\cdot\|_\infty) \times N(\varepsilon_2, \overline{\mathcal{F}}_{M,lcd}, \|\cdot\|_\infty) \\
&= N\left(8^{-1}\varepsilon/\|\xi_{\theta_1}\|_{2,f_0}, \Theta, \|\cdot\|_\infty\right) \times N\left(8^{-1}\varepsilon/\|\xi_{\theta_2}\|_{2,f_0}, \Theta, \|\cdot\|_\infty\right) \times N(8^{-1}\varepsilon/\|\xi_1\|_{2,f_0}, \overline{\mathcal{F}}_{M,lcd}, \|\cdot\|_\infty) \times N(8^{-1}\varepsilon/\|\xi_2\|_{2,f_0}, \overline{\mathcal{F}}_{M,lcd}, \|\cdot\|_\infty) \\
&\leq N_{[]}\left(4^{-1}\varepsilon/\|\xi_{\theta_1}\|_{2,f_0}, \Theta, \|\cdot\|_\infty\right) \times N_{[]}\left(4^{-1}\varepsilon/\|\xi_{\theta_2}\|_{2,f_0}, \Theta, \|\cdot\|_\infty\right) \times N_{[]}(4^{-1}\varepsilon/\|\xi_1\|_{2,f_0}, \overline{\mathcal{F}}_{M,lcd}, \|\cdot\|_\infty) \times N_{[]}(4^{-1}\varepsilon/\|\xi_2\|_{2,f_0}, \overline{\mathcal{F}}_{M,lcd}, \|\cdot\|_\infty) \\
&\leq N_{[]}\left(4^{-1}\varepsilon/\|\xi_{\theta_1}\|_{2,f_0}, \Theta, \|\cdot\|_\infty\right) \times N_{[]}\left(4^{-1}\varepsilon/\|\xi_{\theta_2}\|_{2,f_0}, \Theta, \|\cdot\|_\infty\right) \times N_{[]}(4^{-1}\varepsilon/\|\xi_1\|_{2,f_0}, \mathcal{F}_{M,lcd}, d_H) \times N_{[]}(4^{-1}\varepsilon/\|\xi_2\|_{2,f_0}, \mathcal{F}_{M,lcd}, d_H).
\end{aligned}$$

It was shown in Doss and Wellner [2016, Theorem 3.1] that

$$\log N_{[]}(\varepsilon, \mathcal{F}_{M,lcd}, d_H) \leq A_M \varepsilon^{-1/2},$$

where the constant A_M depends on M. On the other hand, from the well-known result

$$N_{[]}(\varepsilon, \Theta, \|\cdot\|_\infty) \leq \left(\frac{2\,\mathrm{diam}(\Theta)}{\varepsilon}\right)^k,$$

where k is the dimension of Θ. Therefore, returning to the general case of dimension

k and d, we get

$$\begin{aligned}
\log N_{[]}(\varepsilon, \mathcal{F}_{M,c}, \|\cdot\|_{2,f_0}) &\leq \sum_{l=1}^{k} \log\left(\frac{2(d+k)\operatorname{diam}(\Theta)\|\xi_{\theta_l}\|_{2,f_0}}{\varepsilon}\right) + \sum_{j=1}^{d} A_M \left\{\frac{\varepsilon}{(d+k)\|\xi_j\|_{2,f_0}}\right\}^{-1/2} \\
&= \sum_{l=1}^{k} 2\log\left(\frac{2(d+k)\operatorname{diam}(\Theta)\|\xi_{\theta_l}\|_{2,f_0}}{\varepsilon}\right)^{1/2} + A_M\sqrt{d+k}\left\{\sum_{j=1}^{d}\sqrt{\|\xi_j\|_{2,f_0}}\right\}\varepsilon^{-1/2} \\
&\leq 2\sqrt{2(d+k)\operatorname{diam}(\Theta)}\left\{\sum_{l=1}^{k}\sqrt{\|\xi_{\theta_l}\|_{2,f_0}}\right\}\varepsilon^{-1/2} + A_M\sqrt{d+k}\left\{\sum_{j=1}^{d}\sqrt{\|\xi_j\|_{2,f_0}}\right\}\varepsilon^{-1/2} \\
&= A\varepsilon^{-1/2}
\end{aligned}$$

with $A = 4\sqrt{\operatorname{diam}(\Theta)}(d+k)^{1/2}\left\{\sum_{l=1}^{k}\sqrt{\|\xi_{\theta_l}\|_{2,f_0}}\right\} + 2(d+k)^{1/2}A_M\left\{\sum_{j=1}^{d}\sqrt{\|\xi_j\|_{2,f_0}}\right\}$. □

Proof of Theorem 5.2. We need to show that all four conditions from Nan and Wellner [2013, Corollary 1, page 1159] are satisfied. First, we show that

(i) (stochastic equicontinuity)

$$\frac{|\sqrt{n}(\Psi_n - \Psi)(\widehat{\theta}_n, \widehat{F}_n) - \sqrt{n}(\Psi_n - \Psi)(\theta_0, F_0)|}{1 + \sqrt{n}|\Psi_n(\widehat{\theta}_n, \widehat{F}_n)| + \sqrt{n}|\Psi(\widehat{\theta}_n, \widehat{F}_n)|} = o_{p^*}(1).$$

We will show the slightly stronger version of this condition which is

$$|\sqrt{n}(\Psi_n - \Psi)(\widehat{\theta}_n, \widehat{F}_n) - \sqrt{n}(\Psi_n - \Psi)(\theta_0, F_0)| = o_{p^*}(1).$$

To prove this condition, we need to show that the class of $\partial_\theta \log c(F(x);\theta)$ is Donsker and appealing to van der Vaart and Wellner [1996, Corollary 2.3.12, page 115]. Let $g_{F,\theta}(x) = \partial_\theta \log c(F(x);\theta)$ and $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$ we have that

$$\mathbb{G}_n\left(g_{F,\theta} - g_{F_0,\theta_0}\right) = \sqrt{n}(\Psi_n - \Psi)(\theta, F) - \sqrt{n}(\Psi_n - \Psi)(\theta_0, F_0).$$

Recall that under the assumptions of the theorem, $\widehat{F}_n, \widehat{\theta}_n$ are both consistent. Since F_0 is such that each density is log-concave, we can assume that for n large enough each $\widehat{f}_{nj} \in \mathcal{F}_{M,lcd}$, where

$$\mathcal{F}_{M,lcd} \;=\; \{f \in \mathcal{F}_{lcd} : \sup_{s\in\mathbb{R}} f(s) \le M \text{ and } 1/M \le f(s) \text{ if } s \in [-1,1]\}.$$

$\overline{\mathcal{F}}_{M,lcd}$ is the class of associated cumulative distribution functions. Therefore, to prove this result, it is sufficient to show that $\|\mathbb{G}_n\|_{\mathcal{F}^{\delta}_{M,c}}$, where $\mathcal{F}^{\delta}_{M,c} = \{g_{F_1,\theta_1} - g_{F_2,\theta_2} : g_{F_1,\theta_1}, g_{F_2,\theta_2} \in \mathcal{F}_{M,c}, \|g_{F_1,\theta_1} - g_{F_2,\theta_2}\| < \delta\}$ and

$$\mathcal{F}_{M,c} = \{\partial_\theta \log c(F(x);\theta) : \theta \in \Theta, F \in \overline{\mathcal{F}}_{M,lcd}\}.$$

By van der Vaart and Wellner [1996, Corollary 2.3.12, page 115], it is enough to show that $\mathcal{F}_{M,c}$ is P-Donsker. By (R7), an envelope function for $\mathcal{F}_{M,c}$ exists. Next, by Lemma 5.5, we have that

$$\int_0^\infty \sqrt{\log N_{[]}(\varepsilon, \mathcal{F}_{M,c}, L_2(P))}\, d\varepsilon \;\le\; \sqrt{A}\int_0^B \varepsilon^{-1/4}\, d\varepsilon = (4/3)B^{3/4}\sqrt{A} \;< \infty,$$

since $N_{[]}(\varepsilon, \mathcal{F}_{M,c}, L_2(P)) = 1$ for ε large enough. Thus, by van der Vaart and Wellner [1996, Theorem 2.5.6, page 130], $\mathcal{F}_{M,c}$ is P-Donsker. Then, the result follows.

(ii) $\sqrt{n}\Psi_n(\theta_0, F_0) = O_{p^*}(1)$

To prove this condition, we use central limit theorem (CLT). From assumption (U), we have $\Psi(\theta_0, F_0) = E_{f_0}[\partial_\theta \log c(F_0(X);\theta_0)] = 0$ and also $\Psi_n(\theta_0, F_0) =$

$n^{-1}\sum_{i=1}^{n}\partial_\theta \log c(F_0(x_i);\theta_0)$. Thus,

$$\begin{aligned}\sqrt{n}\Psi_n(\theta_0,F_0) &= n^{-1/2}\left\{\frac{1}{n}\sum_{i=1}^{n}\partial_\theta \log c(F_0(x_i);\theta_0) - E_{f_0}[\partial_\theta \log c(F_0(X);\theta_0)]\right\}\\ &= O_{p^*}(1),\end{aligned}$$

since $E_{f_0}[\|\partial_\theta \log c(F_0(X);\theta_0)\|^2]$ is finite by (R4).

(iii) (smoothness)

For some $\beta_2 > 5/4$, we have

$$\begin{aligned}&|\Psi(\theta,F) - \Psi(\theta_0,F_0) - \dot{\Psi}_\theta(\theta_0,F_0)(\theta-\theta_0) - \dot{\Psi}_F(\theta_0,F_0)[F-F_0]|\\ &\qquad = o(\|\theta-\theta_0\|_\infty) + O(\|F-F_0\|_\infty^{\beta_2}).\end{aligned}$$

To prove this condition, we need Lemma A.13 for interchanging between absolute value and integration. We use Taylor's expansion around (θ_0, F_0). Then,

$$
\begin{aligned}
&|\Psi(\theta,F)-\Psi(\theta_0,F_0)-\dot{\Psi}_\theta(\theta_0,F_0)(\theta-\theta_0)-\dot{\Psi}_F(\theta_0,F_0)[F-F_0]| \\
&= |R(\theta^*,F^*)| \\
&\le \Big|2^{-1}\int \sum_{q,s=1}^{k} \partial_{\theta_q}\partial_{\theta_s}\partial_\theta \log c(F^*(x);\theta^*)(\theta_q-\theta_{0q})(\theta_s-\theta_{0s})f_0(x)dx \\
&\quad +2^{-1}\int \sum_{s=1}^{k}\sum_{j=1}^{d} \partial_{u_j}\partial_{\theta_s}\partial_\theta \log c(F^*(x);\theta^*)(\theta_s-\theta_{0s})(F_j-F_{0j})(x_j)f_0(x)dx \\
&\quad +2^{-1}\int \sum_{j=1}^{d}\sum_{q=1}^{k} \partial_{u_j}\partial_{\theta_q}\partial_\theta \log c(F^*(x);\theta^*)(\theta_q-\theta_{0q})(F_j-F_{0j})(x_j)f_0(x)dx \\
&\quad +2^{-1}\int \sum_{j,l=1}^{d} \partial_{u_j}\partial_{u_l}\partial_\theta \log c(F^*(x);\theta^*)(F_j-F_{0j})(x_j)(F_l-F_{0l})(x_l)f_0(x)dx\Big| \\
&\le 2^{-1}\sum_{q,s=1}^{k}\left\{\int \left|\partial_{\theta_q}\partial_{\theta_s}\partial_\theta \log c(F^*(x);\theta^*)\right| f_0(x)dx\right\}|\theta_q-\theta_{0q}||\theta_s-\theta_{0s}| \\
&\quad +2^{-1}\sum_{s=1}^{k}\sum_{j=1}^{d}\left\{\int \left|\partial_{u_j}\partial_{\theta_s}\partial_\theta \log c(F^*(x);\theta^*)\right| f_0(x)dx\right\}|\theta_s-\theta_{0s}|\|F_j-F_{0j}\|_\infty \\
&\quad +2^{-1}\sum_{j=1}^{d}\sum_{q=1}^{k}\left\{\int \left|\partial_{u_j}\partial_{\theta_q}\partial_\theta \log c(F^*(x);\theta^*)\right| f_0(x)dx\right\}|\theta_q-\theta_{0q}|\|F_j-F_{0j}\|_\infty \\
&\quad +2^{-1}\sum_{j,l=1}^{d}\left\{\int \left|\partial_{u_j}\partial_{u_l}\partial_\theta \log c(F^*(x);\theta^*)\right| f_0(x)dx\right\}\|F_j-F_{0j}\|_\infty\|F_l-F_{0l}\|_\infty.
\end{aligned}
$$

Applying (R2) and (R5), therefore we get

$$
\begin{aligned}
&|\Psi(\theta,F)-\Psi(\theta_0,F_0)-\dot{\Psi}_\theta(\theta_0,F_0)(\theta-\theta_0)-\dot{\Psi}_F(\theta_0,F_0)[F-F_0]| \\
&\quad\leq \frac{1}{2}k^2M_3|\theta_q-\theta_{0q}||\theta_s-\theta_{0s}|+\frac{1}{2}kdM_4|\theta_s-\theta_{0s}|\|F_j-F_{0j}\|_\infty \\
&\qquad+\frac{1}{2}kdM_4|\theta_q-\theta_{0q}|\|F_j-F_{0j}\|_\infty+\frac{1}{2}d^2M_2\|F_j-F_{0j}\|_\infty\|F_l-F_{0l}\|_\infty \\
&\quad\leq \frac{1}{2}k^2M_3\|\theta-\theta_0\|_\infty^2+kdM_4\|\theta-\theta_0\|_\infty\|F-F_0\|_\infty+\frac{1}{2}d^2M_2\|F-F_0\|_\infty^2 \\
&\quad\leq \left(\frac{1}{2}k^2M_3+\frac{kd}{2}M_4\right)\|\theta-\theta_0\|_\infty^2+\left(\frac{kd}{2}M_4+\frac{1}{2}d^2M_2\right)\|F-F_0\|_\infty^2 \\
&\quad\leq o(\|\theta-\theta_0\|_\infty)+O(\|F-F_0\|_\infty^2).
\end{aligned}
$$

Therefore, $\beta_2 = 2$ which greater than $5/4$. Then, condition (iii) has been proved.

(iv) $\sqrt{n}\dot{\Psi}_F(\theta_0,F_0)[\widehat{F}_n-F_0]=O_{p^*}(1)$

Letting $\psi_0(u)=\partial_{u_j}\partial_\theta\log c(u;\theta_0)$, for each $j=1,\ldots,d$, we get

$$
\begin{aligned}
&\left[\sqrt{n}\dot{\Psi}_F(\theta_0,F_0)[\widehat{F}_n-F_0]\right]_j \\
&\quad= n^{-1/2}\sum_{i=1}^n\psi_0(F_0(x_i))(\widehat{F}_{nj}-F_{0j})(x_{ij}) \\
&\quad= \sqrt{n}(\mathbb{P}_n-P_0)[\psi_0(F_0(x))(\widehat{F}_{nj}-F_{0j})(x_j)]+\sqrt{n}\int\psi_0(F_0(x))(\widehat{F}_{nj}-F_{0j})(x_j)f_0(x)dx.
\end{aligned}
$$

For the first term, we show that the class of $\{\psi_0(F_0(x))F(x):F\in\overline{\mathcal{F}}_{M,lcd}\}$ is P-Donsker. Since $|F|\leq 1$ and by (R6), the envelope function of $\psi_0(F_0(x))F(x)$ exists. Since, we have the relationship between the supremum norm and Hellinger metric. Moreover, from the bracketing entropy $\log N_{[]}(\varepsilon,\mathcal{F}_{M,lcd},d_H)\leq A_M\varepsilon^{-1/2}$.

The relationship of the bracketing entropy with respect to these two metrics are

$$\log N_{[]}(\varepsilon, \overline{\mathcal{F}}_{M,lcd}, \|\cdot\|_\infty) \leq \log N_{[]}(\varepsilon, \mathcal{F}_{M,lcd}, d_H) \leq A_M \varepsilon^{-1/2}.$$

Therefore,

$$\int_0^\infty \sqrt{\log N_{[]}(\varepsilon, \mathcal{F}_{M,lcd}, L_2(P))}\, d\varepsilon \;\leq\; \sqrt{A_M} \int_0^D \varepsilon^{-1/4}\, d\varepsilon \;=\; (4/3) D^{3/4} \sqrt{A_M} \;<\; \infty,$$

with $D < \infty$. Since for big enough ε, $N_{[]}(\varepsilon, \mathcal{F}_{M,lcd}, L_2(P))$ is 1. Then, the class of $\{\psi_0(F_0(x))F(x) : F \in \overline{\mathcal{F}}_{M,lcd}\}$ is P-Donsker. Hence,

$$\sqrt{n}(\mathbb{P}_n - P_0)[\psi_0(F_0(x))(\widehat{F}_{nj} - F_{0j})(x_j)] = o_{p^*}(1).$$

For the second term, we will show that it is $O_{p^*}(1)$.

$$\sqrt{n} \int \psi_0(F_0(x))(\widehat{F}_{nj} - F_{0j})(x_j) f_0(x) dx \;\leq\; \sqrt{n} \left\{ \int \left| \psi_0(F_0(x)) f_0(x) \right| dx \right\} \|\widehat{F}_{nj} - F_{0j}\|_\infty$$

By (R1), $\int \left|\psi_0(F_0(x)) f_0(x)\right| dx$ is bounded. The left of this proof is to show that $\|\widehat{F}_{nj} - F_{0j}\|_\infty = O_p(n^{-1/2})$. The proof follows Marshall's inequality in Kim et al. [2018, Lemma 2, page 2284] and it satisfies under $f_{0j}(x_j) = e^{\alpha_0 x_j} h_{0j}(x_j)$, for all $x_j \in [X_{j,(1)}, X_{j,(n)}]$, for some $\alpha_0 \in \mathbb{R}$, and $h_0 : [X_{j,(1)}, X_{j,(n)}] \mapsto \mathbb{R}$ is concave. Then, we can show that

$$\sqrt{n} \int \psi_0(F_0(x))(\widehat{F}_{nj} - F_{0j})(x_j) f_0(x) dx \;\leq\; \sqrt{n} \left\{ \int \left| \psi_0(F_0(x)) f_0(x) \right| dx \right\} \rho(|\kappa|) \|\mathbb{F}_{n,j} - F_{0,j}\|_\infty$$

where $\kappa = \alpha_0(X_{j,(n)} - X_{j,(1)})$. When $\alpha_0 \neq 0$, the support of f_{0j} is bounded with (LC), say $[a_0, b_0]$, then in particular we have that $X_{j,(n)} \leq b_0$ and $X_{j,(1)} \geq a_0$ so that

$$|X_{j,(n)} - X_{j,(1)}| \;=\; X_{j,(n)} - X_{j,(1)} \leq b_0 - a_0 = O_p(1).$$

Therefore,

$$\sqrt{n}\int \psi_0(F_0(x))(\widehat{F}_{nj} - F_{0j})(x_j) f_0(x)dx \;=\; \sqrt{n}O_p^*(n^{-1/2}) \;=\; O_p^*(1).$$

The limitation of this proof under $f_{0j}(x_j) = e^{\alpha_0 x_j} h_{0j}(x_j)$ is that not all log-concave densities can be written in this form. However, we will show by the simulation studies of $f_{0j} \sim \Gamma(5,1)$ and also Exp(1.5), and $\beta(5,2)$ that $\sqrt{n}$ convergence of $\widehat{\theta}_n$ still satisfies even if the true marginal densities are not follow the form of f_{0j}.

