

Copula density estimation by total variation penalized likelihood with constraints

Leming Qu

Department of Mathematics

Boise State University

Boise, ID.

Presented at Interface 2008 Conference

Outline

2

- Introduction
- Methodology
- Selection of the smoothing parameter
- Simulation Studies
- Another formulation
- Remarks

Introduction

3

- Quantifying the dependence among two or more variables has been an enduring task in many applications of statistics.
- Copula is a device that fully quantifies the dependence among random variables.
- It has gain popularity in recent years, especially in economics and finance, where dependence across different financial markets plays a key role in studying asset allocations and investment strategies.

What is a copula?

4

- A copula is a function that connects the marginal distributions to the joint distribution.
- (X, Y) : a bivariate random variable,
 F : the distribution function with marginal distributions F_X for X and F_Y for Y .
- The Sklar's Theorem assures the existence of a bivariate distribution function C on $[0, 1]^2$ such that
$$F(x, y) = C(F_X(x), F_Y(y)).$$
- The function C is called the copula associated with (X, Y)

What is the task?

5

- The copula is simply the joint distribution function of $(U, V) = (F_X(x), F_Y(y))$
a bivariate standard uniform random variable.

- We observe independent sample data:

$$(X_1, Y_1), \dots, (X_n, Y_n) .$$

We wish to estimate aspects of the joint distribution of X and Y , in particular, the copula density function.

$$c(u, v) = \frac{\partial}{\partial u \partial v} C(u, v) \quad \text{is the copula density.}$$

Some existing literature

6

For copular density estimation:

- Gijbels & Mielniczuk (1990): Kerkel estimator;
Fermanian (2005) : a goodness-of-fit test.
- Sancetta & Satchell (2004) : techniques based on Bernstein polynomials.
- Hall & Neumeyer (2006) : a wavelet estimators.

Total Variation (TV) penalty based

7 penalized likelihood estimation

- Capable of capturing sharp changes in the target function, suffering less from edge effects when the copula density can be unbounded at boundaries in some statistically important cases.

Existing literature on density estimation by TV penalty

8

- Koenker and Mizera (2006) uses the TV of the derivative of the log density as the penalty in the univariate case and TV of the log density defined in a triogram in the bivariate case.
- Sardy and Tseng (2006) uses the TV of the density itself as the penalty.
- We use the TV of the log density defined in a regular grid in this paper.

Methodology

9

- As copulas are not directly observable, a nonparametric copula density estimator has to be formed in two stages: obtaining the observations for (U, V) first and then estimate the copula density based on these observations.

First stage of the estimation

10

- The original data set (X_i, Y_i) 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.

- If models are available for the marginal distributions of X and Y but not for the joint distribution, then a technique such as maximum likelihood would be used to estimate the marginal distribution functions.
- Otherwise one can either use some nonparametric univariate distribution estimation methods or simply use the empirical distribution functions.

Second stage of the estimation

11

- To estimate the copula density $c(u,v)$ based on the $\{(\hat{U}_i, \hat{V}_i)\}_{i=1}^n$ observations.

Without assuming any parametric form for the copula density $c(u,v)$, we wish to estimate it over the set of rectangles defined by the grids on $[0,1]^2$ with density condition.

- This amounts to assume $c(u,v)$ piecewise constant over fine rectangles and we are treating it as a two-dimensional (2-D) digital image.

Gridding Strategy

12

- Following O'Sullivan (1992) in the density estimation setting, a reasonable gridding strategy is to let the number of grid points m in each direction be cn^α , where the constant c is chosen so that there are 64 grid points when sample size $n = 1,000$.

According to O'Sullivan (1992):

“It could be argued on asymptotic grounds that for a density whose second derivative is square integrable, $\alpha = 1/4$ would be reasonable. Too fine a discretization will unnecessarily compromise computational efficiency.”

