Copula Density Estimation by Finite Mixture of Parametric Copula Densities

Leming Qu *

Abstract

As a stochastic dependence modeling tool beyond the classical normal distribution model, Copula is widely used in Economics, Finance, and Engineering. A Copula density estimation method that is based on finite mixture of parametric Copula densities is proposed here. More specifically, one component of the mixture model is the mixture of Gaussian, Clayton and Gumbel Copulas (termed GCG component) which are capable of capturing symmetrical, lower, and upper tail dependence, respectively. The entire copula density is a mixture of k GCG components. The model parameters are estimated by interior-point algorithm for the resulting constrained maximum likelihood estimation problem, where the gradient of the objective function is not required. The interior-point algorithm is compared with the commonly used expectation-maximization (EM) algorithm in mixture models. Mixture components with small weights can be removed by a thresholding rule. The number of components k is selected by the model selection criterion AIC. Simulation and real data application show the effectiveness of the proposed approach.

Key Words: Copula, dependence modeling, mixture model, maximum likelihood estimation, interior-point algorithm

1. Introduction

Dependence modeling consists of finding a model that describes dependencies between variables, which is a fundamental task of multivariate statistics (Cox and Wermuth (1996)). Statistical approaches to dependence modeling describes an underlying random process in terms of a multivariate distribution. Multivariate probability density estimation based on observed data from a random process is a long standing and active research area in statistics (Scott (1992)). In a linear, Gaussian world stochastic dependencies are captured by correlations. In more general settings, copula (otherwise known as dependence function) has emerged as a useful tool for modeling stochastic dependence. In essence, a copula is a multivariate probability distribution with uniform marginals. One of the main advantages of copula over full probability function is that copula allows the separation of dependence modeling from the marginal distributions.

The copula density estimation can be categorized into parametric, semiparametric, and nonparametric methods. A parametric estimation method assumes copula density belongs to a copula family determined by a few parameters (for example, Shih and Louis (1995)). The parametric copula density estimation problem is then essentially reduced to estimate the few parameters that determine the copula.

Nonparametric estimation of copula density does not assume a specific parametric form for the copula and the marginals and thus provides great flexibility and generality.

Semiparametric copula density estimation method assumes part of the data distribution - such as copula density - follows a parametric model, while the rest - such as the univariate marginal distributions - follow nonparametric models. This includes the following two stage estimation method (Genest et al. (1995)): in the first stage, the univariate marginal distributions are estimated nonparametrically, e.g., by the $n/(n+1)$ times the empirical

*Department of Mathematics, Boise State University, Boise, Idaho 83725-1555, USA. Email: lqu@boisestate.edu

marginal distributions. In the second stage, the copula parameters are estimated by maximizing the pseudo log-likelihood using the data generated in the first stage. The resulting semi-parametric estimator of the dependence parameter is consistent and asymptotically normal under suitable regularity conditions. Genest et al. (1995) proposed a consistent estimator of the limiting variance-covariance matrix of the estimator.

Choros et al. (2010) provided a brief survey of parametric, semiparametric and non-parametric estimation procedures for copula models.

We propose here to estimate a bivariate copula density by finite mixture of parametric copulas. Mixture models provide an interesting alternative to non-parametric modeling, while being less restrictive as opposed to standard parametric distributional assumption (Diebolt and Robert (1994)). We estimate the marginals by their empirical distributions in the first stage. Hence our approach is semiparametric in nature.

Hu (2006) uses a mixture of three copulas to capture various symmetric and asymmetric dependence structures in financial markets. The mixture is composed of a Gaussian copula, a Gumbel copula and a Gumbel survival copula. The Gaussian copula in the mixture relates to traditional approaches based on the Gaussian assumption. Gumbel copula and its survival copula model extreme co-movements in market returns. The former models positive right tail dependence while the latter is its mirror image and models left tail dependence. In Hu (2006), the mixture model is estimated using a two-stage semi-parametric procedure, i.e. the marginals are estimated by the empirical distributions. EM algorithm is used to maximize the pseudo log-likelihood. Kauermann and Meyer (2014) proposes finite mixtures of different Archimedean copula families as a flexible tool for modeling the dependence structure in multivariate data. The parameters in this mixture model are estimated by maximizing the penalized marginal likelihood via iterative quadratic programming. Arakelian and Karlis (2014) uses finite mixture of different copulas for clustering purposes, with parametric marginal distributions. The model parameters are estimated by an EM algorithm based on the standard approach for mixture models.

The rest of the paper is organized as follows: In section 2, we present the finite mixture of parametric copulas model. In section 3, we discuss the interior point algorithm and compare it with classical expectation-maximization (EM) algorithm. Section 4 shows the experimental results. We apply the method to a real data set in section 5. Finally, section 6 concludes the paper.

2. Finite Mixture of Parametric Copulas Model

A bivariate copula density $c(u, v),\ [u, v] \in [0, 1]^2$ can be regarded as the joint probability density function (PDF) of a bivariate standard uniform random variable (U, V).

A bivariate copula $C(u, v)$ defined on the unit square $[0, 1]^2$ is a bivariate cumulative distribution function (CDF) with univariate standard uniform margins:

$$C(u, v) = \int_0^u \int_0^v c(s, t)\mathrm{d}s\mathrm{d}t.$$

Sklar's Theorem (Sklar (1959)) states that the joint CDF $F(x, y)$ of a bivariate random variable (X, Y) with marginal CDF $F_X(x)$ and $F_Y(y)$ can be written as $F(x, y) = C(F_X(x), F_Y(y))$, where copula C is the joint CDF of $(U, V) = (F_X(X), F_Y(Y))$. This indicates a copula connects the marginal distributions to the joint distribution and justifies the use of copulas for building bivariate distributions.

