Analysis of the copula correlation matrix for meta-elliptical distributions

Yue Zhao

Department of Statistical Science
Cornell University

December 9, 2013

Joint work with
Marten Wegkamp

Supported by
NSF Grant DMS 1310119

Outline

Outline

Copula

- Let the continuous random vector $X = (X_1, \dots, X_d)' \in \mathbb{R}^d$ have marginal distribution functions $F_j(x) = \mathbb{P}\left\{X_j \le x\right\}$.
- The *copula function* $C(u_1, \dots, u_d)$ is the joint cumulative distribution function of the transformation

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

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

$$C(u_1, \dots, u_d) = \mathbb{P}\left\{F_1(X_1) \le u_1, \dots, F_d(X_d) \le u_d\right\}.$$

- 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

$$e^{it'\mu}\Psi(t'\overline{\Sigma}t)$$

  with parameters $\mu \in \mathbb{R}^d$, covariance matrix $\overline{\Sigma} \in \mathbb{R}^{d\times d}$, and characteristic generator $\Psi : [0, \infty) \to \mathbb{R}$.
- For instance, when $\Psi(t) = e^{-t/2}$, $X \sim N(\mu, \overline{\Sigma})$.

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*: Ψ and the *copula correlation matrix* Σ, with

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

Kendall's tau

- Let $\widetilde{X} = (\widetilde{X}_1, \ldots, \widetilde{X}_d)' \in \mathbb{R}^d$ be an independent copy of $X = (X_1, \ldots, X_d)'$.
- (The population version of) Kendall's tau between the kth and ℓth coordinates is

$$\tau_{k\ell} = \mathbb{P}\left\{(X_k - \widetilde{X}_k)(X_\ell - \widetilde{X}_\ell) > 0\right\} - \mathbb{P}\left\{(X_k - \widetilde{X}_k)(X_\ell - \widetilde{X}_\ell) < 0\right\}$$

- (The empirical) Kendall's tau statistic is, for samples $X^1, \ldots, X^n$ that are independent copies of X,

$$\widehat{\tau}_{k\ell} = \frac{2}{n(n-1)} \sum\sum_{1 \le i < j \le d} \operatorname{sgn}\left((X_k^i - X_k^j)(X_\ell^i - X_\ell^j)\right).$$

- Let T be the matrix of (the population version of) Kendall's tau, and $\widehat{T}$ the (empirical) matrix of Kendall's tau statistics:

$$[T]_{k\ell} = \tau_{k\ell}, \quad [\widehat{T}]_{k\ell} = \widehat{\tau}_{k\ell}.$$

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 Σ is

$$\widehat{\Sigma} = \sin\left(\frac{\pi}{2} \widehat{T}\right).$$

Outline

1. Introduction
   - Copula, elliptical copula, and the copula correlation matrix
   - Proposed research
2. Bounding the operator norm of $\widehat{\Sigma} - \Sigma$
   - Bounding $\|\widehat{T} - T\|_2$
   - Bounding $\|\widehat{\Sigma} - \Sigma\|_2$
3. Analyzing a factor model for the copula correlation matrix
   - The factor model for Σ
   - 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 $\widehat{\Sigma} = \widehat{\Sigma}_n$ of Σ is constructed from n independent copies of X.

Some recent advanced in elliptical/Gaussian copulas

- Klüppelberg & Kuhn 2009 study the property of $\widehat{\Sigma}$ in the asymptotic regime, i.e., d fixed, $n \to \infty$.
- Liu et al 2012, Xue & Zou 2012 study precision matrix (i.e., Σ^{-1}) estimation for Gaussian copula in high dimensions.
  - Sharp bound on $\|\widehat{\Sigma} - \Sigma\|_{\max}$ is a key step of the study.

Proposed Research: Part I

First, we establish a sharp bound of the operator norm $\|\widehat{\Sigma} - \Sigma\|_2$. (See also Han & Liu 2013.)

- It has often been observed (e.g., Demarta & McNeil 2004, Klüppelberg & Kuhn 2009) that the plug-in estimator $\widehat{\Sigma}$ is not always positive semidefinite.
- A bound on $\|\widehat{\Sigma} - \Sigma\|_2$ 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 Σ.

- The factor model assumes that $\Sigma = \Theta + V$ for a low-rank (or nearly low-rank) matrix Θ 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 $\widetilde{\Sigma}$ for Σ using a penalized least square procedure.
- The bound on $\|\widehat{\Sigma} - \Sigma\|_2$ serves to scale the penalty term.

Outline for bounding $\|\widehat{\Sigma} - \Sigma\|_2$

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

Outline

Theorem

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

$$
\begin{aligned}
\|\widehat{T} - T\|_2 &< \max\left\{\sqrt{\|T\|_2} f_{n,d,\alpha}, f^2_{n,d,\alpha}\right\} \\
&\leq \sqrt{\|\widehat{T}\|_2 f^2_{n,d,\alpha} + f^2_{n,d,\alpha}/4} + f^2_{n,d,\alpha}/2 \\
&< \max\left\{\sqrt{\|T\|_2} f_{n,d,\alpha}, f^2_{n,d,\alpha}\right\} + f^2_{n,d,\alpha}
\end{aligned}
$$

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, α*.*

Outline

Theorem

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

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

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

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

- *In addition,*

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

Outline

A factor model for Σ

From now on, we assume that the copula correlation matrix Σ satisfies a factor model:

$$\Sigma = \Theta + V$$

for some low-rank (or nearly low-rank), positive semidefinite, symmetric matrix Θ and diagonal matrix V.

