Biometrika (2018), **00**, 0, *pp.* 1–14
Printed in Great Britain
doi: 10.1093/biomet/asy011

Robust estimation of high-dimensional covariance and precision matrices

BY MARCO AVELLA-MEDINA

Sloan School of Management, Massachusetts Institute of Technology, 30 Memorial Drive, Cambridge, Massachusetts 02142, U.S.A.

mavella@mit.edu

HEATHER S. BATTEY

Department of Mathematics, Imperial College London, 545 Huxley Building, South Kensington Campus, London SW7 2AZ, U.K.

h.battey@imperial.ac.uk

JIANQING FAN

Department of Operations Research and Financial Engineering, Princeton University, 205 Sherred Hall, Princeton, New Jersey 08540, U.S.A.

jqfan@princeton.edu

AND QUEFENG LI

Department of Biostatistics, University of North Carolina at Chapel Hill, 3105D McGavran-Greenberg Hall, Chapel Hill, North Carolina 27599, U.S.A.

quefeng@email.unc.edu

SUMMARY

High-dimensional data are often most plausibly generated from distributions with complex structure and leptokurtosis in some or all components. Covariance and precision matrices provide a useful summary of such structure, yet the performance of popular matrix estimators typically hinges upon a sub-Gaussianity assumption. This paper presents robust matrix estimators whose performance is guaranteed for a much richer class of distributions. The proposed estimators, under a bounded fourth moment assumption, achieve the same minimax convergence rates as do existing methods under a sub-Gaussianity assumption. Consistency of the proposed estimators is also established under the weak assumption of bounded $2 + \varepsilon$ moments for $\varepsilon \in (0, 2)$. The associated convergence rates depend on ε.

Some key words: Constrained ℓ_1-minimization; Leptokurtosis; Minimax rate; Robustness; Thresholding.

1. INTRODUCTION

Covariance and precision matrices play a central role in summarizing linear relationships among variables. Our focus is on estimating these matrices when their dimension is large relative to the number of observations. Besides being of interest in themselves, estimates of covariance

and precision matrices are used for numerous procedures from classical multivariate analysis, including linear regression.

Consistency is achievable under structural assumptions provided regularity conditions are met. For instance, under the assumption that all rows or columns of the covariance matrix belong to a sufficiently small ℓ_q-ball around zero, thresholding (Bickel & Levina, 2008; Rothman et al., 2008) or its adaptive counterpart (Cai & Liu, 2011) gives consistent estimators of the covariance matrix in the spectral norm for data from a distribution with sub-Gaussian tails. For precision matrix estimation, the same sparsity assumption on the precision matrix motivates the use of the constrained ℓ_1-minimizer of Cai et al. (2011) or its adaptive counterpart (Cai et al., 2016), both of which are consistent in spectral norm under the same sub-Gaussianity condition. Under sub-Gaussianity, Cai & Liu (2011) and Cai et al. (2016) showed that in high-dimensional regimes the adaptive thresholding estimator and adaptive constrained ℓ_1-minimization estimator are minimax optimal within the classes of covariance or precision matrices satisfying their sparsity constraint.

Since sub-Gaussianity is often too restrictive in practice, we seek new procedures that can achieve the same minimax optimality when data are leptokurtic. Inspection of the proofs of Bickel & Levina (2008), Cai & Liu (2011) and Cai et al. (2016) reveals that sub-Gaussianity is needed because their methods are built on the sample covariance matrix, which requires the assumption to guarantee its optimal performance. Here we show that minimax optimality is achievable within a larger class of distributions if the sample covariance matrix is replaced by a robust pilot estimator, thus providing a unified theory for covariance and precision matrix estimation based on general pilot estimators. We also show how to construct pilot estimators that have the required elementwise convergence rates of (1) and (2) below. Within a much larger class of distributions with bounded fourth moment, it is shown that an estimator obtained by regularizing a robust pilot estimator attains the minimax rate achieved by existing methods under sub-Gaussianity. The analysis is extended to show that when only bounded $2+\varepsilon$ moments exist for $\varepsilon \in (0,2)$, matrix estimators with satisfactory convergence rates are still attainable.

Some related work includes that of Liu et al. (2012) and Xue & Zou (2012), who considered robust estimation of graphical models when the underlying distribution is elliptically symmetric, Fan et al. (2015, 2016a,b), who studied robust matrix estimation in the context of factor models, and Chen et al. (2015) and Loh & Tan (2015), who investigated matrix estimation when the data are contaminated by outliers. The present paper is concerned with efficient estimation of general sparse covariance and precision matrices when only certain moment conditions are assumed.

For a p-dimensional random vector X with mean μ, let $\Sigma^* = E\{(X-\mu)(X-\mu)^{\mathrm{T}}\}$ and let $\tilde{\Sigma} = (\tilde{\sigma}_{uv})$ denote an arbitrary pilot estimator of $\Sigma^* = (\sigma^*_{uv})$, where $u, v \in [p]$ with $[p]$ standing for $\{1, \ldots, p\}$. The key requirement on $\tilde{\Sigma}$ for optimal covariance estimation is that

$$\mathrm{pr}\Big[\max_{u,v} |\tilde{\sigma}_{uv} - \sigma^*_{uv}| \leqslant C_0\{(\log p)/n\}^{1/2}\Big] \geqslant 1 - \varepsilon_{n,p}, \tag{1}$$

where C_0 is a positive constant and $\varepsilon_{n,p}$ is a deterministic sequence converging to zero as $n, p \to \infty$ such that $n^{-1}\log p \to 0$. This delivers rates of convergence that match the minimax rates of Cai & Liu (2011) even under violations of their sub-Gaussianity condition, which entails the existence of $b > 0$ such that $E(\exp[t\{X_u - E(X_u)\}]) \leqslant \exp(b^2t^2/2)$ for every $t \in \mathbb{R}$ and every $u \in [p]$. Introduce the sample covariance matrix

$$\hat{\Sigma} = (\hat{\sigma}_{uv}) = n^{-1}\sum_{i=1}^{n}\big(X_i - \bar{X}\big)\big(X_i - \bar{X}\big)^{\mathrm{T}},$$

where $X_1, \ldots, X_n$ are independent and identically distributed copies of X and $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$. Proposition 1 shows that $\hat{\Sigma}$ violates (1) when X is not sub-Gaussian. In other words, the sample covariance does not concentrate exponentially fast in an elementwise sense if sub-Gaussianity is violated.

Similarly, for estimation of the precision matrix $\Omega^* = (\Sigma^*)^{-1}$, the optimality of the adaptive constrained ℓ_1-minimization estimator is retained under a pilot estimator satisfying

$$\mathrm{pr}\Big[\max_{u,v} \big|(\tilde{\Sigma}\Omega^* - I_p)_{uv}\big| \leqslant C_0\{(\log p)/n\}^{1/2}\Big] \geqslant 1 - \varepsilon_{n,p}, \tag{2}$$

where C_0 and $\varepsilon_{n,p}$ are as in (1) and I_p denotes the $p \times p$ identity matrix. While (2) holds with $\tilde{\Sigma} = \hat{\Sigma}$ under sub-Gaussianity of X, it fails otherwise.

The following proposition provides a more formal illustration of the unsuitability of $\tilde{\Sigma} = \hat{\Sigma}$ as a pilot estimator in the absence of sub-Gaussianity.

