Submitted to Bernoulli

Sparse covariance matrix estimation in high-dimensional deconvolution

DENIS BELOMESTNY[1,2], MATHIAS TRABS[3] and ALEXANDRE B. TSYBAKOV[4]

[1] *Duisburg-Essen University, Faculty of Mathematics*
Thea-Leymann-Str. 9 D-45127 Essen, Germany

[2] *National Research University Higher School of Economics*
Shabolovka, 26, 119049 Moscow, Russia,

E-mail: denis.belomestny@uni-due.de

[3] *Universität Hamburg, Faculty of Mathematics*
Bundesstraße 55, 20146 Hamburg, Germany

E-mail: mathias.trabs@uni-hamburg.de

[4] *CREST, ENSAE, Université Paris-Saclay*
5, avenue Henry Le Chatelier, 91120 Palaiseau, France

E-mail: alexandre.tsybakov@ensae.fr

We study the estimation of the covariance matrix Σ of a p-dimensional normal random vector based on n independent observations corrupted by additive noise. Only a general nonparametric assumption is imposed on the distribution of the noise without any sparsity constraint on its covariance matrix. In this high-dimensional semiparametric deconvolution problem, we propose spectral thresholding estimators that are adaptive to the sparsity of Σ. We establish an oracle inequality for these estimators under model miss-specification and derive non-asymptotic minimax convergence rates that are shown to be logarithmic in $\log p/n$. We also discuss the estimation of low-rank matrices based on indirect observations as well as the generalization to elliptical distributions. The finite sample performance of the threshold estimators is illustrated in a numerical example.

MSC 2010 subject classifications: Primary 62H12; secondary 62F12, 62G05.
Keywords: Thresholding, minimax convergence rates, Fourier methods, severely ill-posed inverse problem.

1. Introduction

One of the fundamental problems of multivariate data analysis is to estimate the covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$ of a random vector $X \in \mathbb{R}^p$ based on independent and identically distributed (i.i.d.) realizations $X_1, \dots, X_n$ of X. An important feature of data sets in modern applications is high dimensionality. Since it is well known that classical procedures fail if the dimension p is large, various novel methods of high-dimensional matrix estimation have been developed in the last decade. However, an important question has not yet been settled: How can Σ be estimated in a high-dimensional regime if the observations are corrupted by noise?

Let $X_1, \dots, X_n$ be i.i.d. random variables with multivariate normal distribution $\mathcal{N}(0, \Sigma)$. The maximum likelihood estimator of Σ is the sample covariance estimator

$$\Sigma_X^* := \frac{1}{n}\sum_{j=1}^{n} X_j X_j^\top.$$

The estimation error of Σ_X^* explodes for large p. To overcome this problem, sparsity assumptions can be imposed on Σ, reducing the effective number of parameters. The first rigorous studies of this idea go back to Bickel and Levina [3, 4] and El Karoui [21] who have assumed that most

entries of Σ are zero or very small. This allows for the construction of banding, tapering and thresholding estimators based on Σ_X^*, for which the dimension p can grow exponentially in n. Subsequently, a rich theory has been developed in this direction including Lam and Fan [34] who proposed a penalized pseudo-likelihood approach, Cai et al. [11] who studied minimax optimal rates, Cai and Zhou [12] studying the ℓ_1 loss as well as Rothman et al. [46] and Cai and Liu [8] for more general threshold procedures and adaptation, to mention only the papers most related to the present contribution. For current reviews on the theory of large covariance estimation, we refer to [9, 24]. Heading in a similar direction as noisy observations, covariance estimation in the presence of missing data has been recently investigated by Lounici [37] as well as Cai and Zhang [10].

Almost all estimators in the afore mentioned results build on the sample covariance estimator Σ_X^*. In this paper, we assume that only the noisy observations

$$Y_j = X_j + \varepsilon_j, \quad j = 1, \dots, n,$$

are available, where the errors $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. random vectors in $\mathbb{R}^p$ independent of $X_1, \dots, X_n$. Then the sample covariance estimator Σ_Y^* is biased:

$$\mathbb{E}[\Sigma_Y^*] = \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^n Y_i Y_i^\top\Big] = \Sigma + \Gamma$$

where $\Gamma = \mathbb{E}[\varepsilon_1 \varepsilon_1^\top]$ is the covariance matrix of the errors. Assuming Γ known to correct the bias is not very realistic. Moreover, for heavy tailed the errors ε_j that do not have finite second moments, Γ is not defined and the argument based on Σ_Y^* makes no sense. Several questions arising in this context will be addressed below:

(i) How much information on the distribution of ε_j do we need to consistently estimate Σ?
(ii) Do we need finite second moments of ε_j and/or sparsity restrictions on Γ to estimate Σ?
(iii) What is the minimax optimal rate of estimating Σ based on noisy observations?

If the covariance matrix Γ of the errors exists and is known, the problem does not differ from the direct observation case, since Γ can be simply subtracted from Σ_Y^*. If Γ can be estimated, for instance from a separate sample of the error distribution or from repeated measurements, we can proceed similarly. However, in the latter case, we need to assume that Γ is sparse, since otherwise we cannot find a good estimator for large dimensions. Reducing our knowledge about ε_j further, we may only assume that the distribution of ε_j belongs to a given nonparametric class. This leads to a high-dimensional deconvolution model. The difference from standard deconvolution problems is that the density of X_j's is a parametric object known up to a high-dimensional matrix parameter Σ. A related model in the context of stochastic processes has been recently studied by Belomestny and Trabs [2]. Obviously, we need some assumption on the distribution of errors since otherwise Σ is not identifiable as, for example, in the case of normally distributed ε_j. It turns out that we do not need a sparse covariance structure for the error distribution and we can allow for heavy tailed errors without any moments.

From the deconvolution point of view, it might seem surprising that Σ and thus the distribution of X_j can be estimated consistently without knowing or estimating the distribution of errors ε_j, but as we will show it is possible. The price to pay for this lack of information is in the convergence rates that turn out to be very slow - logarithmic in the sample size. In the pioneering works in one-dimensional case, Matias [40], Butucea and Matias [5] have constructed a variance estimator in deconvolution model with logarithmic convergence rate and a corresponding lower bound. In this paper, we provide a general multidimensional analysis of the minimax rates on the class of sparse covariance matrices.

To replace the sample covariance matrix Σ_Y^* by a deconvolution counterpart, we use some ideas from the literature on density deconvolution. Starting with Carroll and Hall [13] and Fan [22], the deconvolution problem have been extensively studied. In particular, unknown (but inferable) error distributions have been analysed by Neumann [42], Delaigle et al. [18], Johannes [31] and Delaigle

and Hall [17] among others. For adaptive estimation with unknown error distribution we refer to Comte and Lacour [14], Kappus and Mabon [32], Dattner et al. [16] and references therein. Almost all contributions to the deconvolution literature are restricted to a univariate model. Hence, our study contributes to the deconvolution theory by treating the multivariate case; in particular, our techniques for the lower bounds might be of interest. To our knowledge, only Masry [39], Eckle et al. [20], and Lepski and Willer [35, 36] have studied the setting of multivariate deconvolution. They deal with a different problem, namely that of nonparametric estimation of the density of X_j or its geometric features when the distribution of ε_j is known.

Applying a spectral approach, we construct an estimator for the covariance matrix assuming that X_j are normally distributed and that the characteristic function ψ of the distribution of ε_j decays slower than the Gaussian characteristic function. A similar idea in a one-dimensional deconvolution problem has been developed by Butucea et al. [6]. The assumption $\big|\log|\psi(u)|\big| = o(|u|^2)$ as $|u| \to \infty$ implies identifiability of Σ and allows us to construct an estimator $\hat{\Sigma}$, which is consistent in the maximal entry norm. Based on $\hat{\Sigma}$, we then construct hard and soft thresholding estimators $\hat{\Sigma}_\tau^H$ and $\hat{\Sigma}_\tau^S$, respectively, for sparse matrices. The sparsity is described by an upper bound S on the ℓ_q-norm, $q \in [0,2)$, of entries of Σ. We establish sparsity oracle inequalities for $\hat{\Sigma}_\tau^H$ and $\hat{\Sigma}_\tau^S$ when the estimation error is measured in the Frobenius norm. This choice of the norm is naturally related to the distance between two multivariate normal distributions. The oracle bounds reveal that the thresholding estimators adapt to the unknown sparsity S. For the soft thresholding estimator we present an oracle inequality, which shows that the estimator adapts also to approximate sparsity.

Assuming that the characteristic function ψ of ε_j satisfies $\big|\log|\psi(u)|\big| = \mathcal{O}(|u|^\beta)$ for large $u \in \mathbb{R}^p$ and some $\beta \in [0,2)$, we prove the following upper bound on the estimation error in the Frobenius norm:

$$\|\hat{\Sigma}_\tau^H - \Sigma\| \leqslant CS^{1/2}\Big(\log\frac{n}{\log p}\Big)^{-(1-\beta/2)(1-q/2)} \tag{1}$$

for some constant $C > 0$ and with high probability. The dependence of this bound on the sparsity S is the same as found by Bickel and Levina [3] for the case direct observations; furthermore the well-known quotient $n/\log p$ drives the rate. However, the severely ill-posed nature of the inverse problem causes the logarithmic dependence of the rate on $n/\log p$. We also see that the estimation problem is getting harder if β gets closer to 2 where it is more difficult to distinguish the signal from the noise. Furthermore, we establish a lower bound showing that the rate in (1) cannot be improved in a minimax sense for $q = 0$. Let us emphasise that our observations Y_j are by definition not normally distributed. Therefore, the proof of the lower bound differs considerably from the usual lower bounds in high-dimensional statistics, which rely on Gaussian models.

Covariance estimation is crucial in many applications where also observation errors appear. For instance, many portfolio optimization approaches rely on the covariance matrix of a possibly high number of assets where the financial data are typically perturbed due to bid-ask spreads, micro-structure noise etc. [23, 49]. While in a high-frequency regime the observation noise can be handled by local averages, in a low-frequency situation, as daily closing prices, the denoising is more difficult and our deconvolution approach can be applied, cf. [2]. Note that the dimension dependence in [2] can be improved with our analysis for low-rank matrices. As another application the spatial empirical covariance matrices of climate data and their eigenvectors, called empirical orthogonal functions, are important spatio-temporal statistics. Naturally recordings of climate data, e.g. sea surface temperatures, may suffer from measurement errors [15] and should be taken into account. Especially, sparse covariance structures appear in the problem of spatio-temporal wind speed forecasting taking into account the time series data of a target station and data of surrounding stations, see [47].

This paper is organized as follows. In Section 2 we construct and analyze the spectral covariance matrix estimator. In Section 3 the resulting thresholding procedures are defined and analyzed. In Section 4 we investigate upper and lower bounds on the estimation error. In Section 5 some extensions of our approach are discussed including the estimation of low-rank matrices based

on indirect observations as well as the generalization to elliptical distributions. The numerical performance of the procedure is illustrated in Section 6. Longer and more technical proofs are postponed to Section 7 and to the appendix.

Notation: For any $x \in \mathbb{R}^p$ and $q \in (0,\infty]$, the ℓ_q-norm of x is denoted by $|x|_q$ and we write for brevity $|x| := |x|_2$. For $x, y \in \mathbb{R}^p$ the Euclidean scalar product is written as $\langle x, y\rangle$. We denote by I_p the $p \times p$ identity matrix, and by $\mathbb{1}_{\{\cdot\}}$ the indicator function. For two matrices $A, B \in \mathbb{R}^{p\times p}$ the Frobenius scalar product is given by $\langle A, B\rangle := \mathrm{tr}(A^\top B)$ inducing the Frobeninus norm $\|A\| := \sqrt{\langle A, A\rangle}$. The nuclear norm is denoted by $\|A\|_1 := \mathrm{tr}(\sqrt{A^\top A})$ and the spectral norm by $\|A\|_\infty := \sqrt{\lambda_{max}(A^\top A)}$, where $\lambda_{max}(\cdot)$ stands for the maximal eigenvalue. For $A \in \mathbb{R}^{p\times p}$ and $q \in [0,\infty]$ we denote by $|A|_q$ the ℓ_q-norm of the entries of the matrix if $q > 0$ and the number of non-zero entries for $q = 0$. We write $A > 0$ or $A \geqslant 0$ if the matrix $A \in \mathbb{R}^{p\times p}$ is positive definite or semi-definite. We denote by $\mathbb{P}_{\Sigma,\psi}$ the joint distribution of $Y_1, \dots, Y_n$ when the covariance matrix of X_j is Σ and the characteristic function of the noise ε_j is ψ. We will write for brevity $\mathbb{P}_{\Sigma,\psi} = \mathbb{P}$ if there is no ambiguity.