Data and notation

13

- By projecting the converted scattered data onto the grids on $[0,1]^2$, we obtain a_{ij} for $i, j = 1, \dots, m$, with a_{ij} simply the count of observations falling in the (i, j) th rectangle.
- Denote c_{ij} as the value of $c(u,v)$ when (u,v) belongs to the (i, j) th rectangle.
- Let $x_{ij} = \log c_{ij}$.

The penalized log likelihood estimate with TV penalty

14

$$\square \text{ (P)} \quad \min - \sum_{ij} a_{ij} x_{ij} + \lambda TV(x)$$

$$\text{s.t.} \quad \frac{1}{N} \sum_{ij} \exp(x_{ij}) = 1$$

λ : a smoothing parameter controlling the smoothness of the estimate.

$$\square TV(\mathbf{x}) = \sum_{ij} \sqrt{(x_{i+1,j} - x_{ij})^2 + (x_{i,j+1} - x_{ij})^2}$$

Equivalence of two optimization problem

15

- The integrability constraint can be conveniently incorporated into the object function using Lemma 1 of Koenker and Mizera (2006) which is based on the discretized version of Silverman (1982).
- The constrained minimization problem (P) is equivalent to the following unconstrained minimization problem:
- (Q)

$$\min \sum_{ij} \left[-a_{ij} x_{ij} + \frac{n}{N} \exp(x_{ij}) \right] + \lambda TV(x)$$

An iteratively reweighted algorithm

16

- The problem (Q) can be solved by an iteratively reweighted algorithm following O'Sullivan, F. (1992).
- The derivation of this algorithm provides an estimate of the degrees of freedom(df) based on the last Newton-Raphson update.

A second order cone program (SOCP)

17

- Alternatively, the modern interior point methods for convex optimization problems can be used to solve (Q).
- Introducing the artificial barrier t_{ij} , the TV(x) contribution can be reformulated slightly, and we can write (Q) as

$$\min \sum_{ij} \left[-a_{ij} x_{ij} + \frac{n}{N} \exp(x_{ij}) + \lambda t_{ij} \right]$$

$$\text{S.t. } (x_{i,j+1} - x_{ij})^2 + (x_{i,j+1} - x_{ij})^2 - t_{ij}^2 \leq 0$$

A second order cone program (SOCP)

18

- The above problem is a second order cone program (SOCP).
The log-barrier method
(Boyd and Vandenberghe 2004)
is conceptually straightforward to solve the SOCP.
- At the core of the log-barrier method is solving for a series of Newton steps.
- Fortunately, for our particular problem, the Hessian matrix with dimension $N \times N$ has a simple form and is very sparse.
The sparse linear equation solver can be used to speed up the Newton step.

SOCP



- The SOCP is a special case of semidefinite programming (SDP).
- The SOCP can be solved more efficiently than the SDP using the primal-dual, interior-point method.
- The SOCP was studied by many authors in applied mathematics and operation research areas.
- However, there is very little literature in statistics on the use of SOCP to estimate parameters or make inference.

Selection of the smoothing parameter

20

- The smoothing parameter $\lambda > 0$ indexes a continuous class of models. With $\lambda = 0$, no penalty is imposed on the log likelihood, then the estimate deteriorates to the empirical density, i.e. the histogram. With too large a λ , the $TV(x)$ is forced to be zero, then the x has to be a constant.
- The right λ balances the goodness-of-fit and the smoothness of the estimate.

The degrees of freedom (df)

21

- In most model selection strategies, a key step is to determine the degrees of freedom (df) of an estimate.
- We estimate the df by computing the trace of the hat matrix in the final Newton-Raphson update.