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from the unknown distribution F of (X, Y). We wish to estimate aspects of the joint distribution of X and Y, in particular, the copula density function $c(u, v)$.

When the two marginal distributions are continuous, the copula density $c(u, v)$ is the unique bivariate density of $(U, V) = (F_X(X), F_Y(Y))$ as implied by Sklar's theorem. As copulas are not directly observable, a copula density estimator is usually formed in two stages: obtaining the observations for (U, V) first and then estimating the copula density based on these observations.

In the first stage, the original data set (X_i, Y_i) for $i = 1, \ldots, n$ is converted to $(\hat{U}_i, \hat{V}_i) = (\hat{F}_X(X_i), \hat{F}_Y(Y_i))$, where $\hat{F}_X$ and $\hat{F}_Y$ are conventional estimators of F_X and F_Y. If models are available for the marginal distributions of X and Y but not for the joint distribution, one can use a technique such as maximum likelihood to estimate the marginal distribution functions. Otherwise, some nonparametric univariate distribution estimation methods or simply the following empirical CDFs (ECDFs) can be used:

$$\hat{F}_X(x) = \frac{1}{n+1} \sum_{i=1}^{n} I(X_i \leq x), \;\; \hat{F}_Y(y) = \frac{1}{n+1} \sum_{i=1}^{n} I(Y_i \leq y), \tag{1}$$

where $I(\cdot)$ is the indicator function. Scaling by $(n + 1)$ instead of conventional n avoids difficulties arising from the potential unboundedness of copula density as some of the $\hat{U}_i$ or $\hat{V}_i$ tend to one (Genest et al. (1995)). When ECDFs are used as the marginal CDF estimators, $\{(\hat{U}_i, \hat{V}_i)\}_{i=1}^n$ is nothing but the rescaled ranks. In the second stage, we estimate the copula density $c(u, v)$ based on the observations $\{(\hat{U}_i, \hat{V}_i)\}_{i=1}^n$.

Gaussian copula can capture symmetrical dependence structure, while Clayton and Gumbel copula can capture lower and upper tail dependence, respectively. A mixture of Gaussian, Clayton and Gumbel Copulas is capable of capturing symmetrical, lower, and/or upper tail dependence. A single component of Gaussian, Clayton and Gumbel Copula mixture - GCG component - may not be flexible enough to capture a complicated dependence structure, we therefore model a bivariate density $c(u, v)$ as a mixture of a number, say k, of GCG components in some unknown proportions:

$$c(u, v; \theta) = \sum_{j=1}^{k} \left[p_{nj} c_{nj}(u, v; \rho_j) + p_{yj} c_{yj}(u, v; \alpha_j) + p_{mj} c_{mj}(u, v; \beta_j) \right], \tag{2}$$

where p_{nj}, p_{yj}, p_{mj} denotes the proportions; c_{nj}, c_{yj}, c_{mj} denotes the densities; and $\rho_j, \alpha_j, \beta_j$ denotes the parameters of Gaussian, Clayton, and Gumbel copula respectively. Note that: $-1 \leq \rho_j \leq 1$, $\alpha_j \geq 0$, and $\beta_j \geq 1$. Mixture proportions are nonnegative and sum to one. Copula parameters are restricted within the parameter spaces. It is beneficial to put the copula parameters of the same copula type in descending order in the model specification (see section 3.3 for an example).

The maximum pseudo log-likelihood estimator $\hat{\theta}$ in constrained parameter spaces max-

imizes the pseudo log-likelihood

$$\max_{\theta} L(\theta) = \sum_{i=1}^{n} \log \left\{ c(\hat{U}_i, \hat{V}_i; \theta) \right\}$$
$$= \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{k} \left[p_{nj} c_{nj}(\hat{U}_i, \hat{V}_i; \rho_j) + p_{yj} c_{yj}(\hat{U}_i, \hat{V}_i; \alpha_j) + p_{mj} c_{mj}(\hat{U}_i, \hat{V}_i; \beta_j) \right] \right\},$$

Subject to

$$\begin{aligned}
& p_{nj} \geq 0,\ p_{yj} \geq 0,\ p_{mj} \geq 0, \ \text{ for } \ j = 1, \ldots, k; \\
& \sum_{j=1}^{k} (p_{nj} + p_{yj} + p_{mj}) = 1 \\
& -1 \leq \rho_j \leq 1, \ \ \alpha_j \geq 0, \ \ \beta_j \geq 1, \ \text{ for } \ j = 1, \ldots, k. \\
& \rho_1 > \rho_2 > \ldots > \rho_k; \\
& \alpha_1 > \alpha_2 > \ldots > \alpha_k; \\
& \beta_1 > \beta_2 \ldots > \beta_k.
\end{aligned} \tag{3}$$

In the next section, we discuss algorithms for solving this optimization problem.

3. Constrained Maximum Likelihood Estimation by Interior Point Algorithm

We first briefly review the interior point algorithm for solving problems like (3) in this section, then discuss some specifics when applying this algorithm to our problem (3). There is a rich body of literature on this topic in mathematical programming (Wright (1992); Byrd et al. (1999, 2000); Waltz et al. (2006); Wright (1997)). Problem (3) is a special case of the following constrained nonlinear optimization (or programming) problem:

$$\begin{aligned}
& \min_{x} f(x), \\
& \text{Subject to} \\
& \qquad h(x) = 0 \\
& \qquad g(x) \leq 0,
\end{aligned} \tag{4}$$

where $f : R^n \Rightarrow R$, $h : R^n \Rightarrow R^l$ and $g : R^n \Rightarrow R^m$ are twice continuously differentiable functions (Waltz et al. (2006)).