PROPOSITION 1. *Let $E(|X_u X_v - \sigma^*_{uv}|^{1+\gamma}) \leqslant 2$ for $u \neq v$ and some $\gamma > 0$. For all distributions of X satisfying this assumption, there is a distribution* pr *such that for some $\varepsilon < 1/2$,*

$$\mathrm{pr}\big\{|\hat{\sigma}_{uv} - \sigma^*_{uv}| > 2^{-1}\varepsilon^{-1/(1+\gamma)} n^{-\gamma/(1+\gamma)}\big\} \geqslant \varepsilon.$$

This implies that the choice to take the sample covariance as the pilot estimator $\tilde{\Sigma}$ results in a polynomial rate of convergence, which is slower than the exponential rate of concentration in (1). Instead, we introduce robust pilot estimators in § 4 that satisfy the conditions (1) and (2). These estimators only require $\max_{1 \leqslant u \leqslant p} E(|X_u|^4) < \infty$.

Throughout the paper, for a vector $a = (a_1, \ldots, a_p)^{\mathrm{T}} \in \mathbb{R}^p$, $\|a\|_1 = \sum_{v=1}^{p} |a_v|$, $\|a\|_2 = (\sum_{v=1}^{p} a_v^2)^{1/2}$ and $\|a\|_\infty = \max_{1 \leqslant v \leqslant p} |a_v|$. For a matrix $A = (a_{uv}) \in \mathbb{R}^{p \times q}$, $\|A\|_{\max} = \max_{1 \leqslant u \leqslant p, 1 \leqslant v \leqslant q} |a_{uv}|$ is the elementwise maximum norm, $\|A\|_2 = \sup_{\|x\|_2 \leqslant 1} \|Ax\|_2$ is the spectral norm, and $\|A\|_1 = \max_{1 \leqslant v \leqslant q} \sum_{u=1}^{p} |a_{uv}|$ is the matrix ℓ_1-norm. We let I_p denote the $p \times p$ identity matrix; $A \succ 0$ and $A \succeq 0$ mean that A is positive definite and positive semidefinite, respectively. For a square matrix A, we denote its maximum and minimum eigenvalues by $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$, respectively. We also assume that $E(X) = 0$.

2. BROADENING THE SCOPE OF THE ADAPTIVE THRESHOLDING ESTIMATOR

Let $\tau_\lambda(\cdot)$ be a general thresholding function for which:

(i) $|\tau_\lambda(z)| \leqslant |y|$ for all z and y that satisfy $|z - y| \leqslant \lambda$;
(ii) $\tau_\lambda(z) = 0$ for $|z| \leqslant \lambda$;
(iii) $|\tau_\lambda(z) - z| \leqslant \lambda$ for all $z \in \mathbb{R}$.

Similar properties are set forth in Antoniadis & Fan (2001) and were proposed in the context of covariance estimation via thresholding in Rothman et al. (2009) and Cai & Liu (2011). Some examples of thresholding functions satisfying these three conditions are the soft thresholding rule $\tau_\lambda(z) = \mathrm{sgn}(z)(z - \lambda)_+$, the adaptive lasso rule $\tau_\lambda(z) = z(1 - |\lambda/z|^\eta)_+$ with $\eta \geqslant 1$, and the smoothly clipped absolute deviation thresholding rule (Rothman et al., 2009). Although the hard thresholding rule $\tau_\lambda(z) = z\,\mathbb{1}(|z| > \lambda)$ does not satisfy (i), the results presented in this section also hold for hard thresholding. The adaptive thresholding estimator is defined as

$$\hat{\Sigma}^{\mathcal{T}} = (\hat{\sigma}^{\mathcal{T}}_{uv}) = \big\{\tau_{\lambda_{uv}}(\hat{\sigma}_{uv})\big\},$$

where $\hat{\sigma}_{uv}$ is the (u, v)th entry of $\hat{\Sigma}$ and the threshold λ_{uv} is entry-dependent. Equipped with these adaptive thresholds, Cai & Liu (2011) established optimal rates of convergence of the resulting estimator under sub-Gaussianity of X. To accommodate data drawn from distributions violating sub-Gaussianity, we replace the sample covariance matrix $\hat{\Sigma}$ by a pilot estimator $\tilde{\Sigma}$ satisfying (1). The resulting adaptive thresholding estimator is denoted by $\tilde{\Sigma}^{\mathcal{T}}$. As suggested by Fan et al. (2013), the entry-dependent threshold

$$\lambda_{uv} = \lambda \left(\frac{\tilde{\sigma}_{uu}\tilde{\sigma}_{vv} \log p}{n} \right)^{1/2} \tag{3}$$

is used, where $\lambda > 0$ is a constant. This is simpler than the threshold used by Cai & Liu (2011), as it does not require estimation of $\mathrm{var}(\tilde{\sigma}_{uv})$ and achieves the same optimality.

Let $\mathcal{S}^+(\mathbb{R}, p)$ denote the class of positive-definite symmetric matrices with elements in $\mathbb{R}$. Theorem 1 relies on the following conditions on the pilot estimator and the sparsity of Σ^*.

Condition 1. The pilot estimator $\tilde{\Sigma} = (\tilde{\sigma}_{uv})$ satisfies (1).

Condition 2. The matrix $\Sigma^* = (\sigma^*_{uv})$ belongs to the class

$$\mathcal{U}_q = \mathcal{U}_q\{s_0(p)\} = \left\{ \Sigma : \Sigma \in \mathcal{S}^+(\mathbb{R}, p),\ \max_u \sum_{v=1}^{p} (\sigma^*_{uu}\sigma^*_{vv})^{(1-q)/2} |\sigma^*_{uv}|^q \leqslant s_0(p) \right\}.$$

The class of weakly sparse matrices $\mathcal{U}_q$ was introduced by Cai & Liu (2011). The columns of a covariance matrix in $\mathcal{U}_q$ are required to lie in a weighted ℓ_q-ball, where the weights are determined by the variance of the entries of the population covariance.

THEOREM 1. *Suppose that Conditions 1 and 2 hold,* $\log p = o(n)$, *and* $\min_u \sigma^*_{uu} = \gamma > 0$. *There exists a positive constant* C_0 *such that*

$$\inf_{\Sigma^* \in \mathcal{U}_q} \mathrm{pr}\left\{ \|\tilde{\Sigma}^{\mathcal{T}} - \Sigma^*\|_2 \leqslant C_0 s_0(p) \left(\frac{\log p}{n} \right)^{(1-q)/2} \right\} \geqslant 1 - \varepsilon_{n,p},$$

where $\varepsilon_{n,p}$ *is a deterministic sequence that decreases to zero as* $n, p \to \infty$.

The constant C_0 in Theorem 1 depends on q, λ and the unknown distribution of X, and so do the constants appearing in Theorem 2 and Propositions 2–4.

Our result generalizes Theorem 1 of Cai & Liu (2011); the minimax lower bound of our Theorem 1 matches theirs, implying that our procedure is minimax optimal for a wider class of distributions containing the sub-Gaussian distributions.

Cai & Liu (2011) also give convergence rates under bounded moments in all components of X. In that case, a much more stringent scaling condition on n and p is required, as shown in Theorem 1(ii) of Cai & Liu (2011). Their result does not cover the high-dimensional case where $p > n$ when fewer than $8 + \varepsilon$ finite moments exist. If a larger number of finite moments is assumed, p is allowed to increase polynomially with n. However, we allow $\log p = o(n)$.

For the three pilot estimators to be given in § 4, even universal thresholding can achieve the same minimax optimal rate given the bounded fourth moments assumed there. However, adaptive thresholding as formulated in (3) results in better numerical performance.