2. Spectral covariance estimators

Let ψ denote the characteristic function of error distribution:

$$\psi(u) = \mathbb{E}\big[e^{i\langle u,\varepsilon_1\rangle}\big], \quad u \in \mathbb{R}^p.$$

Then the characteristic function of Y_j is given by

$$\varphi(u) := \mathbb{E}\big[e^{i\langle u,Y_j\rangle}\big] = \exp\big(-\frac{1}{2}\langle u, \Sigma u\rangle + \log\psi(u)\big), \quad u \in \mathbb{R}^p.$$

Here and throughout we assume that $\psi(u) \neq 0$ and we use the distinguished logarithm, cf. [48, Lemma 7.6]. This assumption is standard in the literature on deconvolution. Allowing for some zeros of ψ has been studied in [41, 19]. Note that our estimation procedure defined below does not rely on all u in $\mathbb{R}^d$, but uses only u with a certain radius $|u|$.

The canonical estimator for the characteristic function φ is the empirical characteristic function

$$\varphi_n(u) := \frac{1}{n}\sum_{j=1}^n e^{i\langle u,Y_j\rangle}, \quad u \in \mathbb{R}^p.$$

Since $\varphi_n(u)$ concentrates around $\varphi(u)$ with rate $\sqrt{n}$, we have $\varphi_n(u) \neq 0$ with overwhelming probability for sufficiently large frequencies u ensuring $|\varphi(u)| \geqslant C\sqrt{(\log(ep))/n}$ for some constant $C > 1$ (see Lemma 13 and Corollary 14). In this case $\log\varphi_n(u)$ is well defined. On the unlikely event $\{\varphi_n(u) = 0\}$, we may set $\log\varphi_n(u) := 0$.

Arguing similarly to Belomestny and Trabs [2], we consider the identity

$$\frac{\log\varphi_n(u)}{|u|^2} = -\frac{\langle u, \Sigma u\rangle}{|u|^2} + \frac{\log\psi(u)}{|u|^2} + \frac{\log\varphi_n(u) - \log\varphi(u)}{|u|^2}, \quad u \in \mathbb{R}^p \setminus \{0\}. \tag{2}$$

Both sides are normalized by $|u|^2$ being the order of the leading term $\langle u, \Sigma u\rangle$. While the left-hand side of (2) is a statistic based on the observations $Y_1, \dots, Y_n$, the first term on the right-hand side encodes the parameter of interest, namely the covariance matrix Σ. The second term is a deterministic error due to the unknown distribution of ε_j. If $|\log\psi(u)| = o(|u|^2)$, i.e., the error distribution is less smooth than the normal distribution, the deterministic error vanishes as $|u| \to \infty$. The third term in (2) is a stochastic error term. Using the first order approximation we get

$$\log\varphi_n(u) - \log\varphi(u) = \log\Big(\frac{\varphi_n(u) - \varphi(u)}{\varphi(u)} + 1\Big) \approx \frac{\varphi_n(u) - \varphi(u)}{\varphi(u)}. \tag{3}$$

The latter expression resembles the estimation error in classical deconvolution problems. However, there is a difference since here in the denominator we have $\varphi(u)$ rather than the characteristic function of the distribution of errors. A similar structure was detected in the statistical analysis of low-frequently observed Lévy processes by Belomestny and Reiß [1]. Following [1], one can call this type of problems *auto-deconvolution* problems. Since $|\varphi(u)| = e^{-\langle u,\Sigma u\rangle/2}|\psi(u)|$, and we assume that $|\log\psi(u)| = o(|u|^2)$, the stochastic error grows exponentially in $|u|$. Thus, the estimation problem is severely ill-posed even in one-dimensional case.

These remarks lead us to the conclusion that Σ can be estimated consistently without any particular knowledge of the error distribution as soon as $|\log\psi(u)| = o(|u|^2)$, and the spectral radius $|u|$ in (2) is chosen to achieve a trade-off between the stochastic and deterministic errors. To specify more precisely the condition $|\log\psi(u)| = o(|u|^2)$, it is convenient to consider, for any $\beta \in (0,2)$ and $T > 0$, the following nonparametric class of functions ψ:

$$\mathcal{H}_\beta(T) := \big\{\psi \text{ characteristic function on } \mathbb{R}^p : \big|\log|\psi(u)|\big| \leqslant T\big(1+|u|_\beta^\beta\big),\, u \in \mathbb{R}^p\big\}.$$

Note that $\big|\log|\psi(u)|\big| = \log(1/|\psi(u)|)$ since $|\psi(u)| \leqslant 1$. Therefore, the condition that determines the class $\mathcal{H}_\beta(T)$ can be written as the lower bound $|\psi(u)| \geqslant \exp\big(-T\big(1+|u|_\beta^\beta\big)\big)$. If the characteristic function of ε_j belongs to $\mathcal{H}_\beta(T)$, the decay $|u|_\beta^\beta$ for some $\beta < 2$ of the characteristic exponent allows for separating the normal distribution of X_j from error distribution for large $|u|$. The decay rate β determines the ill-posedness of the estimation problem. Noteworthy, we require neither sparsity restrictions on the joint distribution of $(\varepsilon_1,\dots,\varepsilon_n)$ nor moment conditions of these random variables.

A typical representative in the class $\mathcal{H}_\beta$ is a characteristic function of a vector of independent β-stable random variables. In the case of identically distributed marginals, it has the form $\psi(u) = \exp(-\sigma|u|_\beta^\beta), u \in \mathbb{R}^p$, for some parameter $\sigma > 0$. A related example with correlated coefficients is a p-dimensional stable distribution with characteristic function $\psi(u) = \exp(-\sigma|u|_2^\beta)$ (note that $|u|_2^\beta \leqslant |u|_\beta^\beta$). Recalling that stable distributions can be characterized as limit distributions of normalized sums of independent random variables and interpreting ε_j as accumulation of many small measurement errors, suggests that these examples are indeed quite natural.

If $\psi \in \mathcal{H}_\beta(T)$, the deterministic error term in (2) is small for large values of $|u|$. We will choose u in (2) in the form $Uu^{(i,j)}$ where $U > 0$ is large, and $u^{(i,j)}$ are p-dimensional unit vectors defined by

$$u^{(i,i)} := u^{(i)} := (\mathbb{1}_{\{i=k\}})_{k=1,\dots,p} \quad \text{and} \quad u^{(i,j)} := \frac{1}{\sqrt{2}}\big(u^{(i)} + u^{(j)}\big) \text{ for } i \neq j. \tag{4}$$

Using the symmetry of $\Sigma = (\sigma_{i,j})_{i,j=1,\dots,p}$, we obtain

$$\langle u^{(i)}, \Sigma u^{(i)}\rangle = \sigma_{i,i} \quad \text{and} \quad \langle u^{(i,j)}, \Sigma u^{(i,j)}\rangle = \sigma_{i,j} + \frac{\sigma_{i,i}+\sigma_{j,j}}{2}$$

for any $i,j \in \{1,\dots,p\}$ with $i \neq j$. Motivated by (2) applied to $Uu^{(i,j)}$ for some spectral radius $U > 0$, we introduce the *spectral covariance estimator*:

$$\hat\Sigma = (\hat\sigma_{i,j})_{i,j=1,\dots p} \quad \text{with} \quad \hat\sigma_{i,j} := \begin{cases} -\frac{1}{U^2}\operatorname{Re}\big(\log\varphi_n(Uu^{(i)})\big), & \text{if } i = j, \\ -\frac{1}{U^2}\operatorname{Re}\big(\log\varphi_n(Uu^{(i,j)})\big) - \frac{1}{2}(\hat\sigma_{i,i}+\hat\sigma_{j,j}), & \text{if } i \neq j. \end{cases} \tag{5}$$

Equivalently, we can write $\operatorname{Re}(\log\varphi_n(u)) = \log|\varphi_n(u)|$ for any $u \in \mathbb{R}^p$ with $|\varphi_n(u)| \neq 0$. Since $\varphi_n(u)$ concentrates around $\varphi(u)$, cf. Lemma 13, we have $\varphi_n(u) \neq 0$ with high probability if $\varphi(u) \neq 0$.

The spectral covariance estimator $\hat\Sigma$ can be viewed as a counterpart of the classical sample covariance matrix for the case of indirect observations. The entries $\hat\sigma_{i,j}$ of $\hat\Sigma$ enjoy the following concentration property.

Theorem 1. *Assume that* $|\Sigma|_\infty \leqslant R$, *and* $\psi \in \mathcal{H}_\beta(T)$ *for some* $\beta, R, T > 0$. *Let* $\gamma > \sqrt{2}$ *and* $U \geqslant 1$ *satisfy* $8\gamma\sqrt{(\log(ep))/n} < e^{-RU^2-3TU^\beta}$. *Set*

$$\tau(U) = 6\gamma \frac{e^{RU^2+3TU^\beta}}{U^2}\Big(\frac{\log(ep)}{n}\Big)^{1/2} + 3TU^{-2+\beta}. \tag{6}$$

Then, for any $\tau \geqslant \tau(U)$,

$$\mathbb{P}_{\Sigma,\psi}\big(|\hat\sigma_{i,j} - \sigma_{i,j}| < \tau\big) \geqslant 1 - 12(ep)^{-\gamma^2} \quad \text{and} \quad \mathbb{P}_{\Sigma,\psi}\Big(\max_{i,j=1,\dots,p} |\hat\sigma_{i,j} - \sigma_{i,j}| < \tau\Big) \geqslant 1 - c_* p^{2-\gamma^2}$$

where $c_* = 12e^{-\gamma^2}$.

Proof. Set $S(u) = \mathrm{Re}(\log\varphi_n(u) - \log\varphi(u))$. Using (2) we obtain, for all $i, j = 1, \dots, p$,

$$\begin{aligned}|\hat\sigma_{i,i} - \sigma_{i,j}| &\leqslant U^{-2}|S(Uu^{(i,j)})| + U^{-2}\big|\log|\psi(Uu^{(i,j)})|\big| \\ &\leqslant U^{-2}|S(Uu^{(i,j)})| + U^{-2}\max_{i\in\{1,\dots,p\}}\big|\log|\psi(Uu^{(i,j)})|\big|.\end{aligned}$$

For $U \geqslant 1$ the last summand in this display is bounded uniformly by $3TU^{-2+\beta}$ on the class $\mathcal{H}_\beta(T)$. This remark and Corollary 14 in Section 7.1 imply that

$$\mathbb{P}\Big(|\hat\sigma_{i,j} - \sigma_{i,j}| \geqslant \frac{6\gamma\sqrt{\log(ep)}}{\sqrt{n}U^2\min_{i,j\in\{1,\dots,p\}}|\varphi(Uu^{(i,j)})|} + 3TU^{-2+\beta}\Big) \leqslant 12(ep)^{-\gamma^2}$$

if the condition $\gamma\sqrt{(\log(ep))/n} < |\varphi(Uu^{(i,j)})|/8$ is satisfied for all i, j. Note that for any $i, j = 1, \dots, p$, and any $\psi \in \mathcal{H}_\beta(T)$,

$$\begin{aligned}|\varphi(Uu^{(i,j)})| &= \exp\Big(-\frac{U^2\langle u^{(i,j)}, \Sigma u^{(i,j)}\rangle}{2} + \mathrm{Re}\log\psi(Uu^{(i,j)})\Big) \\ &\geqslant \exp\big(-U^2\big(|\Sigma|_\infty + 3TU^{\beta-2}\big)\big).\end{aligned}$$

Therefore, for γ and U satisfying the conditions of the theorem,

$$\mathbb{P}\Big(|\hat\sigma_{i,j} - \sigma_{i,j}| \geqslant 6\gamma\frac{e^{RU^2+3TU^\beta}}{U^2}\Big(\frac{\log(ep)}{n}\Big)^{1/2} + 3TU^{-2+\beta}\Big) \leqslant 12(ep)^{-\gamma^2}.$$

A union bound concludes the proof. □

The first term in $\tau(U)$ is an upper bound for the stochastic error. We recover the familiar factor $\sqrt{(\log p)/n}$ which is due to a sub-Gaussian bound on the maximum of the p^2 entries $(\hat\sigma_{i,j})$. The term $\exp(RU^2 + 3TU^\beta)$ is an upper bound for $\varphi(u)^{-1}$ appearing in the linearization (3). Note that for $\beta < 2$ this bound can be written as $\exp(RU^2(1 + o(1)))$ for $U \to \infty$. This suggests the choice of spectral radius in the form $U_* = c\sqrt{\log(n/\log(ep))}$ for some sufficiently small constant $c > 0$. The second term in (6) bounds the deterministic error and determines the resulting rate $U_*^{-2+\beta} = \mathcal{O}((\log(n/\log(ep)))^{-1+\beta/2})$, cf. Theorem 5.