Therefore,

$$\begin{aligned}\sqrt{n}\dot{\Psi}_F(\theta_0, F_0)[\widehat{F}_n - F_0] &= \sum_{j=1}^{d} \left[\sqrt{n}\dot{\Psi}_F(\theta_0, F_0)[\widehat{F}_n - F_0]\right]_j \\ &= O_{p^*}(1).\end{aligned}$$

Hence, $\widehat{\theta}_n$ is $\sqrt{n}-$consistent.

□

Proof of Theorem 5.3. From the consistency proof of the joint density estimator, we have

$$\begin{aligned}
&\int \left|\widehat{f}_n(x) - f_0(x)\right| dx \\
&= \int \left|c(\widehat{F}_n(x);\widehat{\theta}_n) - c(F_0(x);\theta_0)\right| \prod_{j=1}^{d} f_{0j}(x_j)dx + \sum_{j=1}^{d} \int \left|\widehat{f}_{nj}(x_j) - f_{0j}(x_j)\right| \Big\{ \prod_{i<j} \widehat{f}_{ni}(x_i) \Big\} \Big\{ \prod_{i>j} f_{0i}(x_i) \Big\} c(\widehat{F}_n(x);\widehat{\theta}_n)\ dx.
\end{aligned}$$

For the first term,

$$\int \left|c(\widehat{F}_n(x);\widehat{\theta}_n) - c(F_0(x);\theta_0)\right| \prod_{j=1}^{d} f_{0j}(x_j)dx \quad \leq \quad B_3 \left\{ \|\widehat{F}_n - F_0\|_\infty + \|\widehat{\theta}_n - \theta_0\|_\infty \right\}.$$

From the result that $\|\widehat{F}_n - F_0\|_\infty = O_{p^*}(n^{-2/5})$ and $\|\widehat{\theta}_n - \theta_0\|_\infty = O_{p^*}(n^{-1/2})$, we can conclude that

$$\int \left|c(\widehat{F}_n(x);\widehat{\theta}_n) - c(F_0(x);\theta_0)\right| \prod_{j=1}^{d} f_{0j}(x_j)dx \quad = \quad O_{p^*}(n^{-2/5}).$$

For the second term,

$$\begin{aligned}
&\sum_{j=1}^{d} \int \left|\widehat{f}_{nj}(x_j) - f_{0j}(x_j)\right| \Big\{ \prod_{i<j} \widehat{f}_{ni}(x_i) \Big\} \Big\{ \prod_{i>j} f_{0i}(x_i) \Big\} c(\widehat{F}_n(x);\widehat{\theta}_n)\ dx \\
&\quad \leq \quad B^{d-1} \left\{ \int \left|\widehat{f}_{nj}(x_j) - f_{0j}(x_j)\right|^2 dx \right\}^{1/2} \left\{ \int \left(c(\widehat{F}_n(x);\widehat{\theta}_n)\right)^2 dx \right\}^{1/2} \\
&\quad \leq \quad B^{d-1} \left\{ \int \left|\widehat{f}_{nj}(x_j) - f_{0j}(x_j)\right|^2 dx \right\}^{1/2} \left\{ \sup_{(F,\theta)\in\mathcal{N}} \int \left(c(F(x);\theta)\right)^2 dx \right\}^{1/2}.
\end{aligned}$$

Since, $\int \left|\widehat{f}_{nj}(x_j) - f_{0j}(x_j)\right| dx = O_{p^*}(n^{-2/5})$. Then, $\int \left|\widehat{f}_{nj}(x_j) - f_{0j}(x_j)\right|^2 dx = O_{p^*}(n^{-4/5})$. Therefore, $\left\{ \int \left|\widehat{f}_{nj}(x_j) - f_{0j}(x_j)\right|^2 dx \right\}^{1/2} = O_{p^*}(n^{-2/5})$ and the second term is bounded with (R8). Hence, we can conclude that

$$\sum_{j=1}^{d} \int \left|\widehat{f}_{nj}(x_j) - f_{0j}(x_j)\right| \Big\{ \prod_{i<j} \widehat{f}_{ni}(x_i) \Big\} \Big\{ \prod_{i>j} f_{0i}(x_i) \Big\} c(\widehat{F}_n(x);\widehat{\theta}_n)\ dx \quad = \quad O_{p^*}(n^{-2/5}).$$

Therefore,

$$\int \left|\widehat{f}_n(x) - f_0(x)\right| dx \quad = \quad O_{p^*}(n^{-2/5}).$$

□

5.2.3 Support for the proofs

5.2.3.1 Example of unique solution for Gaussian copula

From assumption of unique solution (U), we will show that it is true for Gaussian copula. Let $x = (x_1, \ldots, x_d)$ be a d-variate data with Gaussian copula and marginals $u_1 = F_1(x_1), \ldots, u_d = F_d(x_d)$. The d-dimensional Gaussian copula density can be represented as

$$c(u_1, \ldots, u_d) = \frac{1}{\sqrt{\det R}} \exp\left(-\frac{1}{2} \begin{bmatrix} \Phi^{-1}(u_1) \\ \vdots \\ \Phi^{-1}(u_d) \end{bmatrix}^T \left[R^{-1} - I \right] \begin{bmatrix} \Phi^{-1}(u_1) \\ \vdots \\ \Phi^{-1}(u_d) \end{bmatrix} \right)$$

where R and $R^{-1} - I$ are given by

$$R = \begin{bmatrix} 1 & \theta & \cdots & \theta \\ \theta & 1 & \cdots & \theta \\ \vdots & \cdots & \cdots & \vdots \\ \theta & \cdots & \theta & 1 \end{bmatrix}, \quad R^{-1} - I \quad = \quad \frac{1}{\det R} \begin{bmatrix} A & B & \cdots & B \\ B & A & B & B \\ \vdots & \cdots & \cdots & \vdots \\ B & \cdots & B & A \end{bmatrix}$$

with A and B depend on θ. To simplify notations, we denote $u_{0j} = F_{0j}(x_j)$,and $y_{0j} = \Phi^{-1}(u_{0j})$ for each $j = 1, \ldots, d$. Then,

$$\begin{aligned}
&\log c(u_{01}, \ldots, u_{0d}; \theta) \\
&= -\frac{1}{2}\log(\det R) - \frac{1}{2\det R}\left\{y_{01}^2 A + y_{01}(y_{02} + \cdots + y_{0d})B + \cdots + y_{0d}^2 A + y_{0d}(y_{01} + \cdots + y_{0,d-1})B\right\}.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\partial_\theta \log c(u_{01}, \ldots, u_{0d}; \theta) \\
&= -\frac{1}{2}\left(\frac{\partial_\theta \det R}{\det R}\right) - \frac{1}{2\det R}\left\{y_{01}^2 \partial_\theta A + y_{01}(y_{02} + \cdots + y_{0d})\partial_\theta B + \cdots + y_{0d}^2 \partial_\theta A + y_{0d}(y_{01} + \cdots + y_{0,d-1})\partial_\theta B\right\} \\
&\quad + \left\{y_{01}^2 A + y_{01}(y_{02} + \cdots + y_{0d})B + \cdots + y_{0d}^2 A + y_{0d}(y_{01} + \cdots + y_{0,d-1})B\right\}\frac{\partial_\theta \det R}{2(\det R)^2}.
\end{aligned}$$

Since for each j, u_{0j} are distributed as $U(0,1)$, and $\Phi \sim N(0,1)$. Then, $\Phi^{-1}(u_{0j}) \sim N(0,1)$. Then, $E_{\theta_0}(y_{0j}) = 0, E_{\theta_0}(y_{0j}^2) = \mathrm{var}(y_{0j}) = 1$, and $E(y_{0j}y_{0l}) = \mathrm{corr}(y_{0j}, y_{0l}) = \theta_0$ for $j, l = 1.\ldots, d$. Therefore,

$$\begin{aligned}
&\Psi(\theta, u_{01}, \ldots, u_{0d}) \\
&= \int \left(\partial_\theta \log c(u_{01}, \ldots, u_{0d}; \theta)\right) c(F_0; \theta_0) g_0(x)\ dx \\
&= -\frac{1}{2}\left(\frac{\partial_\theta \det R}{\det R}\right) - \frac{1}{2\det R}\left\{d\partial_\theta A + d(d-1)\theta_0 \partial_\theta B\right\} + \left\{dA + d(d-1)\theta_0 B\right\}\left(\frac{\partial_\theta \det R}{2(\det R)^2}\right).
\end{aligned}$$

We will show an explicit formula for $d = 2$ and $d = 3$. First, we show each part of $\Psi(\theta, u_{01}, u_{02})$.

$$\det R = 1 - \theta^2, \quad \partial_\theta \det R = -2\theta$$

$$A = 1, \quad \partial_\theta A = 0$$

$$B = -\theta, \quad \partial_\theta B = -1.$$

Hence,

$$\begin{aligned} &\Psi(\theta, u_{01}, u_{02}) \\ &= -\frac{1}{2}\left(\frac{-2\theta}{1-\theta^2}\right) - \frac{1}{2(1-\theta^2)}(-2\theta_0) + (2 - 2\theta_0\theta)\left(\frac{-2\theta}{2(1-\theta^2)^2}\right). \end{aligned}$$

Then, we set $\Psi(\theta, u_{01}, u_{02}) = 0$. We get

$$\theta^3 - \theta^2\theta_0 + \theta - \theta_0 = 0$$

$$\theta^2(\theta - \theta_0) + (\theta - \theta_0) = 0$$

$$(\theta - \theta_0)(\theta^2 + 1) = 0$$

$$\theta = \theta_0.$$

Thus, θ_0 is a unique solution of $\Psi(\theta, u_{01}, u_{02}) = 0$.

Next, each term of $\Psi(\theta, u_{01}, u_{02}, u_{03})$ can be expressed as

$$\det R = 1 + 2\theta^3 - 3\theta^2, \quad \partial_\theta \det R = 6\theta^2 - 6\theta$$

$$A = -2\theta^3 + 2\theta^2, \quad \partial_\theta A = -6\theta^2 + 4\theta$$

$$B = \theta^2 - \theta, \quad \partial_\theta B = 2\theta - 1.$$

Hence,

$$\begin{aligned}
&\Psi(\theta, u_{01}, u_{02}, u_{03}) \\
&= -\frac{1}{2}\left(\frac{6\theta^2 - 6\theta}{1 + 2\theta^3 - 3\theta^2}\right) - \frac{1}{2(1 + 2\theta^3 - 3\theta^2)}\left\{3(-6\theta^2 + 4\theta) + 6\theta_0(2\theta - 1)\right\} \\
&\quad + \left\{3(-2\theta^3 + 2\theta^2) + 6\theta_0(\theta^2 - \theta)\right\}\left(\frac{6\theta^2 - 6\theta}{2(1 + 2\theta^3 - 3\theta^2)^2}\right).
\end{aligned}$$

Then, we set $\Psi(\theta, u_{01}, u_{02}, u_{03}) = 0$. We get

$$\left[12\theta^4 - 24\theta^3 + 18\theta^2 - 12\theta + 6\right](\theta - \theta_0) = 0.$$

Thus, θ_0 is a unique solution of $\Psi(\theta, u_{01}, u_{02}, u_{03}) = 0$.

5.2.3.2 Study of $\sqrt{n}$ rate of convergence for $\widehat{\theta}_n$

In this chapter, we already proved that the copula estimator is $\sqrt{n}$ consistent. However, it is good to know how the copula estimator performs in the simulation study. We expect that the simulation results should provide approximately the $\sqrt{n}$ rate of convergence, which represents in the same way as the theory does. We work

on three methods, which are parametric IFM, log-concave IFM, and kernel IFM. Moreover, we focus on $d = 2, \tau = 0, 0.2, 0.6$, and both margins have the same marginal densities. The simulation results in Figure 5.1 to 5.3 are from 50 sets of simulation. For finding the copula estimator, we use a "Newton's-Raphson" method, which needs the first and second derivatives of the log-likelihood function. In order to make the simulation simple, we choose the Gaussian copula as our interested copula because its derivatives can be found easily. A log-likelihood function of the Gaussian copula is given by

$$\begin{aligned}\ell_c &= \sum_{i=1}^{n}\Bigg[-\frac{1}{2}\log(1-\theta^2) - \frac{\theta^2}{2(1-\theta^2)}(\Phi^{-1}(u))^2 \\ &\quad + \frac{\theta}{1-\theta^2}\Phi^{-1}(u)\Phi^{-1}(v) - \frac{\theta^2}{2(1-\theta^2)}(\Phi^{-1}(v))^2\Bigg].\end{aligned}$$

Then, the corresponding first derivative is

$$\begin{aligned}\ell_c' &= \sum_{i=1}^{n}\Bigg[\frac{\theta}{1-\theta^2} - \frac{\theta}{(1-\theta^2)^2}\left((\Phi^{-1}(u))^2 + (\Phi^{-1}(v))^2\right) \\ &\quad + \frac{1+\theta^2}{(1-\theta^2)^2}\Phi^{-1}(u)\Phi^{-1}(v)\Bigg].\end{aligned}$$

The expression of the second derivative is

$$\begin{aligned}\ell_c'' &= \sum_{i=1}^{n}\Bigg[\frac{1+\theta^2}{(1-\theta^2)^2} - \frac{1+3\theta^2}{(1-\theta^2)^3}\left((\Phi^{-1}(u))^2 + (\Phi^{-1}(v))^2\right) \\ &\quad + \frac{6\theta+2\theta^3}{(1-\theta^2)^3}\Phi^{-1}(u)\Phi^{-1}(v)\Bigg].\end{aligned}$$

Hence, the updating scheme for Newton's-Raphson method is given by

$$\theta^{(t+1)} = \theta^{(t)} - \frac{\ell_c'(\theta^{(t)})}{\ell_c''(\theta^{(t)})},$$

where (t) represents set tth of the simulation study.

Next, we plot graphs between log-scale of standard deviations (y-axis) and sample sizes (x-axis). Then, we fit logarithm of standard deviations of $\widehat{\theta}$ with logarithm of sample sizes by using a least square estimation. After that, we plot slopes, which are shown as the straight lines in Figure 5.1 to 5.3. These lines are approximately $-1/2$, which represent the rate of convergence for the estimator $\widehat{\theta}$.