The interior point approach to this constrained minimization is to replace the inequality constraints by log barrier (Lagrangian) penalty functions that introduce a smooth contribution to the objective function. This leads to the replacement of the nonlinear program (4) by a sequence of approximate barrier subproblems (MATLAB (2014)).

3.1 Initialization Strategy

In practice, the number of mixture components k is unknown. The single GCG component model with $k = 1$ is the simplest model, while a dozen GCG component model where $k = 12$ looks very complex. To choose an appropriate model order k, we start with $k = 1$, fit the model with appropriate initial values, then increase K by 1, fit the model with appropriate initial values again, until a model selection criterion such as AIC no longer improves.

Choosing appropriate initial or starting values for an algorithm that solves a multivariate optimization problem is crucial to the success of the algorithm, especially for non-convex problems. For our problem (3), when $k = 1$, a natural choice to initialize the proportions is the equal values:

$$p_{n1\_\text{init}} = p_{y1\_\text{init}} = p_{m1\_\text{init}} = \frac{1}{3}.$$

The initial value of a copula parameter is chosen as the the maximum likelihood estimate for the single component model without mixtures. For example, the initial value for Gaussian copula parameter ρ_{1_init} is the maximum likelihood estimate for the Gaussian copula model, which is simply the sample correlation coefficient.

When k is increased by 1, the initial value for the $(k+1)$ GCG components model will use the fitted value for the k GCG components model. For the proportion parameters, the idea is to keep the sum of initial proportions for $(k+1)$ GCG components model the same as the sum of fitted proportions for k components model in the Gaussian, Clayton, and Gumbel copula components respectively.

For the copula parameters, we set the initial copula parameter values for the first k components of the $(k+1)$ GCG components model equal to the fitted copula parameter values for the k GCG components model, while the initial copula parameter value for the $(k+1)$th component to be a value near the lower bound of the parameter space.

3.2 Thresholding and Model Selection Criterion

A mixture component with small weight such as 0.01 implies small contribution to the dependence structure, therefore should not be included in the copula model (Cai and Wang (2014)). We set the threshold for weight at 0.05 due to its good performance in simulation studies. Any component with its fitted proportion less than or equal to this threshold will be discarded from the model, leading to reduction of model complexity.

A model selection criterion offers a trade-off between the goodness of fit of the model and the complexity of the model. We use the AIC as the model selection criterion:

$$\text{AIC}(\hat{\theta}_k, k) = L(\hat{\theta}_k) - \text{DF}(k),$$

where $L(\hat{\theta}_k)$ is the log likelihood evaluated at the fitted θ for the k GCG components model, and DF(k) is its Degrees of Freedom. The $L(\hat{\theta}_k)$ measures the goodness of fit of the model, while DF(k) measures the complexity of the model.

$$\text{DF}(k) = 2\sum_{j=1}^{k}[I(\hat{p}_{nj} > 0.05) + I(\hat{p}_{yj} > 0.05) + I(\hat{p}_{mj} > 0.05)] - 1,$$

where $I(\cdot)$ is an indicator function. Here DF(k) - Degrees of Freedom - equals the number of effective parameters in the model. Each kept component whose estimated weight is above the threshold 0.05 has a weight parameter and a copula parameter, hence adding 2 to DF(k). The term -1 in DF(k) is due to the constraint that all the weights sum to 1.

Another well known model selection criterion BIC performs similarly in our simulation study.

So our model selection strategy starts with $k = 1$ with equal weights, fits the model parameter θ, and evaluates AIC$(\hat{\theta}_k, k)$. We then increase k by 1, fit the model with appropriate initial values using the initialization strategy, until AIC no longer increases. If AIC$(\hat{\theta}_k, k) \geq$ AIC$(\hat{\theta}_{(k+1)}, k+1)$, then k is selected as the number of GCG components.

Table 1: A mixture of two Gaussian copulas for simulation

Component	Proportion	ρ
Gaussian 1	0.40	0.3
Gaussian 2	0.60	-0.7

Table 2: Fitted Model with the ordered copula parameter specification for the mixture of two Gaussian copulas: AIC = 83.74

Component	Proportion	ρ
Gaussian 1	0.3704	0.3980
Gaussian 2	0.6296	-0.6869

3.3 Ordered vs Unordered Copula Parameters

One obvious benefit of putting the copula parameters of the same copula type in descending order in the model specification (2) is that we have a naturally good choice for the initial value of the copula parameter of the newly entered copula component - a value near the lower bound of the parameter space. Another benefit is that it tends to improve model fitting. We illustrate this by an example.

We generated $n = 1000$ observations from a mixture of two Gaussian copula model with the parameters specified in table 1. We then fit models with the ordered and unordered copula parameter specifications respectively, starting with $k = 1$ and using the same initial values chosen by the initialization strategy.

The fitted model with the ordered copula parameter specification is given in table 2, where we omitted the components whose weights are below or equal to the threshold 0.05. The fitted model with the unordered copula parameter specification is given in table 3, where we included the components whose weights are below or equal to the threshold 0.05.

In this example, the fitted model with the ordered copula parameter specification has a slightly bigger AIC (83.74 vs 82.39) and identified the correct model.