3. Thresholding

Based on the spectral covariance estimator, we can now propose estimators of high-dimensional sparse covariance matrices. We consider the following sparsity classes of matrices:

$$\begin{aligned}\mathcal{G}_0(S, R) &:= \Big\{\Sigma > 0 : \Sigma = \Sigma^\top, |\Sigma|_0 \leqslant S,\ |\Sigma|_\infty \leqslant R\Big\} \quad \text{and} \\ \mathcal{G}_q(S, R) &:= \Big\{\Sigma > 0 : \Sigma = \Sigma^\top, |\Sigma|_q^q \leqslant S,\ |\Sigma|_\infty \leqslant R\Big\} \quad \text{for } q \in (0, 2),\end{aligned} \tag{7}$$

where $S > 0$ denotes the sparsity parameter and $R > 0$ bounds the largest entry of Σ. We also consider larger classes $\mathcal{G}_q^*(S,R)$ that differ from $\mathcal{G}_q(S,R)$ only in that the condition $\Sigma > 0$ is dropped. Note that $S \geqslant p$ for the classes $\mathcal{G}_q(S,R)$, since otherwise the condition $\Sigma > 0$ does not hold. This restriction on S does not apply to the classes $\mathcal{G}_q^*(S,R)$, for which the unknown effective dimension of Σ can be smaller than p. However, for the classes $\mathcal{G}_q^*(S,R)$, the overall model remains, in general, p-dimensional since the distribution of the noise can be supported on the whole space $\mathbb{R}^p$.

The sparsity classes considered by Bickel and Levina [3] and in many subsequent papers are given by

$$\mathcal{U}_q(s,R) := \Big\{\Sigma > 0 : \Sigma = \Sigma^\top, \max_i \sum_{j=1}^p |\sigma_{i,j}|^q \leqslant s, \max_i \sigma_{i,i} \leqslant R\Big\}$$

for $s, R > 0$, $q \in (0,1)$ and with the usual modification for $q = 0$. We have $\mathcal{U}_q(s,R) \subseteq \mathcal{G}_q(sp,R)$, so that our results can be used to obtain upper bounds on the risk for the classes $\mathcal{U}_q(s,R)$.

Based on the spectral covariance estimator, we define the *spectral hard thresholding estimator* for Σ as

$$\hat\Sigma_\tau^H := (\hat\sigma_{i,j}^H)_{i,j=1,\dots,p} \quad \text{with} \quad \hat\sigma_{i,j}^H := \hat\sigma_{i,j}\mathbb{1}_{\{|\hat\sigma_{i,j}|>\tau\}}, \tag{8}$$

for some threshold value $\tau > 0$. The following theorem gives an upper bound on the risk of this estimator in the Frobenius norm.

Theorem 2. *Let $R, T, S > 0$, $\beta \in [0,2)$, and $q \in [0,2)$. Let $\tau(U)$ be defined in* (6) *with parameters $\gamma > \sqrt{2}$ and $U \geqslant 1$ satisfying $8\gamma\sqrt{(\log(ep))/n} \leqslant e^{-RU^2-3TU^\beta}$. Then*

$$\sup_{\Sigma\in\mathcal{G}_q^*(S,R),\psi\in\mathcal{H}_\beta(T)} \mathbb{P}_{\Sigma,\psi}(\|\hat\Sigma_\tau^H - \Sigma\| \geqslant 3S^{1/2}\tau^{1-q/2}) \leqslant c_* p^{2-\gamma^2}$$

provided that $\tau \geqslant \tau(U)$ for $q = 0$, and $\tau \geqslant 2\tau(U)$ for $q \in (0,2)$. Here, $c_* = 12e^{-\gamma^2}$.

Proof. First, consider the case $q = 0$ and $\tau \geqslant \tau(U)$. In view of Theorem 1, the event $\mathcal{A} = \big\{\max_{i,j=1,\dots,p}|\hat\sigma_{i,j} - \sigma_{i,j}| < \tau\big\}$ is of probability at least $1 - c_* p^{2-\gamma^2}$ for all $\tau \geqslant \tau(U)$. On $\mathcal{A}$ we have the inclusion $\big\{j : |\hat\sigma_{i,j}| > \tau\big\} \subseteq \big\{j : \sigma_{i,j} \neq 0\big\}$, so that $|\hat\Sigma^H|_0 \leqslant |\Sigma|_0$. Therefore, on the event $\mathcal{A}$,

$$\|\hat\Sigma_\tau^H - \Sigma\|^2 \leqslant |\hat\Sigma_\tau^H - \Sigma|_0\,|\hat\Sigma_\tau^H - \Sigma|_\infty^2 \leqslant 2|\Sigma|_0\,|\hat\Sigma_\tau^H - \Sigma|_\infty^2 \leqslant 2S|\hat\Sigma_\tau^H - \Sigma|_\infty^2.$$

Note that, again on $\mathcal{A}$, we have $|\hat\Sigma_\tau^H - \Sigma|_\infty \leqslant |\hat\Sigma_\tau^H - \hat\Sigma|_\infty + |\hat\Sigma - \Sigma|_\infty \leqslant 2\tau$. Combining this with the last display implies the assertion of the theorem for $q = 0$.

Consider now the case $q \in (0,2)$ and $\tau \geqslant 2\tau(U)$. We use the following elementary fact: If $|y - \vartheta| \leqslant r$ for some $y, \vartheta \in \mathbb{R}$ and $r > 0$, then $|y\mathbb{1}_{\{|y|>2r\}} - \vartheta| \leqslant 3\min\{|\vartheta|, r\}$ (cf.[51]). Taking $y = \hat\sigma_{i,j}$, $\vartheta = \sigma_{i,j}$, and $r = \tau/2$, and using Theorem 1 we obtain that, on the event of probability at least $1 - c_* p^{2-\gamma^2}$,

$$|\hat\sigma_{i,j}^H - \sigma_{i,j}| \leqslant 3\min\{|\sigma_{i,j}|, \tau/2\}, \quad i,j = 1,\dots,p.$$

Thus, for any $q \in (0,2)$, with probability at least $1 - c_* p^{2-\gamma^2}$,

$$\|\hat\Sigma_\tau^H - \Sigma\|^2 = \sum_{i,j}(\hat\sigma_{i,j}^H - \sigma_{i,j})^2 \leqslant 9\sum_{i,j}\min\{\sigma_{i,j}^2, \tau^2/4\} \leqslant 9(\tau/2)^{2-q}|\Sigma|_q^q \leqslant 9\tau^{2-q}S.$$

Since all bounds hold uniform in $\Sigma \in \mathcal{G}_q^*(S,R)$ and $\psi \in \mathcal{H}_\beta(T)$, the theorem is proven. □

In the direct observation case where $\varepsilon_j = 0$ we have $\psi(u) = 1$ for all $u \in \mathbb{R}^p$, so that the deterministic error term in (6) disappears. In this case, U can be fixed and the threshold can be

chosen as a multiple of $\sqrt{(\log p)/n}$, analogously to [3]. Together with the embedding $\mathcal{U}_q(s,R) \subseteq \mathcal{G}_q(sp,R)$, we recover Theorem 2 from Bickel and Levina [3]. In Section 4 we will discuss in detail the optimal choice of the spectral radius and the threshold in the presence of noise.

The *spectral soft thresholding estimator* is defined as

$$\hat{\Sigma}_\tau^S := (\hat{\sigma}_{i,j}^S)_{i,j=1,\dots,p} \quad \text{with} \quad \hat{\sigma}_{i,j}^S := \operatorname{sign}(\hat{\sigma}_{i,j})\big(|\hat{\sigma}_{i,j}| - \tau\big)_+$$

with some threshold $\tau > 0$. It is well known, cf., e.g. [51], that

$$\hat{\Sigma}_\tau^S = \operatorname*{arg\,min}_{A\in\mathbb{R}^{p\times p}} \big\{|A - \hat{\Sigma}|_2^2 + 2\tau|A|_1\big\}. \tag{9}$$

Adapting the proof of Theorem 2 in Rigollet and Tsybakov [44], we obtain the following oracle inequality, which is sharp for $q = 0$ and looses a factor 2 otherwise.

Theorem 3. *Assume that $|\Sigma|_\infty \leqslant R$, and $\psi \in \mathcal{H}_\beta(T)$ for some $\beta, R, T > 0$. Let $\tau \geqslant \tau(U)$ where $\tau(U)$ is defined in (6) with parameters $\gamma > \sqrt{2}$ and $U \geqslant 1$ such that $8\gamma\sqrt{(\log(ep))/n} \leqslant e^{-RU^2-3TU^\beta}$. Then,*

$$\|\hat{\Sigma}_\tau^S - \Sigma\|^2 \leqslant \min_{A\in\mathbb{R}^{p\times p}} \Big\{\|A-\Sigma\|^2 + (1+\sqrt{2})^2\tau^2|A|_0\Big\} \tag{10}$$

with probability at least $1 - c_* p^{2-\gamma^2}$ where $c_* = 12e^{-\gamma^2}$. For any $q \in (0,2)$ we have, with probability at least $1 - c_* p^{2-\gamma^2}$,

$$\|\hat{\Sigma}_\tau^S - \Sigma\|^2 \leqslant \min_{A\in\mathbb{R}^{p\times p}} \Big\{2\|A-\Sigma\|^2 + c(q)\tau^{2-q}|A|_q^q\Big\} \tag{11}$$

where $c(q) > 0$ is a constant depending only on q.

Proof. Starting from the characterization (9), we use Theorem 2 by Koltchinskii et al. [33]. To this end, we write $\hat{\sigma}_{i,j} = \sigma_{i,j} + \xi_{i,j},\ i,j \in \{1,\dots,p\}$, where $\xi_{i,j}$ are random variables with exponential concentration around zero due to Theorem 1. Observing $\hat{\sigma}_{i,j}$ is thus a sequence space model in dimension p^2 and a special case of the trace regression model $Y_j = \operatorname{tr}(Z_{i,j}^\top A_0) + \xi_{i,j}$ considered in [33]. Namely, A_0 is the diagonal matrix with diagonal entries $\sigma_{i,j}$ and $Z_{i,j}$ are diagonalisations of the canonical basis in $\mathbb{R}^{p\times p}$. In particular, Assumption 1 in [33] is satisfied for $\mu = p$, i.e., $\|B\|_{L_2(\Pi)}^2 = p^{-2}|B|_2^2$ where we use the notation of [33]. Note also that the rank of a diagonal matrix B is equal to the number of its non-zero elements. Consequently, Theorem 2 in [33] yields with $\lambda = \frac{2\tau}{p^2}$ that

$$|\hat{\Sigma}_\tau^S - \Sigma|_2^2 \leqslant \min_{A\in\mathbb{R}^{p\times p}} \Big\{|A-\Sigma|_2^2 + (1+\sqrt{2})^2\tau^2|A|_0\Big\}$$

on the event that $\mathcal{A} = \{\max_{i,j}|\hat{\sigma}_{i,j} - \sigma_{i,j}| < \tau\}$. To estimate the probability of $\mathcal{A}$, we apply Theorem 1. Inequality (11) follows from (10) using the same argument as in Corollary 2 of [44]. □

This theorem shows that the soft thresholding estimator allows for estimating matrices that are a not exactly sparse but can be well approximated by a sparse matrix. Choosing $A = \Sigma$ in the oracle inequalities (10) and (11) we obtain the following corollary analogous to Theorem 2.

Corollary 4. *Let $R, T, S > 0$, $\beta \in (0,2)$, and $q \in [0,2)$. Let $\tau \geqslant \tau(U)$ where $\tau(U)$ is defined in (6) with parameters $\gamma > \sqrt{2}$ and $U \geqslant 1$ such that $8\gamma\sqrt{(\log(ep))/n} \leqslant e^{-RU^2-3TU^\beta}$. Then*

$$\sup_{\Sigma\in\mathcal{G}_q^*(S,R),\psi\in\mathcal{H}_\beta(T)} \mathbb{P}_{\Sigma,\psi}(\|\hat{\Sigma}_\tau^S - \Sigma\| \geqslant CS^{1/2}\tau^{1-q/2}) \leqslant c_* p^{2-\gamma^2}$$

where $C = 1+\sqrt{2}$ for $q = 0$, and $C = \sqrt{c(q)}$ for $q \in (0,2)$.

4. Minimax optimality

In this section, we study minimax optimal rates for the estimation of Σ on the class $\mathcal{G}_q(S,R)\times\mathcal{H}_\beta(T)$. We first state an upper bound on the rate of convergence of the hard thresholding estimator in this high-dimensional semiparametric problem. It is an immediate consequence of Theorem 2. Due to Corollary 4, the result directly carries over to the soft thresholding estimator.