Unfortunately, $\tilde{\Sigma}^{\mathcal{T}}$ may not be positive semidefinite, but it can be projected onto the cone of positive-semidefinite matrices through the convex optimization

$$\tilde{\Sigma}_{+}^{\mathcal{T}} = \arg\min_{\Sigma \succeq 0} \|\Sigma - \tilde{\Sigma}^{\mathcal{T}}\|_2. \tag{4}$$

By definition, $\|\tilde{\Sigma}_{+}^{\mathcal{T}} - \tilde{\Sigma}^{\mathcal{T}}\|_2 \leqslant \|\Sigma^* - \tilde{\Sigma}^{\mathcal{T}}\|_2$, so the triangle inequality yields

$$\|\tilde{\Sigma}_{+}^{\mathcal{T}} - \Sigma^*\|_2 \leqslant \|\tilde{\Sigma}_{+}^{\mathcal{T}} - \tilde{\Sigma}^{\mathcal{T}}\|_2 + \|\tilde{\Sigma}^{\mathcal{T}} - \Sigma^*\|_2 \leqslant 2\,\|\tilde{\Sigma}^{\mathcal{T}} - \Sigma^*\|_2.$$

Hence, the price to pay for projection is no more than a factor of two, which does not affect the convergence rate. The projection is easily made by linear programming (Boyd & Vandenberghe, 2004).

3. Broadening the scope of the adaptively constrained ℓ_1-minimization estimator

We consider a robust modification of the adaptively constrained ℓ_1-minimization estimator of Cai et al. (2016). In a spirit similar to § 2, our robust modification relies on the existence of a pilot estimator $\tilde{\Sigma}$ satisfying (2). Construction of the robust adaptive constrained ℓ_1-minimizer relies on a preliminary projection, resulting in the positive-definite estimator

$$\tilde{\Sigma}^{+} = \arg\min_{\Sigma \succeq \varepsilon I_p} \|\Sigma - \tilde{\Sigma}\|_{\max} \tag{5}$$

for an arbitrarily small positive number ε. The minimization problem in (5) can be rewritten as minimizing ϑ such that $|(\tilde{\Sigma} - \Sigma)_{uv}| \leqslant \vartheta$ and $\Sigma \succeq \varepsilon I_p$ for all $1 \leqslant u, v \leqslant p$. This problem can be solved in Matlab using the cvx solver (Grant & Boyd, 2014).

Given $\tilde{\Sigma}^{+} = (\tilde{\sigma}_{uv}^{+})$, our estimator of Ω^* is constructed by replacing $(\hat{\Sigma} + n^{-1} I_p)$ with $\tilde{\Sigma}^{+}$ in the original constrained ℓ_1-minimization procedure. For ease of reference, the steps are reproduced below. Define the first-stage estimator $\check{\Omega}^{(1)}$ of Ω^* through the vectors

$$\check{\omega}_v^{\dagger} = \arg\min_{\omega_v \in \mathbb{R}^p} \left\{ \|\omega_v\|_1 : \|\tilde{\Sigma}^{+}\omega_v - e_v\|_\infty \leqslant \delta_{n,p} \max_v(\tilde{\sigma}_{vv}^{+})\, \omega_{vv},\ \omega_{vv} > 0 \right\}, \quad v \in [p], \tag{6}$$

with $\omega_v = (\omega_{1v}, \ldots, \omega_{pv})^{\mathrm{T}}$, $\delta_{n,p} = \delta\{(\log p)/n\}^{1/2}$ for $\delta > 0$, and e_v being the vector that has value 1 in the vth coordinate and zeros elsewhere. More specifically, define $\check{\omega}_v^{(1)}$ as an adjustment of $\check{\omega}_v^{\dagger}$ such that the vth entry is

$$\check{\omega}_{vv}^{(1)} = \check{\omega}_{vv}^{\dagger}\, \mathbb{1}\left\{\tilde{\sigma}_{vv}^{+} \leqslant (n/\log p)^{1/2}\right\} + \{(\log p)/n\}^{1/2}\, \mathbb{1}\left\{\tilde{\sigma}_{vv}^{+} > (n/\log p)^{1/2}\right\}, \tag{7}$$

and define the first-stage estimator as $\check{\Omega}^{(1)} = (\check{\omega}_1^{(1)}, \ldots, \check{\omega}_p^{(1)})$. A second-stage adaptive estimator $\check{\Omega}^{(2)} = (\check{\omega}_1^{(2)}, \ldots, \check{\omega}_p^{(2)})$ is defined by solving, for each column,

$$\check{\omega}_v^{(2)} = \arg\min_{\omega_v \in \mathbb{R}^p} \left\{ \|\omega_v\|_1 : \left|\left(\tilde{\Sigma}^{+}\omega_v - e_v\right)_u\right| \leqslant \lambda_{n,p}(\tilde{\sigma}_{uu}^{+}\check{\omega}_{vv}^{(1)})^{1/2}\ (u = 1, \ldots, p) \right\}, \tag{8}$$

where $\lambda_{n,p} = \lambda\{(\log p)/n\}^{1/2}$ for $\lambda > 0$. In practice, the optimal values of δ and λ are chosen by crossvalidation. The final estimator, $\tilde{\Omega}$, of Ω^* is a symmetrized version of $\check{\Omega}^{(2)}$ constructed as

$$\tilde{\Omega} = (\tilde{\omega}_{uv}), \quad \tilde{\omega}_{uv} = \tilde{\omega}_{vu} = \check{\omega}_{uv}^{(2)} \mathbb{1}(|\check{\omega}_{uv}^{(2)}| \leqslant |\check{\omega}_{vu}^{(2)}|) + \check{\omega}_{vu}^{(2)} \mathbb{1}(|\check{\omega}_{uv}^{(2)}| > |\check{\omega}_{vu}^{(2)}|). \tag{9}$$

The theoretical properties of $\tilde{\Omega}$ are derived under Conditions 3 and 4.

Condition 3. The pilot estimator $\tilde{\Sigma} = (\tilde{\sigma}_{uv})$ satisfies (2).

Condition 4. The matrix $\Omega^* = (\omega_{uv}^*)$ belongs to the class

$$\mathcal{G}_q = \mathcal{G}_q(c_{n,p}, M_{n,p}) = \Big\{ \Omega \in \mathcal{S}^+(\mathbb{R}, p) : \max_v \sum_{u=1}^{p} |\omega_{uv}|^q \leqslant c_{n,p}, \ \|\Omega\|_1 \leqslant M_{n,p}, \\ \frac{1}{M_1} \leqslant \lambda_{\min}(\Omega) \leqslant \lambda_{\max}(\Omega) \leqslant M_1 \Big\},$$

where $0 \leqslant q \leqslant 1$, $M_1 > 0$ is a constant, and $M_{n,p}$ and $c_{n,p}$ are positive deterministic sequences that are bounded away from zero and allowed to diverge as n and p grow.

In this class of precision matrices, sparsity is imposed by restricting the columns of Ω^* to lie in an ℓ_q-ball of radius $c_{n,p}$ $(0 \leqslant q < 1)$.

THEOREM 2. *Suppose that Conditions 1, 3 and 4 are satisfied with $c_{n,p} = o(n/\log p)$. Under the scaling condition $\log p = O(n^{1/2})$ we have, for a positive constant C_0,*

$$\inf_{\Omega^* \in \mathcal{G}_q} \text{pr}\left\{ \|\tilde{\Omega} - \Omega^*\|_2 \leqslant C_0 M_{n,p}^{1-q} c_{n,p} \left(\frac{\log p}{n} \right)^{(1-q)/2} \right\} \geqslant 1 - \varepsilon_{n,p},$$

where $\varepsilon_{n,p}$ is a deterministic sequence that decreases to zero as $n, p \to \infty$ and $\tilde{\Omega}$ is the robust adaptively constrained ℓ_1-minimization estimator described in (6)–(9).

