Analysis of the copula correlation matrix for meta-elliptical distributions

Yue Zhao

Department of Statistical Science
Cornell University

December 9, 2013
Bounding the operator norm of $\hat{\Sigma} - \Sigma$

Analyzing a factor model for the copula correlation matrix

Joint work with Marten Wegkamp

Supported by NSF Grant DMS 1310119
Introduction

Bounding the operator norm of $\hat{\Sigma} - \Sigma$

Analyzing a factor model for the copula correlation matrix

Outline

1. Introduction
   - Copula, elliptical copula, and the copula correlation matrix
   - Proposed research

2. Bounding the operator norm of $\hat{\Sigma} - \Sigma$
   - Bounding $\|\hat{T} - T\|_2$
   - Bounding $\|\hat{\Sigma} - \Sigma\|_2$

3. Analyzing a factor model for the copula correlation matrix
   - The factor model for $\Sigma$
   - Study of an elementary factor model
   - Analysis of the refined estimator
Introduction

Bounding the operator norm of $\hat{\Sigma} - \Sigma$

Analyzing a factor model for the copula correlation matrix

Outline

1. Introduction
   - Copula, elliptical copula, and the copula correlation matrix
   - Proposed research

2. Bounding the operator norm of $\hat{\Sigma} - \Sigma$
   - Bounding $\|\hat{T} - T\|_2$
   - Bounding $\|\hat{\Sigma} - \Sigma\|_2$

3. Analyzing a factor model for the copula correlation matrix
   - The factor model for $\Sigma$
   - Study of an elementary factor model
   - Analysis of the refined estimator
Let the continuous random vector $X = (X_1, \ldots, X_d)' \in \mathbb{R}^d$ have marginal distribution functions $F_j(x) = \mathbb{P}\{X_j \leq x\}$.

The copula function $C(u_1, \ldots, u_d)$ is the joint cumulative distribution function of the transformation

$$U = (F_1(X_1), \ldots, F_d(X_d))',$$

i.e., for $u = (u_1, \ldots, u_d)' \in [0, 1]^d$,

$$C(u_1, \ldots, u_d) = \mathbb{P}\{F_1(X_1) \leq u_1, \ldots, F_d(X_d) \leq u_d\}.$$

Copula is invariant under strictly increasing transformation of the marginals — allows separate specifications of the marginals and the dependence structure.
Elliptical distribution

A random vector \( X \in \mathbb{R}^d \) has an elliptical distribution if its characteristic function \( \mathbb{E}(\exp(it'X)) \) can be written as

\[
e^{it'\mu}\Psi(t'\Sigma t)
\]

with parameters \( \mu \in \mathbb{R}^d \), covariance matrix \( \Sigma \in \mathbb{R}^{d \times d} \), and characteristic generator \( \Psi : [0, \infty) \to \mathbb{R} \).

For instance, when \( \Psi(t) = e^{-t/2} \), \( X \sim \mathcal{N}(\mu, \Sigma) \).
Meta-elliptical distributions and the copula correlation matrix

- We call the set of distributions that can be obtained through strictly increasing transformations of the marginals of elliptical distributions the *meta-elliptical distributions*.
- Elliptical copulas are the copulas of the meta-elliptical distributions.
- Elliptical copula is characterized by *copula parameters*: $\Psi$ and the *copula correlation matrix* $\Sigma$, with

$$
\Sigma_{ij} = \frac{\bar{\Sigma}_{ij}}{\sqrt{\bar{\Sigma}_{ii} \bar{\Sigma}_{jj}}}.
$$