Table 3: Fitted Model with the unordered copula parameter specification for the mixture of two Gaussian copulas: AIC = 82.39

Component		Proportion	Copula Parameter
Gaussian	1	0.4537	-0.7107
	2	0.0000	-0.4120
	3	0.1138	-0.7107
Clayton	1	0.1048	1.5585
	2	0.0000	0.1393
	3	0.0000	3.7960
Gumbel	1	0.1296	1.1000
	2	0.1981	1.1000
	3	0.0000	1.3182

Table 4: A single GCG component model for simulation

Component	Proportion	parameter
Gaussian	0.35	0.5
Clayton	0.40	3
Gumbel	0.25	10

Table 5: Comparison of EM and interior point solutions: Iteration counts, log likelihoods, and CPU times (in seconds) for three EM variants and the interior point solver

Algorithm	EM1	EM2	EM3	IP
Iterations	50	100	500	16
Time	1.07	2.16	11.12	0.30
$L(\theta)$-550	0.8193	0.8987	0.8990	11.3148

3.4 Interior Point Algorithm vs the EM Algorithm

EM algorithm has dominated the literature on maximum likelihood estimation of mixture models. For the problem of Kiefer-Wolfowitz nonparametric maximum likelihood estimator for mixtures, Koenker and Mizera (2014) compared the modern interior point methods with EM algorithm. Their experience was that modern interior point methods are vastly superior, both in terms of accuracy and computational effort.

Here we compare the interior point algorithm with the EM Algorithm for a simulated data set. We generated $n = 1000$ observations from a single GCG component model with the parameters sepcified in table 4.

We used MATLAB optimization toolbox's *fmincon()* function for the implementation of the interior point algorithm (MATLAB (2014)) to solve problem (3). As in Koenker and Mizera (2014), in Table 5 we report timing information and the values of $L(\theta)$ achieved for the interior point algorithm and EM algorithms with various number of iterations.

The EM algorithm makes little progress from 50 to 500 iterations. By contrast, the interior point algorithm as implemented in MATLAB is both quicker and more accurate.

Table 6 reports initial values and fitted values of the algorithms for the component proportion parameters, where the initial values for proportions are all equal. The interior point algorithm's fitted proportions are closer to the true proportions than the ones by the EM algorithms.

Table 7 reports initial values and fitted values of the algorithms for the component copula parameters, The initial value of a copula parameter is the maximum likelihood estimate for the single component model without mixtures. The interior point algorithm's fitted copula parameters are closer to the true copula parameters than the ones by the EM algorithms.

Table 6: Comparison of EM and interior point solutions: proportion estimates

Component	True Proportion	Initial Value	Estimated Proportion			
			EM1	EM2	EM3	IP
Gaussian	0.35	0.3333	0.3367	0.3367	0.3367	0.3717
Clayton	0.40	0.3333	0.3375	0.3375	0.3375	0.3916
Gumbel	0.25	0.3333	0.3259	0.3259	0.3259	0.2367

Table 7: Comparison of EM and interior point solutions: copula parameter estimates

Component	True Parameter	Initial Value	Estimated copula parameter			
			EM1	EM2	EM3	IP
Gaussian	0.5	0.7610	0.5279	0.5284	0.5284	0.5804
Clayton	3	2.0737	3.5718	3.5592	3.5592	3.2684
Gumbel	10	2.1133	6.1807	6.1974	6.1974	10.7149

4. Estimation of Mutual Information Using Copula Density - Simulations

Mutual information (MI) is an information-theoretic quantity that measures the stochastic dependence between random variables. The mutual information $MI(X,Y)$ for the random variables X and Y is the Kullback-Leibler (KL) distance from the joint probability density $f(x,y)$ to the product of the marginal probability densities $f_X(x)$ and $f_Y(y)$:

$$MI(X,Y) \equiv E\left[\log\frac{f(X,Y)}{f_X(X)f_Y(Y)}\right] = \iint f(x,y)\log\frac{f(x,y)}{f_X(x)f_Y(y)}dxdy.$$

$MI(X,Y)$ can be expressed in terms of the copula density $c(u,v)$ as:

$$MI(X,Y) = E\left[\log c(U,V)\right] = \iint\limits_{[0,1]^2} c(u,v)\log c(u,v)dudv.$$

Therefore a byproduct of a copula density estimator $\hat{c}(u,v)$ is a MI estimator given by:

$$\widehat{MI}(X,Y)_c = \iint\limits_{[0,1]^2} \hat{c}(u,v)\log\hat{c}(u,v)dudv. \tag{5}$$

A high quality copula density estimator $\hat{c}(u,v)$ will lead to a high quality MI estimator $\widehat{MI}(X,Y)_c$.

There exists numerous methods for MI estimation, see Wang et al. (2009) for a detailed review. The classical histogram method (Moddemeijer (1989b)) estimates a continuous density by a histogram with fixed bin size. Recently, MLMI - maximum likelihood mutual information (Suzuki et al. (2009d)) method is proposed as an estimator of MI based on the density-ratio estimation method KLIEP. Kullback-Leibler Importance Estimation Procedure (KLIEP) (Sugiyama et al. (2008)) is an algorithm to directly estimate the ratio of two density functions without going through density estimation. The optimization problem involved with KLIEP is convex so the unique global optimal solution can be obtained efficiently. Furthermore, the KLIEP solution tends to be sparse, which contributes to reducing the computational time. The default basis functions are chosen as Gaussian kernels and the Gaussian width is chosen by cross-validation (CV). In a numerical experiment, MLMI with CV performs well. Squared-loss Mutual Information (SMI) is a squared-loss variant of mutual information (Sakai and Sugiyama (2014)). Therefore a byproduct of a copula density estimator $\hat{c}(u,v)$ is a SMI estimator given by:

$$\widehat{SMI}(X,Y)_c = \frac{1}{2}\iint\limits_{[0,1]^2}\left[\hat{c}(u,v)-1\right]^2 dudv. \tag{6}$$

LSMI - Least-Squares Mutual Information (Sakai and Sugiyama (2014)) - is an estimator of SMI based on the density-ratio estimation method uLSIF. Good SMI estimator was

Table 8: Copula Models for Simulation

Model	Component	Proportion	Parameter
1	Gaussian	0.6	$[-0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75]$
	Gaussian	0.4	0.5
2	Gaussian	0.6	$[-0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75]$
	Clayton	0.4	2
3	Clayton	0.6	$[0.22, 0.67, 1.33, 2, 3, 6, 18]$
	Gumbel	0.4	4

demonstrated to be useful in performing various machine learning tasks such as dimension reduction, clustering, and causal inference Suzuki et al. (2009c).