5.2.3.3 $\sqrt{n}$ convergence for $\widehat{\theta}_n$ when condition (LC) is not satisfied

This section is to show that when f_{0j} does not follow the conditions in Kim et al. [2018, Lemma 2, page 2284], then $\sqrt{n}$ rate of convergence is still satisfied. We will show by three log-concave distributions, which are $\Gamma(5,1)$, Exp(1.5), and $\beta(5,2)$. For $\Gamma(a,b)$, when we take $\alpha_0 = -b$, and $h_{0j}(x_j) = (b^a/\Gamma(a))x_j^{a-1}$, $h_{0j}(x_j)$ is not concave for all $x_j \in [X_{j,(1)}, X_{j,(n)}]$ when $a \notin (1,2)$. For Exp(λ), we set $\alpha_0 = -\lambda$ and $h_{0j}(x_j) = \lambda$. We can clearly see that $h_{0j}(x_j)$ is not concave for all λ. For $\beta(a,b)$, we take $\alpha_0 = 0$ and $h_{0j}(x_j) = \left(x_j^{a-1}(1-x_j)^{b-1}\right)/(B(a,b))$, where $B(a,b) = (a-1)!(b-1)!/(a+b-1)!$. Since the second derivative of $h_{0j}(x_j)$ is complicated, so it is hard to find the values of a and b which $h_{0j}(x_j)$ is concave. In the simulation study, we choose $\beta(5,2)$ as an

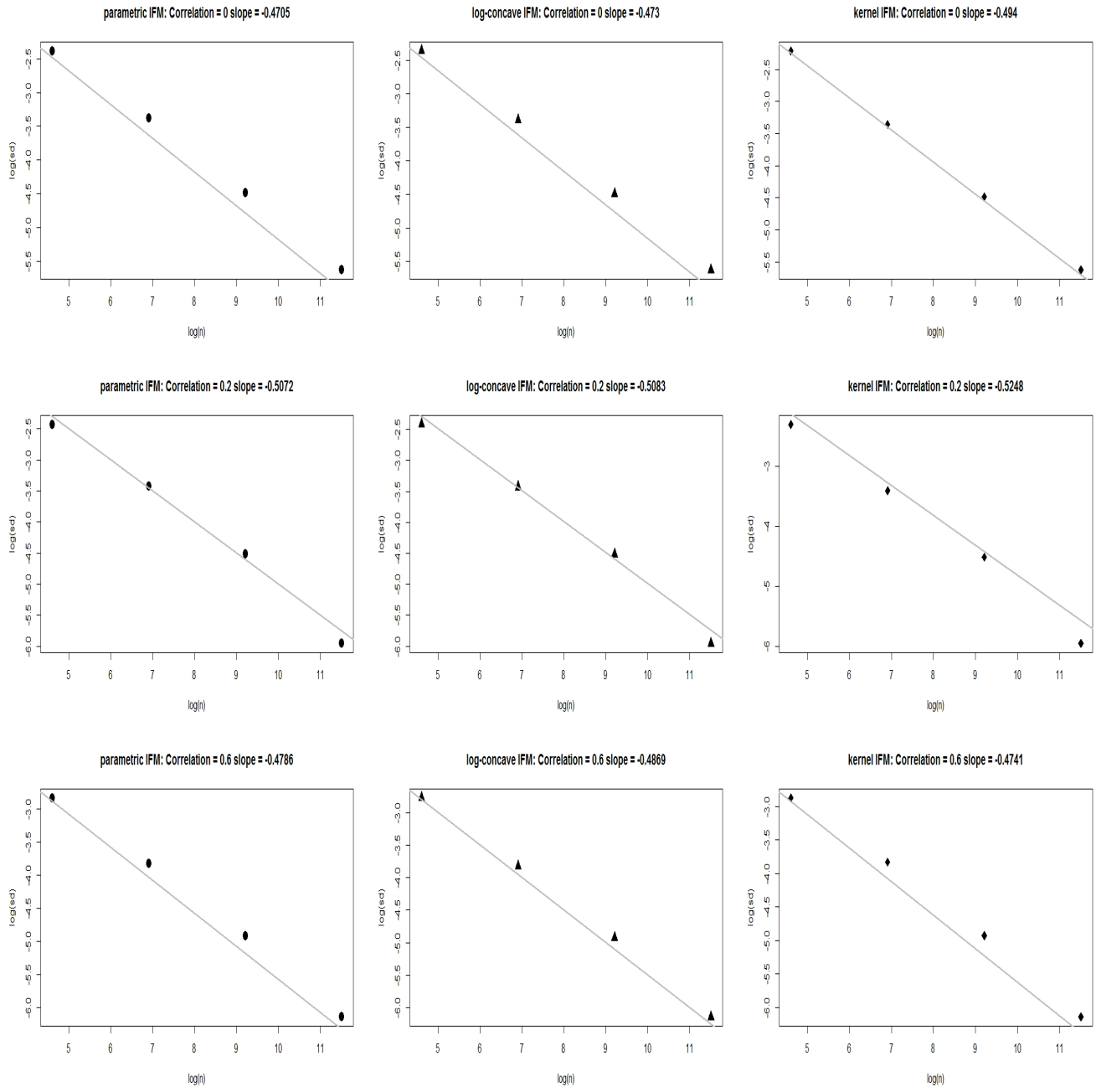


Figure 5.1: Study $\sqrt{n}$ rate of convergence for $N(0,1)$

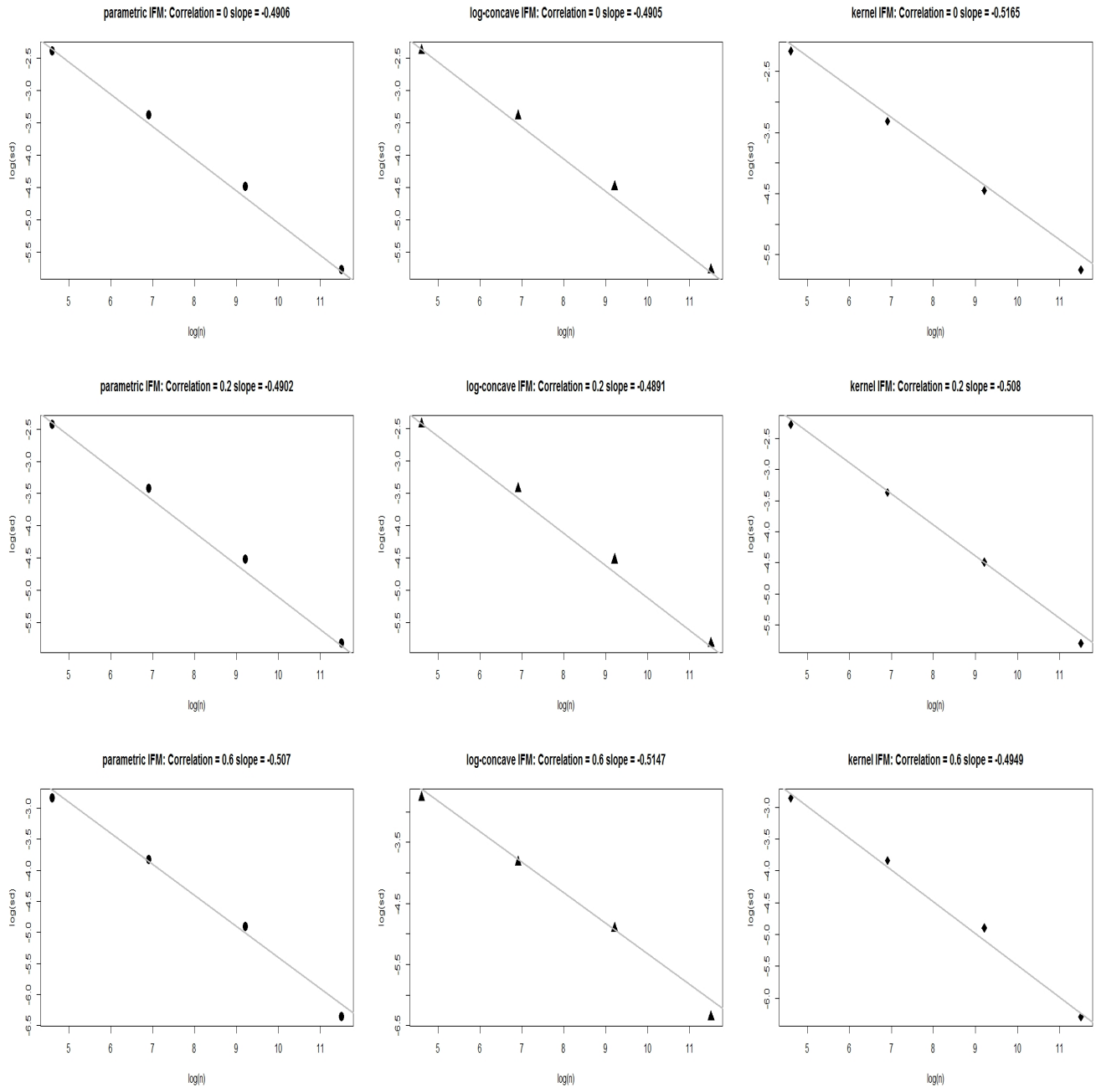


Figure 5.2: Study $\sqrt{n}$ rate of convergence for $\Gamma(2,1)$

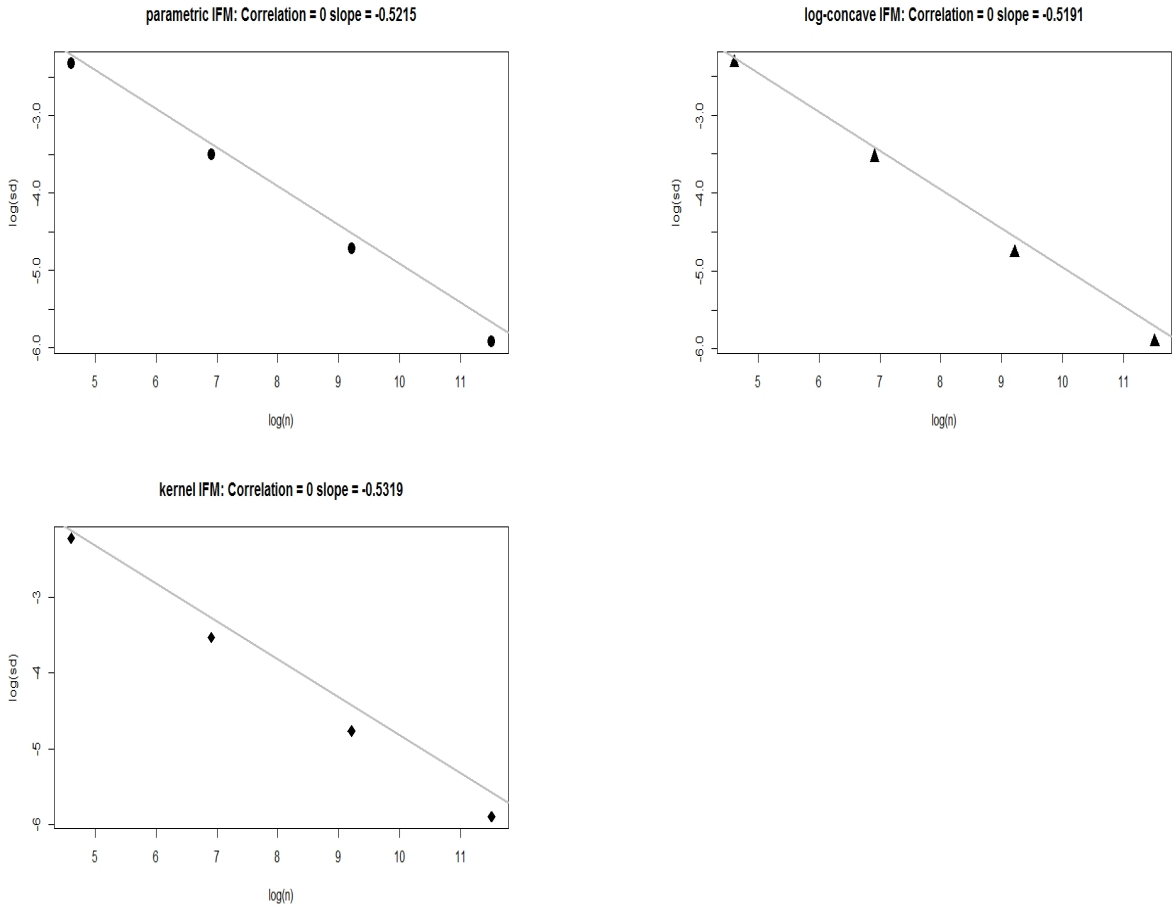
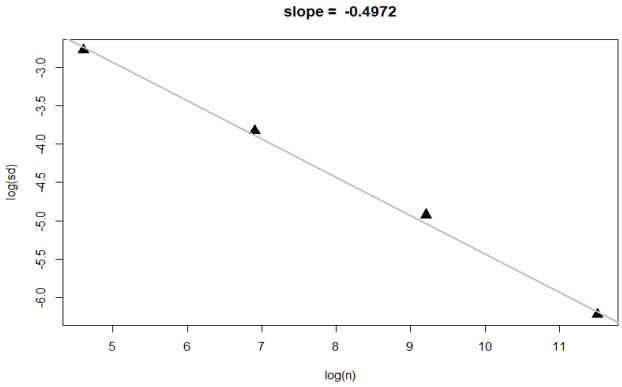
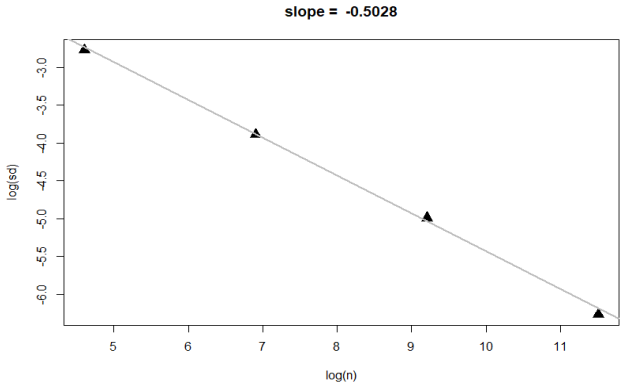


Figure 5.3: Study $\sqrt{n}$ rate of convergence for t_5

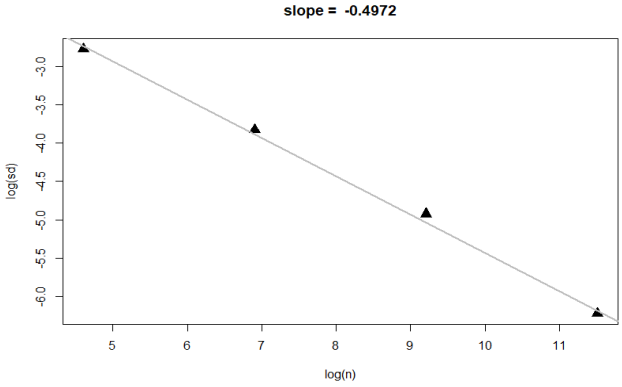
example of $\beta(a, b)$ which is not satisfy the form of f_{0j}.



$\tau = 0.6$, $\Gamma(5,1)$, $\Gamma(5,1)$



$\tau = 0.6$, Exp(1.5), Exp(1.5)



$\tau = 0.6$, $\beta(5,2)$, $\beta(5,2)$

Figure 5.4: Estimated rate of convergence for $\hat{\theta}$ when $f_0(x) = e^{\alpha_0 x} h_0(x)$

5.2.4 Regularity Conditions

(U) There is a unique solution θ_0 to $\Psi(\theta, F_0) = 0$.

(B) The state space of the copula parameter Θ is a bounded subset of $\mathbb{R}^k$.

(LC) The true density has the form $f_0 = e^{\alpha_0 x} h_0(x)$ and the support of f_0 is bounded if $\alpha_0 \neq 0$.

(R1) There exists a finite constant M_1 such that

$$\max_{j=1,\ldots,d} \sup_{\theta \in \Theta} E_{f_0}\left[\left|\partial_{u_j} \partial_\theta \log c(F_0(X);\theta)\right|\right] \leq M_1.$$

(R2) There exists a neighbourhood $\mathcal{N}$ such that $F_0 \in \mathcal{N}$ and a finite constant M_2 such that for all $F \in \mathcal{N}$

$$\max_{j,l=1,\ldots,d} \sup_{\theta \in \Theta} E_{f_0}\left[\left|\partial_{u_j} \partial_{u_l} \partial_\theta \log c(F(X);\theta)\right|\right] \leq M_2.$$

(R3) There exists a function $m(x)$ such that

$$\sup_{\theta \in \Theta} |\partial_\theta^2 \log c(F_0(x);\theta)| \leq m(x),$$

and $m(x)$ has a bounded moment $E[m^k(X)] < \infty$ for some integer $k \geq 1$.

(R4) $E_{f_0}[\|\partial_\theta \log c(F_0(X);\theta_0)\|^2] < \infty$

(R5) **(i)** There exists a neighbourhood $\mathcal{N}$ such that $(F_0, \theta_0) \in \mathcal{N}$ and a finite constant M_3 such that for all $(F, \theta) \in \mathcal{N}$

$$\max_{q,s=1,\ldots,k} E_{f_0}\left[\left|\partial_{\theta_q} \partial_{\theta_s} \partial_\theta \log c(F(X);\theta)\right|\right] \leq M_3.$$

(ii) There exists a neighbourhood $\mathcal{N}$ such that $(F_0, \theta_0) \in \mathcal{N}$ and a finite constant M_4 such that for all $(F, \theta) \in \mathcal{N}$

$$\max_{q=1,\ldots,k} \max_{j=1,\ldots,d} E_{f_0}\left[\left|\partial_{u_j}\partial_{\theta_q}\partial_\theta \log c(F(X);\theta)\right|\right] \leq M_4.$$

(R6) For $j = 1, \ldots, d$,

$$\sup_{F\in\overline{\mathcal{F}}_{M,lcd}} \sup_{\theta\in\Theta} |\partial_{\theta_l}\partial_\theta \log c(F(x);\theta)| \leq \xi_{\theta_l}(x)$$
$$\sup_{F\in\overline{\mathcal{F}}_{M,lcd}} \sup_{\theta\in\Theta} |\partial_{u_j}\partial_\theta \log c(F(x);\theta)| \leq \xi_j(x)$$

where $E_{f_0}[\xi_{\theta_l}(X)^2] < \infty$ and $E_{f_0}[\xi_j(X)^2] < \infty$.

(R7)

$$\sup_{F\in\overline{\mathcal{F}}_{M,lcd}} \sup_{\theta\in\Theta} |\partial\theta \log c(F(x);\theta)| \leq G(x),$$

where $E_{f_0}[G^2(X)] < \infty$.

(R8) There exists a neighbourhood $\mathcal{N}$ such that $(F_0, \theta_0) \in \mathcal{N}$ and a finite constant B_2 such that for all $(F, \theta) \in \mathcal{N}$

$$\sup_{(F,\theta)\in\mathcal{N}} \int \left(c(F(x);\theta)\right)^2 dx \leq B_2.$$

(R9) **(i)** There exists a neighbourhood $\mathcal{N}$ such that $(F_0, \theta_0) \in \mathcal{N}$ and a finite constant D_1 such that for all $(F, \theta) \in \mathcal{N}$

$$\max_{j=1,\ldots,d} \int \left|\partial_{u_j} c(F(x);\theta)\right| dx \leq D_1.$$

(ii) There exists a neighbourhood $\mathcal{N}$ such that $(F_0, \theta_0) \in \mathcal{N}$ and a finite constant D_2 such that for all $(F, \theta) \in \mathcal{N}$

$$\max_{l=1,\ldots,k} \int |\partial_{\theta_l} c(F(x); \theta)| \, dx \quad \leq \quad D_2.$$

6 Finite mixture models

In several applications, mixture models take part in data analysis; for example, in finance, customers are segmented into several groups where the financial institutions can treat them accordingly to their needs. In biology, mixture models help us categorizing flu strains into groups. We can see the evolution over the influenza seasons which leads to the developing of a vaccine, see Li et al. [2016]. Moreover, nowadays the trend of big data is booming. Several techniques of clustering and classification with mixture models have been used. Suppose we observe n random variables $X = \{X_1, \ldots, X_n\} \in \mathbb{R}^d$ with the observations represent an existence of subpopulations. Then, a finite mixture model is given by

$$f(x) = \sum_{j=1}^{k} \pi_j f_j(x), \tag{6.1}$$

where k denote a number of finite subpopulations, π_j are the mixing proportions for subpopulation jth, where $\pi_l = 1 - \sum_{j \neq l} \pi_j$, and $f_1, \ldots, f_k$ are the densities of subpopulation kth.