Theorem 5. *Let $R,T,S>0$, $\beta\in(0,2)$, and $q\in[0,2)$. For $\gamma>\sqrt{2}$, set*

$$U_*=\sqrt{\frac{1}{4R}\log\frac{n}{64\gamma^2\log(ep)}}. \tag{12}$$

Let n be large enough such that $U_*\geqslant(\frac{3T}{R})^{1/(2-\beta)}\vee(\bar c/T)^{1/\beta}\vee 1$ for some numerical constant $\bar c>0$. Then for any $\tau\geqslant\tau(U_*)$ where $\tau(\cdot)$ is defined in (6) we have

$$\sup_{(\Sigma,\psi)\in\mathcal{G}_q(S,R)\times\mathcal{H}_\beta(T)}\mathbb{P}_{\Sigma,\psi}\Big(\|\hat\Sigma_\tau^H-\Sigma\|\geqslant\bar C_1\bar r_{n,p}\Big)\leqslant\bar C_0p^{2-\gamma^2}\qquad\text{with}$$
$$\bar r_{n,p}:=S^{1/2}\Big(R^{1-\beta/2}T\Big(\log\frac{n}{\log(ep)}\Big)^{-1+\beta/2}\Big)^{1-q/2} \tag{13}$$

for some numerical constants $\bar C_0,\bar C_1>0$.

Proof. It follows from the assumption on U_* that $3TU_*^\beta\leqslant RU_*^2$. This and the definition of U_* imply that $8\gamma\sqrt{(\log(ep))/n}\leqslant e^{-RU_*^2-3TU_*^\beta}$. Therefore, we can apply Theorem 2, which yields the result since

$$\tau(U_*)\leqslant 6\gamma\frac{e^{2RU_*^2}}{U_*^2}\Big(\frac{\log(ep)}{n}\Big)^{1/2}+3TU_*^{-2+\beta}\leqslant\Big(\frac{2\bar c}{3}+3\Big)TU_*^{-2+\beta}.$$

□

It is interesting to compare Theorem 5 with the result of Butucea and Matias [5] corresponding to $p=1,S=1$, and establishing a logarithmic rate for estimation of the variance in deconvolution model under exponential decay of the Fourier transform of ε_j. Butucea and Matias [5] have shown that, if $\log|\psi(u)|=\mathcal{O}(|u|^\beta)$, their estimator achieves asymptotically a mean squared error of the order $(\log n)^{-1+\beta/2}$. This coincides with the case $p=1$ and $q=0$ of the non-asymptotic bound in (13). A similar rate for $p=1$ has been obtained by Matias [40] under the assumptions on the decay of the Laplace transform.

We now turn to the lower bound matching (13) for $q=0$. Intuitively, the slow rate comes from the fact that the error distribution can mimic the Gaussian distribution up to some frequency in the Fourier domain. A rigorous application of this observation to the construction of lower bounds goes back to Jacod and Reiß [30], though in quite a different setting. For the multidimensional case that we consider here the issue becomes particularly challenging.

Theorem 6. *Let $\beta\in(0,2)$ and assume that $C_1p\leqslant S\leqslant C_2p$, $T(\log n)^{-1+\beta/2}\leqslant C_3R^{\beta/2}$, $T(\log n)^{c_*}\geqslant 1\vee R^{\beta/2}$ for some constants $C_1,C_2,C_3>0$, and $c_*>0$. Then, there are constants $c_1,c_2>0$ such that*

$$\inf_{\tilde\Sigma}\sup_{(\Sigma,\psi)\in\mathcal{G}_0(S,R)\times\mathcal{H}_\beta(T)}\mathbb{P}_{\Sigma,\psi}\big(\|\tilde\Sigma-\Sigma\|\geqslant c_1\underline{r}_{n,p}\Big)>c_2\quad\text{with}$$
$$\underline{r}_{n,p}:=S^{1/2}R^{1-\beta/2}T\big(\log n\big)^{-1+\beta/2}$$

where the infimum is taken over all estimators $\tilde\Sigma$.

The proof of this theorem is postponed to Section 8. We use the method of reduction to testing of many hypotheses relying on a control of the χ^2-divergence between the corresponding distributions, cf. Theorem 2.6 in [50]. The present high-dimensional setting introduces some additional difficulties. When the dimension p of the sample space is growing, an increasing number of derivatives of the characteristic functions has to be taken into account for the χ^2-bound. Achieving bounds of the correct order in p causes difficulty when p is arbitrarily large. We have circumvented this problem by introducing a block structure to define the hypotheses. The construction of the family of covariance matrices of X_j used in the lower bounds relies on ideas from Rigollet and Tsybakov [44], while the error distributions are chosen as perturbed β-stable distributions. To bound the χ^2-divergence, we need a lower bound on the probability density of Y_j. It is shown by Butucea and Tsybakov [7] that the tails of the density of a one-dimensional stable distribution are polynomially decreasing. We generalize this result to the multivariate case (cf. Lemma 15 below) using properties of infinitely divisible distributions.

We now give some comments on the lower bound of Theorem 6. Assuming S of order p means that we consider quite a sparse regime. We always have $S \leqslant p^2$. Recall also that $S \geqslant p$ as the diagonal of the covariance matrix is included in the definition of S for the class $\mathcal{G}_0(S,R)$. An alternative strategy pursued in the literature is to estimate a correlation matrix, i.e., to assume that all diagonal entries are known and equal to one. However, this seems not very natural in the present noisy observation scheme. On the other hand, Theorem 6 shows that even in the sparse regime $S = \mathcal{O}(p)$ the estimation error tends to ∞ as $n \to \infty$ for dimensions p growing polynomially in n. The logarithmic in n rate reflects the fact that the present semiparametric problem is severely ill-posed.

Comparing the lower bound $\underline{r}_{n,p}$ with the upper bound $\bar{r}_{n,p}$ from Theorem 5, we see that they coincide if the dimension satisfies $p = \mathcal{O}(\exp(cn^\gamma))$ for some $\gamma \in [0,1)$ and some $c > 0$. Thus, we have established the minimax optimal rate under this condition. Note also that we only loose a factor of order $\log\log p$ for very large p, for instance, if $p = e^{n/\log n}$.

5. Discussion and extensions

5.1. The adaptivity issue

Since the threshold $\tau(U_*)$ in Theorem 5 depends on unknown parameters R, T, and β, a natural question is whether it is possible to construct an adaptive procedure independent of these parameters that achieves the same rate. One possibility to explore consists in selecting τ in a data-driven way. Another option would be to construct estimators corresponding to values of R, T, and β on a grid, and then to aggregate them.

For direct observations an adaptive choice of the threshold, more precisely a cross-validation criterion, has been proposed by Bickel and Levina [3] and was further investigated by Cai and Liu [8]. For noisy observations that we consider here, the adaptation problem turns out to be more delicate since not only an optimal constant has to be selected but also the order of magnitude of $\tau(U)$ depends on the unknown parameter β.

Often an upper bound R on the maximal entry of Σ is known, so that one does not need considering adaptation to R. Ignoring the issue of unknown R, the choice of the spectral radius U_* of the order $\sqrt{R^{-1}\log(n/\log(ep))}$ is universal, which reflects the fact that the estimation problem is severely ill-posed with dominating bias. Indeed, U_* in Theorem 5 corresponds to undersmoothing such that the deterministic estimation error dominates the stochastic error without deteriorating the convergence rates. To construct an adaptive counterpart of τ, we need either an estimator of the error of an optimal procedure for estimating Σ under the $|\cdot|_\infty$-loss or an estimator of the "regularity" β. Therefore, extrapolating the argument of Low [38] to our setting, it seems plausible that an adaptive choice of τ cannot, in general, lead to the optimal rate. This does not exclude that optimal adaptive estimators can be constructed by other type of procedure, such as aggregation of estimators on the grid as mentioned above.

5.2. Low-rank covariance matrix

Alternatively to the above setting where the covariance matrix Σ is sparse, we can consider a low-rank matrix Σ. This is of particular interest in the context of factor models where, as discussed by Fan et al. [25, 26], an additional observation error should be taken into account. While [25, 26] estimate the covariance matrix of the noisy observations assuming that the errors have a sparse covariance structure, a spectral approach analogous to the one developed above allows for estimating directly the low-rank covariance matrix of X without sparsity restrictions on the error distribution.

Such an approach, which is at first sight quite natural, would be to use the spectral covariance estimator $\hat{\Sigma}$ from (5) together with a nuclear norm penalization. The following oracle inequality is an easy consequence of Theorem 1 in Koltchinskii et al. [33].

Proposition 7. *Assume that $\mathbb{M} \subseteq \mathbb{R}^{p\times p}$ is convex and let $\tau > 0$. On the event $\{2\|\hat{\Sigma}-\Sigma\|_\infty \leqslant \tau\}$, the estimator $\hat{\Sigma}^R_\tau := \arg\min_{S\in\mathbb{M}}\{\|S-\hat{\Sigma}\|^2 + \tau\|S\|_1\}$ satisfies*

$$\|\hat{\Sigma}^R_\tau - \Sigma\|^2 \leqslant \inf_{S\in\mathbb{M}} \Big\{\|S-\Sigma\|^2 + \big(\frac{1+\sqrt{2}}{2}\big)^2\tau^2\,\mathrm{rank}(S)\Big\}.$$

To use this proposition, we need to find a bound on the spectral norm $\|\hat{\Sigma}-\Sigma\|_\infty$ that hold with high probability. The techniques from Cai et al. [11] designed for the case of direct observations allow us to obtain an upper bound on this quantity of order p up to a logarithmic in $n/\log(p)$ factor. Thus, the convergence rate of this estimator is rather slow.

Let us show now that another estimator can be constructed based the approach from Belomestny and Trabs [2], which allows for a better dependence on p. To this end, we write

$$-\frac{\langle u, \Sigma u\rangle}{|u|^2} = \langle \Theta(u), \Sigma\rangle \quad \text{with design matrix } \Theta(u) := -\frac{uu^\top}{|u|^2}, \quad u \in \mathbb{R}^p \setminus \{0\}.$$

For a weight function $w\colon \mathbb{R}^p \to \mathbb{R}_+$ supported on the annulus $\{u \in \mathbb{R}^p : \frac{1}{4} \leqslant |u| \leqslant \frac{1}{2}\}$ and a spectral radius $U \geqslant 1$, we set $w_U(u) := U^{-p}w(u/U), u \in \mathbb{R}^p$. Motivated by (2), we define the weighted Lasso-type estimator

$$\tilde{\Sigma}_\lambda := \operatorname*{arg\,min}_{M\in\mathbb{M}} \Big\{ \int_{\mathbb{R}^p} \Big(\frac{\mathrm{Re}\log\varphi_n(u)\mathbb{1}_{\{|\varphi_n(u)|\geqslant\iota\}}}{|u|^2} - \langle\Theta(u), M\rangle\Big)^2 w_U(u)\mathrm{d}u + \lambda\|M\|_1\Big\} \tag{14}$$

for a convex set $\mathbb{M} \subseteq \{M \in \mathbb{R}^{p\times p} : M \geqslant 0\}$ and with nuclear norm penalisation for some $\lambda > 0$. We have inserted a truncation function $\mathbb{1}_{\{|\varphi_n(u)|\geqslant\iota\}}$ for some threshold $\iota > 0$ which increases the stability of the estimator by cutting off frequencies with too small point estimates $\varphi_n(u)$. Under the universal choice $\iota = 1/(2\sqrt{n})$ this indicator function will be one with high probability. The estimator $\tilde{\Sigma}_\lambda$ is associated to the weighted scalar product which replaces the classical empirical scalar product:

$$\langle A, B\rangle_U := \int_{\mathbb{R}^d} \langle\Theta(u), A\rangle\langle\Theta(u), B\rangle w_U(u)\mathrm{d}u \quad \text{and} \quad \|A\|^2_U := \langle A, A\rangle_U,$$

for matrices $A, B \in \mathbb{R}^{p\times p}$. As in [2, Lemma 3.2] we have for any for any positive semi-definite matrix $A \in \mathbb{R}^{p\times p}$ an isometry with respect to the Frobenius norm

$$\underline{\varkappa}_w\|A\|^2 \leqslant \|A\|^2_U \leqslant \overline{\varkappa}_w\|A\|^2 \quad \text{with} \quad \underline{\varkappa}_w := \int_{\mathbb{R}^p} \frac{|v_1|^4}{|v|^4} w(v)\mathrm{d}v, \quad \overline{\varkappa}_w := \|w\|_{L^1}.$$

Adapting slightly the proof of Theorem 1 in [33], we obtain the following oracle inequality.