Remark 1. Our class of precision matrices is slightly more restrictive than that considered in Cai et al. (2016), since we require $1/M_1 \leqslant \lambda_{\min}(\Omega^*) \leqslant \lambda_{\max}(\Omega^*) \leqslant M_1$ instead of $\lambda_{\max}(\Omega^*)/\lambda_{\min}(\Omega^*) \leqslant M_1$. The difference is marginal since $\sigma_{uu}^* = e_u^{\mathrm{T}} \Sigma^* e_u = \|\Sigma^{*1/2} e_u\|_2^2 \leqslant \lambda_{\max}(\Sigma^*) = 1/\lambda_{\min}(\Omega^*)$ and $\lambda_{\max}(\Omega^*)/\lambda_{\min}(\Omega^*) \leqslant M_1$ implies that $0 < M_1^{-1}\lambda_{\max}(\Omega^*) \leqslant \lambda_{\min}(\Omega^*) \leqslant \lambda_{\max}(\Omega^*) \leqslant M_1 \lambda_{\min}(\Omega^*) < \infty$. We therefore only exclude precision matrices associated with either exploding or imploding covariance matrices, i.e., we exclude $\sigma_{uu}^* \to 0$ and $\sigma_{uu}^* \to \infty$ for all $u \in [p]$. Ren et al. (2015) also require $1/M_1 \leqslant \lambda_{\min}(\Omega^*) \leqslant \lambda_{\max}(\Omega^*) \leqslant M_1$.

A positive-semidefinite estimator with the same convergence rate as $\tilde{\Omega}$ can be constructed by projecting the symmetric matrix $\tilde{\Omega}$ onto the cone of positive-semidefinite matrices, as in (4).

Next, we present three pilot estimators whose performance is favourable with respect to the sample covariance matrix when the sub-Gaussianity assumption is violated. We verify Conditions 1 and 3 for these estimators. Condition 1 will be verified for all three pilot estimators. When $\|\Omega^*\|_1$ is bounded, Condition 1 implies Condition 3 because $\|\tilde{\Sigma}\Omega^* - I_p\|_{\max} = \|(\tilde{\Sigma} - \Sigma^*)\Omega^*\|_{\max} \leqslant \|\Omega^*\|_1 \|\tilde{\Sigma} - \Sigma^*\|_{\max}$. When $\|\Omega^*\|_1 \to \infty$, Condition 3 is verified for the adaptive Huber estimator. We emphasize that Condition 3 is only needed if the goal is to obtain a minimax optimal estimator of Ω. A consistent estimator is still attainable if only Condition 1 holds when $\|\Omega^*\|_1 \to \infty$. A more thorough discussion appears in the Supplementary Material.

4. Robust pilot estimators

4·1. *A rank-based estimator*

The rank-based estimator requires only the existence of the second moment. However, it makes arguably more restrictive assumptions, as it requires the distribution of X to be elliptically symmetric.

Definition 1. *A random vector $Z = (Z_1, \ldots, Z_p)^{\mathrm{T}}$ follows an elliptically symmetric distribution if and only if $Z = \mu + \xi AU$, where $\mu \in \mathbb{R}^p$, $A \in \mathbb{R}^{d \times q}$ with $q = \mathrm{rank}(A)$, U is uniformly distributed on the unit sphere in $\mathbb{R}^q$, and ξ is a positive random variable independent of U.*

Observe that $\Sigma^* = D^* R^* D^*$ where $R = (r^*_{uv})$ denotes the correlation matrix and $D^* = \mathrm{diag}\{(\sigma^*_{11})^{1/2}, \ldots, (\sigma^*_{pp})^{1/2}\}$. Liu et al. (2012) and Xue & Zou (2012) both proposed rank-based estimation of R^*, exploiting a bijective mapping between Pearson correlation and Kendall's tau or Spearman's rho dependence measures that hold for elliptical distributions. More specifically, Kendall's tau concordance between X_u and X_v is defined as

$$\tau^*_{uv} = \mathrm{pr}\big\{(X_u - Y_u)(X_v - Y_v) > 0\big\} - \mathrm{pr}\big\{(X_u - Y_u)(X_v - Y_v) < 0\big\},$$

where Y is an independent copy of X. With $X_i = (X_{i,1}, \ldots, X_{i,p})^{\mathrm{T}}$, the empirical analogue of τ^*_{uv} is

$$\tilde{\tau}_{uv} = \binom{n}{2}^{-1} \sum_i \sum_{j<i} \Big[\mathbb{1}\big\{(X_{iu} - Y_{iu})(X_{iv} - Y_{iv}) > 0\big\} - \mathbb{1}\big\{(X_{iu} - Y_{iu})(X_{iv} - Y_{iv}) < 0\big\}\Big].$$

Since $\tau^*_{uv} = 2\pi^{-1} \arcsin(r^*_{uv})$, an estimator of R^* is $\tilde{R} = (\tilde{r}_{uv}) = \{\sin(\pi \tilde{\tau}_{uv}/2)\}$. An analogous bijection exists between Spearman's rho and Pearson's correlation; see Xue & Zou (2012) for details. We propose to estimate the elements of the diagonal matrix D^* using a median absolute deviation estimator, $\tilde{D} = \mathrm{diag}\{(\tilde{\sigma}_{11})^{1/2}, \ldots, (\tilde{\sigma}_{pp})^{1/2}\}$, where $\tilde{\sigma}_{uu} = C_u \, \mathrm{med}_{i \in [n]}\{|X_{iu} - \mathrm{med}_{j \in [n]}(X_{ju})|\}$. Here, $\mathrm{med}_{i \in [n]}(\cdot)$ denotes the median within the index set $[n]$ and $C_u = F_u^{-1}(3/4)$ is the Fisher consistency constant, where F_u is the distribution function of $X_{1u} - \nu_u$ and ν_u is the median of X_{1u}. Finally, the rank-based estimator is defined as $\tilde{\Sigma}_R = (\tilde{\sigma}^R_{uv}) = \tilde{D}\tilde{R}\tilde{D}$.

Proposition 2. *Let $X_1, \ldots, X_n$ be independent and identically distributed copies of the elliptically symmetric random vector X with covariance matrix $\Sigma^* = D^* R^* D^*$. Assume that $\max_u \sigma^*_{uu} = \varsigma < \infty$ and $\min_u \sigma^*_{uu} > C_2\{(\log p)/n\}^{1/2} + 1/U^4$, where $U < C_1 \pi \sqrt{2}/(4\varsigma)$ for $C_1 > 0$ and $C_2 > 0$. Then*

$$\mathrm{pr}\Big[\max_{u,v} |\tilde{\sigma}^R_{uv} - \sigma^*_{uv}| \leqslant c\{(\log p)/n\}^{1/2}\Big] \geqslant 1 - \varepsilon_{n,p},$$

with $\varepsilon_{n,p} \leqslant C_0 p^{-L}$ for positive constants c, C_0 and L.

In estimating marginal variances, we use median absolute deviation estimators to avoid higher moment assumptions. This assumes knowledge of $F_u^{-1}(3/4)$, without which these marginal variances can be estimated by using the adaptive Huber estimator or the median of means estimator given in the next two subsections. This requires existence of a fourth moment; see Propositions 3 and 5.