Since (6.1) has missing values, which is the mixing proportions π_j for all $j = 1, \ldots, k$, so we will use an expectation-maximization (EM) algorithm to find the MLE. First, we present a concept of EM algorithm. Then, we will focus on a Gaussian mixture model (GMM). After that, we will present the EM algorithm for log-concave mixture model (LCMM).

6.1 Concept of EM algorithm

The EM algorithm is an iterative method for finding MLE of a model with unobserved variables. It has two main steps, which are from its short notation E and M. First, we do the expectation (E-step), which is for filling in the missing group labels. Then, we maximize the model from E-step with the maximum likelihood estimation, so we call this step as the maximization (M-step). Suppose each f_j in (6.1) has corresponding parameters α_j. We denote $\lambda = \{\alpha_1, \ldots, \alpha_k, \pi_1, \ldots, \pi_k\}$ as a set of parameters for (6.1). Then, a likelihood function is given by

$$L(\lambda|x) = \prod_{i=1}^{n} \sum_{j=1}^{k} \pi_j f_j(x_i).$$

Also, a log-likelihood function can be represented as

$$\ell(\lambda|x) = \sum_{i=1}^{n} \log\left(\sum_{j=1}^{k} \pi_j f_j(x_i)\right). \tag{6.2}$$

However, we cannot find the ML estimator from (6.2) because of their missing values, so we introduce a new variable which is

$$w_{ij} = \begin{cases} 1, & \text{if } x_i \text{ is from } f_j, \\ 0, & \text{otherwise,} \end{cases}$$

where $1 \le i \le n$ and $1 \le j \le k$. Then, the log-likelihood function of x and w can be represented as

$$\begin{aligned} \ell(\lambda|x,w) &= \sum_{i=1}^{n} \log\left(\sum_{j=1}^{k} w_{ij} f_j(x_i)\right) \\ &\ge \sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij} \log f_j(x_i). \end{aligned} \tag{6.3}$$

The last inequality is from using Jensen's inequality and it is a lower bound of $\ell(\lambda|x,w)$. To simplify notation, we denote $\ell(\lambda|x,w)$ as $\ell(\lambda)$. In E-step, the algorithm will choose a lower bound that clings to $\ell(\lambda)$ and in M-step, that lower bound will be maximized. Since $\ell(\lambda)$ is bigger than the lower bound, $\ell(\lambda)$ will be increased too. We describe details of E and M-steps as follows.

E-step is to take an expectation of (6.3) with respect to the parameter λ that is

$$E_\lambda\left[\sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij} \log f_j(x_i)\Big|\lambda, x\right]. \tag{6.4}$$

In (6.4), only w_{ij} are random variables, so we can reduce (6.4) to

$$E_\lambda\left(w_{ij}|\lambda, x\right).$$

Since w_{ij} follow Bernoulli distribution. The expectation of Bernoulli distribution is its probability. Hence,

$$E_\lambda\left(w_{ij}|\lambda,x\right) \;=\; p \;=\; P_\lambda\left(w_{ij}=1|\lambda,x\right) \;=\; \frac{P_\lambda(w_{ij}=1)P_\lambda\left(X_i|w_{ij}=1,\lambda\right)}{\sum_{j=1}^{k} P_\lambda(w_{ij}=1)P_\lambda\left(X_i|w_{ij}=1,\lambda\right)}.$$

Therefore, the membership weights for each data point x_i in cluster j are given by

$$\widehat{w}_{ij} = \frac{\widehat{\pi}_j f_j(x_i|\lambda)}{\sum_{j=1}^{k} \widehat{\pi}_j f_j(x_i|\lambda)}. \tag{6.5}$$

Next, we use the membership weights in (6.5) to estimate $\widehat{\lambda}$. We do the maximum likelihood estimation, which maximizes the objective function Q in (6.6) with respect to $\alpha_1,\ldots,\alpha_k$. This is the **M-step** where function Q is given by

$$Q(\lambda|\widehat{w},x) = \sum_{i=1}^{n}\sum_{j=1}^{k} \widehat{w}_{ij}\log f_j(x_i|\lambda). \tag{6.6}$$

Therefore,

$$\widehat{\alpha}_j = \text{argmax } Q(\lambda|\widehat{w},x) \qquad \text{for } j=1,\ldots,k.$$

To summarize, the EM algorithm follows these steps:

1. Start with an initial guess for λ, say $\widehat{\lambda}^{(0)}$

2. For iteration rth and $j=1,\ldots,k$,

   E-step:

$$\widehat{w}_{ij}^{(r)} = \frac{\widehat{\pi}_j^{(r-1)} f_j(x_i|\widehat{\lambda}^{(r-1)})}{\sum_{j=1}^{k} \widehat{\pi}_j^{(r-1)} f_j(x_i|\widehat{\lambda}^{(r-1)})},$$

$$\widehat{\pi}_j^{(r)} = \frac{\sum_{i=1}^n \widehat{w}_{ij}^{(r)}}{n},$$

M-step:

$$\widehat{\alpha}_j^{(r)} = \text{argmax } Q(\lambda^{(r-1)}|\widehat{w}^{(r)}, x).$$

3. Step 2 is iterated until $\left|\frac{\ell(\widehat{\lambda}^{(r)})-\ell(\widehat{\lambda}^{(r-1)})}{\ell(\widehat{\lambda}^{(r-1)})}\right| < \epsilon$ for some small positive values of ϵ.

6.2 EM algorithm in Gaussian mixture models (GMM)

When f_j in (6.1) are estimated with Gaussian distributions, the mixture model is called GMM. A set of parameters for k subpopulations is $\lambda = \{\mu_j, \sigma_j^2, \pi_j\}$ for $j = 1, \ldots, k$. To estimate λ, we use EM algorithm where its M-step is to maximize the Q function, which is

$$Q(\lambda|\widehat{w}, x) = \sum_{i=1}^n \sum_{j=1}^k \widehat{w}_{ij} \log f_j(x_i|\mu_j, \sigma_j^2).$$

The explicit formulas for $\widehat{\mu}_j$ and $\widehat{\sigma}_j^2$ can be derived by setting the first derivative of the Q function with respect to each parameter equals to zero. Therefore, for each j we get

$$\widehat{\mu}_j = \frac{\sum_{i=1}^n \widehat{w}_{ij} x_i}{\sum_{i=1}^n \widehat{w}_{ij}}$$
$$\widehat{\sigma}_j^2 = \frac{\sum_{i=1}^n \widehat{w}_{ij}(x_i - \widehat{\mu}_j)^2}{\sum_{i=1}^n \widehat{w}_{ij}}.$$

As we can see, for GMM, the estimators $\widehat{\mu}_j$ and $\widehat{\sigma}_j^2$ are just the weighted mean and variance. Since we get $\widehat{w}_{ij}$ from E-step, the mixing proportions are given by

$$\widehat{\pi}_j = \frac{\sum_{i=1}^n \widehat{w}_{ij}}{n}.$$

6.3 EM algorithm in log-concave mixture models

Estimating (6.1) by using GMM does not perform well when each true density f_j is skewed. Hence, we will estimate f_j with log-concave distribution instead of Gaussian distribution. We call this model a log-concave mixture model (LCMM), that is given by

$$f(x) = \sum_{j=1}^{k} \pi_j \exp \varphi_j(x).$$

Steps of EM algorithm for LCMM are similar to the EM algorithm of GMM. The only difference is that each f_j is estimated with log-concave distribution instead of Gaussian distribution. Hence, in the maximization step, the estimator $\widehat{f_j}$ can be found by doing the log-concave MLE, which is available in the logcondens packgage for $d = 1$.

Note that the mixture models may be unbounded in some cases, for example, when f_j is a Gaussian distribution with $\mu_1 = x_1$ and $\sigma^2 \to 0$. The likelihood function will be infinite which is an unboundness issue for the mixture model. Similarly,

the mixtures of log-concave distributions may have this problem too. Hu et al. [2016] presented the algorithm, which defines a log-concave MLE on a constrained parameter space. Moreover, they claimed that if the algorithm starts with a good initial value, the unboundness problem will be very rare.

6.4 Simulation study

We perform a simulation study for multivariate LCMM by focusing on the perfomance of classification results. Chang and Walther [2007] studied multivariate LCMM via the copula model. They used a Gaussian copula to model dependence between two dimensions. The EM algorithm has been used where its M-step computes the log-concave ML estimators for each subpopulation. Similarly, our simulation study also use the Gaussian copula with both symmetric and skew marginal distributions. As we mentioned before, the copula model can solve the computationally intensive problem that occurs when we find the log-concave ML estimators with Shor's r-algorithm in the LogConcDEAD package. Our study will show percentages of misclassification cases of our proposed model compare with GMM, and multivariate log-concave with Shor's r-algorithm. Suppose we observe n random variables $X = \{X_1, \ldots, X_n\} \in \mathbb{R}^d$. The joint density function of LCMM is given by

$$f(x) = \sum_{j=1}^{k} \pi_j f_j(x),$$

where each $f_j(x)$ is the d-dimensional data with Gaussian copula and log-concave marginals. Its joint density function with correlation matrix R can be represented as

$$f_j(x) = c(F_1(x_1), \ldots, F_d(x_d); \theta) \prod_{l=1}^{d} \exp \varphi_l(x_l),$$

where

$$c(F_1(x_1), \ldots, F_d(x_d); \theta) = \frac{1}{\sqrt{\det R}} \exp\left(-\frac{1}{2} \begin{bmatrix} \Phi^{-1}(F_1(x_1)) \\ \vdots \\ \Phi^{-1}(F_d(x_d)) \end{bmatrix}^T \left(R^{-1} - I\right) \begin{bmatrix} \Phi^{-1}(F_1(x_1)) \\ \vdots \\ \Phi^{-1}(F_d(x_d)) \end{bmatrix} \right).$$

The random vectors $\Phi^{-1}(F_1(x_1)), \ldots, \Phi^{-1}(F_d(x_d))$ are quantile functions of cumulative distribution function of $X_1, \ldots, X_d$, respectively. Moreover, we consider the parameter of Gaussian copula when $\theta = 0$. In our study, we compare the simulation results of three methods, which are

1. Gaussian copula with log-concave marginals (proposed method),

2. GMM, and

3. multivariate log-concave with Shor's r-algorithm in LogConcDEAD package.

We compare the performance of each method by using percentages of average misclassification cases in which a formula can be represented as

$$\frac{1}{n}\left(\frac{\sum_{i=1}^{t} m_i}{t}\right) \times 100, \tag{6.7}$$

where m_i is the number of misclassification cases from simulation set ith for $i = 1, \ldots, t$. t and n denote sets of simulation and sample sizes, respectively. Details of the simulation study are in Table 6.1 and the corresponding results are shown in Table 6.2 and 6.3.

Table 6.1: Details for the simulation study

Dimensions (d)	2	
Number of subpopulations (k)	2	
Mixing proportions (π_1, π_2)	0.6, 0.4	
Distributions	**1st subpopulation**	**2nd subpopulation**
case I	N(2,2) and N(2,2)	N(7,2) and N(7,2)
case II	γ(2,2) and γ(5,2)	Beta$(2, 8)$ and Beta$(6, 8)$
Dependence	0	
Sample sizes (n)	50, 100, 300, 500, 1000	
Sets of simulation (t)	100	

Table 6.2: Classification results of case I: average number of misclassification cases from 100 simulation sets where the number in brackets are percentages from (6.7)

n	**GMM**	**Copula+log-concave margins**	**Multivariate log-concave**
50	0.49 (0.98%)	0.49 (0.98%)	0.55 (1.1%)
100	0.87 (0.87%)	0.95 (0.95%)	0.94 (0.94%)
300	1.91 (0.64%)	2.01 (0.67%)	2.55 (0.85%)
500	2.94 (0.59%)	3.16 (0.632%)	3.74 (0.75%)
1000	6.08 (0.61%)	6.17 (0.617%)	7.4 (0.74%)

As expected, the proposed method performs better than multivariate log-concave especially when the sample size is large. On the contrary, GMM performs the best in this case but not too far ahead of our proposed method. Hence, GMM is the best for symmetric distribution. However, the proposed method still performs good results.

Table 6.3: Classification results of case II: average number of misclassification cases from 100 simulation sets where the number in brackets are percentages from (6.7)

n	**GMM**	**Copula+log-concave margins**	**Multivariate log-concave**
50	0.55 (1.10%)	0.64 (1.28%)	0.51 (1.02%)
100	0.9 (0.90%)	0.74 (0.74%)	0.73 (0.73%)
300	2.42 (0.81%)	1.81 (0.60%)	2.04 (0.68%)
500	3.84 (0.77%)	3.1 (0.62%)	3.38 (0.68%)
1000	7.95 (0.79%)	5.71 (0.57%)	6.12 (0.61%)

Similar to case I, in case II the proposed method performs better than multivariate log-concave when the sample size is large. In contrast, GMM for skew distributions performs the worst unlike GMM for Gaussian distribution in case I. However, for small sample size, GMM still works well and performs better than the copula model.

7 Breast cancer data example

For a real data set, we study the performance of our proposed method in classification problem. We choose the Wisconsin breast cancer data set from `http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29` as an example. This is also an example that can be found in Cule et al. [2010]. This data set has 30 dimensions with 2 subpopulations, which are benign and malignant. The sample size is 569 where 357 are benign and 212 are malignant. The study of breast cancer data set follows these steps.

1. We do principal component analysis (PCA) of 30 dimensions and choose the first two components for the classification problem. The first two components can capture 63% of variability of the whole data set (44% in the 1st component, 19% in the 2nd component). A data plot can be found in Figure 7.1. The joint density function for breast cancer data set can be expressed as

$$f(x_1, x_2) = \pi_1 f_1(x_1, x_2) + \pi_2 f_2(x_1, x_2).$$

For our proposed method, f_j are modeled as the copula density with log-concave marginals, which can be represented as

$$f_j(x_1, x_2, \theta) = c(F_1(x_1), F_2(x_2); \theta)\Pi_{l=1}^{2} \exp \varphi_l(x_l); \quad j = 1, 2.$$

2. Before doing an EM algorithm, we choose a copula density for both subpopulations f_1 and f_2 with Bayesian information criterion (BIC). It turns out that Frank copula has been chosen for both subpopulations but with different parameters θ. Frank copula is one of the Archimedean copulas which its copula density function is given by

$$c(F_1(x_1), F_2(x_2)) = \frac{\theta e^{-\theta F_1(x_1) - \theta F_2(x_2)}(1 - e^{-\theta})}{\left[(e^{-\theta} - 1) + (e^{-\theta F_1(x_1)} - 1)(e^{-\theta F_2(x_2)} - 1)\right]^2}; \qquad \theta \in (-\infty, \infty) \backslash \{0\}.$$

We use BiCopSelect package for copula selection and the corresponding estimator $\widehat{\theta}$ is obtained from MLE. We get $\widehat{\theta} = -2.44$ and 0.18 for the first and second subpopulations, respectively. We use Frank copula for every iteration in EM algorithm with updated copula estimators $\widehat{\theta}$ from the M step.

3. Then, we do EM algorithm. E-step is for finding w_{ij} to estimate proportions π_1, π_2. M-step is for estimating the log-concave ML estimators and also copula estimator $\widehat{\theta}$. Hence, we get the group number for each individual observations.

4. Then, we compare our proposed method with GMM and multivariate log-concave EM.

We measure the performance of our proposed method by using percentage of misclassification cases. The results show that the misclassification cases of our proposed method, GMM, and multivariate log-concave EM are 49, 59 and 46, respectively. As expected, our proposed method performs much better than GMM but it is three more misclassification cases than multivariate log-concave that is 0.5% of the total 569 observations. However, to fit the mixture model, our proposed method uses only 30 seconds which much less computational time than multivariate log-concave that uses almost 30 minutes. All corresponding plots are given below.