The MATLAB code for mutual information estimation by histogram is publicly available (Moddemeijer (1989a)). The MATLAB implementation of MLMI and LSMI for multiplicative kernel models are publicly available (Suzuki et al. (2009b,a)) - which is the main reason that we choose them for comparison.

We compare the estimator $\widehat{MI}(X,Y)_c$ in (5) with MI by histogram and MLMI, and compare $\widehat{SMI}(X,Y)_c$ in (6) with LSMI in terms of the root mean squared error(RMSE) between the estimator and the true value of the parameter $MI(X,Y)$ or $SMI(X,Y)$ through Monte Carlo simulation. Three different copula models are used for the simulation with their parameters given in Table 8.

In model 1 and 2, the correlation coefficient ρ for the Gaussian Copula was -0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75. In model 3, the parameter θ for the Clayton copula was chosen to be 0.22, 0.67, 1.33, 2, 3, 6, 18, corresponding to bivariate Kendall's τ being equal to 0.1, 0.25, 0.4, 0.5, 0.6, 0.75, 0.9, respectively. The correct values of $MI(X,Y)$ and $SMI(X,Y)$ were computed by numerical approximation for the given copula density function $c(u,v)$. The sample size n was set to 500 and 1000 with 100 Monte Carol replication each.

For each copula model, independent and identically distributed (i.i.d.) standard uniform bivariate random variables $\{(U_i, V_i)\}_{i=1}^n$ were generated from the specified mixture copula model. That was, $\{U_i\}_{i=1}^n$ was a sample from a Uniform(0,1) distribution, and so was the $\{V_i\}_{i=1}^n$. The joint density of (U,V) was the specified copula density $c(u,v)$. In the simulation, the marginal distributions were both standard Gaussian distributions. But this knowledge of marginal distributions is not used for copula density estimation, because they were estimated by ECDFs (1).

RMSEs of $\widehat{MI}(X,Y)_c$ vs. Histogram and MLMI, and $\widehat{SMI}(X,Y)_c$ vs LSMI for Model 1 through 3 are presented in Figures 1 through 3, respectively.

The simulation results show that the MI estimator based on copula density estimator outperforms those by histogram based and MLMI for the simulated data in most cases. The SMI estimator based on copula outperforms the LSMI for the simulated data in most cases as well. Only in a few cases of the Gaussian-Gaussian copula model with moderate correlation coefficients, copula based estimator slightly under-performs. It appears that both histogram based and MLMI estimators have difficulties when the data is highly dependent in the tails.

5. Application to real data

We apply our GCG components mixture model to a subset of the Framingham Heart study data (`http://www.framingham.com/heart/`). We focus on the dependence structure underlying the diastolic (DBP) and the systolic (SBP) blood pressures (in mmHg)

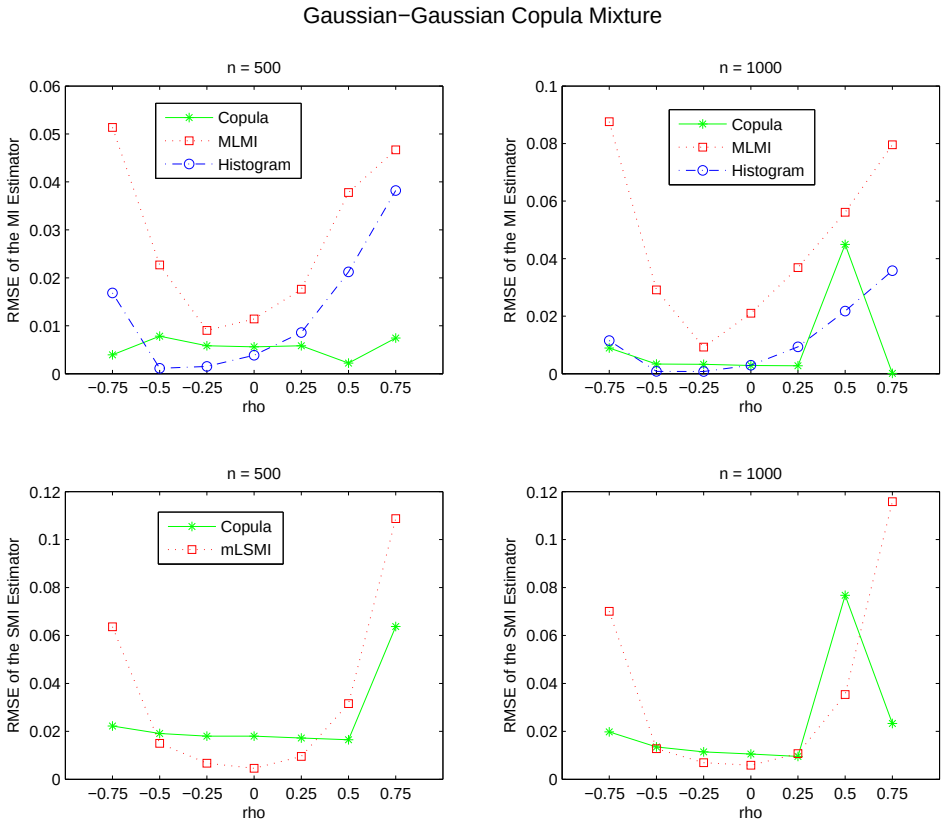

Figure 1: RMSE of $\widehat{MI}(X,Y)_c$ vs. Histogram and MLMI, $\widehat{SMI}(X,Y)_c$ vs. LSMI for the Gaussian-Gaussian copula mixture model. Computations are based on sample size $n = 500, 1000$ and 100 Monte Carlo repetitions.