4·2. *An adaptive Huber estimator*

The Huber-type M-estimator only requires the existence of fourth moments. Let $Y_{uv} = (X_u - \mu_u^*)(X_v - \mu_v^*)$. Then $\sigma_{uv}^* = E(Y_{uv}) = \mu_{uv}^* - \mu_u^*\mu_v^*$ where $\mu_u^* = E(X_u)$, $\mu_v^* = E(X_v)$ and $\mu_{uv}^* = E(X_uX_v)$. We propose to estimate σ_{uv}^* robustly through robust estimators of μ_u^*, μ_v^* and μ_{uv}^*. For independent and identically distributed copies $Z_1, \ldots, Z_n$ of a real random variable Z with mean μ, Huber's (1964) M-estimator of μ is defined as the solution to

$$\sum_{i=1}^{n} \psi_H(Z_i - \mu) = 0, \tag{10}$$

where $\psi_H(z) = \min\{H, \max(-H, z)\}$ is the Huber function. Replacing Z_i in (10) by $X_{i,u}$, $X_{i,u}$ and $X_{i,u}X_{i,v}$ gives the Huber estimators $\tilde{\mu}_u^H$, $\tilde{\mu}_v^H$ and $\tilde{\mu}_{uv}^H$ of μ_u^*, μ_v^* and μ_{uv}^*, respectively, from which the Huber-type estimator of Σ^* is defined as $\tilde{\Sigma}_H = (\tilde{\sigma}_{uv}^H) = (\tilde{\mu}_{uv}^H - \tilde{\mu}_u^H\tilde{\mu}_v^H)$.

We depart from Huber (1964) by allowing H to grow to infinity as n increases, as our objectives differ from those of Huber (1964) and of classical robust statistics (Huber & Ronchetti, 2009). There, the distribution generating the data is assumed to be a contaminated version of a given parametric model, where the contamination level is small, and the objective is to estimate features of the parametric model as if no contamination were present. Our goal is instead to estimate the mean of the underlying distribution, allowing departures from sub-Gaussianity. In related work, Fan et al. (2017) have shown that when H is allowed to diverge at an appropriate rate, the Huber estimator of the mean concentrates exponentially fast around the true mean when only a finite second moment exists. In a similar spirit, we allow H to grow with n in order to alleviate the bias. An appropriate choice of H trades off bias and robustness. We build on Fan et al. (2017) and Catoni (2012), showing that our proposed Huber-type estimator satisfies Conditions 1 and 3.

PROPOSITION 3. *Assume* $\max_{1\leqslant u\leqslant p} E(X_u^4) = \kappa^2 < \infty$. *Let* $\tilde{\Sigma}_H = (\tilde{\sigma}_{uv}^H)$ *be the Huber-type estimator with* $H = K(n/\log p)^{1/2}$ *for* $K \geqslant 4\kappa(2+L)^{1/2}$ *and* $L > 0$ *satisfying* $(2+L)(\log p)/n < 1/8$. *Under the scaling condition* $(\log p)/n \to 0$ *we have, for large n and a constant* $C > \kappa(1 + 2\max_u |\mu_u^*|)$,

$$\mathrm{pr}\Big[\max_{u,v} |\tilde{\sigma}_{uv}^H - \sigma_{uv}^*| \leqslant C\{(\log p)/n\}^{1/2}\Big] \geqslant 1 - \varepsilon_{n,p},$$

where $\varepsilon_{n,p} \leqslant C_0 p^{-L}$ *for positive constants* C_0 *and* L.

Proposition 3 verifies Condition 1 for $\tilde{\Sigma}_H$, provided H is chosen to diverge at the appropriate rate. As quantified in Proposition 4, $\tilde{\Sigma}_H$ also satisfies Condition 3 when H is of the same rate as in Proposition 3. The proof of this result entails extending a large deviation result of Petrov (1995).

PROPOSITION 4. *Assume that* $\max_{1\leqslant u\leqslant p} E(X_u^4) = \kappa^2 < \infty$ *and* $\Omega^* \in \mathcal{G}_q(c_{n,p}, M_{n,p})$ *with* $c_{n,p} = O\{n^{(1-q)/2}/(\log p)^{(3-q)/2}\}$. *Let* $\tilde{\Sigma}_H = (\tilde{\sigma}_{uv}^H)$ *be the Huber-type estimator defined below* (10) *with* $H = K(n/\log p)^{1/2}$ *for* $K \geqslant 4\kappa(2+L)^{1/2}$ *and* $L > 0$. *Assume that the truncated population covariance matrix* $\Sigma_H = E\{\mathbb{1}(|X_uX_v| \leqslant H)X_uX_v\}$ *satisfies* $\|\Sigma_H\Omega^* - I_p\|_{\max} = O[\{(\log p)/n\}^{1/2}]$. *Under the scaling condition* $(\log p)/n^{1/3} = O(1)$ *we have, for large n and a constant* $C > \kappa(1 + 2\max_u |\mu_u^*|)$,

$$\mathrm{pr}\Big[\max_{u,v} \big|(\tilde{\Sigma}_H\Omega^* - I_p)_{uv}\big| \leqslant C\{(\log p)/n\}^{1/2}\Big] \geqslant 1 - \varepsilon_{n,p},$$

where $\varepsilon_{n,p} \leqslant C_0(\log p)^{-1/2}p^{-L}$ *for positive constants* C_0 *and* L.

4·3. *A median of means estimator*

The median of means estimator was proposed by Nemirovsky & Yudin (1983) and has been further studied by Lerasle & Oliveira (2011), Bubeck et al. (2013) and Joly & Lugosi (2016). It is defined as the median of M means obtained by partitioning the data into M subsamples. A heuristic explanation for its success is that taking means within subsamples results in a more symmetric sample while the median makes the solution concentrate faster.

Our median of means estimator for Σ^* is constructed as $\tilde{\Sigma}_M = (\tilde{\sigma}_{uv}^M) = (\tilde{\mu}_{uv}^M - \tilde{\mu}_u^M \tilde{\mu}_v^M)$, where $\tilde{\mu}_{uv}^M$, $\tilde{\mu}_u^M$ and $\tilde{\mu}_v^M$ are median of means estimators of $(X_{iu}X_{iv})_{i=1}^n$, $(X_{iu})_{i=1}^n$ and $(X_{iv})_{i=1}^n$, respectively; in each case, each of the M means is computed on an regular partition $B_1, \ldots, B_M$ of $[n]$. It is assumed that M is a factor of n.

The value of M is a tuning parameter that affects the accuracy of the median of means estimator. The choice of M involves a compromise between bias and variance. For the extreme cases, $M = n$ and $M = 1$, we obtain respectively the sample median and the sample mean. The latter is asymptotically unbiased but does not concentrate exponentially fast in the presence of heavy tails, while the former concentrates exponentially fast but not to the population mean under asymmetric distributions. Proposition 5 gives the range of M for which both goals are achieved simultaneously.

PROPOSITION 5. *Assume* $\max_{1 \leqslant u \leqslant p} E(X_u^4) = \kappa^2 < \infty$. *Let* $\tilde{\Sigma}_M = (\tilde{\sigma}_{uv}^M)$ *be the median of means estimator described above based on a regular partition* $B_1, \ldots, B_M$ *with* $M = \lceil (2 + L) \log p \rceil$ *for a positive constant* L. *Under the scaling condition* $(\log p)/n \to 0$ *we have, for large* n *and a constant* $C > 2(6\mathrm{e})^{1/2}\{\kappa + 2 \max_{u,v} \mu_u^* (\sigma_{vv}^*)^{1/2}\}$,

$$\mathrm{pr}\Big[\max_{u,v} |\tilde{\sigma}_{uv}^M - \sigma_{uv}^*| \leqslant C\{(\log p)/n\}^{1/2} \Big] \geqslant 1 - \varepsilon_{n,p},$$

where $\varepsilon_{n,p} \leqslant C_0 p^{-L}$ *for positive constants* C_0 *and* L.