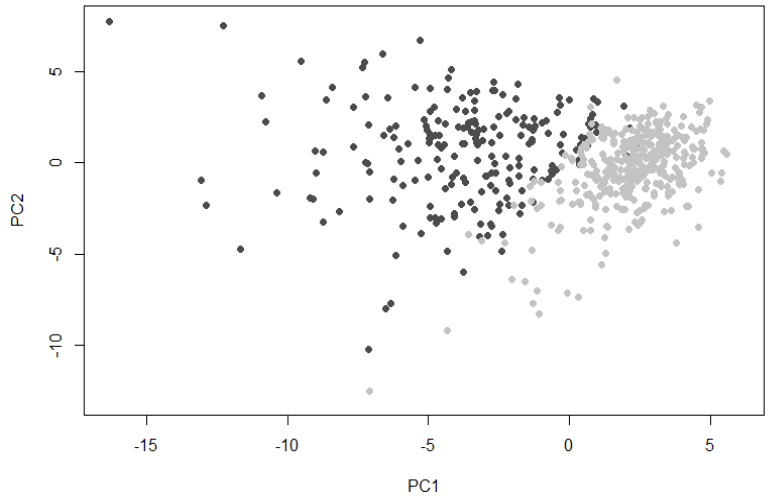
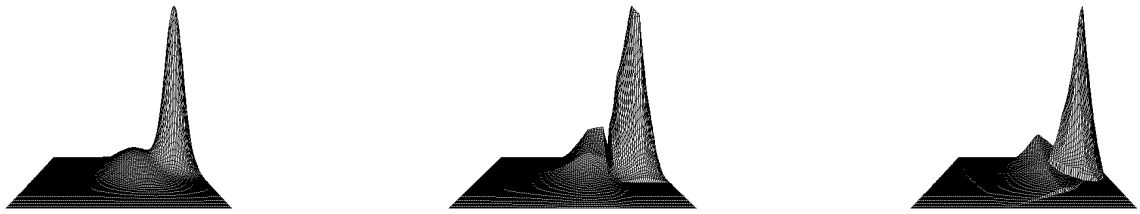
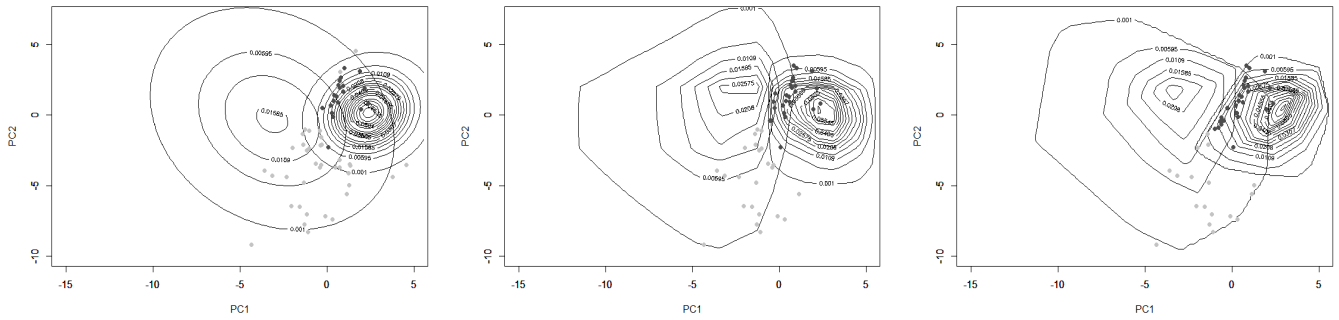


Figure 7.1: Breast cancer data with benign as light grey dots and malignant as dark grey dots



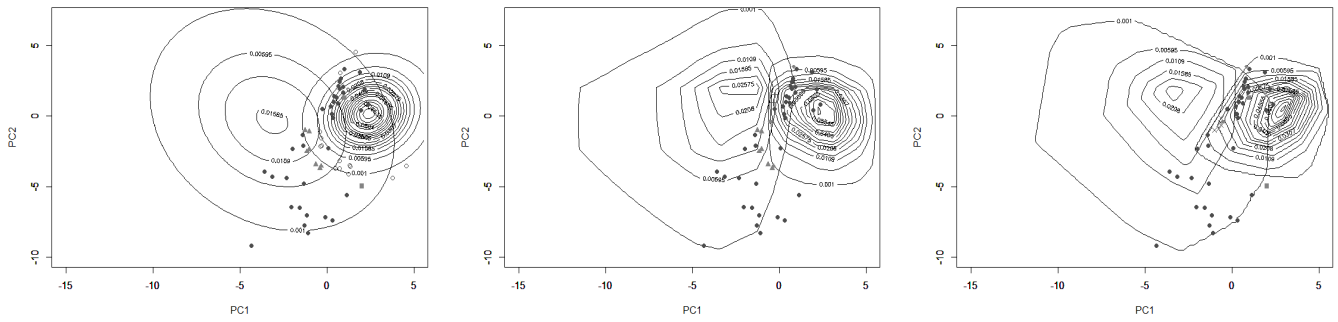
(a) (b) (c)

Figure 7.2: Surface plots for Breast cancer data set from (a) Gaussian mixture model (b) mixture of Frank copula with log-concave marginals (c) multivariate log-concave mixture



(a) (b) (c)

Figure 7.3: Contour plots with misclassification cases (benign as light grey dots and malignant as dark grey dots) for Breast cancer data set from (a) Gaussian mixture model (b) mixture of Frank copula with log-concave marginals (c) multivariate log-concave mixture



(a) (b) (c)

Figure 7.4: Contour plots with misclassification cases for Breast cancer data set from (a) Gaussian mixture model (b) mixture of Frank copula with log-concave marginals (c) multivariate log-concave mixture; each symbol is for misclassification cases in • all methods, ▪ both GMM & multivariate log-concave mixture, ▲ GMM & copula model, ◆ copula model & multivariate log-concave mixture, × only multivariate log-concave mixture, + only copula model, ○ only GMM

8 Further research

8.1 Vine copulas

Along the thesis, we model dependencies between dimensions with the same copula family. However, the dependencies can be modeled differently for each pair of dimensions, which is the idea of vine copula. The first regular vine is introduced from Joe [1994]. The concept of vine copulas relates to a mixture of both unconditional and conditional distribution functions. First, we look at the joint density function in terms of conditional density functions. Suppose we consider the d-dimensional data, the joint density function is given by

$$f(x_1, \ldots, x_d) = f(x_1|x_2, \ldots, x_d) f(x_2|x_3, \ldots, x_d) \cdots f(x_{d-1}|x_d) f(x_d). \tag{8.1}$$

From Sklar's theorem, the joint density in (8.1) can be written in terms of the copula density times the marginal distributions, which is

$$f(x_1, \ldots, x_d) = c\{F_1(x_1), \ldots, F_d(x_d)\} f_1(x_1) \cdots f_d(x_d).$$

Moreover, the conditional distribution functions can split into a set of pair-copula times marginal distribution functions. We present an example of three dimensional data where their joint density function is

$$f(x_1, x_2, x_3) = f(x_1|x_2, x_3)f(x_2|x_3)f(x_3). \tag{8.2}$$

Each term of conditional density function in (8.2) is given by

$$\begin{aligned}
f(x_2|x_3) &= \frac{f(x_2, x_3)}{f(x_3)} \\
&= c_{23}\{F_2(x_2), F_3(x_3)\}f_2(x_2), \\
f(x_1|x_3) &= \frac{f(x_1, x_3)}{f(x_3)} \\
&= c_{13}\{F_1(x_1), F_3(x_3)\}f_1(x_1), \\
f(x_1|x_2, x_3) &= \frac{f(x_1, x_2|x_3)}{f(x_2|x_3)} \\
&= \frac{c_{12|3}\{F_1(x_1|x_3), F_2(x_2|x_3)\}f(x_1|x_3)f(x_2|x_3)}{f(x_2|x_3)} \\
&= c_{12|3}\{F_1(x_1|x_3), F_2(x_2|x_3)\}f(x_1|x_3) \\
&= c_{12|3}\{F_1(x_1|x_3), F_2(x_2|x_3)\}c_{13}\{F_1(x_1), F_3(x_3)\}f_1(x_1).
\end{aligned}$$

Plug in all conditional density functions into (8.2), we gets

$$f(x_1, x_2, x_3) = c_{13}\{F_1(x_1), F_3(x_3)\}c_{23}\{F_2(x_2), F_3(x_3)\}c_{12|3}\{F_1(x_1|x_3), F_2(x_2|x_3)\}f_1(x_1)f_2(x_2)f_3(x_3) \tag{8.3}$$

As we can see, the joint density function can be written as the product of pair copula densities and marginal densities where the pair-copula densities consist of

unconditional pairs such as c_{13}, c_{23} and conditional pair such as $c_{12|3}$. More than that, each pair of copulas can come from different copula families. However, the decomposition of (8.3) is not unique. There are some other ways of writing the joint density function for three variables, for instance,

$$f(x_1, x_2, x_3) = c_{12}\{F_1(x_1), F_2(x_2)\}c_{23}\{F_2(x_2), F_3(x_3)\}c_{13|2}\{F_1(x_1|x_2), F_3(x_3|x_2)\}f_1(x_1)f_2(x_2)f_3(x_3).$$

For higher dimensions, the number of pair-copula is $d(d-1)/2$, which depends on the dimension d. Bedford and Cooke [2001] introduced a graphical model called "regular vine" (R-vine) that helps to organize the complexity of the pair-copula construction. Moreover, regular vine consists of several ways of writing. We will concentrate on two famous types, which are drawable vine (D-vine) and canonical vine (C-vine). In Figure 8.1 and 8.2, there are three notations to be clarified, which are variables, trees, and edges. The variables are data for each dimension, trees are represented as T_i and edges are line connected between dimensions. Figure 8.1 shows an example of six dimensions for the D-vine copula with 5 trees and 15 edges. In addition, Figure 8.2 represents the C-vine diagram with the same number of trees and edges as D-vine. Note that when the first tree is decided, the following trees are straightforward and there is only one way to be assigned.

Definition 8.1. Dibmanna et al. [2012] Let $\mathcal{V} = (T_1, \ldots, T_{n-1})$ is an R-vine on n elements if

1. T_1 is a tree with nodes $N_1 = \{1, \ldots, n\}$ and a set of edges denoted E_1.

2. For $i = 2, \ldots, n-1$, T_i is a tree with nodes $N_i = E_{i-1}$ and edge set E_i.

3. For $i = 2, \ldots, n-1$ and $\{a, b\} \in E_i$ with $a = \{a_1, a_2\}$ and $b = \{b_1, b_2\}$ it must hold that $\#(a \cap b) = 1$ (proximity condition).

Steps for using vine copulas can be summarized as follows:

1. model selection, i.e., selecting which conditioned and unconditioned pairs to use,

2. choosing bivariate copula family for each pair from the first step, where details of the copula families and their corresponding parameters are in Chapter 3, and

3. estimating all parameters corresponding to the copula family that has been chosen from the previous step.

For model selection in the first step, it is done tree by tree via some model selection methods such as optimal C-vines structure selection, see Czado et al. [2012], or the traveling salesman problem for D-vines. An objective of model selection is to choose the model that can capture the most dependence of the lower trees. For the copula selection in the second step, AIC, BIC, or goodness of fit tests can be used. Finally,

we do the parameter estimation. The MLE can also be done. However, sequential estimation is another choice where the parameters are estimated sequentially from the top of the tree. Moreover, sometimes the sequential estimation is used for the starting values of the MLE.

8.2 Asymptotic normality for copula estimator

We already show that the copula estimator has the $\sqrt{n}$ rate of convergence. However, the asymptotic normality of the copula estimator under the log-concave marginals still be an open problem to study.

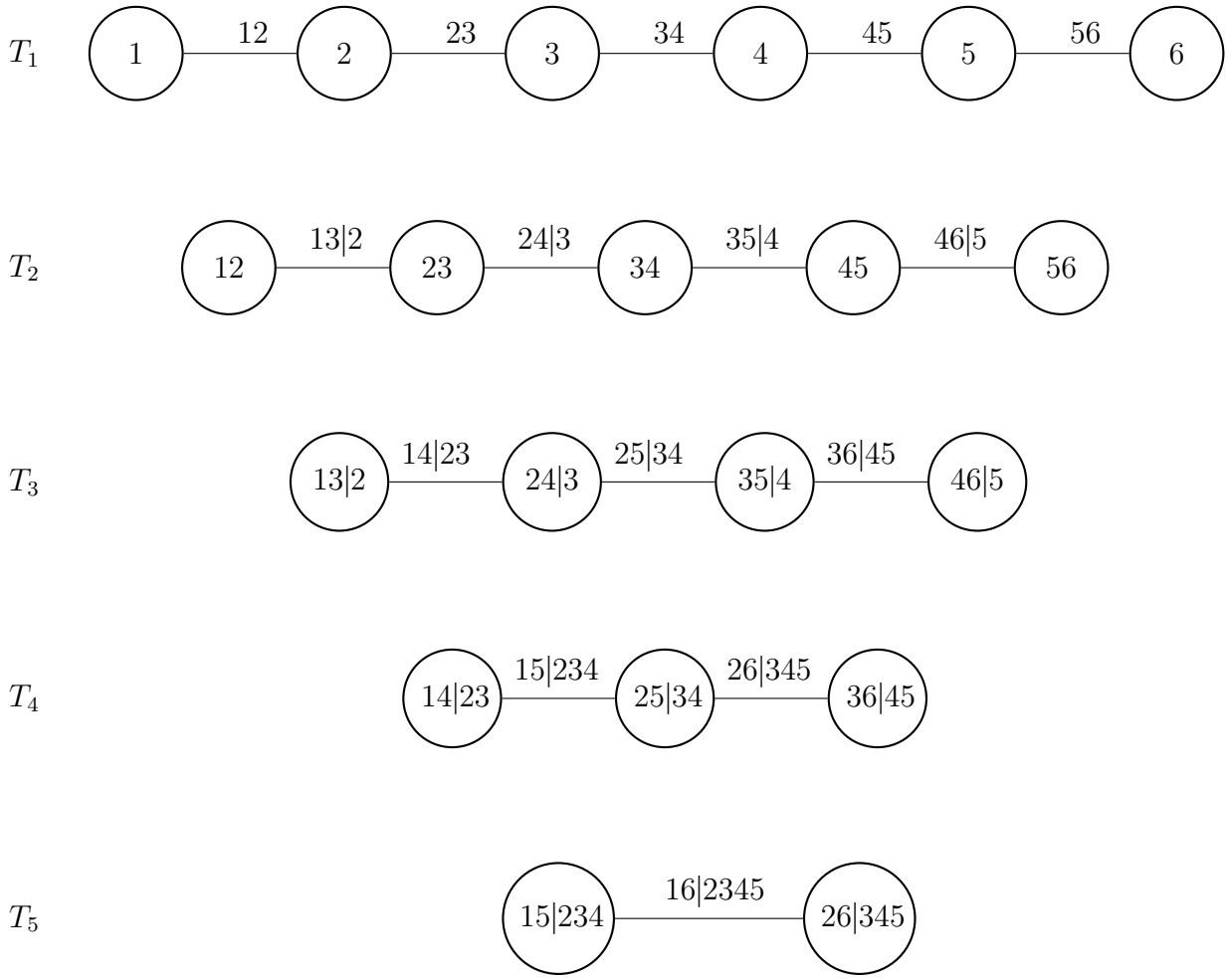


Figure 8.1: D-vine with 6 dimensions, 5 trees and 15 edges

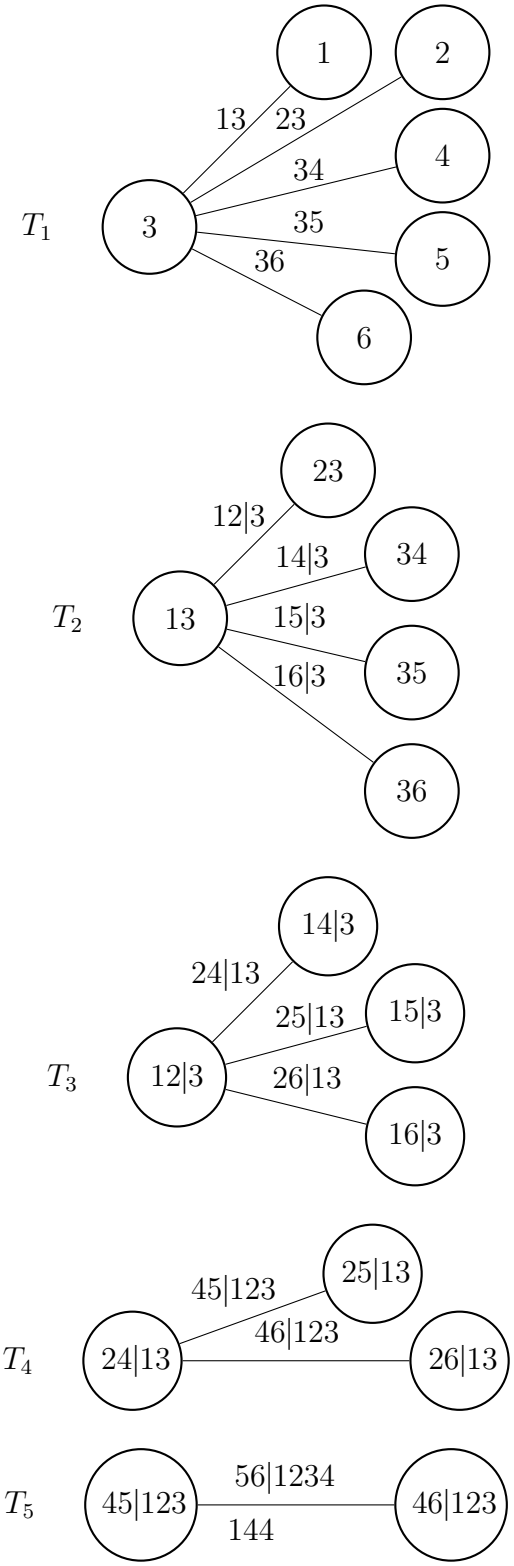


Figure 8.2: C-vine with 6 dimensions, 5 trees and 15 edges

9 Clustering using log-concave densities in $d = 1$

In this chapter, we focus on a clustering problem with one-dimensional log-concave densities. Our objective is to study the performance of using the log-concave MLE for estimating the density of each subpopulation compare with the classical GMM. We also propose a new criterion for selecting the number of subpopulations. Our proposed criterion derives from the concept of Bayesian approach which is similar to the derivation of BIC. We will present the derivation of our proposed criterion, which is called "proposed BIC". Then, we will do some simulation studies by using the log-concave mixture model (LCMM) with this proposed criterion.

9.1 Derivation of proposed BIC under LCMM

Let λ be a set of parameters in a model M, the Bayesian approach for a model selection is to maximize a posterior probability of a model M given data $x = \{x_1, \ldots, x_n\}$.