Yue Zhao
Copula correlation matrix for elliptical copulas
Kendall’s tau

- Let \( \tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_d)' \in \mathbb{R}^d \) be an independent copy of \( X = (X_1, \ldots, X_d)' \).
- (The population version of) Kendall’s tau between the \( k \)th and \( \ell \)th coordinates is
  \[
  \tau_{k\ell} = \mathbb{P}\{ (X_k - \tilde{X}_k)(X_\ell - \tilde{X}_\ell) > 0 \} - \mathbb{P}\{ (X_k - \tilde{X}_k)(X_\ell - \tilde{X}_\ell) < 0 \}
  \]
- (The empirical) Kendall’s tau statistic is, for samples \( X^1, \ldots, X^n \) that are independent copies of \( X \),
  \[
  \hat{\tau}_{k\ell} = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq d} \text{sgn} \left( (X_k^i - X_k^j)(X_\ell^i - X_\ell^j) \right).
  \]
- Let \( T \) be the matrix of (the population version of) Kendall’s tau, and \( \hat{T} \) the (empirical) matrix of Kendall’s tau statistics:
  \[
  [T]_{k\ell} = \tau_{k\ell}, \quad [\hat{T}]_{k\ell} = \hat{\tau}_{k\ell}.
  \]
Kendall’s tau matrix as plug-in estimator

- Kendall’s tau is a rank correlation and hence depends only on the copula parameters.
- For elliptical copulas,

\[ \Sigma = \sin \left( \frac{\pi}{2} T \right) \]

with the sine function acting component-wise (Kruskal 1958, Kendall & Gibbons 1990, Fang, Fang & Kotz 2002).

- Hence, the natural plug-in estimator for \( \Sigma \) is

\[ \hat{\Sigma} = \sin \left( \frac{\pi}{2} \hat{T} \right). \]
Introduction

Bounding the operator norm of $\hat{\Sigma} - \Sigma$

Analyzing a factor model for the copula correlation matrix

Proposed research

Outline

1. Introduction
   - Copula, elliptical copula, and the copula correlation matrix
   - Proposed research

2. Bounding the operator norm of $\hat{\Sigma} - \Sigma$
   - Bounding $\|\hat{T} - T\|_2$
   - Bounding $\|\hat{\Sigma} - \Sigma\|_2$

3. Analyzing a factor model for the copula correlation matrix
   - The factor model for $\Sigma$
   - Study of an elementary factor model
   - Analysis of the refined estimator
For the rest of the talk, we assume

- $X \in \mathbb{R}^d$ follows a meta-elliptical distribution,
- The copula correlation matrix of $X$ is $\Sigma \in \mathbb{R}^{d \times d}$,
- The plug-in estimator $\hat{\Sigma} = \hat{\Sigma}_n$ of $\Sigma$ is constructed from $n$ independent copies of $X$. 
Some recent advanced in elliptical/Gaussian copulas

- Klüppelberg & Kuhn 2009 study the property of $\hat{\Sigma}$ in the asymptotic regime, i.e., $d$ fixed, $n \to \infty$.
- Liu et al 2012, Xue & Zou 2012 study precision matrix (i.e., $\Sigma^{-1}$) estimation for Gaussian copula in high dimensions.
  - Sharp bound on $\|\hat{\Sigma} - \Sigma\|_{\text{max}}$ is a key step of the study.
First, we establish a sharp bound of the operator norm $\|\hat{\Sigma} - \Sigma\|_2$. (See also Han & Liu 2013.)

- It has often been observed (e.g., Demarta & McNeil 2004, Klüppelberg & Kuhn 2009) that the plug-in estimator $\hat{\Sigma}$ is not always positive semidefinite.
- A bound on $\|\hat{\Sigma} - \Sigma\|_2$ quantifies the extent to which the non-positive semidefinite problem may happen.
- The bound may also have potential applications in copula testing, copula classification, etc.
Proposed Research: Part II

Later, we study a factor model for $\Sigma$.