Outline

Prelude: an elementary factor model for Σ

In an elementary factor model for $\Sigma = \Theta + V$, we assume that $\mathrm{rank}(\Theta) = r < d$, and

$$V = \sigma^2 I_d.$$

Prelude: an elementary factor model for Σ

- Let the plug-in estimator $\widehat{\Sigma}$ have the eigen-decomposition

$$\sum_{k=1}^{d} \widehat{\lambda}_k \widehat{u}_k \widehat{u}_k', \quad \widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_d.$$

- Construct the following closed-form estimators for r, σ^2 and Θ, based on some threshold μ:

$$\begin{aligned} \widehat{r} &= \sum_{k=1}^{d} \mathbb{1}\left\{\widehat{\lambda}_k \geq \widehat{\lambda}_d + \mu\right\}, \\ \widehat{\sigma}^2 &= \frac{1}{d-\widehat{r}} \sum_{j>\widehat{r}} \widehat{\lambda}_j, \\ \widehat{\Theta} &= \sum_{k=1}^{\widehat{r}} (\widehat{\lambda}_k - \widehat{\sigma}^2) \widehat{u}_k \widehat{u}_k'. \end{aligned}$$

Prelude: an elementary factor model for Σ, continued

Theorem

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

$$\widehat{r} = r,$$
$$\|\widehat{\Theta} - \Theta\|_F^2 \leq 8r\|\widehat{\Sigma} - \Sigma\|_2^2,$$
$$|\widehat{\sigma}^2 - \sigma^2| \leq \|\widehat{\Sigma} - \Sigma\|_2.$$

Here, $\|\cdot\|_F$ *denotes the Frobenius norm.*

Prelude: an elementary factor model for Σ, continued

Corollary

Set

$$\mu = 8\sqrt{\|\widehat{T}\|_2 f_{n,d,\alpha}^2 + f_{n,d,\alpha}^4/4} + 8f_{n,d,\alpha}^2$$

$$\bar{\mu} = 8\sqrt{\|T\|_2} f_{n,d,\alpha} + 12 f_{n,d,\alpha}^2.$$

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

$$\widehat{r} = r,$$

$$\|\widehat{\Theta} - \Theta\|_F^2 \leq 2r\bar{\mu}^2,$$

$$|\widehat{\sigma}^2 - \sigma^2| \leq \bar{\mu}/2.$$

hold with probability exceeding $1 - 2\alpha$.

Outline

The refined estimator $\widetilde{\Sigma}$ of Σ

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

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

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

$$\widetilde{\Theta} = \underset{\Theta' \in \mathbb{R}^{d \times d}}{\operatorname{argmin}} \left\{ \frac{1}{2} \|\Theta'_o - \widehat{\Sigma}_o\|_F^2 + \mu \|\Theta'\|_* \right\}.$$

Here,

- A_o denotes $A - \operatorname{diag}(A)$,
- $\|\cdot\|_*$ denotes the nuclear norm.

Comparison of $\widetilde{\Sigma}$ with existing estimators

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

$$(\Theta, V) = \operatorname{argmin} \operatorname{tr}(\Theta') \text{ subject to } \begin{cases} \Sigma = \Theta' + V \\ \Theta' \succsim 0 \\ V \text{ diagonal} \end{cases}$$

- If Θ has column space $\mathcal{U}$ such that its *coherence*

$$\max_{1 \leq i \leq d} \|P_{\mathcal{U}} e_i\|_2 < 1/\sqrt{2},$$

then the above algorithm recovers Θ, V exactly.

Comparison of $\widetilde{\Sigma}$ with existing estimators, continued

- For Θ with effective low-rank and without sparse corruption, Lounici 2013 proposes

$$\widetilde{\Theta} = \operatorname*{argmin}_{\Theta' \in \mathbb{R}^{d \times d}} \left\{ \frac{1}{2} \|\Theta' - \widehat{\Theta}\|_F^2 + \mu \|\Theta'\|_* \right\}$$

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

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

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

This formulation requires $\lambda > 0$.

Oracle inequality for the refined estimator $\widetilde{\Sigma}$

Theorem

Let Θ_r be best rank-r approximation to Θ, with reduced SVD $\Theta_r = U_r D_r U_r'$. Suppose $\gamma_r = \|U_r U_r'\|_{\max} \le 1/9$. Set

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

(recall μ is the regularization parameter)

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

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

$$\|\widetilde{\Sigma} - \Sigma\|_F^2 \le \sum_{j>r} \lambda_j^2(\Theta) + 18 r \bar{\mu}^2.$$

Summary

- We have obtained a sharp bound on the operator norm of $\widehat{\Sigma} - \Sigma$, for the plug-in estimator $\widehat{\Sigma}$ of the copula correlation matrix Σ.
- We have applied the above bound to the setting of a factor model of Σ to obtain a refined estimator $\widetilde{\Sigma}$; a sharp oracle inequality for $\widetilde{\Sigma}$ is established.

Thanks!