Theorem 8. *Let $\mathbb{M}$ be convex. Define*

$$\mathcal{R}_n := \int_{\mathbb{R}^p} \Big(\frac{\operatorname{Re}\log\varphi_n(u)\mathbb{1}_{\{|\varphi_n(u)|\geqslant\iota\}}}{|u|^2} - \langle\Theta(u),\Sigma\rangle\Big)\Theta(u)w_U(u)\mathrm{d}u.$$

The estimator $\tilde\Sigma_\lambda$ from (14) satisfies on the event $\{\|\mathcal{R}_n\|_\infty \leqslant \lambda\}$

$$\|\tilde\Sigma_\lambda - \Sigma\|_U^2 \leqslant \inf_{M\in\mathbb{M}} \big\{\|M-\Sigma\|_U^2 + C_*^2\lambda^2\operatorname{rank}(M)\big\}$$

for the constant $C_* = (1+\sqrt{2})/(2\underline{\varkappa}_w)$ depending only w.

We omit the proof of this theorem as it is analogous to Theorem 3.4 in [2]. In combination with the isometry property we obtain an oracle inequality with respect to the Frobenius norm:

$$\|\tilde\Sigma_\lambda - \Sigma\|^2 \leqslant \inf_{M\in\mathbb{M}} \big\{C_1^*\|M-\Sigma\|^2 + C_2^*\lambda^2\operatorname{rank}(M)\big\}$$

with $C_1^* = \overline{\varkappa}_w/\underline{\varkappa}_w$ and $C_2^* = (1+\sqrt{2})^2/(4\underline{\varkappa}_w^3)$. The best leading constant in this oracle inequality can be obtained by minimizing C_1^* with respect to w. We do not detail it here.

To apply Theorem 8, we need a sharp probabilistic bound for $\|\mathcal{R}_n\|_\infty$. At first sight, this might look similar to bounding $\|\hat\Sigma - \Sigma\|_\infty$ in Proposition 7. However, the dependence on the dimension is much better because the design matrix satisfies $\|\Theta(u)\|_\infty = 1$.

Consider the error distributions in the subclass of $\mathcal{H}_\beta(T)$ defined as follows:

$$\mathcal{H}'_\beta(T) := \Big\{\psi \text{ characteristic function} : \big|\log|\psi(u)|\big| \leqslant T\big(1+|u|_2^\beta\big),\ u\in\mathbb{R}^p\Big\} \subseteq \mathcal{H}_\beta(T).$$

Theorem 9. *Let $T>0$, $\beta\in[0,2)$ and $\psi\in\mathcal{H}'_\beta(T)$ and choose $\iota = \frac{1}{2\sqrt{n}}$. Then there are constants $C_i = C_i(w) > 0, i=1,2$, depending only on w, such that for any $\gamma\geqslant 1$ and any $U\geqslant 1$ satisfying $e^{\|\Sigma\|_\infty U^2/8+2TU^\beta} \leqslant \sqrt{n}$ we have $\mathbb{P}(\|\mathcal{R}_n\|_\infty \geqslant \lambda) \leqslant 3e^{-\gamma^2}$ if*

$$\lambda \geqslant C_1\gamma^2\frac{e^{\|\Sigma\|_\infty U^2/4+4TU^\beta}}{U^2\sqrt{n}} + C_2TU^{-2+\beta}. \tag{15}$$

The proof is given in the appendix. The right-hand side of (15) is similar to the threshold (6), but without $\sqrt{\log p}$. Hence, this upper bounds depends on the dimension p only via spectral norm $\|\Sigma\|_\infty$. In the well-specified case, $\Sigma\in\mathbb{M}$ and optimizing over the spectral radius yields U of the order $\sqrt{(\log n)/\|\Sigma\|_\infty}$ and the corresponding λ of the order $(\|\Sigma\|_\infty^{-1}\log n)^{-1+\beta/2}$. The error bound takes the form

$$\|\tilde\Sigma_\lambda - \Sigma\| \leqslant C\sqrt{\operatorname{rank}(\Sigma)}\,\|\Sigma\|_\infty^{1-\beta/2}(\log n)^{-1+\beta/2}$$

with high probability. Here, $C>0$ is a constant depending only on w and T. Note that this bound for the estimation error improves a corresponding result in [2]. In the direct observation case, we can choose $U = \|\Sigma\|_\infty^{-1/2}$ and obtain $\|\tilde\Sigma_\lambda - \Sigma\| \leqslant C\|\Sigma\|_\infty\sqrt{\operatorname{rank}(\Sigma)/n}$ with high probability.

5.3. Elliptical distributions

Most of the literature on high-dimensional covariance estimation relies on a sub-Gaussian assumption on the distribution of X_j. To relax the moment assumption and allow for heavy-tailed distributions, the rich class of elliptical distributions has been studied, see the review paper by Fan et al. [24]. We refer to Fang et al. [27] for an introduction to the theory of elliptical distributions.

We will now outline how our approach can be generalized to the case where X_j follow a centered elliptical distribution, that is the characteristic function of X_j is of the form

$$\mathbb{E}[e^{i\langle u,X_j\rangle}] = \Phi(u^\top\Sigma u), \qquad u\in\mathbb{R}^p,$$

for some scalar function $\Phi\colon \mathbb{R} \to \mathbb{R}$ and some positive definite matrix Σ, which is proportional to the covariance matrix. The function Φ is called the *characteristic generator*. It is easy to see that $\mathbb{E}[X_j X_j^\top] = -2\Phi'(0)\Sigma$ provided that Φ is differentiable. We impose the mild assumption that $\Phi(\cdot) = \exp(-\eta(\cdot))$ for some function $\eta\colon \mathbb{R}_+ \to \mathbb{R}_+$. Then, the characteristic function of the observations Y_j has the form

$$\varphi(u) = \exp\big(-\eta(u^\top \Sigma u) + \log \psi(u)\big), \quad u \in \mathbb{R}^p.$$

We recover the Gaussian case with $\eta(x) = \frac{x}{2}$. Other important examples are multivariate α-stable distributions where $\eta(x) = x^{\alpha/2}$ for $\alpha \in (0, 2]$ or normal mixtures. To adapt the estimation strategy from Section 2, we assume that $|\operatorname{Re} \log \psi(u)|$ decays slower than $\eta(u^\top \Sigma u)$. If η is differentiable and strictly monotone with inverse function η^{-1}, a first order Taylor approximation and the fact that $(\eta^{-1})' = 1/(\eta' \circ \eta^{-1})$ yield

$$\eta^{-1}\big(-\log|\varphi(u)|\big) = \eta^{-1}\big(\eta\big(u^\top \Sigma u\big) - \log|\psi(u)|\big) \approx u^\top \Sigma u - \frac{\log|\psi(u)|}{\eta'(u^\top \Sigma u)}.$$

If the last term is of smaller order than $u^\top \Sigma u = \langle u, \Sigma u\rangle$ for $|u| \to \infty$, we can use these heuristics to estimate Σ. The argument is made rigorous by the following lemma proved in the appendix.

Lemma 10. *Let $\mathbb{E}[e^{i\langle u, X_j\rangle}] = \exp(-\eta(u^\top \Sigma u))$ for a positive-definite matrix Σ and a strictly monotone function $\eta\colon \mathbb{R}_+ \to \mathbb{R}_+$ which is twice continuously differentiable outside a neighbourhood of the origin. Assume further that*

$$\frac{\big|\log|\psi(u)|\big|}{\eta'(\langle u, \Sigma u\rangle)} \leqslant T(1+|u|)^\beta \qquad \text{and} \qquad |x\eta''(x)| \leqslant T|\eta'(x)|, \qquad \text{for all } u \in \mathbb{R}^p, x \in \mathbb{R}_+,$$

for some $\beta < 2$ and $T > 0$. For all $u \in \mathbb{R}^p$ with $|u| \geqslant (2^{\beta+1}T^2/\lambda_{min})^{1/(2-\beta)} \vee 1$ we then have

$$\Big|\eta^{-1}\big(-\log|\varphi(u)|\big) - \langle u, \Sigma u\rangle - \frac{\log|\psi(u)|}{\eta'(\langle u, \Sigma u\rangle)}\Big| \leqslant \frac{4T^2}{\lambda_{min}}|u|^{2\beta-2},$$

where $\lambda_{min} > 0$ is the smallest eigenvalue of Σ.

A major consequence of this lemma for our purposes is that $|u|^{-2}\eta^{-1}\big(-\log|\varphi(u)|\big) = \frac{\langle u, \Sigma u\rangle}{|u|^2} + \mathcal{O}(|u|^{-2+\beta})$ as $|u| \to \infty$. Thus, we can act as in Section 2. This leads to the estimator $\hat\Sigma^\Phi = (\hat\sigma^\Phi_{i,j})_{i,j=1,\dots p}$ for Σ where

$$\hat\sigma^\Phi_{i,i} := \frac{1}{U^2}\eta^{-1}\big(-\operatorname{Re}\big(\log\varphi_n(Uu^{(i)})\big)\big),$$
$$\hat\sigma^\Phi_{i,j} := \frac{1}{U^2}\eta^{-1}\big(-\operatorname{Re}\big(\log\varphi_n(Uu^{(i,j)})\big)\big) - \frac{\hat\sigma^\Phi_{i,i} + \hat\sigma^\Phi_{j,j}}{2} \quad \text{for} \quad i \neq j.$$

Applying an argument as in Lemma 10 together with the linearization for $\log\varphi_n$, we can bound the stochastic error of the estimators $\hat\sigma^\Phi_{i,j}$. We obtain the following proposition analogous to Theorem 1. The proof is again postponed to the appendix.

Proposition 11. *Let the assumptions of Lemma 10 be satisfied. Let $\gamma > \sqrt{2}$ and suppose that $U \geqslant (2^{2+\beta}T^2/\lambda_{min})^{1/(2-\beta)} \vee 1$ satisfies $8\gamma\sqrt{(\log(ep))/n} < \Delta_{\Sigma,U}$ for*

$$\Delta_{\Sigma,U} := \min_{i,j} \eta'(U^2\langle u^{(i,j)}, \Sigma u^{(i,j)}\rangle)|\varphi(Uu^{(i,j)})|.$$

Set

$$\tau(U) = \frac{12\gamma}{U^2\Delta_{\Sigma,U}}\sqrt{\frac{\log(ep)}{n}} + 4(T+1)U^{-2+\beta}.$$

Then, for $c_* = 12e^{-\gamma^2}$,

$$\mathbb{P}_{\Sigma,\psi}\Big(\max_{i,j=1,\dots,p}|\hat\sigma^\Phi_{i,j} - \sigma_{i,j}| < \tau(U)\Big) \geqslant 1 - c_* p^{2-\gamma^2}.$$

Under more specific assumptions on η it is possible to derive a uniform bound for $\Delta_{\Sigma,U}$. Since $|\varphi(u)| \geqslant \exp(-c \operatorname{Re} \eta(u^\top \Sigma u))$ for some constant $c > 0$, the stochastic error may not explode as fast as for normal distributions resulting in possibly faster convergence rates depending on η. Relying on $\hat{\Sigma}^\Phi$, hard and soft thresholding estimators can be constructed with similar behaviour as for the Gaussian case.

For the estimator $\hat{\Sigma}^\Phi$, the function η is assumed to be known. It would be interesting to extend the approach of this section to the case where η belongs to a parametric family introducing an additional nuisance parameter.

6. Numerical example

In this section we numerically analyse the performance of the soft thresholding estimator for the convolution model $Y = X + \varepsilon$, where X follows a p-dimensional normal distribution with zero mean and covariance matrix Σ and ε is independent of X and has an elliptical distribution. Specifically, we study the model

$$\varepsilon \stackrel{d}{=} \sqrt{W} A Z,$$

where $Z \sim \mathcal{N}(0, I_p)$ has a standard p-dimensional normal distribution, A is a $p \times p$ matrix and W is a nonnegative random variable with a Laplace transform $\mathcal{L}$. As can be easily seen, the characteristic function of ε is given by

$$\psi(u) = \mathbb{E}\big[e^{i\langle u, \varepsilon\rangle}\big] = \mathcal{L}\left(\frac{u^\top A A^\top u}{2}\right).$$

Thus ε has indeed an elliptical distribution. We assume that W follows a Gamma distribution with the density $p_W(x) = \Gamma(\vartheta)^{-1} x^{\vartheta - 1} e^{-x}, x \geq 0$, for some $\vartheta > 0$. Then we have

$$\psi(u) = \left(1 + \frac{u^\top A A^\top u}{2}\right)^{-\vartheta}.$$

Our aim is to compare several estimators of the covariance matrix Σ based on n independent copies $Y_1, \dots, Y_n$ of Y. In the direct observations case where $\varepsilon = 0$ we may apply the sample covariance matrix

$$\Sigma^{\mathrm{cov}} := \Sigma_Y^* = \frac{1}{n} \sum_{j=1}^n Y_j Y_j^\top. \tag{16}$$

Adapting to sparsity in a high-dimensional framework, a soft thresholding estimator based on Σ^{cov} is given by the solution of the optimisation problem, cf. Rothman et al. [46],

$$\Sigma_\tau^{\mathrm{s}} := \operatorname*{arg\,min}_{S \in \mathbb{R}^{p \times p}} \big\{ |S - \Sigma^{\mathrm{cov}}|_2^2 + 2\tau |S|_1 \big\}, \tag{17}$$

with threshold parameter $\tau > 0$. In some situations positive definiteness of the covariance matrix estimate is desirable when the covariance estimator is, for example, applied to supervised learning or if one needs to generate samples from the underlying normal distribution. In order to achieve positive definiteness, Rothman [45] proposed to use the following modification of (17):

$$\Sigma_\tau^{\mathrm{pds}} := \operatorname*{arg\,min}_{S \in \mathbb{R}^{p \times p},\, S \succ 0} \big\{ |S - \Sigma^{\mathrm{cov}}|_2^2 + 2\tau |S|_1 - \lambda \log |S| \big\}, \tag{18}$$

where $|S|$ denotes the determinant of the matrix S and λ is a fixed small number. The logarithmic barrier term in (18) ensures the existence of a positive definite solution, since $\log |S| = \sum_{j=1}^p \log(\sigma_j(S))$, where $\sigma_j(S)$ is the jth largest eigenvalue of $S > 0$. In order to solve (18), an algorithm similar to the graphical lasso algorithm can be applied, see Friedman et al. [28].