The Hat Matrix

22

- The penalty term in O'Sullivan (1992) is quadratic, hence the gradient of the penalty term is linear.
- Our penalty term $TV(x)$ is non-quadratic, its gradient is nonlinear.
- In essence, we seek the linear approximation to the gradient of $TV(x)$ in order to find the approximate Hat Matrix

The degrees of freedom (df)

23

- After some derivations, the trace of the hat matrix :

$$df(\lambda) = T \left\{ \left[\lambda \frac{N}{n} \text{diag}(\exp(-x_\lambda)) B(x_\lambda) + I \right]^{-1} \right\}$$

Where $B(x_\lambda)$ is a sparse matrix

AIC & BIC

24

- $AIC(\lambda) = l(x_\lambda) + df(\lambda)$.
- $BIC(\lambda) = l(x_\lambda) + df(\lambda) \log(n)/2$
- In our simulation study, it is found that the λ selected by AIC tends to be smaller than the optimum λ which minimize the integrated square error (ISE) of our density estimate. This is consistent with the well-publicized negative correlation between the optimal and the CV or AIC smoothing parameters.
- The smoothing parameter selected by BIC mimics the optimum one well in our simulation study.

Simulation Study

25

- In this simulation, we only consider the 2nd stage of the estimation procedure by assuming marginals being known.
That is, we are estimating the joint density of a bivariate Uniform random variables.
- This will not confound the estimation error in the first and second stage of the copula density estimation.

Clayton and Gumbel copula

26

- The Clayton copula:

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

we use $\theta = 0.6$;

- The Gumbel copula:

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

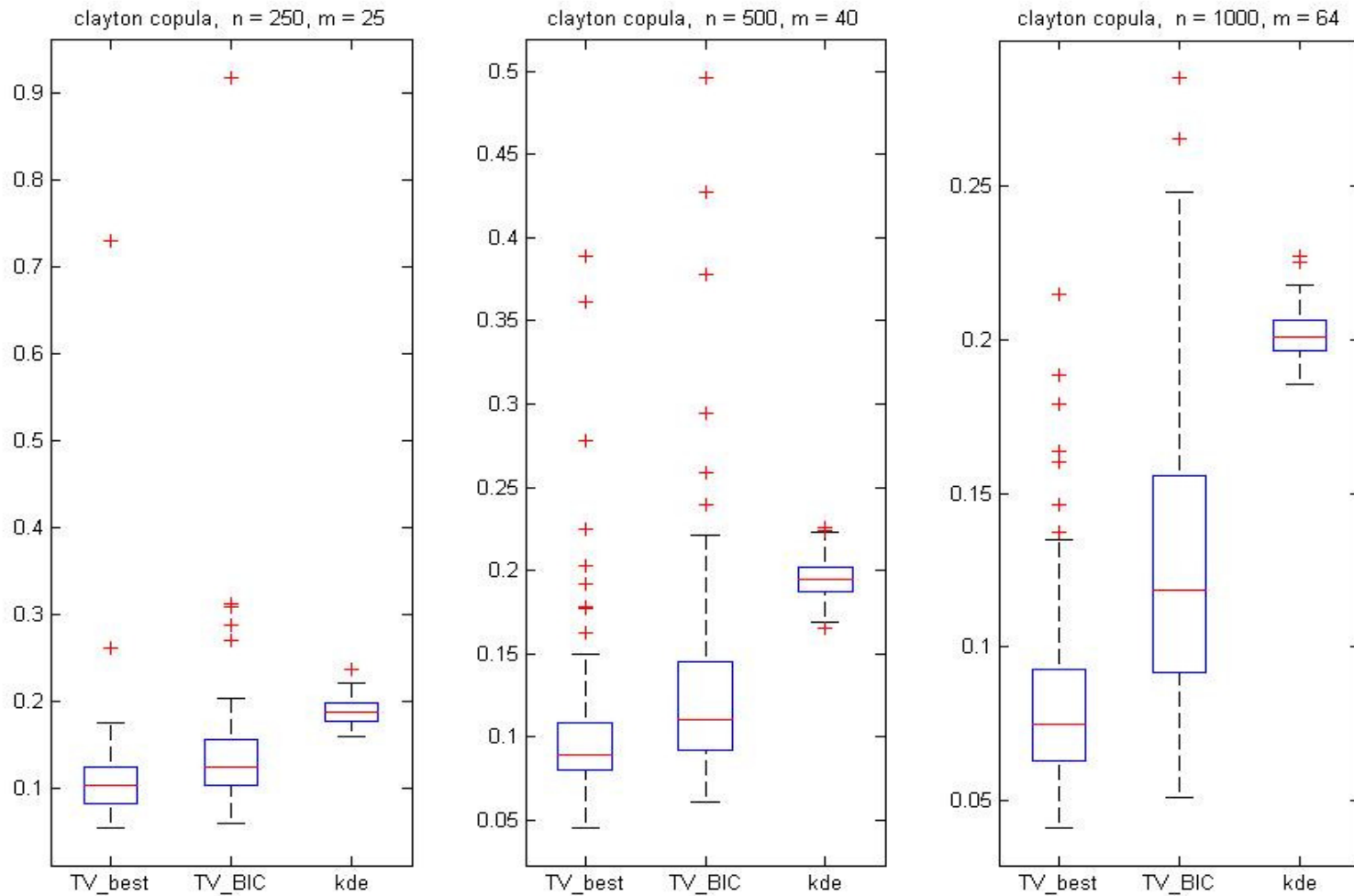
we use $\theta = 2$;