- The factor model assumes that $\Sigma = \Theta + V$ for a low-rank (or nearly low-rank) matrix $\Theta$ and a diagonal matrix $V$.
- Within this context, the effective number of parameters is $O(r \cdot d)$, for $r = \text{rank}(\Theta)$, instead of $O(d^2)$ under the generic model.
- We propose a refined estimator $\tilde{\Sigma}$ for $\Sigma$ using a penalized least square procedure.
- The bound on $\|\hat{\Sigma} - \Sigma\|_2$ serves to scale the penalty term.
Outline for bounding $\|\hat{\Sigma} - \Sigma\|_2$

- Bounding $\|\hat{T} - T\|_2$ (matrix counterpart of Hoeffding 1963’ classical bound for scalar U-statistic).
- Bounding $\|\hat{\Sigma} - \Sigma\|_2$ in terms of $\|\hat{T} - T\|_2$ (matrix counterpart of the Lipschitz property).
Outline

1. Introduction
   - Copula, elliptical copula, and the copula correlation matrix
   - Proposed research

2. Bounding the operator norm of $\hat{\Sigma} - \Sigma$
   - Bounding $\|\hat{T} - T\|_2$
   - Bounding $\|\hat{\Sigma} - \Sigma\|_2$

3. Analyzing a factor model for the copula correlation matrix
   - The factor model for $\Sigma$
   - Study of an elementary factor model
   - Analysis of the refined estimator
Theorem

With probability at least $1 - \alpha$, we have

$$\| \hat{T} - T \|_2 < \max \left\{ \sqrt{\| T \|_2^2 f_{n,d,\alpha}}, f_{n,d,\alpha}^2 \right\}$$

$$\leq \sqrt{\| \hat{T} \|_2^2 f_{n,d,\alpha}^2 + f_{n,d,\alpha}^2 / 4 + f_{n,d,\alpha}^2 / 2}$$

$$< \max \left\{ \sqrt{\| T \|_2^2 f_{n,d,\alpha}^2}, f_{n,d,\alpha}^2 \right\} + f_{n,d,\alpha}^2$$

with

$$f_{n,d,\alpha} = \sqrt{\frac{16}{3} \cdot \frac{d \cdot \log(2\alpha^{-1}d)}{n}}.$$ 

The above bounds hold for all $n, d, \alpha$. 
Outline

1. Introduction
   - Copula, elliptical copula, and the copula correlation matrix
   - Proposed research

2. Bounding the operator norm of $\hat{\Sigma} - \Sigma$
   - Bounding $\|\hat{T} - T\|_2$
   - Bounding $\|\hat{\Sigma} - \Sigma\|_2$

3. Analyzing a factor model for the copula correlation matrix
   - The factor model for $\Sigma$
   - Study of an elementary factor model
   - Analysis of the refined estimator
Theorem

- We have, with probability at least $1 - \alpha$, that

$$
\|\hat{\Sigma} - \Sigma\|_2 \leq \pi \|\hat{T} - T\|_2 + \frac{3}{16} \pi^2 \cdot f_{n,d,\alpha}^2.
$$

- Combining earlier bound for $\|\hat{T} - T\|_2$, we have, with probability at least $1 - 2\alpha$, that

$$
\|\hat{\Sigma} - \Sigma\|_2 < \pi \cdot \max \left\{ \sqrt{\|T\|_2 f_{n,d,\alpha}}, f_{n,d,\alpha}^2 \right\} + \frac{3}{16} \pi^2 \cdot f_{n,d,\alpha}^2.
$$

- In addition,

$$
\frac{2}{\pi} \|\Sigma\|_2 \leq \|T\|_2 \leq \|\Sigma\|_2.
$$
Introduction

Bounding the operator norm of $\hat{\Sigma} - \Sigma$

Analyzing a factor model for the copula correlation matrix

Outline

1. Introduction
   - Copula, elliptical copula, and the copula correlation matrix
   - Proposed research

2. Bounding the operator norm of $\hat{\Sigma} - \Sigma$
   - Bounding $\|\hat{T} - T\|_2$
   - Bounding $\|\hat{\Sigma} - \Sigma\|_2$

3. Analyzing a factor model for the copula correlation matrix
   - The factor model for $\Sigma$
   - Study of an elementary factor model
   - Analysis of the refined estimator
From now on, we assume that the copula correlation matrix $\Sigma$ satisfies a factor model:

$$\Sigma = \Theta + V$$

for some low-rank (or nearly low-rank), positive semidefinite, symmetric matrix $\Theta$ and diagonal matrix $V$. 
Introduction