measured on 663 male subjects at their first visit. The scatter plot of the log-blood pressures and the scatter plot of the standardized ranks of the log-blood pressures can be found in Fig. 4. It is evident that there is a strong positive dependence between the two responses. Lambert (2007) analyzed this data set assuming the copula density of the log-blood pressures was a sub-family of copulas named Archimedean with unknown (strict) generator. Lambert (2007) proposed a ratio approximation of the Archimedean copula generator and of its first derivative using B-splines, estimated the associated parameters using Markov chains Monte Carlo methods, and found that Gumbel copula was appropriate for this data without being fully satisfactory.

We applied our estimation procedure to this data set. For comparison, we also estimated parametric copula densities by assuming Gaussian, Clayton and Gumbel copula respectively for the data. Table 9 lists AICc along with the fitted parameters for these models. We found that Gaussian-Gumbel mixture copula model selected by our model fitting procedure slightly outperformed the Gumbel copula model in terms of AIC. Fig. 5 plots the fitted Gaussian-Gumbel copula density on the left and Gumbel copula density on the right panel. They look similar, with small difference in the front and back corners.

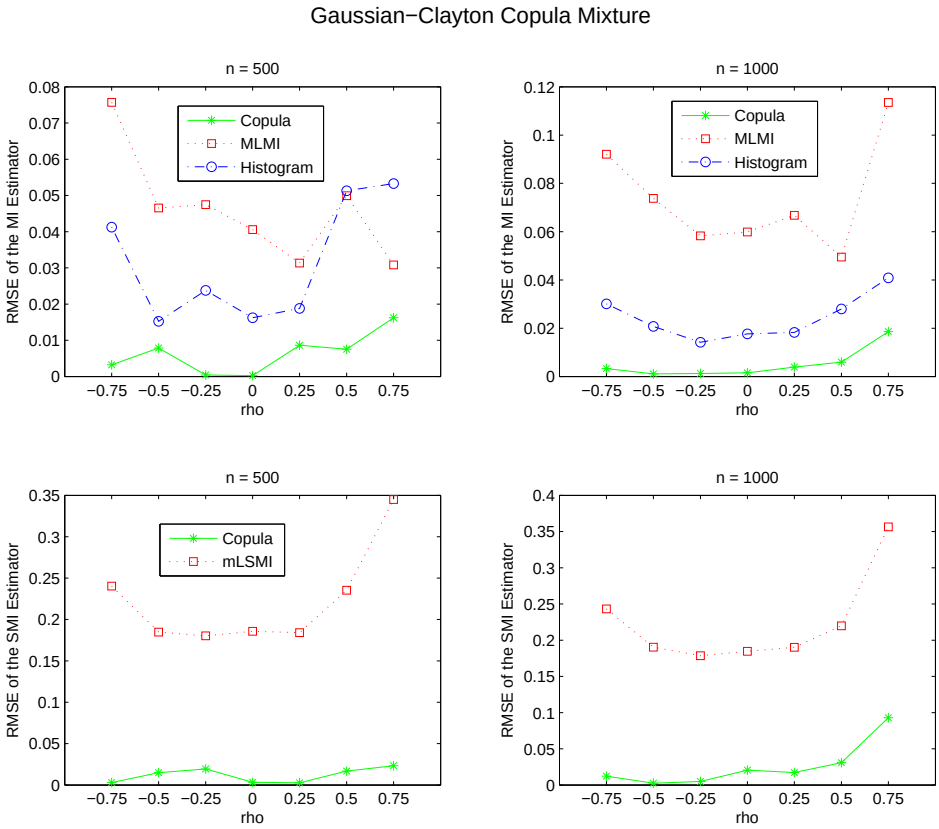

Figure 2: RMSE of $\widehat{MI(X,Y)}_c$ vs. Histogram and MLMI, $\widehat{SMI(X,Y)}_c$ vs. LSMI for the Gaussian-Clayton copula mixture model. Computations are based on sample size $n = 500, 1000$ and 100 Monte Carlo repetitions.

Table 9: AICs of the fitted models for the blood pressures data with $n = 663$

Model	AIC	Fitted parameters
Gaussian-Gumbel	289.1358	$\hat{p} = [0.5607, 0.4393], \hat{\theta} = [0.8357, 1.8513]$
Gumbel	282.2860	$\hat{\theta} = 2.1179$
Gaussian	273.3293	$\hat{\rho} = 0.7494$
Clayton	176.5958	$\hat{\theta} = 1.4467$

6. Concluding remarks

We presented a finite mixture of Gaussian, Clayton and Gumbel copula components model subject to the constraints that the model parameters for the same parametric components are ordered. The model parameters are estimated by interior-point algorithm for the resulting constrained maximum likelihood estimation problem, where the gradient of the objective function is not required.