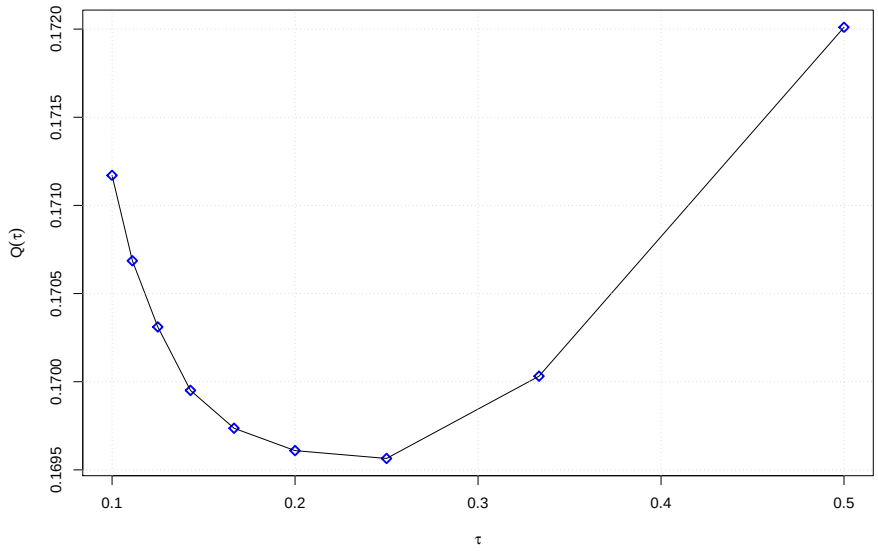

Figure 1. The objective function $\mathcal{Q}_{100}(\tau)$ for the choice of the tuning parameter τ

Turning back to our deconvolution problem, we have already seen that the estimators (16), (17) and (18) fail to deliver a consistent estimator for Σ unless ε is zero. Hence, we finally introduce the positivity preserving version of the spectral soft thresholding estimator from (9):

$$\Sigma_\tau^{\mathrm{sps}} := \underset{S\in\mathbb{R}^{p\times p},\, S\succ 0}{\arg\min} \big\{|S-\hat{\Sigma}|_2^2 + 2\tau|S|_1 - \lambda\log|S|\big\}. \tag{19}$$

The tuning parameter τ can be chosen using a method introduced in [3]. The data is randomly partitioned N times into a training set of size n_1 and a validation set of size n_2 with $n_2 = \lfloor n/\log(n)\rfloor$ and $n_1 = n - n_2$. The tuning parameter is then selected as $\widehat{\tau} = \arg\min_\tau \mathcal{Q}_N(\tau)$, where

$$\mathcal{Q}_N(\tau) = \sum_{m=1}^{N} \|\Sigma_\tau^{\mathrm{sps},(m,n_1)} - \hat{\Sigma}^{(m,n_2)}\|^2,$$

where $\Sigma_\tau^{\mathrm{sps},(m,n_1)}$ is the estimator, with penalty parameter τ, computed with the training set of the mth split and $\hat{\Sigma}^{(m,n_2)}$ is the estimator (5) computed with the validation set of the mth split.

First, we consider a tridiagonal model where the population covariance matrix Σ has entries $\sigma_{ij} = 0.4\cdot 1(|i-j|=1) + 1(i=j)$, $i,j\in\{1,2,\ldots,p\}$. Using this covariance model with $p=20$, we generate $n=50$ realizations of independent normal random vectors with mean zero and the covariance matrix Σ. Adding an independent noise ε with the above elliptical distribution with $A = I_d$, depending on the parameter ϑ, we compute three estimates Σ^{cov}, $\Sigma_\tau^{\mathrm{pds}}$ and $\Sigma_\tau^{\mathrm{sps}}$. This procedure was repeated 500 times. The parameters of the algorithms are $\tau = 0.25$, $\lambda = 10^{-4}$, where the parameter τ is selected as a minimum of the function $\mathcal{Q}_{100}(\tau)$ shown in Figure 1.

The results are presented in Figure 2 for the case of direct observations and for three different noise specifications corresponding to the values $\vartheta\in\{0.5,1,2\}$. The used values of the tuning parameter U are $1,3,3$, respectively. While in the case of direct observations, the estimator $\Sigma_\tau^{\mathrm{sps}}$ has no advantages over Σ^{cov} and $\Sigma_\tau^{\mathrm{pds}}$, it significantly outperforms these two estimators in the case of non-zero noise. We do not only observe a strong bias for Σ^{cov} and $\Sigma_\tau^{\mathrm{pds}}$ in the presence of noise, but also a much better concentration of the spectral estimator $\Sigma_\tau^{\mathrm{sps}}$ compared to the other two procedures. The higher is the variance of the noise, the stronger are these bias and variance effects.

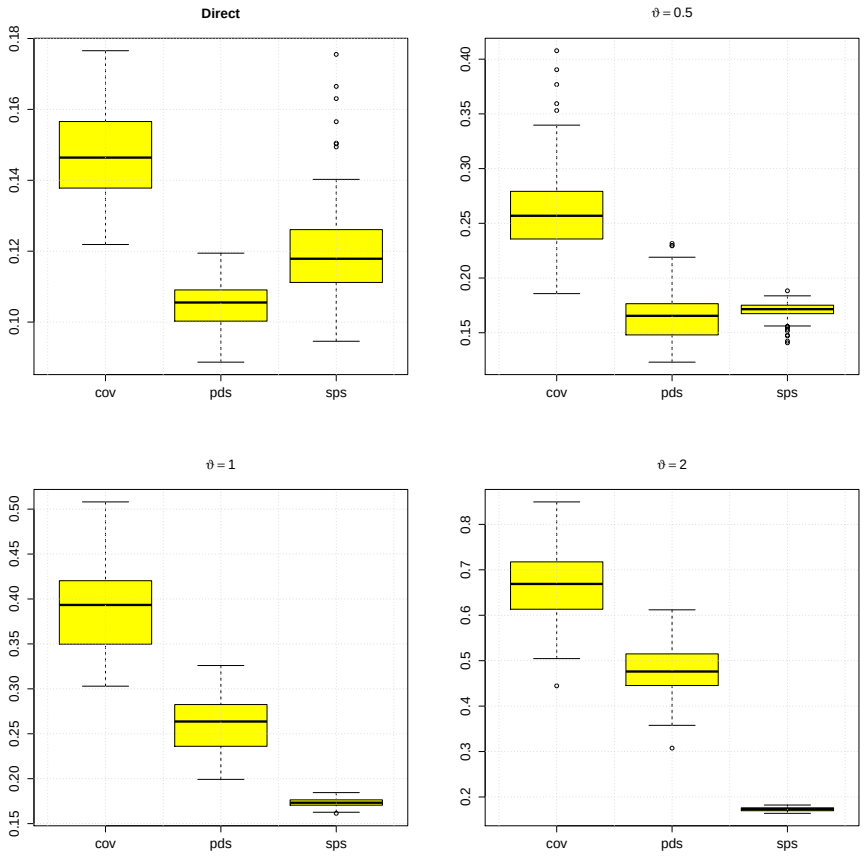

Figure 2. Tridiagonal Σ : box plots of the estimation errors $\|\Sigma_\tau^o - \Sigma\|$ for $o \in \{\mathrm{cov}, \mathrm{pds}, \mathrm{sps}\}$ in the case of the convolution model $Y = X + \varepsilon$ with $\varepsilon \stackrel{d}{=} \sqrt{W}Z$, where $Z \sim \mathcal{N}_{20}(0, I_{20})$ and $W \sim \mathrm{Gamma}(\vartheta)$.

Now, let us consider the case of normal noise. Note that this situation corresponds to $\beta = 2$ and is not covered (at least formally) by our theoretical study. Specifically we generate samples from the model $Y = X + \varepsilon$, where X follows a p-dimensional normal distribution with zero mean and covariance matrix Σ and ε is independent of X and has also normal distribution with zero mean and covariance matrix $\rho^2 I$. We again consider tridiagonal model where the population covariance matrix Σ has entries $\sigma_{ij} = 0.4 \cdot 1(|i-j| = 1) + 1(i = j)$, $i, j \in \{1, 2, \ldots, p\}$. In Figure 3 the corresponding estimation errors for three methods are presented in the case of $p = 20$, $\tau = 0.4$, $n = 50$ and $\rho \in \{0.1, 0.5\}$. As one can see, even in the case of misspecified models the spectral estimator continues to perform reasonably well.

Finally, we study the situation where the matrix Σ is block diagonal with the elliptical error distribution from above. In particular, we generate positive definite matrix with randomly-signed, non-zero elements. A shift is added to the diagonal of the matrix so that its condition number equals p. Using this covariance model, we generated $n = 100$ realizations of independent 20-dimensional normal random vectors with mean zero and covariance Σ. We then proceed as before considering the case of direct observations and $\vartheta \in \{0.5, 1, 2\}$. The tuning parameter U was taken to be 3 for all three cases. The errors show a similar behaviour as in the first case, see Figure 4.

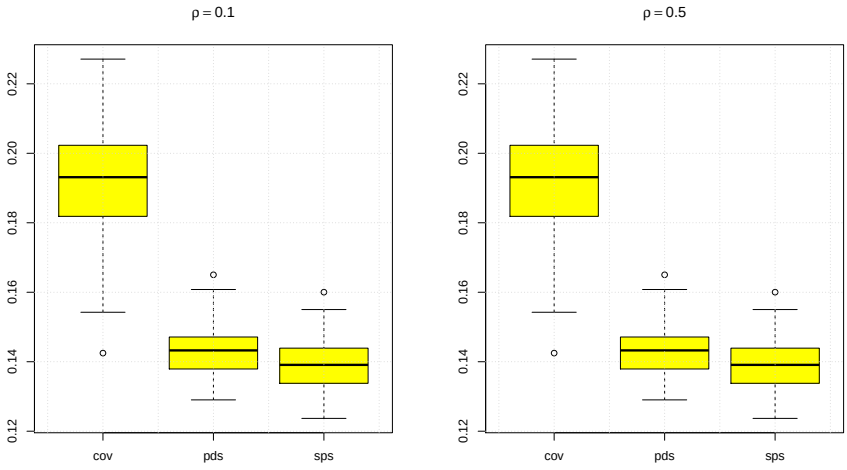

Figure 3. Tridiagonal Σ : box plots of the estimation errors $\|\Sigma_\tau^o - \Sigma\|$ for $o \in \{\mathrm{cov}, \mathrm{pds}, \mathrm{sps}\}$ in the case of the convolution model $Y = X + \varepsilon$ with $\varepsilon \stackrel{d}{=} Z$, where $Z \sim \mathcal{N}_{20}(0, \rho I_{20})$.

7. Proofs

7.1. Concentration of the spectral estimator

For the proof of Theorem 1, we need the following lemmas. Set $S(u) = \mathrm{Re}(\log \varphi_n(u) - \log \varphi(u))$.