We get the posterior probability of the model as

$$\mathrm{P}(M \mid x) \;=\; \frac{P(x \mid M)P(M)}{P(x)}.$$

If we assume that all candidate models are equally likely, maximizing the posterior probability will be the same as maximizing the marginal likelihood $P(x \mid M)$. Let $p(\lambda)$ denote a prior density. Hence,

$$\begin{aligned} \mathrm{P}(x \mid M) &= \int f(x \mid \lambda)p(\lambda)\; d\lambda \\ &= \int \exp[\log f(x \mid \lambda)]p(\lambda)\; d\lambda. \end{aligned} \tag{9.1}$$

Let $\ell(\lambda) = \log f(x \mid \lambda)$ denote a log-likelihood function. Then, we do the Taylor series around the maximum likelihood estimator $\widehat{\lambda}$. Therefore, we get

$$\ell(\lambda) \;=\; \ell(\widehat{\lambda}) + (\lambda - \widehat{\lambda})^T \nabla_{\widehat{\lambda}} \ell(\widehat{\lambda}) + \frac{1}{2}(\lambda - \widehat{\lambda})^T H_{\widehat{\lambda}} (\lambda - \widehat{\lambda}) + o_p(1),$$

where $H_{\widehat{\lambda}} = \partial_\lambda \partial_\lambda^T \ell(\widehat{\lambda})$. We know that at the ML etimator $\nabla_{\widehat{\lambda}} \ell(\widehat{\lambda}) = 0$. Thus,

$$\ell(\lambda) \;=\; \ell(\widehat{\lambda}) + \frac{1}{2}(\lambda - \widehat{\lambda})^T H_{\widehat{\lambda}} (\lambda - \widehat{\lambda}) + o_p(1). \tag{9.2}$$

Similarly, Taylor's expansion of the prior density at $\widehat{\lambda}$ is given by

$$p(\lambda) = p(\widehat{\lambda}) + (\lambda - \widehat{\lambda})^T \partial_\lambda p(\widehat{\lambda}) + o_p(1). \tag{9.3}$$

Then, substituting (9.2) and (9.3) in (9.1) and ignoring the term of $o_p(1)$, the marginal likelihood function can be expressed as

$$\begin{aligned} P(x \mid M) &\approx \int \exp\left(\ell(\widehat{\lambda}) + \frac{1}{2}(\lambda - \widehat{\lambda})^T H_{\widehat{\lambda}}(\lambda - \widehat{\lambda})\right)\left(p(\widehat{\lambda}) + (\lambda - \widehat{\lambda})^T \partial_\lambda p(\widehat{\lambda})\right) d\lambda \\ &\approx \exp \ell(\widehat{\lambda}) p(\widehat{\lambda}) \int \exp\left(\frac{1}{2}(\lambda - \widehat{\lambda})^T H_{\widehat{\lambda}}(\lambda - \widehat{\lambda})\right) d\lambda. \end{aligned}$$

Let $J_{\widehat{\lambda}} = -H_{\widehat{\lambda}}$ denote the Fisher information matrix and let $Y = (\lambda - \widehat{\lambda})$. Thus,

$$P(x \mid M) \approx \exp \ell(\widehat{\lambda}) p(\widehat{\lambda}) \int \exp\left(-\frac{1}{2} Y^T J_{\widehat{\lambda}} Y)\right) d\lambda.$$

Since $J_{\widehat{\lambda}}$ is a symmetric matrix, we can diagonalize it as $J_{\widehat{\lambda}} = S^T \Lambda S$ for some orthogonal matrices S where Λ is a diagonal matrix. S is a unitary matrix where $S^T S = S S^T$, which $\det(S) = 1$. Then, we make substitution $Y = S^T U$. The Jacobian matrix $J_{mn}(U) = \frac{\partial Y_m}{\partial U_n}$, then $J(U) = S^T$ and $\det J(U) = 1$. Hence,

$$\begin{aligned} P(x \mid M) &\approx \exp \ell(\widehat{\lambda}) p(\widehat{\lambda}) \int \exp\left\{-\frac{1}{2}(U^T \Lambda U)\right\} \det J(U) du \\ &= \exp \ell(\widehat{\lambda}) p(\widehat{\lambda}) \int \exp\left\{-\frac{1}{2} \sum_{j=1}^{|\lambda|} \lambda_j u_j^2\right\} du \\ &= \exp \ell(\widehat{\lambda}) p(\widehat{\lambda}) \prod_{j=1}^{|\lambda|} \int \exp\left\{-\frac{1}{2} \lambda_j u_j^2\right\} du, \end{aligned}$$

where λ_j is the jth eigenvalue of the matrix $J_{\widehat{\lambda}}$.

Next, we use Laplace's method for estimating an integral. Laplace's method hold when $f(x)$ has a unique global maximum at x_0 and it decays to zero away from its

maximum where M is a large number and $f''(x_0) < 0$. Then,

$$\int_a^b e^{Mf(x)}dx \approx \sqrt{\frac{2\pi}{M|f''(x_0)|}}e^{Mf(x_0)} \quad \text{as } M \to \infty.$$

Therefore,

$$\begin{aligned} P(x \mid M) &\approx \exp \ell(\widehat{\lambda})p(\widehat{\lambda}) \prod_{j=1}^{|\lambda|} \sqrt{\frac{2\pi}{\frac{1}{2}2\lambda_j}} \qquad (\text{by } \ M = \frac{1}{2} \text{ and } f(u) = -\lambda u^2) \\ &= \exp \ell(\widehat{\lambda})p(\widehat{\lambda}) \frac{(2\pi)^{\frac{|\lambda|}{2}}}{\prod_{j=1}^{\lambda} \lambda_j^{\frac{1}{2}}} \\ &= \exp \ell(\widehat{\lambda})p(\widehat{\lambda}) \frac{(2\pi)^{\frac{|\lambda|}{2}}}{|J_{\widehat{\lambda}}|^{\frac{1}{2}}}. \end{aligned} \tag{9.4}$$

When we take logarithm to the equation (9.4) and multiply by -2, we get

$$-2\log P(x \mid M) \approx -2\ell(\widehat{\lambda}) - 2\log p(\widehat{\lambda}) - |\lambda|\log(2\pi) + \log|J_{\widehat{\lambda}}|.$$

Moreover, when $p(\lambda)$ is a uniform prior, $p(\widehat{\lambda}) = 1$. Hence,

$$-2\log P(x \mid M) \approx -2\ell(\widehat{\lambda}) - |\lambda|\log(2\pi) + \log|J_{\widehat{\lambda}}|. \tag{9.5}$$

Each element in the observed Fisher information matrix at the ML estimators $\widehat{\lambda}$ can be expressed as

$$\begin{aligned} J_{pq} &= -\partial_{\lambda_p}\partial_{\lambda_q}\ell\left(\widehat{\lambda}|x\right) \\ &= -\sum_{i=1}^{n} \partial_{\lambda_p}\partial_{\lambda_q}\ell\left(\widehat{\lambda}|x_i\right) \\ &= -\frac{1}{n}\sum_{i=1}^{n} n\partial_{\lambda_p}\partial_{\lambda_q}\ell\left(\widehat{\lambda}|x_i\right). \end{aligned}$$

Since the data are i.i.d and n is large, we can apply the weak law of large numbers on the random variable $n\ell(\widehat{\lambda}|x_i)$. Hence, we get

$$\frac{1}{n}\sum_{i=1}^{n} n\ell(\widehat{\lambda}|x_i) \xrightarrow{P} E\big[n\ell(\widehat{\lambda}|X)\big].$$

By using the weak law of large numbers, each element in the observed Fisher information matrix is given by

$$\begin{aligned} J_{pq} &= -\partial_{\lambda_p}\partial_{\lambda_q} E\big[n\ell(\widehat{\lambda}|X)\big] \\ &= -n\partial_{\lambda_p}\partial_{\lambda_q} E\big[\ell(\widehat{\lambda}|X)\big] \\ &= -n\partial_{\lambda_p}\partial_{\lambda_q} E\big[\ell(\widehat{\lambda}|X_1)\big] \\ &= nI_{pq}. \end{aligned}$$

Therefore,

$$|J_{\widehat{\lambda}}| = n^{|\lambda|}|I_{\widehat{\lambda}}|. \tag{9.6}$$

Then, we plug in the result from (9.6) into (9.5). Hence, we get

$$\begin{aligned} -2\log P(x \mid M) &\approx -2\ell(\widehat{\lambda}) - |\lambda|\log(2\pi) + |\lambda|\log n + \log|I_{\widehat{\lambda}}| \\ &= -2\ell(\widehat{\lambda}) + |\lambda|\{\log n - \log(2\pi)\} + \log|I_{\widehat{\lambda}}|. \end{aligned} \tag{9.7}$$

For a large n, a classical BIC ignores the terms that do not depend on the sample size. Hence, the classical BIC can be expressed as

$$BIC = -2\ell(\widehat{\lambda}) + |\lambda|\log n. \tag{9.8}$$

On the contrary, LCMM has the form

$$f(x) = \sum_{j=1}^{k} \pi_j \exp \varphi_j(x).$$

The unknown parameters are $\{k, \pi_1, \ldots, \pi_k, \varphi_1, \ldots, \varphi_k\}$ where $\varphi_j : \mathbb{R} \to [-\infty, \infty)$ are concave functions. Since $\widehat{\varphi}_j$ are the ML estimators for each $j = 1, \ldots, k$, $\widehat{\varphi}_j$ are piecewise linear on $[X_{(1)}, X_{(n)}]$ and contains m_j knots. Recall that knots occurs at some points of data $[X_{(1)}, X_{(n)}]$ and $X_{(1)}, X_{(n)}$ always be knots, then $2 < m_j < n$. Moreover, Balabdaoui et al. [2009] showed that m_j also depends on the sample size. Let $\mathcal{K}_{ji} \in [X_{(1)}, X_{(n)}]$ denote knot ith of subpopulation jth, then a set of knots for each subpopulation j is $\{\mathcal{K}_{j1}, \mathcal{K}_{j2}, \ldots, \mathcal{K}_{jm_j}\}$. At each piecewise linear $\widehat{\varphi}_j$, the slopes can also be denoted as $\{S_{j1}, S_{j2}, \ldots, S_{j,m_j-1}\}$ where S_{ji} denote the slope between $\widehat{\varphi}_j(\mathcal{K}_{ji})$ and $\widehat{\varphi}_j(\mathcal{K}_{j,i+1})$. Hence, the number of knots and the number of slopes for k subpopulations are $p = \sum_{j=1}^{k} m_j$ and $s = \sum_{j=1}^{k}(m_j - 1) = p - k$, respectively. Moreover, we also need one starting point for each subpopulation. Therefore, $|\lambda|$ in (9.7) comes from knots $\mathcal{K}$, slopes S, starting points, and mixing proportions π. Hence, $|\lambda| = p + (p - k) + k + (k - 1) = 2p + k - 1$.

For the last part of (9.7), we need to calculate the $|I_{\widehat{\lambda}}|$. The observed Fisher information matrix for the LCMM consists of three parts which are φ, $\mathcal{K}$, and π. For $j = 1, \ldots, k$, let $I(\widehat{\varphi}_j)$ denote the Fisher information matrix of the jth subpopulation, which contains the information about both φ and $\mathcal{K}$, M denote the observed Fisher

information matrix that contains the information for k mixing proportions which follows a multinomial distribution. Then, $I_{\widehat{\lambda}}$ is given by

$$I_{\widehat{\lambda}} = I(\widehat{\lambda}) = diag\{I(\widehat{\varphi}_1), \ldots, I(\widehat{\varphi}_k), M\}.$$

Then,

$$|I(\widehat{\lambda})| = |M| \prod_{j=1}^{k} |I(\widehat{\varphi}_j)|,$$

where $|M| = (\pi_1 \cdots \pi_k)^{-1}$. Therefore,

$$|I(\widehat{\lambda})| = \frac{\prod_{j=1}^{k} |I(\widehat{\varphi}_j)|}{\prod_{j=1}^{k} \pi_j}.$$

Finally, we get

$$\log |I(\widehat{\lambda})| = \sum_{j=1}^{k} \log |I(\widehat{\varphi}_j)| - \sum_{j=1}^{k} \log \pi_j. \tag{9.9}$$

Substituting (9.9) into (9.7), then

$$-2 \log P(x \mid M) = -2\ell(\widehat{\lambda}) + |\lambda| \{\log n - \log(2\pi)\} - \sum_{j=1}^{k} \log \pi_j + \sum_{j=1}^{k} \log |I(\widehat{\varphi}_j)|. \tag{9.10}$$

An objective of the last three terms in (9.10) is for penalizing the log-likelihood function in the first term. Since when n large, $\log n$ always dominates $\log(2\pi)$. Therefore, we can ignore the $\log(2\pi)$ term. Then, the "proposed BIC" is given by

$$\text{proposed} BIC = -2\ell(\widehat{\lambda}) + |\lambda| \log n - \sum_{j=1}^{k} \log \pi_j + \sum_{j=1}^{k} \log |I(\widehat{\varphi}_j)|, \tag{9.11}$$

where $|\lambda| = 2p + k - 1$.

Next, we discuss how to calculate the observed Fisher information matrix for the LCMM. As mentioned above, $I(\widehat{\varphi}_j)$ contains the information of φ and $\mathcal{K}$. The derivatives for φ are already presented in (2.5) and (2.6) where the derivatives for the $\mathcal{K}$ are stated below. Note that $\mathcal{K} \in [X_{(1)}, X_{(n)}]$.

We use the directional derivative to derive the observed Fisher information matrix. For simplicity, let φ and φ_0 denote the short terms of $\varphi(x)$ and $\varphi(x_0)$, respectively. Then, let $\varphi = \varphi_0 + \epsilon\Delta$ where Δ is a suitable basis function. According to Dümbgen et al. [2011], we use $\Delta_i = \min(x - x_i, 0)$ where $2 \le i \le m$, m is the number of knots, and x_i are locations of the ith knot. Note that $\Delta_1 = 1$ and $\int_{x_1}^{x_m} e^{\varphi_0 + \epsilon\Delta} dx = 1$. Then, the log function of the subpopulation jth is given by

$$\begin{aligned}
\ell(\varphi) &= \log(e^{\varphi_0 + \epsilon\Delta}) \\
&= \log\left(\frac{e^{\varphi_0 + \epsilon\Delta}}{\int_{x_1}^{x_m} e^{\varphi_0 + \epsilon\Delta} dx}\right) \\
&= \varphi_0 + \epsilon\Delta - \log \int_{x_1}^{x_m} e^{\varphi_0 + \epsilon\Delta}\, dx.
\end{aligned}$$

To get rid of subscript j, let $\widehat{\varphi}$ be the log-concave ML estimators of subpopulation jth, then

$$\tilde{\varphi} = \widehat{\varphi} + \epsilon\Delta - \log \int_{x_1}^{x_m} e^{\widehat{\varphi} + \epsilon\Delta}\, dx.$$

Therefore, the first derivative can be calculated from

$$\begin{aligned}
\lim_{\epsilon\to 0}\frac{\ell(\tilde{\varphi})-\ell(\widehat{\varphi})}{\epsilon} &= \lim_{\epsilon\to 0}\left\{\int_{x_1}^{x_m}\Delta d\mathbb{F}_n(x)-\frac{1}{\epsilon}\log\int_{x_1}^{x_m}e^{\widehat{\varphi}+\epsilon\Delta}\,dx\right\}\\
&= \int_{x_1}^{x_m}\Delta d\mathbb{F}_n(x)-\lim_{\epsilon\to 0}\frac{1}{\epsilon}\log\int_{x_1}^{x_m}e^{\widehat{\varphi}+\epsilon\Delta}\,dx\\
&= \int_{x_1}^{x_m}\Delta d\mathbb{F}_n(x)-\lim_{\epsilon\to 0}\frac{1}{\epsilon}\left(\log\int_{x_1}^{x_m}e^{\widehat{\varphi}+\epsilon\Delta}\,dx-\log 1\right)\\
&= \int_{x_1}^{x_m}\Delta d\mathbb{F}_n(x)-\lim_{\epsilon\to 0}\frac{1}{\epsilon}\left(\log\int_{x_1}^{x_m}e^{\widehat{\varphi}+\epsilon\Delta}\,dx-\log\int_{x_1}^{x_m}e^{\widehat{\varphi}}dx\right)\\
&= \int_{x_1}^{x_m}\Delta d\mathbb{F}_n(x)-\lim_{\epsilon\to 0}\frac{\int_{x_1}^{x_m}\Delta e^{\widehat{\varphi}+\epsilon\Delta}dx}{\int_{x_1}^{x_m}e^{\widehat{\varphi}+\epsilon\Delta}\,dx}\\
&= \int_{x_1}^{x_m}\Delta d\mathbb{F}_n(x)-\int_{x_1}^{x_m}\Delta e^{\widehat{\varphi}}\,dx.
\end{aligned}$$

Then, the second derivative can be represented as

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon}\left(\int_{x_1}^{x_m}\Delta_k d\mathbb{F}_n(x)-\int_{x_1}^{x_m}\Delta_k e^{\widehat{\varphi}+\epsilon\Delta_i}dx-\int_{x_1}^{x_m}\Delta_k d\mathbb{F}_n(x)+\int_{x_1}^{x_m}\Delta_k e^{\widehat{\varphi}}dx\right)$$

$$\begin{aligned}
&= \lim_{\epsilon\to 0}\frac{1}{\epsilon}\left(\int_{x_1}^{x_m}\Delta_k e^{\widehat{\varphi}}dx-\int_{x_1}^{x_m}\Delta_k e^{\widehat{\varphi}+\epsilon\Delta_i}\,dx\right)\\
&= \lim_{\epsilon\to 0}\left(-\int_{x_1}^{x_m}\Delta_k\Delta_i e^{\widehat{\varphi}+\epsilon\Delta_i}\,dx\right)\\
&= -\int_{x_1}^{x_m}\Delta_k\Delta_i e^{\widehat{\varphi}}\,dx\\
&= -\sum_{i=2}^{m}\int_{x_{i-1}}^{x_i}\Delta_k\Delta_i e^{\widehat{\varphi}}\,dx.
\end{aligned}$$

Therefore, for $i = k = 1$,

$$\partial_{x_1}^2\ell(\widehat{\varphi}) = -1,$$

for $2 \le i = k \le m$,

$$
\begin{aligned}
-\int_{x_{i-1}}^{x_i} (\Delta_i)^2 e^{\widehat{\varphi}}\, dx &= -\int_{x_{i-1}}^{x_i} (\min(x - x_i, 0))^2 e^{\widehat{\varphi}}\, dx \\
&= -\int_{x_{i-1}}^{x_i} (x - x_i)^2 e^{\widehat{\varphi}}\, dx. \qquad (9.12)
\end{aligned}
$$

By following Dümbgen et al. [2011, page 5], we can rewrite (9.12) as follows:

$$
\partial^2_{x_i} \ell(\widehat{\varphi}) \;=\; \delta^3_{i-1} J_{11}(\varphi_{i-1}, \varphi_i) - \delta^3_{i-1} J_{10}(\varphi_{i-1}, \varphi_i).
$$

For $1 \le i = k - 1 < m$,

$$
\begin{aligned}
-\int_{x_{i-1}}^{x_i} \Delta_i \Delta_{i+1} e^{\widehat{\varphi}}\, dx &= -\int_{x_{i-1}}^{x_i} \min(x_{i-1} - x, 0) \min(x - x_i, 0) e^{\widehat{\varphi}}\, dx \\
&= -\int_{x_{i-1}}^{x_i} (x_{i-1} - x)(x - x_i) e^{\widehat{\varphi}}\, dx. \qquad (9.13)
\end{aligned}
$$

We can rewrite (9.13) by following Dümbgen et al. [2011, page 17]. Hence,

$$
\partial_{x_i} \partial_{x_{i-1}} \ell(\widehat{\varphi}) \;=\; -\delta^3_{i-1} J_{11}(\varphi_{i-1}, \varphi_i).
$$

Finally, we need partial derivatives for knots and locations.

$$\partial_{\varphi_i}\partial_{x_k}\ell(\widehat{\varphi}) = \begin{cases} J_{10}(\varphi_1,\varphi_2) & \text{for } i=k=1, \\ J_{10}(\varphi_i,\varphi_{i+1}) - J_{10}(\varphi_i,\varphi_{i-1}) & \text{for } 2\le i=k<m, \\ -J_{01}(\varphi_{m-1},\varphi_m) & \text{for } i=k=m, \\ -J_{10}(\varphi_i,\varphi_{i+1}) & \text{for } 1\le i=k-1<m, \\ J_{01}(\varphi_{i-1},\varphi_i) & \text{for } 2\le i=k+1\le m, \\ 0 & \text{for } |i-k|>1. \end{cases}$$

9.2 Simulation studies

An objective of this simulation is to study the clustering performance of univariate LCMM by using the proposed BIC in (9.11). We compare our proposed method with the classical BIC in (9.8), where the mixture models have been considered in two ways, which are GMM and LCMM. These three methods can be summarized as follows.