The extension of our method to trivariate or higher dimensional copula density estimation is possible but requires more considerations on dealing with the multiple parameters of a copula model. For example, the parameter of a trivariate Gaussian copula is a 3×3 correlation matrix. The ordering of such correlation matrices is not a straight forward extension

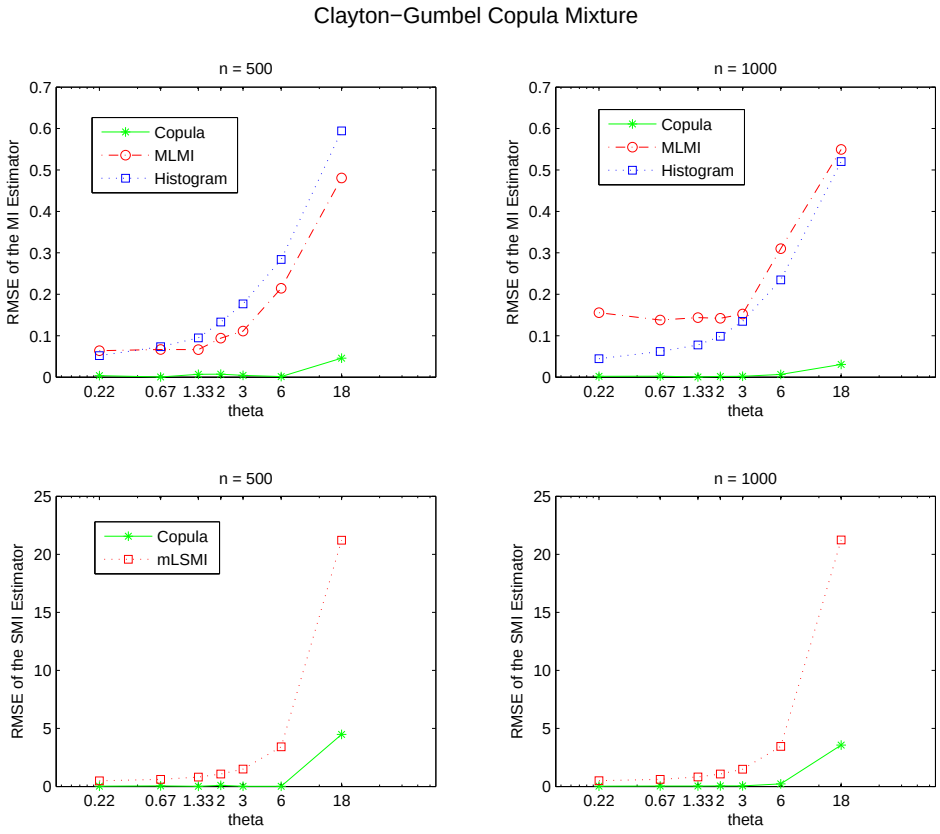

Figure 3: RMSE of $\widehat{MI(X,Y)_c}$ vs. Histogram and MLMI, $\widehat{SMI(X,Y)_c}$ vs. LSMI for the Gaussian-Clayton-Gumbel copula mixture model. Computations are based on sample size $n = 500, 1000$ and 100 Monte Carlo repetitions.

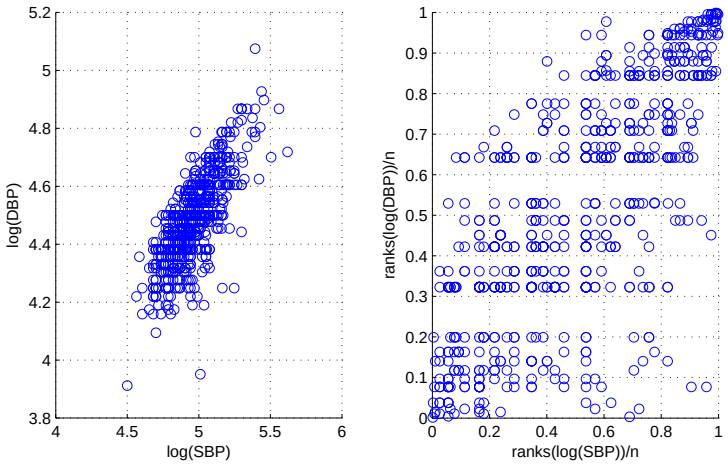

Figure 4: Left: log(DBP) vs. log(SBP); Right: standardized ranks of log(DBP) vs. those of log(SBP)

of the ordering of scalars.