Lemma 12. *For any $x \in (0, 1]$, and any $u \in \mathbb{R}^p$ such that $\varphi(u) \neq 0$,*

$$\mathbb{P}\big(|S(u)| \geqslant x\big) \leqslant 3\mathbb{P}\Big(|\varphi_n(u) - \varphi(u)| \geqslant \frac{x}{2}|\varphi(u)|\Big).$$

Proof. We have

$$S(u) = \log\Big|\frac{\varphi_n(u)}{\varphi(u)}\Big| \leqslant \log\Big(\Big|\frac{\varphi_n(u) - \varphi(u)}{\varphi(u)}\Big| + 1\Big) \leqslant \Big|\frac{\varphi_n(u) - \varphi(u)}{\varphi(u)}\Big|.$$

Thus, $\mathbb{P}\big(S(u) \geqslant x\big) \leqslant \mathbb{P}\big(|\varphi_n(u) - \varphi(u)| \geqslant x|\varphi(u)|\big)$ for all $x > 0$. Next, on the event $\Big\{|\varphi_n(u) - \varphi(u)| \leqslant \frac{1}{2}|\varphi(u)|\Big\}$ we have

$$-S(u) = \log\Big|\frac{\varphi(u)}{\varphi_n(u)}\Big| \leqslant \log\Big(\Big|\frac{\varphi_n(u) - \varphi(u)}{\varphi_n(u)}\Big| + 1\Big) \leqslant \log\Big(2\Big|\frac{\varphi_n(u) - \varphi(u)}{\varphi(u)}\Big| + 1\Big) \leqslant 2\Big|\frac{\varphi_n(u) - \varphi(u)}{\varphi(u)}\Big|.$$

Therefore, for any $x > 0$,

$$\mathbb{P}\big(-S(u) \geqslant x\big) \leqslant \mathbb{P}\Big(2|\varphi_n(u) - \varphi(u)| \geqslant x|\varphi(u)|\Big) + \mathbb{P}\Big(|\varphi_n(u) - \varphi(u)| > \frac{1}{2}|\varphi(u)|\Big).$$

Since $x \in (0, 1]$, we obtain $\mathbb{P}\big(-S(u) \geqslant x\big) \leqslant 2\mathbb{P}\big(|\varphi_n(u) - \varphi(u)| \geqslant (x/2)|\varphi(u)|\big)$ and hence the lemma. □

Lemma 13. *For any $\kappa \in (0, \sqrt{n}/8]$ we have*

$$\mathbb{P}\Big(|\varphi_n(u) - \varphi(u)| \geqslant \frac{3\kappa}{\sqrt{n}}\Big) \leqslant 4e^{-\kappa^2}.$$

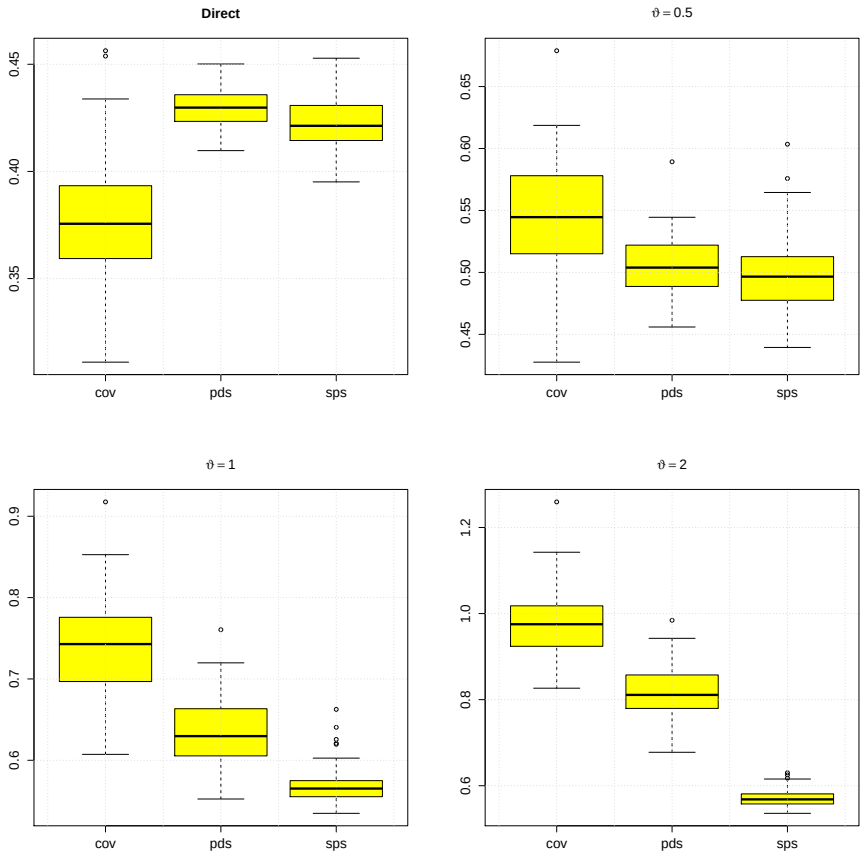

Figure 4. Block diagonal Σ : box plots of the estimation errors $\|\Sigma_\tau^o - \Sigma\|$ for $o \in \{\mathrm{cov}, \mathrm{pds}, \mathrm{sps}\}$ in the case of the convolution model $Y = X + \varepsilon$ with $\varepsilon \stackrel{d}{=} \sqrt{W}Z$, where $Z \sim \mathcal{N}_{20}(0, I_{20})$ and $W \sim \mathrm{Gamma}(\vartheta)$.

Proof. We decompose $\varphi_n - \varphi$ into real and imaginary part. Both can be estimated analogously, such that we consider only the real part. We write

$$\mathrm{Re}\big(\varphi_n(u) - \varphi(u)\big) = \frac{1}{n}\sum_{k=1}^{n} \xi_k(u) \quad \text{with} \quad \xi_k(u) := \mathrm{Re}\left(e^{i\langle u, Y_k\rangle}\right) - \mathrm{Re}\,\varphi(u).$$

The independent and centred random variables $\xi_k(u), k = 1, \dots, n$, satisfy

$$|\xi_k(u)| \leqslant 2 \quad \text{and} \quad \mathrm{Var}(\xi_k(u)) \leqslant 1 - |\varphi(u)|^2 \leqslant 1.$$

Using the fact that $\kappa \in (0, \sqrt{n}/8]$ and then applying Bernstein's inequality we find

$$\mathbb{P}\Big(\big|\,\mathrm{Re}\,\big(\varphi_n(u) - \varphi(u)\big)\big| \geqslant \frac{3\kappa}{2\sqrt{n}}\Big) \leqslant \mathbb{P}\Big(\big|\,\mathrm{Re}\,\big(\varphi_n(u) - \varphi(u)\big)\big| \geqslant \frac{\sqrt{2}\kappa}{\sqrt{n}} + \frac{2\kappa^2}{3n}\Big) \leqslant 2e^{-\kappa^2}.$$

□

Corollary 14. *For any $\gamma > 0$ and $u \in \mathbb{R}^p$ such that $\gamma\sqrt{(\log(ep))/n} \leqslant |\varphi(u)|/8$ we have*

$$\mathbb{P}\Big(|S(u)| \geqslant \frac{6\gamma\sqrt{\log(ep)}}{\sqrt{n}|\varphi(u)|}\Big) \leqslant 12(ep)^{-\gamma^2}.$$

Proof. We use Lemma 12 with $x = \frac{6\gamma\sqrt{\log(ep)}}{\sqrt{n}|\varphi(u)|}$ and then Lemma 13 with $\kappa = \gamma\sqrt{\log(ep)}$. To apply Lemma 12 we need $6\gamma\sqrt{\frac{\log ep}{n}} \leqslant |\varphi(u)|$, while Lemma 13 requires $8\gamma\sqrt{\frac{\log ep}{n}} \leqslant 1$. Since $|\varphi(u)| \leqslant 1$ both conditions are satisfied. □

8. Proof of the lower bound: Theorem 6

Since $C_1 p \leqslant S \leqslant C_2 p$ it is enough to assume that $2p \leqslant S$ (otherwise we consider a $(C_1 p/2)$-dimensional subspace). Furthermore, we will assume without loss of generality that $S = p + 2k$ for some integer $k \geqslant 1$ corresponding to p non-zero diagonal entries and $2k$ non-zero off-diagonal entries of the covariance matrix. Note that under our assumptions, S, k and p are of the same order up to constants:

$$\frac{S}{4} \leqslant k = \frac{S-p}{2} \leqslant \frac{S}{2} \leqslant \frac{C_2 p}{2}. \tag{20}$$

Let $\mathbb{P}_{\Sigma,\psi}$ denote the distribution of Y_j corresponding to the covariance matrix $\Sigma \in \mathcal{G}_q(S,R)$ and to the error distribution with characteristic function $\psi \in \mathcal{H}_\beta(T)$. Set

$$\varphi_{\Sigma,\psi}(u) := \mathbb{E}_{\Sigma,\psi}[e^{i\langle u, Y_j\rangle}] = \exp\Big(-\frac{1}{2}\langle u, \Sigma u\rangle + \log\psi(u)\Big).$$

Applying Theorem 2.6 in [50], it is sufficient to construct a finite number of pairs (Σ_i, ψ_i) with $\Sigma_0 = RI_p, \psi_0 \in \mathcal{H}_\beta(T)$ and $(\Sigma_i, \psi_i) \in \mathcal{G}_q(p+2k, R) \times \mathcal{H}_\beta(T)$ for $i = 1, \dots, M$, such that the following two conditions hold:

(i) $\|\Sigma_i - \Sigma_j\| \geqslant CS^{1/2}T\big(R^{-1}\log n\big)^{-1+\beta/2}$ for all $0 \leqslant i < j \leqslant M$ and some constant $C > 0$,
(ii) $\chi^2(\mathbb{P}^{\otimes n}_{\Sigma_j,\psi_j}, \mathbb{P}^{\otimes n}_{\Sigma_0,\psi_0}) \leqslant M/3$ for all $j = 1, \dots, M$.

Step 1: Constructing the pairs (Σ_i, ψ_i). Without loss of generality, consider p that can be decomposed as $p = Lb$ where b and L are integers. For a block size $b \in \mathbb{N}$ and $L = p/b \in \mathbb{N}$ let $\mathcal{B} \subseteq \mathbb{R}^{p\times p}$ denote the set of symmetric block diagonal matrices $B = \mathrm{diag}(A_1, \dots, A_L)$ satisfying:

- $B = (b_{ij})$ has exactly k non-zero over-diagonal entries, all equal to 1;
- $b_{ii} = 0$ for $i = 1, \dots, n$;
- $A_l \in \mathbb{R}^{b\times b}$ for $l = 1, \dots, L$.

There are $N := Lb(b-1)/2 = p(b-1)/2$ positions over the diagonal of B where the entry 1 can possibly appear. Since $k \leqslant C_2 p/2$, we have $k < N$ for $b > C_2 + 1$. In what follows, we select $b > C_2 + 1$, which is a fixed integer independent of k and p. Lemma A.3 in Rigollet and Tsybakov [43] yields that there is a subset $\{B_1, \dots B_M\} \subseteq \mathcal{B}$ such that for any $i \neq j$ we have $\|B_i - B_j\|^2 \geqslant (k+1)/4$ and for some constants $C_1', c_1' > 0$,

$$\log M \geqslant C_1' k \log\Big(1 + \frac{eN}{4k}\Big) \geqslant C_1' k \log\Big(1 + \frac{c_1' bp}{k}\Big). \tag{21}$$

We consider now the following family of matrices

$$\Sigma_0 = RI_p, \quad \Sigma_j = RI_p + \frac{\rho T}{b}\delta_{n,p}^{2-\beta} B_j, \quad j = 1, \dots, M,$$

where

$$\delta_{n,p} = R^{1/2}\Big(6\log\frac{n}{\rho'\log(1 + c_1' bp/k)}\Big)^{-1/2}, \tag{22}$$

and $\rho, \rho' > 0$ are small enough constants to be chosen later. By construction and using (20) we have

$$\begin{aligned}\|\Sigma_i - \Sigma_j\| &\geqslant \rho\frac{T}{2b}\delta_{n,p}^{2-\beta}k^{1/2} \geqslant \frac{T}{2b}k^{1/2}\Big(6R^{-1}\log\Big(\frac{n}{\rho'\log\big(1 + \frac{c_1' bp}{k}\big)}\Big)\Big)^{-(1-\beta/2)} \\ &\geqslant c_2' T S^{1/2}\big(R^{-1}\log n\big)^{-(1-\beta/2)}\end{aligned}$$

where $c_2' > 0$ is a constant. Moreover, since by assumption of the theorem, $R^{-1}Tb^{-1}\delta_{n,p}^{2-\beta}$ is uniformly bounded, the matrices Σ_i are diagonally dominant and thus positive semi-definite for sufficiently small ρ. We conclude that the Σ_i thus defined are covariance matrices satisfying the lower bound in (i) above.