5. INFINITE KURTOSIS

In the previous discussion we assumed the existence of fourth moments of X for the Huber-type estimator in § 4. We now relax the condition of boundedness of $E(X_u^4)$ to that of $E(|X_u|^{2+\varepsilon})$ for some $\varepsilon > 0$ and all $u \in [p]$. The following proposition lays the foundations for the analysis of high-dimensional covariance or precision matrix estimation with infinite kurtosis. It extends Theorem 5 in Fan et al. (2017) and gives rates of convergence for Huber's estimator of $E(X_u)$ assuming a bounded $1 + \varepsilon$ moment for $\varepsilon \in (0, 1]$. The result is optimal in the sense that our rates match the minimax lower bound given in Theorem 3.1 of Devroye et al. (2016). The rates depend on ε, and when $\varepsilon = 1$ they match those of Catoni (2012) and Fan et al. (2017).

PROPOSITION 6. *Let* $\delta \in (0, 1)$, $\varepsilon \in (0, 1]$ *and* $n > 12 \log(2\delta^{-1})$, *and let* $Z_1, \ldots, Z_n$ *be independent and identically distributed random variables with mean* μ *and bounded* $1 + \varepsilon$ *moment, i.e.,* $E(|Z_1 - \mu|^{1+\varepsilon}) = v < \infty$. *Take* $H = \{vn/\log(2\delta^{-1})\}^{1/(1+\varepsilon)}$. *Then, with probability at least* $1 - \delta$,

$$\tilde{\mu}^H - \mu \leqslant \frac{7 + \sqrt{2}}{2} v^{1/(1+\varepsilon)} \left\{ \frac{\log(2\delta^{-1})}{n} \right\}^{\varepsilon/(1+\varepsilon)},$$

where $\tilde{\mu}^H$ *is as defined in* § 4·2.

COROLLARY 1. *Under the conditions of Proposition 6, the Huber estimator satisfies*

$$\mathrm{pr}\left[|\tilde{\mu}^H - \mu| \leqslant \frac{7+\sqrt{2}}{2} v^{1/(1+\varepsilon)} \left\{\frac{\log(2\delta^{-1})}{n}\right\}^{\varepsilon/(1+\varepsilon)}\right] \geqslant 1 - 2\delta.$$

Corollary 1 allows us to generalize the upper bounds of the Huber-type estimator. The following two theorems establish rates of convergence for the adaptive thresholding and the adaptively constrained ℓ_1-minimization estimators. While we do not prove that these rates are minimax optimal under $2+\varepsilon$ finite moments, the proof expands on the elementwise maximum norm convergence of the pilot estimator, which is optimal by Theorem 3.1 of Devroye et al. (2016), and the resulting rates for adaptive thresholding match the minimax rates of Cai & Liu (2011) when $\varepsilon = 2$. This is a strong indication that the rates are sharp.

THEOREM 3. *Suppose that Condition 2 is satisfied and assume* $\max_{1\leqslant u\leqslant p} E(|X_u|^{2+\varepsilon}) \leqslant \kappa_\varepsilon^2$. *Let* $\hat{\Sigma}_H^{\mathcal{T}}$ *be the adaptive thresholding estimator defined in* § 2 *based on the Huber pilot estimator* $\tilde{\Sigma}_H$ *with* $H = K(n/\log p)^{1/(2+\varepsilon)}$ *for* $K \geqslant 2^{-1}(7+\sqrt{2})\kappa_\varepsilon(2+L)^{\varepsilon/(2+\varepsilon)}$ *and* $L > 0$. *Under the scaling condition* $\log p = O(n^{1/2})$ *and choosing* $\lambda_{uv} = \lambda\{\tilde{\sigma}_{uu}^H \tilde{\sigma}_{vv}^H (\log p)/n\}^{\varepsilon/(2+\varepsilon)}$ *for some* $\lambda > 0$, *we have, for sufficiently large n,*

$$\inf_{\Sigma^* \in \mathcal{U}_q} \mathrm{pr}\left\{\|\hat{\Sigma}_H^{\mathcal{T}} - \Sigma^*\|_2 \leqslant C s_0(p) \left(\frac{\log p}{n}\right)^{\varepsilon(1-q)/(2+\varepsilon)}\right\} \geqslant 1 - \varepsilon_{n,p},$$

where $\varepsilon_{n,p} \leqslant C_0 p^{-L}$ *for positive constants* C_0 *and* L.

THEOREM 4. *Suppose that Condition 4 is satisfied,* $\max_{1\leqslant u\leqslant p} E(|X_u|^{2+\varepsilon}) \leqslant \kappa_\varepsilon^2$ *and* $c_{n,p} = O\{n^{(1-q)/2}/(\log p)^{(3-q)/2}\}$. *Let* $\hat{\Omega}_H$ *be the adaptively constrained* ℓ_1*-minimization estimator defined in* § 3 *based on the Huber pilot estimator* $\tilde{\Sigma}_H$ *with* $H = K(n/\log p)^{1/(2+\varepsilon)}$ *for* $K \geqslant 2^{-1}(7+\sqrt{2})\kappa_\varepsilon(2+L)^{\varepsilon/(2+\varepsilon)}$ *and* $L > 0$. *Assume that the truncated population covariance matrix* $\Sigma_H = E\{\mathbb{1}(|X_u X_v| \leqslant H) X_u X_v\}$ *satisfies* $\|\Sigma_H \Omega^* - I_p\|_{\max} = O[\{(\log p)/n\}^{\varepsilon/(2+\varepsilon)}]$. *Under the scaling condition* $(\log p)/n^{1/3} = O(1)$, *we have, for sufficiently large n,*

$$\inf_{\Omega^* \in \mathcal{G}_q} \mathrm{pr}\left\{\|\tilde{\Omega}_H - \Omega^*\|_2 \leqslant C M_{n,p}^{1-q} c_{n,p} \left(\frac{\log p}{n}\right)^{\varepsilon(1-q)/(2+\varepsilon)}\right\} \geqslant 1 - \varepsilon_{n,p},$$

where $\varepsilon_{n,p} \leqslant C_0 p^{-L}$ *for positive constants* C_0 *and* L.

A result similar to Proposition 6 was obtained in Lemma 2 of Bubeck et al. (2013) for the median of means estimator. Expanding on it, we obtain a result analogous to Theorem 3 for the median of means matrix estimator.