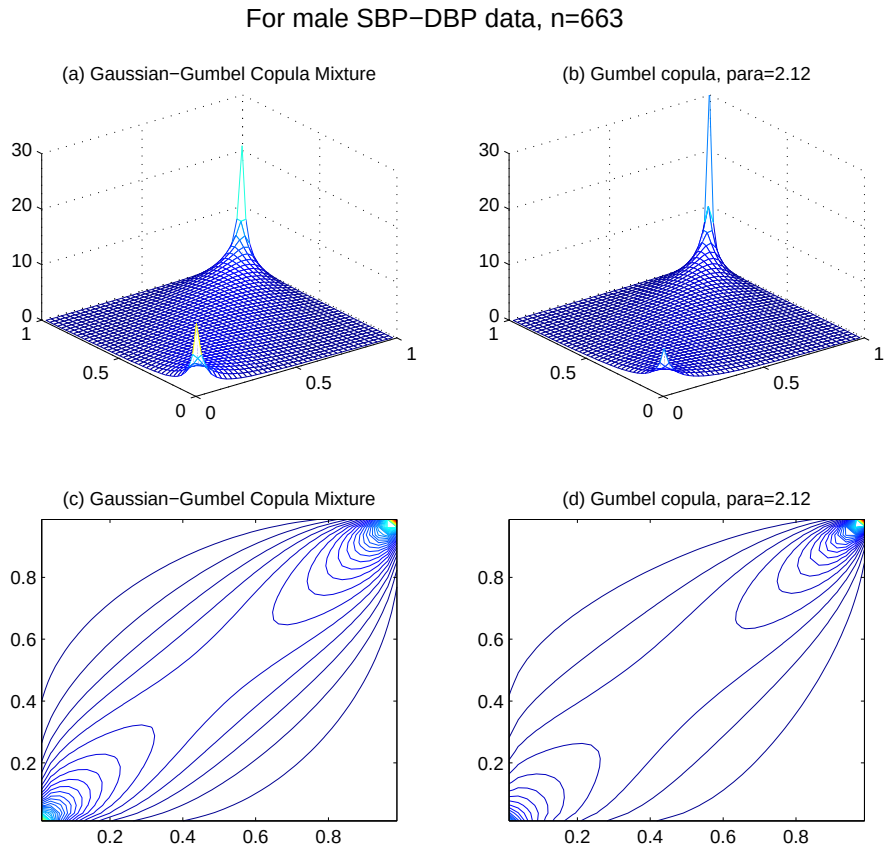

Figure 5: Left: Gaussian-Gumbel mixture copula model selected by the GCG components model fitting procedure; Right: Gumbel copula density with $\hat{\theta} = 2.1179$

References

Arakelian, V. and Karlis, D. "Clustering Dependencies Via Mixtures of Copulas." *Communications in Statistics - Simulation and Computation*, 43(7):1644–1661 (2014).

Byrd, R., Gilbert, J. C., and Nocedal, J. "A Trust Region Method Based on Interior Point Techniques for Nonlinear Programming." *Mathematical Programming*, 89(1):149185 (2000).

Byrd, R., Hribar, E., and Nocedal, J. "An Interior Point Algorithm for Large-Scale Nonlinear Programming." *SIAM Journal on Optimization*, 9(4):877900 (1999).

Cai, Z. and Wang, X. "Selection of Mixed Copula Model via Penalized Likelihood." *Journal of the American Statistical Association*, 109(506):788–801 (2014).

Choros, B., Ibragimov, R., and Permiakova, E. "Copula Estimation." In *Copula Theory and Its Applications, Lecture Notes in Statistics, V198*, 77–91. Springer (2010).

Cox, D. and Wermuth, N. *Multivariate Dependencies: Models, Analysis and Interpretation*. CRC (1996).

Diebolt, J. and Robert, C. P. "Estimation of Finite Mixture Distributions through Bayesian Sampling." *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(2):363–375 (1994).

Genest, C., Ghoudi, K., and Rivest, L. "A semiparametric estimation procedure of dependence parameters in multivariate families of distributions." *Biometrika*, 82(3):543–552 (1995).

Hu, L. "Dependence patterns across financial markets: a mixed copula approach." *Applied Financial Economics*, 16(10):717–729 (2006).

Kauermann, G. and Meyer, R. "Penalized marginal likelihood estimation of finite mixtures of Archimedean copulas." *Computational Statistics*, 29:283–306 (2014).

Koenker, R. and Mizera, I. "Convex Optimization in R." *Journal of Statistical Software*, 60(5):1–23 (2014).

Lambert, P. "Archimedean Copula Estimation Using Bayesian Splines Smoothing Techniques." *Computational Statistics & Data Analysis*, 51(12):6307–6320 (2007).

MATLAB. *R2014a, fmincon function documentation*. Natick, Massachusetts: The MathWorks Inc. (2014).

Moddemeijer, R. *MATLAB implementation of Mutual Information Estimation by Histogram* (1989a). `http://www.cs.rug.nl/~rudy/matlab/`, accessed April 6, 2015.

—. "On estimation of entropy and mutual information of continuous distributions." *Signal Processing*, 16(3):233 – 248 (1989b).

Sakai, T. and Sugiyama, M. "Computationally efficient estimation of squared-loss mutual information with multiplicative kernel models." *IEICE Transactions on Information and Systems*, E97-D(4):968–971 (2014).

Scott, D. *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley: New York (1992).

Shih, J. H. and Louis, T. A. "Inferences on the association parameter in copula models for bivariate survival data." *Biometrics*, 51:1384–1399 (1995).

Sklar, A. *Fonctions de répartition à n dimensions et leurs marges*. Publ. Inst. Stat. Univ. Paris (1959).

Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Bunau, P., and Kawanabe, M. "Direct importance estimation for covariate shift adaptation." *Annals of the Institute of Statistical Mathematics*, 60(4):699–746 (2008).

Suzuki, T., Sugiyama, M., Kanamori, T., and Sese, J. *MATLAB implementation of Least-Squares Mutual Information (LSMI)* (2009a). `http://www.ms.k.u-tokyo.ac.jp/software.html#LSMI`, accessed April 6, 2015.

—. *MATLAB implementation of Maximum Likelihood Mutual Information (MLMI)* (2009b). `http://www.ms.k.u-tokyo.ac.jp/software.html#MLMI`, accessed April 6, 2015.

—. "Mutual information estimation reveals global associations between stimuli and biological processes." *BMC Bioinformatics*, 10(1):S52 (2009c).

Suzuki, T., Sugiyama, M., and Tanaka, T. "Mutual information approximation via maximum likelihood estimation of density ratio." In *Information Theory, 2009. ISIT 2009. IEEE International Symposium on*, 463 –467 (2009d).

Waltz, R. A., Morales, J. L., Nocedal, J., and Orban, D. "An interior algorithm for nonlinear optimization that combines line search and trust region steps." *Mathematical Programming*, 107(3):391408 (2006).

Wang, Q., Kulkarni, S. R., and Verdú, S. "Universal Estimation of Information Measures for Analog Sources." *Found. Trends Commun. Inf. Theory*, 5:265–353 (2009).

Wright, M. H. "Interior Methods for Constrained Optimization." *Acta Numerica*, 1:341–407 (1992).

Wright, S. *Primal-Dual Interior-Point Methods*. Philadelphia: SIAM (1997).