Simulation

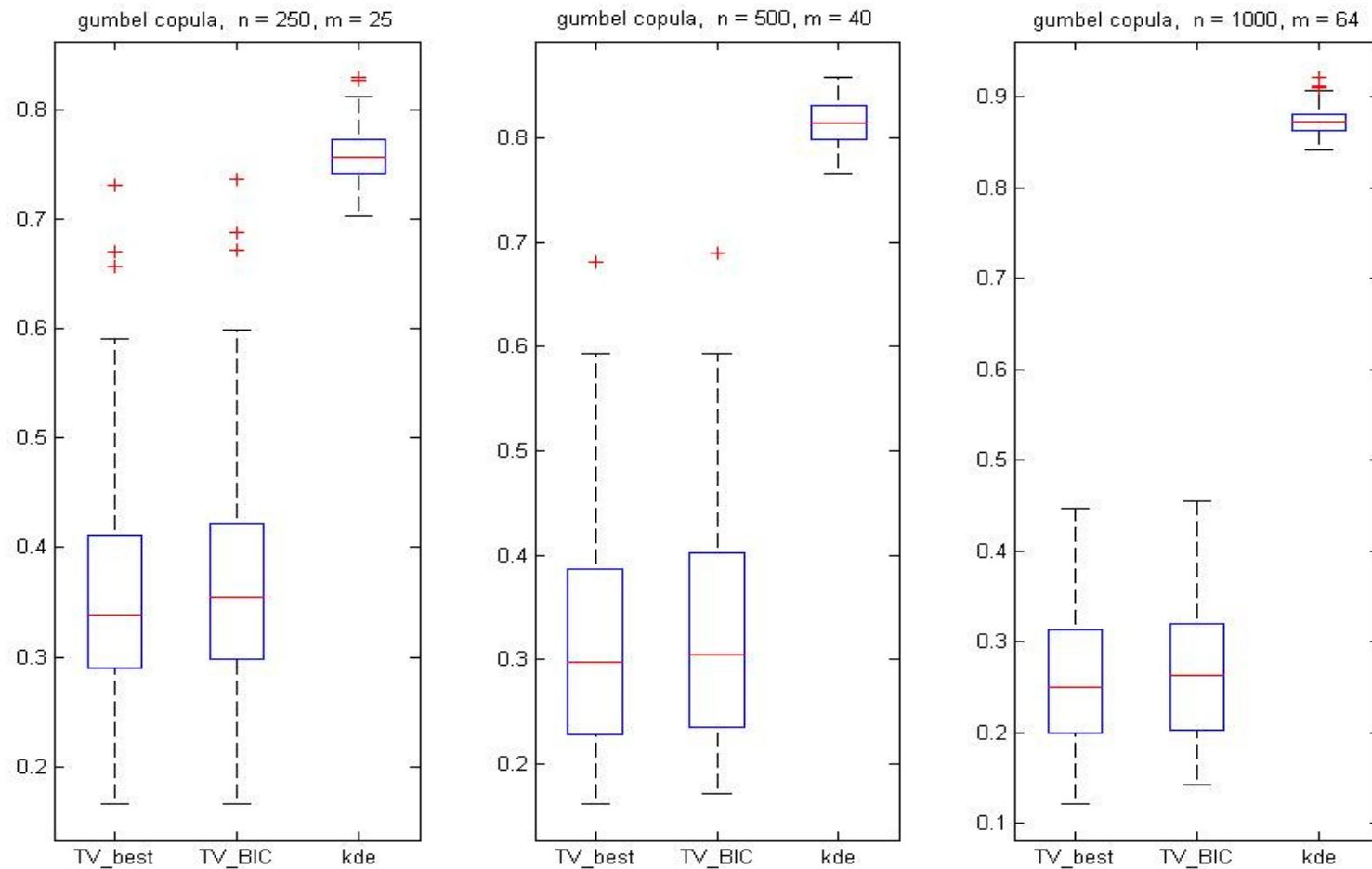
27

- For each copula model, independent and identically distributed uniform random bivariate variables $\{(U_i, V_i)\}_{i=1}^n$ are generated from the specified Copula.