Bounding the operator norm of $\hat{\Sigma} - \Sigma$

Analyzing a factor model for the copula correlation matrix

Outline

1. Introduction
   - Copula, elliptical copula, and the copula correlation matrix
   - Proposed research

2. Bounding the operator norm of $\hat{\Sigma} - \Sigma$
   - Bounding $\|\hat{T} - T\|_2$
   - Bounding $\|\hat{\Sigma} - \Sigma\|_2$

3. Analyzing a factor model for the copula correlation matrix
   - The factor model for $\Sigma$
   - Study of an elementary factor model
   - Analysis of the refined estimator

Yue Zhao
Copula correlation matrix for elliptical copulas
In an elementary factor model for $\Sigma = \Theta + V$, we assume that $\text{rank}(\Theta) = r < d$, and

$$V = \sigma^2 I_d.$$
Prelude: an elementary factor model for $\Sigma$

- Let the plug-in estimator $\hat{\Sigma}$ have the eigen-decomposition
  \[ \sum_{k=1}^{d} \hat{\lambda}_k \hat{u}_k \hat{u}_k', \quad \hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_d. \]

- Construct the following closed-form estimators for $r$, $\sigma^2$, and $\Theta$, based on some threshold $\mu$:
  \[
  \hat{r} = \sum_{k=1}^{d} \mathbb{I} \left\{ \hat{\lambda}_k \geq \hat{\lambda}_d + \mu \right\},
  \]
  \[
  \hat{\sigma}^2 = \frac{1}{d - \hat{r}} \sum_{j>\hat{r}} \hat{\lambda}_j,
  \]
  \[
  \hat{\Theta} = \sum_{k=1}^{\hat{r}} (\hat{\lambda}_k - \hat{\sigma}^2) \hat{u}_k \hat{u}_k'.
  \]
Theorem

Assume $\lambda_r(\Theta) \geq 2\mu$. On the event \( \left\{ 2\|\hat{\Sigma} - \Sigma\|_2 < \mu \right\} \), we have

\[
\hat{r} = r, \\
\|\hat{\Theta} - \Theta\|_F^2 \leq 8r\|\hat{\Sigma} - \Sigma\|_2^2, \\
|\hat{\sigma}^2 - \sigma^2| \leq \|\hat{\Sigma} - \Sigma\|_2.
\]

Here, \( \| \cdot \|_F \) denotes the Frobenius norm.
Corollary

Set

\[
\mu = 8 \sqrt{\| \hat{T} \|_2 f_{n, d, \alpha}^2 + f_{n, d, \alpha}^4 / 4} + 8 f_{n, d, \alpha}^2
\]
\[
\bar{\mu} = 8 \sqrt{\| T \|_2 f_{n, d, \alpha}^2 + 12 f_{n, d, \alpha}^2}.
\]

Assume \( \lambda_r(\Theta) \geq 2 \bar{\mu} \), and \( \| T \|_2 \geq f_{n, d, \alpha}^2 \). Then,

\[
\hat{r} = r,
\]
\[
\| \hat{\Theta} - \Theta \|^2_F \leq 2 r \bar{\mu}^2,
\]
\[
| \hat{\sigma}^2 - \sigma^2 | \leq \bar{\mu} / 2.
\]

hold with probability exceeding \( 1 - 2\alpha \).
Outline

1 Introduction
   - Copula, elliptical copula, and the copula correlation matrix
   - Proposed research

2 Bounding the operator norm of $\hat{\Sigma} - \Sigma$
   - Bounding $\|\hat{T} - T\|_2$
   - Bounding $\|\hat{\Sigma} - \Sigma\|_2$

3 Analyzing a factor model for the copula correlation matrix
   - The factor model for $\Sigma$
   - Study of an elementary factor model
   - Analysis of the refined estimator
The refined estimator $\tilde{\Sigma}$ of $\Sigma$