6. FINITE-SAMPLE PERFORMANCE

We illustrate the performance of the estimators discussed in § § 2 and 3 under a range of data-generating scenarios and for every choice of pilot estimator discussed in § 4. For the adaptive

Table 1. *Estimation errors (with standard errors in parentheses) of the adaptive thresholding estimator of* Σ^* *based on four different pilot estimators; values are averaged over* 500 *replications*

Distribution	Error	Sample covariance	Adaptive Huber	Median of means	Rank-based
MVN	$\|\hat{\Sigma} - \Sigma^*\|_2$	2·88 (0·04)	2·86 (0·04)	3·31 (0·05)	3·01 (0·07)
MVN	$\|\tilde{\Sigma} - \Sigma^*\|_{\max}$	0·98 (0·09)	0·92 (0·09)	1·50 (0·14)	1·61 (0·23)
T	$\|\hat{\Sigma} - \Sigma^*\|_2$	8·95 (0·53)	3·92 (0·06)	4·46 (0·24)	5·02 (0·06)
T	$\|\tilde{\Sigma} - \Sigma^*\|_{\max}$	8·72 (0·55)	1·87 (0·05)	3·35 (0·74)	2·54 (0·04)
ST	$\|\hat{\Sigma} - \Sigma^*\|_2$	7·12 (0·17)	4·88 (0·05)	4·96 (0·06)	5·16 (0·06)
ST	$\|\tilde{\Sigma} - \Sigma^*\|_{\max}$	6·89 (0·18)	2·41 (0·04)	2·43 (0·04)	2·57 (0·04)
CST	$\|\hat{\Sigma} - \Sigma^*\|_2$	5·47 (0·23)	4·14 (0·06)	4·60 (0·05)	5·13 (0·06)
CST	$\|\tilde{\Sigma} - \Sigma^*\|_{\max}$	5·07 (0·27)	2·02 (0·05)	2·27 (0·05)	2·56 (0·04)

MVN, the normal distribution; T, the t distribution; ST, the skewed t distribution; CST, the contaminated skewed t distribution.

thresholding estimator of Σ^*, we use a hard thresholding rule with the entry-dependent thresholds of (3). In each of 500 Monte Carlo replications, $n = 200$ independent copies of a random vector X of dimension $p = 400$ are drawn from a model with either a sparse covariance matrix Σ^* or a sparse precision matrix Ω^*, depending on the experiment. We consider four different scenarios for the distribution of X: the zero-mean multivariate normal distribution; the t distribution with 3·5 degrees of freedom and infinite kurtosis; the skewed t distribution with four degrees of freedom and skew parameter equal to 20; and the contaminated skewed t distribution (Azzalini, 2005) with four degrees of freedom and skew parameter equal to 10. Data in the last scenario are generated as $X = (1 - b)Z_1 + bZ_2$, where $Z_1 \sim P$, $Z_2 \sim Q$ and $b \sim \mathrm{Bi}(1, 0{\cdot}05)$; here P is the t distribution generating most of the data, while Q is a normal distribution with a mean vector of -8 and covariance matrix equal to the identity. Any unspecified tuning parameters from the adaptive thresholding estimator and adaptively constrained ℓ_1-minimization estimator are chosen by crossvalidation to minimize the spectral norm error. Unspecified constants in the tuning parameters of the robust pilot estimators are conservatively chosen to be those that would be optimal if the true distribution was a Student t distribution with five degrees of freedom. We consider the following two structures for Σ^* and Ω^*.

(i) Sparse covariance matrix: similar to Model 2 in the simulation section of Cai & Liu (2011), we take the true covariance model to be the block-diagonal matrix $\Sigma^* = \mathrm{blockdiag}(\Sigma_1^*, \Sigma_2^*)$, where $\Sigma_2^* = 4I_{p/2\times p/2}$, $\Sigma_1^* = A + \varepsilon I_{p/2\times p/2}$, $A = (a_{uv})$ with independent $a_{uv} = \mathrm{Un}(0{\cdot}3, 0{\cdot}8) \times \mathrm{Bi}(1, 0{\cdot}2)$ and $\varepsilon = \max\{-\lambda_{\min}(A), 0\} + 0{\cdot}01$ to ensure that Σ_1^* is positive definite.

(ii) Banded precision matrix: following Cai et al. (2016), we take the true precision matrix to be of the banded form $\Omega_0 = \{\omega_{ij}\}$, where $\omega_{ii} = 1$, $\omega_{i,i+1} = 0{\cdot}6$, $\omega_{i+2,i} = \omega_{i,i+2} = 0{\cdot}3$ and $\omega_{ij} = 0$ for $|i - j| \geqslant 3$.

Table 1 shows that while the sample covariance estimator performs well for the normally distributed case, when the true model departs from normality, thresholding this estimator gives poor performance, reflected by its elevated estimation error in both the maximum norm and the spectral norm. By contrast, thresholding one of our proposed robust pilot estimators does not suffer from these heavy-tailed distributions. Table 2 shows a similar pattern for the precision matrix estimators. The gains are apparent for all robust pilot estimators, as predicted by our theory.

Table 2. *Estimation errors (with standard errors in parentheses) of the adaptively constrained ℓ_1-minimizers to Ω^* based on four different pilot estimators; values are averaged over 500 replications*

Distribution	Error	Sample covariance	Adaptive Huber	Median of means	Rank-based
MVN	$\|\hat{\Omega} - \Omega^*\|_2$	2·62 (0·01)	2·61 (0·01)	2·59 (0·01)	2·59 (0·01)
MVN	$\|\tilde{\Sigma}\Omega^* - I_p\|_{\max}$	1·05 (0·09)	1·02 (0·09)	1·85 (0·28)	2·90 (0·55)
T	$\|\hat{\Omega} - \Omega^*\|_2$	2·54 (0·03)	2·26 (0·02)	2·43 (0·02)	2·41 (0·02)
T	$\|\tilde{\Sigma}\Omega^* - I_p\|_{\max}$	2·66 (3·96)	0·81 (0·03)	1·02 (0·19)	1·01 (0·19)
ST	$\|\hat{\Omega} - \Omega^*\|_2$	2·27 (0·15)	1·97 (0·05)	2·08 (0·08)	2·12 (0·08)
ST	$\|\tilde{\Sigma}\Omega^* - I_p\|_{\max}$	1·40 (1·59)	0·97 (0·02)	1·05 (0·03)	0·96 (0·02)
CST	$\|\hat{\Omega} - \Omega^*\|_2$	2·65 (0·02)	2·01 (0·04)	2·12 (0·06)	2·10 (0·06)
CST	$\|\tilde{\Sigma}\Omega^* - I_p\|_{\max}$	9·65 (3·76)	0·97 (0·04)	2·16 (2·19)	0·92 (0·03)

Table 3. *Number of connections detected by two types of methods*

	Top 100 connections				Equal tuning parameters			
	Within	Between	Total		Within	Between	Total
Huber estimator	60	40	100	Huber estimator	27	15	42
Sample covariance	55	45	100	Sample covariance	55	45	100

7. REAL-DATA EXAMPLE

A gene regulatory network, also known as a pathway, is a set of genes that interact with each other to control a specific cell function. With recent advances in genomic research, many such networks have been discovered and their functions thoroughly studied. Certain pathways are now known and available in public databases such as KEGG (Ogata et al., 2000). One popular way to infer a gene regulatory network is through estimation of the precision matrix associated with gene expression (Wit & Abbruzzo, 2015). However, such data often contain outliers. To assess whether our robust estimator can improve inference on gene regulatory networks, we use a microarray dataset and compare our findings with generally acknowledged scientific truth from the genomics literature. The microarray data come from a study by Huang et al. (2011) on the inflammation process of cardiovascular disease. They identified that the toll-like receptor signalling pathway plays a key role in the inflammation process. Their study involves $n = 48$ patients and the data are available from the Gene Expression Omnibus via the accession name GSE20129. We consider 95 genes from the toll-like receptor signalling pathway and another 62 genes from the peroxisome proliferator-activated receptor signalling pathway, which is known to be unrelated to cardiovascular disease. A good method should discover connections for genes within each of the pathways but not across them. We use both the original version of the adaptively constrained ℓ_1-minimization estimator and our robustified version via the Huber pilot estimator to estimate the precision matrix and therefore the gene regulatory network.