- The sample sizes considered are $n = 250, n=500$ and $n=1000$.
- It is replicated 100 times.

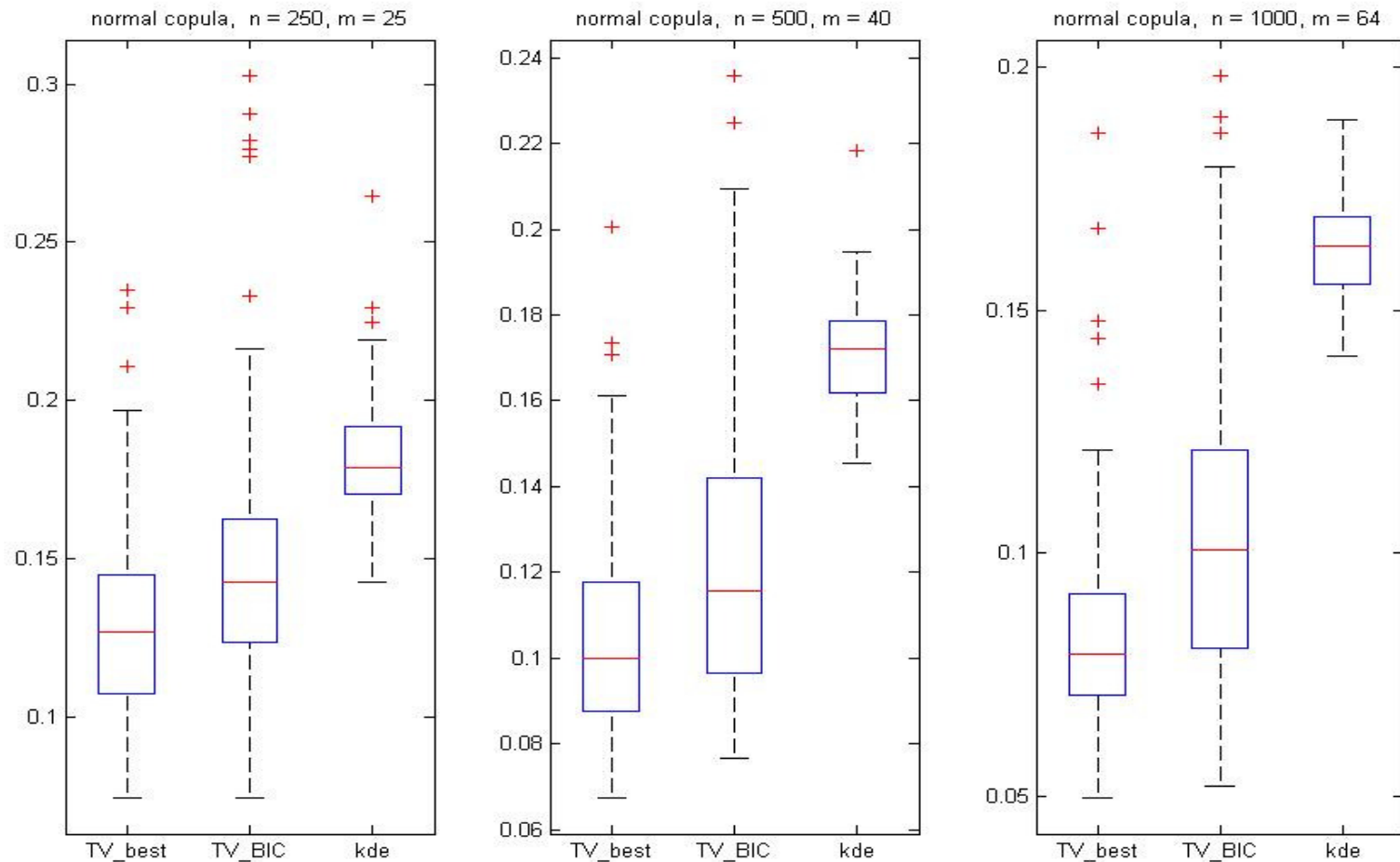
Boxplots of ISEs of the estimate for the Clayton copula