Let the refined estimator $\tilde{\Sigma}$ of the copula correlation matrix $\Sigma$ be

$$\tilde{\Sigma} = \tilde{\Theta}_o + I_d,$$

with $\tilde{\Theta}$ being the solution to the convex program

$$\tilde{\Theta} = \arg\min_{\Theta' \in \mathbb{R}^{d \times d}} \left\{ \frac{1}{2} \| \Theta'_o - \hat{\Sigma}_o \|_F^2 + \mu \| \Theta' \|_* \right\}.$$

Here,

- $A_o$ denotes $A - \text{diag}(A)$,
- $\| \cdot \|_*$ denotes the nuclear norm.
Comparison of $\tilde{\Sigma}$ with existing estimators

- Saunderson et al 2012 study a factor model for $\Sigma$ in the *noiseless* setting using *minimum trace factor analysis*:

\[
(\Theta, V) = \arg\min \text{tr}(\Theta') \text{ subject to } \begin{cases}
\Sigma = \Theta' + V \\
\Theta' \succ 0 \\
V \text{ diagonal}
\end{cases}
\]

- If $\Theta$ has column space $U$ such that its *coherence*

\[
\max_{1 \leq i \leq d} \|Pu e_i\|_2 < 1/\sqrt{2},
\]

then the above algorithm recovers $\Theta, V$ exactly.
Comparison of $\tilde{\Sigma}$ with existing estimators, continued

For $\Theta$ with effective low-rank and without sparse corruption, Lounici 2013 proposes

$$
\tilde{\Theta} = \arg\min_{\Theta' \in \mathbb{R}^{d \times d}} \left\{ \frac{1}{2} \|\Theta' - \hat{\Theta}\|_F^2 + \mu \|\Theta'\|_* \right\}
$$

for an unbiased initial estimator $\hat{\Theta}$ of $\Theta$ (which we lack).

For $\Sigma = \Theta + S$ with low-rank matrix $\Theta$ and generic sparse matrix $S$, Chandrasekaran et al 2012, Hsu et al 2011 propose

$$(\tilde{\Theta}, \tilde{S}) = \arg\min_{\Theta', S \in \mathbb{R}^{d \times d}} \left\{ \frac{1}{2} \|\Theta' + S - \hat{\Sigma}\|_F^2 + \mu \|\Theta'\|_* + \lambda \|S\|_{\ell_1} \right\}.$$  

This formulation requires $\lambda > 0$. 
Oracle inequality for the refined estimator $\tilde{\Sigma}$

Theorem

Let $\Theta_r$ be best rank-$r$ approximation to $\Theta$, with reduced SVD $\Theta_r = U_r D_r U'_r$. Suppose $\gamma_r = \|U_r U'_r\|_{\text{max}} \leq 1/9$. Set

$$\mu = 10 \sqrt{\|\hat{T}\|_2 f^2_{n,d,\alpha} + f^4_{n,d,\alpha}/4 + 10f^2_{n,d,\alpha}},$$

(recall $\mu$ is the regularization parameter)

$$\bar{\mu} = 10 \max \left[ \sqrt{\|\hat{T}\|_2 f_{n,d,\alpha}, f^2_{n,d,\alpha}} \right] + 15f^2_{n,d,\alpha}.$$

Then with probability at least $1 - 2\alpha$,

$$\|\tilde{\Sigma} - \Sigma\|^2_F \leq \sum_{j>r} \lambda^2_j(\Theta) + 18r\bar{\mu}^2.$$
We have obtained a sharp bound on the operator norm of $\hat{\Sigma} - \Sigma$, for the plug-in estimator $\hat{\Sigma}$ of the copula correlation matrix $\Sigma$.

We have applied the above bound to the setting of a factor model of $\Sigma$ to obtain a refined estimator $\tilde{\Sigma}$; a sharp oracle inequality for $\tilde{\Sigma}$ is established.
Introduction
Bounding the operator norm of $\hat{\Sigma} - \Sigma$
Analyzing a factor model for the copula correlation matrix

The factor model for $\Sigma$
Study of an elementary factor model
Analysis of the refined estimator

Thanks!