Table 9.1: Details for simulation study

Number of subpopulations (k)	2	
Mixing proportions (π_1, π_2)	0.6, 0.4	
Distribution	**1st subpopulation**	**2nd subpopulation**
case I	N(2,2)	N(8,2)
case II	Beta$(2,5)$	Beta$(2,5)+5$
Parameters	caseI: $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi_1, \pi_2$	
	caseII: $\alpha_1, \alpha_2, \beta_1, \beta_2, \pi_1, \pi_2$	
Sample sizes (n)	500, 1000, 2000, 3000	
Sets of simulation (t)	200	

1. LCMM with proposed BIC (proposed method)

$$\text{proposed}BIC = -2\ell(\widehat{\lambda}) + |\lambda| \log n - \sum_{j=1}^{k} \log \pi_j + \sum_{j=1}^{k} \log |I(\widehat{\varphi}_j)|$$

2. LCMM with BIC

$$BIC = -2\ell(\widehat{\lambda}) + |\lambda| \log n$$

3. GMM with BIC

$$BIC = -2\ell(\widehat{\lambda}) + |\lambda| \log n$$

Table 9.2 shows percentages of choosing the correct number of subpopulations for each method. The LCMM with our proposed BIC performs similarly to the GMM

Table 9.2: Clustering results for univariate case

Sample sizes	**500**	**1,000**	**2,000**	**3,000**
Case I: Gaussian distribution				
GMM with BIC	100%	99.5%	100%	100%
LCMM with proposed BIC (proposed method)	99.5%	100%	99.5%	99.5%
LCMM with BIC	68.5%	97%	95.5%	97%
Case II: beta distribution				
GMM with BIC	5.5%	0%	0%	0%
LCMM with proposed BIC (proposed method)	99.5%	100%	100%	100%
LCMM with BIC	98%	100%	99.5%	97%

with BIC in Gaussian case but performs much better than GMM with BIC in beta distribution. Moreover, when we compare our proosed BIC with classical BIC in the LCMM, the proposed BIC gives better reults in all cases. Furthermore, GMM with classical BIC works well with Gaussian distribution as it usually does. Unlike skew distributions such as beta distribution, GMM provides the worst results since it is suitable for symmetric distributions.

Bibliography

Fadoua Balabdaoui, Kasper Rufibach, and Jon A. Wellner. Limit distributions theory for the maximum likelihood estimation of a log-concave density. *The Annals of Statistics*, 37(3):1299 – 1331, 2009.

Tim Bedford and Roger M. Cooke. Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial Intelligence*, 32(1-4):245–268, 2001.

Lucien Birgé. Estimation of unimodal densities without smoothness assumptions. *The Annal of Statistics*, 25(3):970–981, 1997.

Eike Christian Brechmann and Ulf Schepsmeier. Modeling dependence with c- and d-vine copulas: The r package cdvine. *Journal of Statistical Software*, 52(3):1–27, 2013.

Ga'elle Chagny. An introduction to nonparametric adaptive estimation. *Graduate J. Math.*, 1:105–120, 2016.

George T. Chang and Guenther Walther. Clustering with the mixture of log-concave distributions. *Computational Statistics and Data Analysis*, 51:6242–6251, 2007.

Arthur Charpentier, Jean-David Fermanian, and Olivier Scaillet. *The estimation of copulas : theory and practice*. Jorn Rank, London: Risk Books, 2007. ISBN 190433945X.

Yining Chen, Madeleine Cule, Robert Gramacy, and Richard Samworth. *Log-concave Density Estimation in Arbitrary Dimensions*, 2015. URL `https://CRAN.R-project.org/package=LogConcDEAD`. R package version 1.5-9.

Zhen Chen. *A Flexible Copula Model for Bivariate Survival Data*. PhD thesis, University of Rochester, 2012.

Madeleine Cule. *Maximum Likelihood Estimation of a Multivariate Log-Concave Density*. PhD thesis, Universities of Cambridge, 2009.

Madeleine Cule, Richard Samworth, and Michael Stewart. Maximum likelihood estimation of a multidimensional logconcave density (long version). *Journal of the Royal Statistical Society*, 72:545–607, 2010.

C. Czado, U. Schepsmeier, and A. Min. Maximum likelihood estimation of mixed c-vine pair copula with application to exchange rate. *Statistical modeling*, 12(3): 229 – 255, 2012.

P. Deheuvels. La fonction de dépendance empirique et ses propriétés. *Académie Royale de Belgique*, 65(5):274–292, 1979.

J. Dibmanna, E. C. Brechmanna, C. Czadoa, and D. Kurowickab. Selecting and estimating regular vine copulae and application to financial returns. *Elsvier*, 2012.

Charles R. Doss and Jon A. Wellner. Global rates of convergence of the mles of log-concave and s-concave densities. *The Annals of Statistics*, 44(3):954–981, 2016.

Lutz Dümbgen and Kaspar Rufibach. *logcondens: Computations Related to Univariate Log-Concave Density Estimation*, 2011. R package version 2.0.3.

Lutz Dümbgen and Kasper Rufibach. Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency. *Bernoulli*, 15(1):40 – 68, 2009.

Lutz Dümbgen, Andre Husler, and Kaspar Rufibach. Active set and em algorithms for log-concave densities based on complete and censored data. Technical report, University of Bern, 2011.

Tarn Duong. *Kernel Smoothing*, 2017. URL `http://www.mvstat.net/tduong`. R package version 1.4.0.

V. Durrleman, A. Nikeghbali, and T. Roncalli. Which copula is the right one?

Technical report, Groupe de Recherche Opérationnelle Créedit Lyonnais France, 2000.

R. Fletcher. *Practical methods of optimization*. A Wiley-Interscience Publication. John Wiley & Sons, Ltd., Chichester, second edition, 1987. ISBN 0-471-91547-5.

Alison L. Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002.

Piet Groeneboom, Geurt Jongbloed, and Jon A. Wellner. A canonical process for estimation of convex functions: The invelope of integrated brownian motion + t^4. *The Annals of Statistics*, 29:1620–1652, 2001a.

Piet Groeneboom, Geurt Jongbloed, and Jon A. Wellner. Estimation of a convex function: Characterizations and asymptotic theory. *The Annals of Statistics*, 29: 1653–1698, 2001b.

Arsalane Chouaib Guidoum. *Kernel Estimator and Bandwidth Selection for Density and Its Derivatives*, 2015. URL `https://cran.r-project.org/web/packages/kedd/kedd.pdf`. R package version 1.0.3.

Marius Hofert, Ivan Kojadinovic, Martin Maechler, and Jun Yan. *Multivariate Dependence with Copulas*, 2017. URL `http://copula.r-forge.r-project.org/`. R package version 3.1.0.

Hao Hu, Yichao Wi, and Weixin Yao. Maximum likelihood estimation of the mixture of log-concave densities. *Computational Statistics and Data Analysis*, 101:137–147, 2016.

I.A. Ibragimov. On the composition of unimodal distributions. *Theory of Probability and its Applications*, 1(2):255–260, 1956.

Harry Joe. Multivariate extreme-value distributions with applications to environmental data. *Canadian Journal of Statistics*, 22:47–64, 1994.

Harry Joe. *Multivariate models and dependence concepts*, volume 73 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1997. ISBN 0-412-07331-5.

Harry Joe. Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94:401–419, 2005.

Jean-Frédéric Jouanin, Gaël Riboulet, and Thierry Roncalli. Financial applications of copula functions. 2011.

Arlene K. H. Kim and Richard J. Samworth. Global rates of convergence in log-concave density estimation. *The Annals of Statistics*, 44(6):2756 – 2779, 2016.

Arlene K. H. Kim, Adityanand Guntuboyina, and Richard J. Samworth. Adaption in log-concave density estimation. *The Annals of Statistics*, 46(5), 2018.

Gunky Kim, Mervyn J. Silvapulle, and Paramsothy Silvapulle. Comparison of semiparametric and parametric methods for estimating copulas. *Computational Statistics and Data Analysis*, 51:2836–2850, 2007.

Xuan Li, Hanna Jankowski, Sawitree Boonpatcharanon, Victoria Tran, X Wang, and Jane Heffernan. Clustering neuraminidase influenza protein sequences. pages 210–220, 06 2016. doi: 10.1142/9789813141919_0014.

Thomas Nagler. *kdecopula: An R Package for the Kernel Estimation of Copula Densities*, 2017. R package version 0.8.0.

Bin Nan and Jon A. Wellner. A general semiparametric z-estimation approach for case-cohort studies. *Statistical Sinica*, 23:1155–1180, 2013.

Roger B. Nelsen. *An introduction to copulas*. Springer Series in Statistics. Springer, New York, second edition, 2006. ISBN 0-387-28659-4.

Andrew J. Patton. A review of copula models for economic time series. *Journal of Multivariate Analysis*, 110:4–18, 2012.

Bruno Rémillard, Nicolas Papageorgiou, and Soustra Frédéric. Copula-based semi-

parametric models for multivariate time series. *Journal of Multivariate Analysis*, 110:30–42, 2012.

Ulf Schepsmeier, Jakob Stoeber, Eike Christian Brechmann, Benedikt Graeler, Thomas Nagler, Tobias Erhardt, and Almeida. *Statistical Inference of Vine Copulas*, 2018. URL `https://github.com/tnagler/VineCopula`. R package version 2.1.8.

B.W. Silverman. On the estimation of a probability density function by the maximum penalized likelihood method. *The Annals of Statistics*, 10(3):795–810, 1982.

Sklar. Fonctions de repartition a n dimensionset leurs marges. *Publ. Inst., Statist. Univ. Paris*, 8:229–231, 1959.

Aad W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998. ISBN 0-521-49603-9.

Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. ISBN 0-387-94640-3. With applications to statistics.

Guenther Walther. Detecting the presence of mixing with multiscale maximum likelihood. *Journal of the American Statistical Association*, 97:508 – 513, 2002.

Guenther Walther. Inference and modeling with log-concave distributions. *Statistical Science*, 24(3):319–327, 2009.

Larry Wasserman. *All of nonparametric statistics*. Springer Texts in Statistics. Springer, New York, 2006. ISBN 0-387-25145-6.

Jun Yan. Enjoy the joy of copulas: With a package copula. *Journal of Statistical Software*, 21(4):1–21, 2007.

A Appendices

A.1 Definitions and Lemmas

Definition A.1. For any probability measure P, let

$$L_r(P) = \|f\|_{r,P} = \left(\int |f|^r dP \right)^{1/r}.$$

For a supremum norm L_∞,

$$\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|.$$

For a maximum norm ℓ_∞,

$$\|f\|_\infty = \max_i |x_i|.$$

Definition A.2. Let y and y_0 are the vector of length d. The pth-order Taylor series expansion around y_0 can be represented as

$$A(y) = \sum_{|\alpha| \le p} \frac{\partial^\alpha A(y_0)}{\alpha!} h^\alpha + R_p(y^*, h) \quad \text{where } h = y - y_0 \quad \text{and}$$

$$R_p(y^*, h) = \sum_{|\alpha| = p+1} \frac{\partial^\alpha A(y^*)}{\alpha!} h^\alpha \quad \text{for some} \quad y^* \in (y, y_0).$$

Definition A.3. (P-Glivenko-Cantelli)[van der Vaart, 1998, page 269] Let $X_1, \ldots, X_n$ is a random sample from a probability distribution P on a measurable space $(\mathcal{X}, \mathcal{A})$. A class $\mathcal{F}$ of measurable functions $f : \mathcal{X} \mapsto \mathbb{R}$ is called $P-$Glivenko-Cantelli if

$$\|\mathbb{P}_n f - Pf\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |\mathbb{P}_n f - Pf| \xrightarrow{a.s.} 0.$$

Lemma A.4. *(Glivenko-Cantelli)[van der Vaart, 1998, Theorem 19.4, page 270] Every class $\mathcal{F}$ of measurable functions such that $N_{[]}(\epsilon, \mathcal{F}, L_1(P)) < \infty$ for every $\epsilon > 0$ is P-Glivenko-Cantelli.*

Lemma A.5. *(Cauchy-Schwarz inequality for integral)*

$$\int f(x)g(x)dx \leq \left(\int f^2(x)dx\right)^{1/2} \left(\int g^2(x)dx\right)^{1/2}$$

Definition A.6. Let state space is $\mathbb{R}$. Kolmogorov metric can be represented as the distance between their distribution functions F and G, which can be represented as

$$d_K(F, G) = \sup_{x \in \mathbb{R}} |F(x) - G(x)|.$$

Definition A.7. (Donsker class)[van der Vaart and Wellner, 1996] A Donsker class is a set of function for which the empirical distribution with independent and identically distributed random variables verifies a uniform central limit theorem, with limiting distribution as a Gaussian process.

Definition A.8. The covering number $N(\varepsilon, \mathcal{F}, d)$ is the minimum number of balls of radius ε needed to cover the set $\mathcal{F}$.

Definition A.9. Given $l, u : \mathcal{X} \mapsto \mathbb{R}$ the bracket $[l, u]$ is the set of all functions f with $l \leq f \leq u$.

Definition A.10. An ε-bracket in $L_r(P)$ is a bracket $[l, u]$ with $P(u - l)^r < \varepsilon^r$.

Definition A.11. The bracketing number $N_{[]}(\varepsilon, \mathcal{F}, d)$ is the minimum number of ε-brackets needed to cover $\mathcal{F}$.

Definition A.12. The bracketing entropy is the logarithm of the bracketing number.