We first choose the tuning parameters that deliver the top 100 connections for each method. Table 3 reports the selection results, also displayed in Fig. 1. Our robust method identifies more connections within each pathway and fewer connections across the pathways.

We tried taking the same tuning parameter in the constrained ℓ_1-minimization step (8) for each procedure. Table 3 gives the results. Our estimator detects fewer connections; however, the percentage of within-pathway connections estimated using the Huber pilot estimator is much higher than that of the sample covariance estimator. If the genomics literature is correct, our

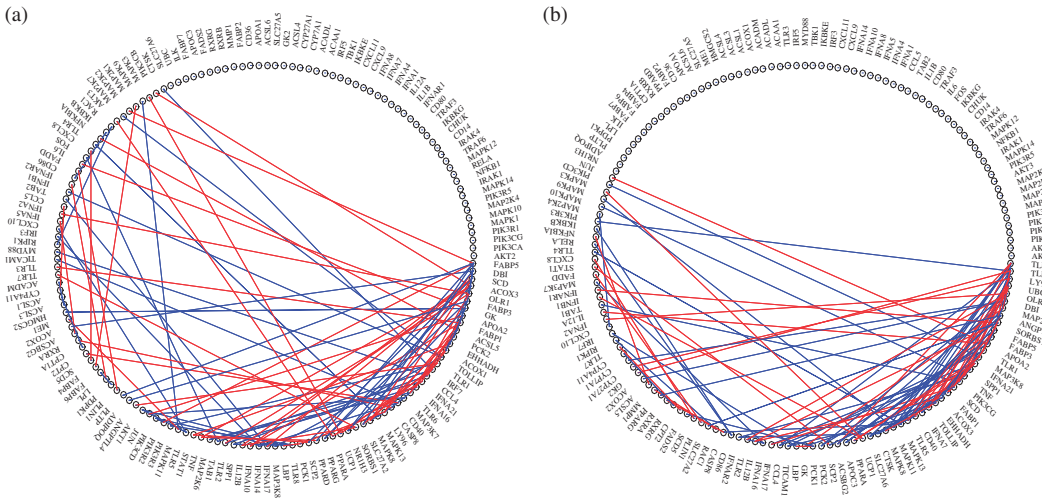

Fig. 1. Connections estimated by the adaptively constrained ℓ_1-minimization estimator using (a) the sample covariance and (b) the Huber-type pilot estimator; blue lines represent within-pathway connections and red lines between-pathway connections.

results show that use of the Huber pilot estimator improves inference for this example, in which heavy tails and skewness are present.

Acknowledgement

This work was partially supported by the U.S. National Science Foundation and National Institutes of Health. Avella-Medina was partially supported by the Swiss National Science Foundation. Battey was partially supported by the U.K. Engineering and Physical Sciences Research Council. The authors thank the editor, the associate editor and three referees for valuable comments.

Supplementary material

Supplementary Material available at *Biometrika* online includes the proofs of all propositions and theorems and additional plots for the real-data example.

References

Antoniadis, A. & Fan, J. (2001). Regularization of wavelet approximations. *J. Am. Statist. Assoc.* **96**, 939–57.
Azzalini, A. (2005). The skew-normal distribution and related multivariate families. *Scand. J. Statist.* **32**, 159–200.
Bickel, P. J. & Levina, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36**, 2577–604.
Boyd, S. & Vandenberghe, L. (2004). *Convex Optimization*. Cambridge: Cambridge University Press.
Bubeck, S., Cesa-bianchi, N. & Lugosi, G. (2013). Bandits with heavy tail. *IEEE Trans. Info. Theory* **59**, 7711–7.
Cai, T. T. & Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Am. Statist. Assoc.* **106**, 672–4.
Cai, T. T., Liu, W. & Luo, X. (2011). A constrained ℓ_1-minimization approach to sparse precision matrix estimation. *J. Am. Statist. Assoc.* **106**, 594–607.
Cai, T. T., Liu, W. & Zhou, H. (2016). Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *Ann. Statist.* **44**, 455–88.
Catoni, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Ann. Inst. Henri Poincaré Prob. Statist.* **48**, 1148–85.
Chen, M., Gao, C. & Ren, Z. (2015). Robust covariance matrix estimation via matrix depth. *arXiv:* 1506.00691.

DEVROYE, L., LERASLE, M., LUGOSI, G. & OLIVEIRA, R. I. (2016). Sub-Gaussian mean estimators. *Ann. Statist.* **44**, 2695–725.

FAN, J., HAN, F. , LIU, H. & VICKERS, B. (2016a). Robust inference of risks of large portfolios. *J. Economet.* **194**, 298–308.

FAN, J., LI, Q. & WANG, Y. (2017). Estimation of high-dimensional mean regression in absence of symmetry and light-tail assumptions. *J. R. Statist. Soc.* B **79**, 247–65.

FAN, J., LIAO, Y. & MINCHEVA, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *J. R. Statist. Soc.* B **75**, 603–80.

FAN, J., LIU, H. & WANG, W. (2015). Large covariance estimation through elliptical factor models. *arXiv:* 1507.08377.

FAN, J., WANG, W. & ZHONG, Y. (2016b). Robust covariance estimation for approximate factor models. *arXiv:* 1602.00719.

GRANT, M. & BOYD, S. (2014). CVX: Matlab software for disciplined convex programming, version 2.1.

HUANG, C.-C., LIU, K., POPE, R. M., DU, P., LIN, S., RAJAMANNAN, N. M., HUANG, Q.-Q., JAFARI, N., BURKE, G. L., POST, W. et al. (2011). Activated TLR signaling in atherosclerosis among women with lower Framingham risk score: The multi-ethnic study of atherosclerosis. *PloS One* **6**, e21067.

HUBER, P. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35**, 73–101.

HUBER, P. & RONCHETTI, E. (2009). *Robust Statistics*. Hoboken, New Jersey: Wiley, 2nd edn.

JOLY, E. & LUGOSI, G. (2016) Robust estimation of U-statistics. *Stoch. Proces. Appl.* **126**, 3760–73.

LERASLE, M. & OLIVEIRA, R. (2011). Robust empirical mean estimators. *arXiv:* 1112.3914.

LIU, H., HAN, F., YUAN, M., LAFFERTY, J. & WASSERMAN, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.* **40**, 2293–326.

LOH, P. L. & TAN, X. L. (2015). High-dimensional robust precision matrix estimation: Cellwise corruption under ε-contamination. *arXiv:* 1509.07229.

NEMIROVSKY, A. S. & YUDIN, D. B. (1983). *Problem Complexity and Method Efficiency in Optimization*. Hoboken, New Jersey: Wiley.

OGATA, H., GOTO, S., SATO, K., FUJIBUCHI, W., BONO, H. & KANEHISA, M. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30.

PETROV, V. (1995). *Limit Theorems of Probability Theory*. Oxford: Clarendon Press.

REN, Z., SUN, T., ZHANG, C.-H. & ZHOU, H. (2015). Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *Ann. Statist.* **43**, 991–1026.

ROTHMAN, A., BICKEL, P., LEVINA, E. & ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Statist.* **2**, 494–515.

ROTHMAN, A., LEVINA, E. & ZHU, J. (2009). Generalized thresholding of large covariance matrices. *J. Am. Statist. Assoc.* **104**, 177–86.

WIT, E. C. & ABBRUZZO, A. (2015). Inferring slowly-changing dynamic gene-regulatory networks. *BMC Bioinformatics* **16**, S5.

XUE, L. & ZOU, H. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Ann. Statist.* **40**, 2541–71.

[*Received on* 4 *February* 2017. *Editorial decision on* 15 *January* 2018]