Boxplots of ISEs of the estimate for the Gumbel copula



Boxplots of ISEs of the estimate for the Normal copula

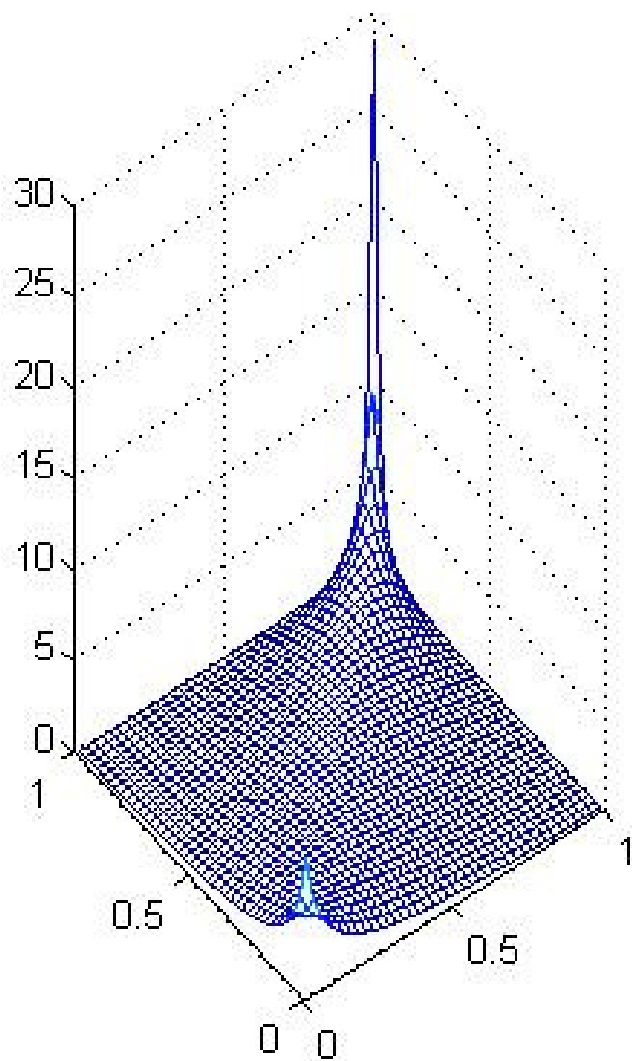


Monte Carlo approximations to MISE

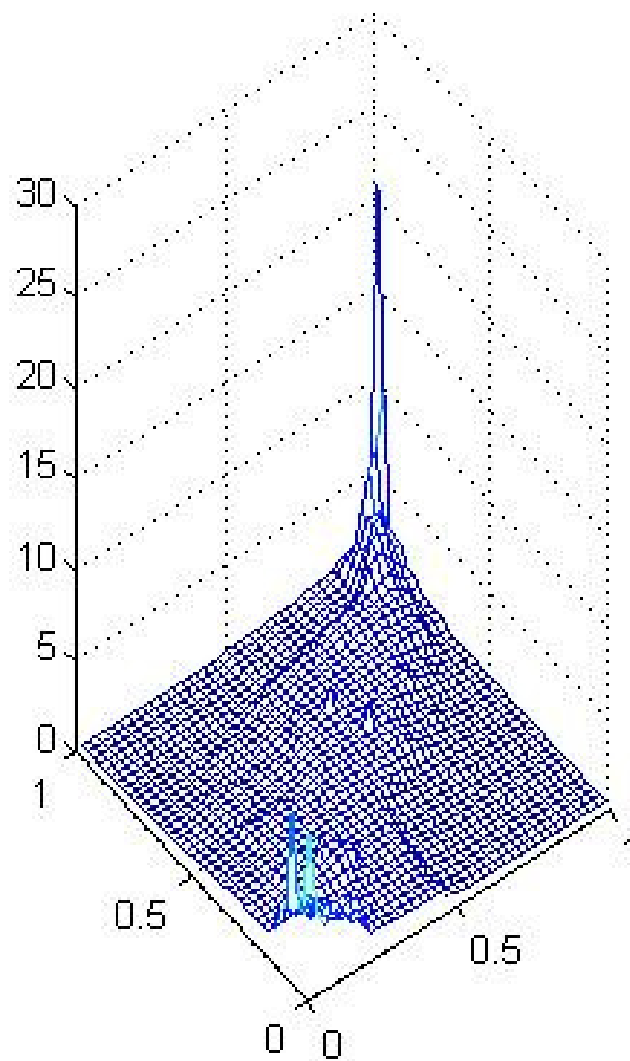
□ Copula	n	m	TV_best	TV_BIC	kde
□	-----				
□	250	25	0.1134	0.1397	0.1894
□ clayton	500	40	0.1044	0.1308	0.1947
□	1000	64	0.0829	0.1291	0.2016
□	-----				
□	250	25	0.3621	0.3723	0.7585
□ gumbel	500	40	0.3123	0.3232	0.8140
□	1000	64	0.2617	0.2705	0.8732
□	-----				
□	250	25	0.1294	0.1494	0.1822
□ normal	500	40	0.1045	0.1251	0.1711
□	1000	64	0.0839	0.1048	0.1627

A typical run for Gumbel copula with $n=1000$

the true copula density



the estimated copula density, λ selected by BIC



The penalized log likelihood estimate with TV penalty: Another formulation

33

$$\square \text{ (P2) } \min - \sum_{ij} a_{ij} \log(c_{ij}) + \lambda TV(c)$$

$$\text{s.t. } \sum_{j=1}^m c_{ij} = 1, \quad \text{for } i=1, \dots, m;$$
$$\sum_{j=1}^m c_{ij} = 1, \quad \text{for } j=1, \dots, m$$

λ : a smoothing parameter controlling the smoothness of the estimate.

$$\square TV(c) = \sum_{ij} \sqrt{(c_{i+1,j} - c_{ij})^2 + (c_{i,j+1} - c_{ij})^2}$$

Another formulation



- The optimization problem is harder
- There are $2m$ constraints
- Still the SOCP ?
- Provide genuine copular density

Some Remarks

35

- Some future efforts: asymptotic analysis (consistency, convergence rates)
- Try different penalty: TV of the derivatives of the copula density
- Higher dimensional data?