Lemma A.13. *Let g be a real function which is continuous on the interval $[a, b]$. Then,*

$$\left| \int_a^b g(x)dx \right| \leq \int_a^b |g(x)dx| .$$

Lemma A.14. *The Taylor series for an exponential function e^x at $a = 0$ is*

$$\sum_{i=0}^{\infty} \frac{x^i}{i!}.$$

A.2 Explicit formulas of J functions

From (2.4), the $J(\varphi_j, \varphi_k)$ function can be represented as

$$J(\varphi_j, \varphi_k) = \frac{\exp(\varphi_k) - \exp(\varphi_j)}{\varphi_k - \varphi_j}.$$

Therefore, the partial derivative of J functions are in the following form.

$$\begin{aligned} J_{10}(\varphi_j, \varphi_k) &= \partial_{\varphi_j} J(\varphi_j, \varphi_k) \\ &= \frac{\exp(\varphi_k) - \exp(\varphi_j) - \exp(\varphi_j)(\varphi_k - \varphi_j)}{(\varphi_k - \varphi_j)^2} \end{aligned}$$

$$\begin{aligned} J_{01}(\varphi_j, \varphi_k) &= \partial_{\varphi_k} J(\varphi_j, \varphi_k) \\ &= \frac{\exp(\varphi_k)(\varphi_k - \varphi_j) - \exp(\varphi_k) - \exp(\varphi_j)}{(\varphi_k - \varphi_j)^2} \end{aligned}$$

$$\begin{aligned} J_{20}(\varphi_j, \varphi_k) &= \partial^2_{\varphi_j} J(\varphi_j, \varphi_k) \\ &= \frac{2\exp(\varphi_k) - \exp(\varphi_j)[(\varphi_k - \varphi_j + 1)^2 + 1]}{(\varphi_k - \varphi_j)^3} \end{aligned}$$

$$\begin{aligned} J_{02}(\varphi_j, \varphi_k) &= \partial^2_{\varphi_k} J(\varphi_j, \varphi_k) \\ &= \frac{2\exp(\varphi_k) - \exp(\varphi_j)[(\varphi_k - \varphi_j + 1)^2 + 1]}{(\varphi_k - \varphi_j)^3} \end{aligned}$$

Clearly see that $J_{20}(\varphi_j, \varphi_k) = J_{02}(\varphi_j, \varphi_k)$.

$$\begin{aligned} J_{11}(\varphi_j, \varphi_k) &= \partial_{\varphi_j}\partial_{\varphi_k} J(\varphi_j, \varphi_k) \\ &= \frac{\exp(\varphi_k)(\varphi_k - \varphi_j - 2) + \exp(\varphi_j)(\varphi_k - \varphi_j + 2)}{(\varphi_k - \varphi_j)^3} \end{aligned}$$

A.3 Sample codes

A.3.1 Sample code for a univariate log-concave MLE (Figure 2.1)

```
> library(logcondens)
> set.seed(1)
> x <- rnorm(1000)
> res <- activeSetLogCon(x)
> plot(res$x, exp(res$phi), type='l', lwd=2, xlab="x", ylab="y", lty=1)
> lines(res$x, dnorm(res$x), lwd=2, lty=2)
> legend("topright", c("estimated␣density", "true␣density"),
> lty=c(1,2), lwd=c(2,2))
> plot(res$x, res$phi, type='l', lwd=2, xlab="x", ylab="phi")
> kn <- res$knots
> abline(v=kn, col="red", lty=2)
```

A.3.2 Sample code for a univariate log-concave MLE showing locations and values of knots (Figure 2.2)

```
> set.seed(1)
> n <- 500
> x <- rnorm(n)
> res <- activeSetLogCon(x)
> res$knots
 [1] -3.0080486 -2.2852355 -1.5235668 -1.4707524 -0.3836321 -0.3672215
 [7]  0.5939013  1.0691615  1.5868335  2.3079784  3.8102767
```

```
> head(res$phi)
[1] -4.828354 -4.587058 -3.986300 -3.603007 -3.480720 -3.372153
> f <- exp(res$phi)
> head(f)
[1] 0.0079997  0.010183  0.018568  0.027242  0.030785  0.034316
> plot(x=res$x, y=res$phi, type="l")
```

A.3.3 Sample code for a multivariate log-concave MLE

```
> library(LogConcDEAD)
> library(mvtnorm)
> set.seed(1)
> d <- 2
> n <- 500
> x <- rmvnorm(n, mean=c(0,0), sigma=diag(1,d,d))
> res <- mlelcd(x)
> head(res$logMLE)
[1] -3.109657 -2.399040 -3.635341 -4.660609 -3.019291 -2.094872
> f <- exp(res$logMLE)
> head(f)
[1] 0.044616  0.090805  0.026375  0.009461  0.048836  0.123086
```

A.3.4 Sample code for breast cancer example

```
> library(logcondens)
> library(copula)
> library(LogConcDEAD)
> library(VineCopula)

> set.seed(12345)
> data <- read.csv("BreastCancer.csv", header=F)
> dat <- data[,3:32]
> pca <- prcomp(x=dat, center = TRUE, scale. = TRUE)
> plot(pca, type = "l", ylim = c(0,15))
> pcadat <- pca$x
> x <- pcadat[,1:2]
> k <- length(table(data[,2]))
> n <- nrow(x)
> d <- ncol(x)
> truecomp <- replace(c <- c(data[,2]),c=="B",1)

> prec     <-   1e-5
> prec1     <-  1e-8
> # starting values
> mc_start  <-hc(modelName="VVV", x)
```

```
> class <- c(hclass(mc_start, k))
> props<-table(class)
> if(min(props)<2){
+   mc_start<-hc(modelName="EEE", x)
+   class <- c(hclass(mc_start, k))
+ }
> cprops <- as.vector(table(class)/length(class))
> q<-matrix(0, nrow=n, ncol=k)
> qcop<-matrix(0, nrow=n, ncol=k)
> mean <- matrix(rep(0,k*d),k,d)
> var <- matrix(rep(0,k*d),k,d)
> corr <- c(rep(0,2))
> for (c in 1:k){
+   mean[c,] <- apply(x[class==c,],2,mean)
+   var[c,] <- apply(x[class==c,],2,var)
+ }
> for(i in 1:k){
+   q[,i]<-dmvnorm(x, mean=mean[i,], sigma=diag(var[i,],d,d))
+   qcop[,i]<-dmvnorm(x, mean=mean[i,], sigma=diag(var[i,],d,d))}
> fit<-as.vector(cprops %*% t(q))
> fitcop<-as.vector(cprops %*% t(qcop))
> likold  <-  -100000000
> likoldcop  <-  -100000000
```

```
> ###### Copula+Log-concave marginals ######
> propscop <- cprops
> classdat <- class
> # 1. copula selection for each component
> par <- c(0,0)
> for(i in 1:k){
+   sel <- BiCopSelect(pobs(x[classdat==i,1]), pobs(x[classdat==i,2]),
+             selectioncrit = "BIC",familyset = 1:10, rotations = F)
+   par[i] <- sel$par
+   if(i==1){
+     copseld1 <- frankCopula(param = par[i], dim = d)
+   }else{
+     copseld2 <- frankCopula(param = par[i], dim = d)
+   }
+ }
> r <- 500
> for(jj in 1:r){
+   ### 2. find copula estimator
+   estpar <- c(0,0)
+   for(i in 1:k){
+     count <- nrow(x[classdat==i,])
+     FFhat <- matrix(rep(0,count*d),count,d)
```

```
+     for(j in 1:d){
+       dx <- x[classdat==i,j]
+       densx <- logConDens(dx, smoothed = F)
+       Fres <- evaluateLogConDens(dx,densx)[,4]
+       FFhat[,j] <- round(Fres,10)
+       FFhat[,j][which(FFhat[,j]==0)] <- 1/(count+1)
+       FFhat[,j][which(FFhat[,j]==1)] <- count/(count+1)
+     }
+     if(i==1){
+       fitcop <- fitCopula(copula=copseld1, data = FFhat,start = par[i],
+                           method = "ml",optim.control = list(maxit=1000))
+       estpar[i] <- fitcop@estimate
+     }else{
+       fitcop <- fitCopula(copula=copseld2, data = FFhat,start = par[i],
+                           method = "ml",optim.control = list(maxit=1000))
+       estpar[i] <- fitcop@estimate
+     }
+   }
+   par <- estpar

+       ### 3. fit marginal density
+   wcop<-t(t(qcop)*propscop)
+   fitcop<-as.vector(propscop %*% t(qcop))
```

```
+   wcop<-wcop/fitcop
+   qcop<-matrix(0, nrow=n, ncol=k)
+   for (i in 1:k){
+     whichxc<-wcop[,i]/sum(wcop[,i]) > prec1/n
+     wcopuse <- wcop[whichxc,i]
+     count <- nrow(x[whichxc,])
+     xx <- x[whichxc,]
+     mle1 <- activeSetLogCon(x=xx[,1], w=wcopuse)
+     mle2 <- activeSetLogCon(x=xx[,2], w=wcopuse)
+     f1hat <- rep(0,count)
+     f2hat <- rep(0,count)
+     for (ii in 1:count){
+       f1hat[ii] <- exp(mle1$phi)[which(xx[ii,1]==mle1$xn)]
+       f2hat[ii] <- exp(mle2$phi)[which(xx[ii,2]==mle2$xn)]
+     }
+     fhat <- cbind(f1hat,f2hat)
+     FF1hat <- rep(0,count)
+     FF2hat <- rep(0,count)
+     for (ii in 1:count){
+       FF1hat[ii] <- mle1$Fhat[which(xx[ii,1]==mle1$xn)]
+       FF2hat[ii] <- mle2$Fhat[which(xx[ii,2]==mle2$xn)]
+     }
+     F1hat <- round(FF1hat,10)
```

```
+     F1hat[which(F1hat==0)] <- 1/(count+1)
+     F1hat[which(F1hat==1)] <- count/(count+1)
+     F2hat <- round(FF2hat,10)
+     F2hat[which(F2hat==0)] <- 1/(count+1)
+     F2hat[which(F2hat==1)] <- count/(count+1)
+     FFhat <- cbind(F1hat, F2hat)

+     ### 4. fit copula density
+     if(i==1){
+       copseld1 <- frankCopula(param = par[i], dim = d)
+       copfit <- dCopula(u=FFhat,copula = copseld1)
+     }else{
+       copseld2 <- frankCopula(param = par[i], dim = d)
+       copfit <- dCopula(u=FFhat,copula = copseld2)
+     }
+       ### 5. fit joint density
+     cqcop <- 1
+     for(l in 1:d){
+       cqcop <- cqcop*fhat[,l]
+     }
+     qcop[whichxc,i] <- copfit*cqcop
+   }
+   #update class of data
```

```
+   classdat <- rep(0,n)
+   for (lc in 1:n){
+     classdat[lc] <- which.max(wcop[lc,])
+   }
+   propscop  <-apply(wcop, 2, sum)/sum(wcop)
+   fitcop<-as.vector(propscop %*% t(qcop))
+   tempcop<-t(log(t(t(qcop)*propscop)))*propscop
+   liknewcop<-sum(tempcop[tempcop >-Inf])
+   changecop  <-abs((liknewcop-likoldcop)/likoldcop)
+   if (changecop < prec) break
+   likoldcop  <-liknewcop
+ }
> llcop    <-sum(log(fitcop))
> copcomp <- rep(0,n)
> for (lc in 1:n){
+   copcomp[lc] <- which.max(wcop[lc,])
+ }
> copcomp[which(copcomp==1)] <- 0
> copcomp[which(copcomp==2)] <- 1
> copcomp[which(copcomp==0)] <- 2
> ind <- which(copcomp != truecomp)
> ncopmis <- length(ind)
> ncopmis
```

```
[1] 49


> copmiscase <- x[ind,]
> res <- cbind(ind,truecomp[ind],copmiscase)
> write.csv(res, "CopCaseMisclassification.csv", row.names = F)


> ###### Gaussian Mixture Model ######
> GMM  <-  Mclust(x, modelNames="VVV", G=k)
> GMMcomp <- GMM$classification
> llGMM <- GMM$loglik
> mGMM <- GMM$parameters$mean[1,1]
> GMMcomp[which(GMMcomp==1)] <- 0
> GMMcomp[which(GMMcomp==2)] <- 1
> GMMcomp[which(GMMcomp==0)] <- 2
> ind <- which(GMMcomp != truecomp)
> nGMMmis <- length(ind)
> nGMMmis
[1] 59


> GMMmiscase <- x[ind,]
> res <- cbind(ind,truecomp[ind],GMMmiscase)
> write.csv(res, "GMMCaseMisclassification.csv", row.names = F)
```

```
> ###### Multivatiate log-concave by LogConcDEAD ######
> props <- cprops
> for (j in 1:r) {
+   w<-t(t(q)*props)
+   fit<-as.vector(props %*% t(q))
+   w<-w/fit
+   q<-matrix(0, nrow=n, ncol=k)
+   for(i in 1:k){
+     whichx<-w[,i]/sum(w[,i]) > prec1/n
+     wuse <- w[whichx,i]/sum(w[whichx,i])
+        if(i==1){
+       mle1 <- mlelcd(x[whichx,], w = wuse)
+       q[whichx,i]  <-  exp(mle1$logMLE)
+     }else{
+       mle2 <- mlelcd(x[whichx,], w = wuse)
+       q[whichx,i]  <-  exp(mle2$logMLE)
+     }
+   }
+   props  <-apply(w, 2, sum)/sum(w)
+   fit<-as.vector(props %*% t(q))
+   temp<-t(log(t(t(q)*props)))*props
+   liknew<-sum(temp[temp>-Inf])
+   change  <-abs((liknew-likold)/likold)
```

```
+   if (change < prec) break
+   likold  <-liknew
+ }

> ll      <-sum(log(fit))
> LCcomp <- rep(0,n)
> for (lc in 1:n){
+   LCcomp[lc] <- which.max(w[lc,])
+ }

> LCcomp[which(LCcomp==1)] <- 0
> LCcomp[which(LCcomp==2)] <- 1
> LCcomp[which(LCcomp==0)] <- 2
> ind <- which(LCcomp != truecomp)
> nLCmis <- length(ind)
> nLCmis
[1] 46

> LCmiscase <- x[ind,]
> res <- cbind(ind,truecomp[ind],LCmiscase)
> write.csv(res, "LCCaseMisclassification.csv", row.names = F)

> ########## Contour and surface plots for each method
```

```
> gr <- grey.colors(10)
> level <- seq(0.001,0.1, length.out=21)
> GG <- 100
> s <- c(min(x[,1]), max(x[,1]),min(x[,2]), max(x[,2]))
> g1 <- seq(s[1],s[2],length.out = GG)
> g2 <- seq(s[3],s[4],length.out = GG)
> grid <- expand.grid(X=g1, Y=g2)
> grid <- as.matrix(grid)

> ###### Copula+Log-concave marginals ######
> xc1 <- x[which(copcomp==1),]
> fitd1 <- logConDens(x=xc1[,1], smoothed = F)
> fitd1res <- evaluateLogConDens(grid[,1],fitd1)
> fd1 <- fitd1res[,3]
> Fd1 <- fitd1res[,4]
> fitd2 <- logConDens(x=xc1[,2], smoothed = F)
> fitd2res <- evaluateLogConDens(grid[,2],fitd2)
> fd2 <- fitd2res[,3]
> Fd2 <- fitd2res[,4]
> copc1 <- frankCopula(param = par[1], dim = d)
> copdens <- dCopula(u=cbind(Fd1,Fd2),copula = copc1)
> zc1 <- matrix(copdens*fd1*fd2,GG,GG)
```

```
> xc2 <- x[which(copcomp==2),]
> fitd1 <- logConDens(x=xc2[,1], smoothed = F)
> fitd1res <- evaluateLogConDens(grid[,1],fitd1)
> fd1 <- fitd1res[,3]
> Fd1 <- fitd1res[,4]
> fitd2 <- logConDens(x=xc2[,2], smoothed = F)
> fitd2res <- evaluateLogConDens(grid[,2],fitd2)
> fd2 <- fitd2res[,3]
> Fd2 <- fitd2res[,4]
> copc2 <- frankCopula(param = par[2], dim = d)
> copdens <- dCopula(u=cbind(Fd1,Fd2),copula = copc2)
> zc2 <- matrix(copdens*fd1*fd2,GG,GG)
> zhat <- (propscop[1]*zc1)+(propscop[2]*zc2)

> miscase <- read.csv("CopCaseMisclassification.csv", header=T)
> xpl <- miscase[,3:4]
> Bdat <- xpl[which(miscase[,2]==1),]
> Mdat <- xpl[which(miscase[,2]==2),]

> contour(g1,g2,zc1,xlim=c(-15,5), ylim=c(-10,7),levels=level)
> par(new=T)
> contour(g1,g2,zc2,xlim=c(-15,5), ylim=c(-10,7),levels=level)
> par(new=TRUE)
```

```
> plot(Bdat, col=gr[7],type="p", pch=20, cex=1.5,xlim=c(-15,5), ylim=c(-10,7))
> par(new=TRUE)
> plot(Mdat, col=gr[1],type="p", pch=20, cex=1.5,xlim=c(-15,5), ylim=c(-10,7))
> persp(g1,g2,zhat,box=F,phi = 0, theta = 0, col=gr[10])

> ###### Gaussian Mixture Model ######
> GMMprop <- GMM$parameters$pro
> mean <- GMM$parameters$mean
> sigmaC1 <- GMM$parameters$variance$sigmasq[1]
> sigmaC2 <- GMM$parameters$variance$sigmasq[2]
> fhat1 <- dmvnorm(expand.grid(g1,g2),mean = c(mean[,1]),
+                       sigma = cov(x[which(GMM$classification==1),]))
> z1 <- matrix(fhat1,GG,GG)
> fhat2 <- dmvnorm(expand.grid(g1,g2),mean = c(mean[,2]),
+                       sigma = cov(x[which(GMM$classification==2),]))
> z2 <- matrix(fhat2,GG,GG)
> zmix <- matrix(GMMprop[1]*fhat1+GMMprop[2]*fhat2,GG,GG)

> miscase <- read.csv("GMMCaseMisclassification.csv", header=T)
> xpl <- miscase[,3:4]
> Bdat <- xpl[which(miscase[,2]==1),]
> Mdat <- xpl[which(miscase[,2]==2),]
```

```
> contour(g1,g2,z1,xlim=c(-15,5), ylim=c(-10,7),levels=level)
> par(new=TRUE)
> contour(g1,g2,z2,xlim=c(-15,5), ylim=c(-10,7),levels=level)
> par(new=TRUE)
> plot(Bdat, col=gr[7],type="p", pch=20, cex=1.5,xlim=c(-15,5), ylim=c(-10,7))
> par(new=TRUE)
> plot(Mdat, col=gr[1],type="p", pch=20, cex=1.5,xlim=c(-15,5), ylim=c(-10,7)
+      ,xlab="", ylab="")
> persp(g1,g2,zmix, box=F,phi = 0, theta = 0, col=gr[10])

> ###### Multivatiate log-concave by LogConcDEAD ######
> flc1 <- dlcd(grid,mle1)
> flc2 <- dlcd(grid,mle2)
> zlc <- matrix(props[1]*flc1+props[2]*flc2,GG,GG)

> miscase <- read.csv("LCCaseMisclassification.csv", header=T)
> xpl <- miscase[,3:4]
> Bdat <- xpl[which(miscase[,2]==1),]
> Mdat <- xpl[which(miscase[,2]==2),]
> cg1 <- interplcd(mle1, gridlen = 100)
> plot(mle1, g=cg1, type="c",
+      xlim=c(-15,5),ylim=c(-10,7),addp=F, main="", col="black",
+      levels=level)
```

```
> par(new=TRUE)
> cg2 <- interplcd(mle2, gridlen = 100)
> plot(mle2, g=cg2, type="c",
+      xlim=c(-15,5), ylim=c(-10,7),addp=F, main="", col="black",
+      levels=level)
> par(new=TRUE)
> plot(Bdat, col=gr[7],type="p", pch=20, cex=1.5,xlim=c(-15,5), ylim=c(-10,7))
> par(new=TRUE)
> plot(Mdat, col=gr[1],type="p", pch=20, cex=1.5,xlim=c(-15,5), ylim=c(-10,7)
+      ,xlab="", ylab="")
> persp(g1,g2,zlc,box=F,phi = 0, theta = 0, col=gr[10])
```