We now turn to the construction of characteristic functions ψ_j. To have an as small as possible L^2-distance between the characteristic functions, we choose ψ_j such that $\log\psi_j(u) - \log\psi_0(u)$ mimics $\langle u, (\Sigma_j - \Sigma_0)u\rangle/2$ for small frequencies, keeping the block structure. In what follows, we denote by $\mathcal{F}$ the Fourier transform operator. On each block of the matrix $B_j = \operatorname{diag}(A_{j,1}, \dots, A_{j,L})$, for $j = 1, \dots, M, l = 1, \dots, L$, we define

$$\begin{aligned}\log\psi_{j,l}(u) :=& \frac{\rho T\delta_{n,p}^{2-\beta}}{2b}\langle u, A_{j,l}u\rangle \mathcal{F}K(\delta_{n,p}u) + \log\psi_{0,l}(u), \qquad u \in \mathbb{R}^b,\\ \log\psi_{0,l}(u) :=& \int_{\mathbb{R}^b} \big(e^{i\langle u,x\rangle} - i\langle u,x\rangle \mathbb{1}_{\{\beta\geqslant 1\}} - 1\big)\frac{T}{\xi_b|x|^{\beta+b}}\mathrm{d}x, \qquad u \in \mathbb{R}^b,\end{aligned}$$

where $\xi_b > 0$ is a constant depending only on b, and $K \in L^1(\mathbb{R}^b) \cap C^2(\mathbb{R}^b)$ is a function satisfying $\mathcal{F}K \in C^\infty(\mathbb{R}^b)$, and

$$\mathcal{F}K(u) = 1 \quad \text{for } |u| \leqslant 1, \quad \mathcal{F}K(u) = 0 \quad \text{for } |u| > 2, \quad \text{and} \quad 0 \leqslant \mathcal{F}K(u) \leqslant 1 \quad \forall\, u.$$

Writing $u^l := (u_{b(l-1)+1}, \dots, u_{bl})$ for $1 \leqslant l \leqslant L$ and $u \in \mathbb{R}^p$, we then set

$$\psi_j(u) := \prod_{l=1}^L \psi_{j,l}(u^l), \qquad j = 0, \dots, M.$$

Note that $\psi_{0,l}$ is the characteristic function of a b-dimensional symmetric stable distribution, cf. Sato [48, Thm. 14.3]. To check that $\psi_0 \in \mathcal{H}_\beta(T)$ is satisfied, we use Theorem 14.10 in [48], which yields

$$\big|\log|\psi_0(u)|\big| \leqslant \sum_{l=1}^L \big|\log|\psi_{0,l}(u)|\big| \leqslant \sum_{l=1}^L C_\beta \frac{T}{\xi_b}\frac{2\pi^{b/2}}{\Gamma(b/2)}|u^l|^\beta \leqslant C_\beta \frac{T}{\xi_b}\frac{2\pi^{b/2}}{\Gamma(b/2)}|u|_\beta^\beta,$$

where $C_\beta > 0$ is a constant depending only on β and where $\frac{2\pi^{b/2}}{\Gamma(b/2)}$ is the surface area of the $(b-1)$-dimensional sphere. Thus, choosing

$$\xi_b = c\frac{2\pi^{b/2}}{\Gamma(b/2)} \tag{23}$$

for some sufficiently large $c > 0$ guarantees that $\psi_0 \in \mathcal{H}_\beta(T)$. Note that ξ_b is bounded uniformly in b.

We also have $\psi_j \in \mathcal{H}_\beta(T)$ for sufficiently small ρ since maximal singular value $\|A_{j,l}\|_\infty \leqslant b$ and

$$\begin{aligned}\sum_{l=1}^L \Big|\frac{\rho T\delta_{n,p}^{2-\beta}}{2b}\langle u^l, A_{j,l}u^l\rangle \mathcal{F}K(\delta_{n,p}u^l)\Big| &\leqslant \frac{\rho\|K\|_{L^1}T}{2b}\delta_{n,p}^{2-\beta}\sum_{l=1}^L \|A_{j,l}\|_\infty |u^l|^2 \mathbb{1}_{\{|u^l|\leqslant 2/\delta_{n,p}\}}\\ &\leqslant \frac{\rho\|K\|_{L^1}T}{2b}\delta_{n,p}^{2-\beta}\sum_{l=1}^L b\big(\frac{2}{\delta_{n,p}}\big)^{2-\beta}|u^l|^\beta \qquad (24)\\ &\leqslant 2\rho\|K\|_{L^1}T|u|_\beta^\beta.\end{aligned}$$

It remains to verify that $\psi_{j,l}, l = 1, \dots, L$ are indeed characteristic functions. Denoting $A_{j,l} = (a_{k,m}^{j,l})_{k,m=1,\dots,b}$ and

$$\nu_{j,l} := \frac{1}{2b}\sum_{k,m} a_{k,m}^{j,l}(\partial_k\partial_m K),$$

where ∂_k stands for the derivative with respect to kth coordinate, we rewrite the characteristic exponent as

$$\begin{aligned}\log\psi_{j,l}(u) &= \frac{\rho T\delta_{n,p}^{2-\beta}}{2b}\sum_{k,l=1}^{b} a_{k,m}^{j,l}u_k u_l \mathcal{F}K(\delta_{n,p}u) + \log\psi_0(u)\\ &= -\mathcal{F}\Big[\rho T\delta_{n,p}^{-\beta-b}\nu_{j,l}(\delta_{n,p}^{-1}\cdot)\Big](u) + \log\psi_0(u)\\ &= \int_{\mathbb{R}^b}\big(e^{i\langle u,x\rangle} - i\langle u,x\rangle\mathbb{1}_{\{\beta>1\}} - 1\big)\Big(\frac{T}{\xi_b|x|^{\beta+b}} - \rho T\delta_{n,p}^{-\beta-b}\nu_{j,l}(x/\delta_{n,p})\Big)\mathrm{d}x\end{aligned}$$

where in the last line we have used the relations $\int_{\mathbb{R}^b}\nu_{j,l}(x)\mathrm{d}x = \mathcal{F}\nu_{j,l}(0) = 0$ and, if $\beta \geqslant 1$, $\int_{\mathbb{R}^b} i\langle u,x\rangle\nu_{j,l}(x)\mathrm{d}x = \langle u, \nabla(\mathcal{F}\nu_{j,l})(0)\rangle = 0$ for any $u\in\mathbb{R}^b$. Consequently, $\psi_{j,l}$ is the characteristic function of an infinitely divisible distribution with Lévy density $T\xi_b^{-1}|x|^{-\beta-b} - \rho T\delta_{n,p}^{-\beta-b}\nu_{j,l}(x/\delta_{n,p})$ provided that the latter is non-negative. To check this, it is enough to verify the equivalent condition $\rho\xi_b\nu_{j,l}(x) \leqslant |x|^{-\beta-b}$ for all $x\in\mathbb{R}^b\setminus\{0\}$ and some sufficiently small ρ. We have

$$\begin{aligned}\||x|^{\beta+b}\nu_{j,l}(x)\|_\infty &\leqslant \|\nu_{j,l}\|_\infty + \||x|^{2\lceil(\beta+b)/2\rceil}\nu_{j,l}(x)\|_\infty\\ &\leqslant \|\mathcal{F}\nu_{j,l}\|_{L^1} + \|\Delta^{\lceil(\beta+b)/2\rceil}\mathcal{F}\nu_{j,l}\|_{L^1}\end{aligned} \tag{25}$$

where Δ denotes the Laplace operator, $\lceil x\rceil$ is the minimal integer greater than x, and $\|\cdot\|_{L^q}$ stands for the $L_q(\mathbb{R}^b)$-norm. By construction, $\mathcal{F}\nu_{j,l}(u) = \frac{1}{2b}\langle u, A_{j,l}u\rangle\mathcal{F}K(u)$, and thus

$$\|\mathcal{F}\nu_{j,l}\|_{L^1} \leqslant \frac{\|A_{j,l}\|_\infty}{2b}\big\||u|^2\mathcal{F}K(u)\big\|_{L^1} \leqslant \frac{1}{2}\big\||u|^2\mathcal{F}K(u)\big\|_{L^1}$$

where we have used the inequality $\|A_{j,l}\|_\infty \leqslant b$. Since the support of $\mathcal{F}K$ is compact the last expression is bounded. The second term in (25) admits an analogous bound.

Step 2: Bounding the χ^2-divergence. Due to the block structure, for any pair (Σ_i,ψ_i) we have

$$\mathbb{P}_{\Sigma_i,\psi_i} = \prod_{l=1}^{L}\mathbb{P}_{i,l}$$

for all $i = 1,\dots,M$, $l = 1,\dots,L$, where $\mathbb{P}_{i,l}$ is the convolution of the normal distribution $\mathcal{N}(0, RI_b + \frac{\rho T}{b}\delta_{n,p}^{2-\beta}A_{i,l})$ on $\mathbb{R}^b$ with a distribution given by the characteristic function $\psi_{i,l}$. We also denote by $\mathbb{P}_0$ the convolution of $\mathcal{N}(0,RI_b)$ with the stable distribution given by $\psi_{0,l}$. We have

$$\begin{aligned}\chi^2(\mathbb{P}_{\Sigma_i,\psi_i}^{\otimes n}, \mathbb{P}_{\Sigma_0,\psi_0}^{\otimes n}) &= \big(1+\chi^2(\mathbb{P}_{\Sigma_i,\psi_i},\mathbb{P}_{\Sigma_0,\psi_0})\big)^n - 1\\ &= \prod_{l=1}^{L}\big(1+\chi^2(\mathbb{P}_{i,l},\mathbb{P}_0)\big)^n - 1.\end{aligned} \tag{26}$$

Thus, to check condition (ii) stated at the beginning of this subsection, we need to bound from above the value

$$\chi^2(\mathbb{P}_{i,l},\mathbb{P}_0) = \int_{f_0(x)>0}\frac{\big(f_{i,l}(x)-f_0(x)\big)^2}{f_0(x)}\mathrm{d}x \tag{27}$$

where $f_{i,l}$ and f_0 are the densities of $\mathbb{P}_{i,l}$ and $\mathbb{P}_0$, respectively. To this end, we first establish a lower bound for f_0, which is the density of the convolution of a normal distribution on $\mathbb{R}^b$ with zero mean and covariance matrix RI_b and a stable distribution on $\mathbb{R}^b$. If there is no Gaussian component, we write $R=0$ referring to a convolution of the stable distribution with a Dirac measure in zero.

Lemma 15. *In the special case of a standard stable density f_0 ($R=0$, $T=1$) and $\beta\in(0,2)$ we have $f_0(x) \geqslant C_b(1+|x|^{\beta+b})^{-1}$ for a constant $C_b>0$ depending only on b. If $R>0$ and $T>0$ are such that $T(\log n)^{-c} \leqslant CR^{\beta/2}$ for some $C,c>0$, we have the lower bound*

$$f_0(x) \geqslant C_b' R^{-b/2}(\log n)^{-cb/\beta}\frac{1}{(1+T^{-1-b/\beta}|x|^{b+\beta})}$$

for another constant $C_b' > 0$.

Proof. *Step 1:* We first consider the case $R = 0, T = 1$ and start with $\beta \in (0,1)$. We have $f_0 = h_c * h_f$ where

$$h_c(x) := \mathcal{F}^{-1}\Big[\exp\Big(\frac{1}{\xi_b}\int_{|y|\leqslant 1}\big(e^{i\langle u,y\rangle}-1\big)\big(\frac{1}{|y|^{\beta+b}}-1\big)\mathrm{d}y\Big)\Big](x),$$
$$h_f(x) := \mathcal{F}^{-1}\Big[\exp\Big(\frac{1}{\xi_b}\int_{\mathbb{R}^b}\big(e^{i\langle u,y\rangle}-1\big)\frac{1}{|y|^{\beta+b}\vee 1}\mathrm{d}y\Big)\Big](x), \quad x\in\mathbb{R}^b,$$

are the densities of an infinitely divisible distribution with Lévy density $\nu_c(x) := \frac{1}{\xi_b}(\frac{1}{|x|^{\beta+b}} - 1)\mathbb{1}_{\{|x|\leqslant 1\}}$ and an infinitely divisible distribution with Lévy density $\nu_f(x) := \frac{1}{\xi_b(|x|^{\beta+b}\vee 1)}$, $x\in\mathbb{R}^b$, respectively. Since ν_f is integrable, h_f is the density of a compound Poisson distribution which can be written as convolution exponential, cf. [48, Remark 27.3],

$$h_f = e^{-\nu_f(\mathbb{R}^b)}\sum_{j=0}^{\infty}\frac{\nu_f^{*j}}{j!} \tag{28}$$

where ν_f^{*j} denotes the j-fold convolution of ν_f, and $\nu_f^{*0} := \delta_0$ is the Dirac measure in zero. Therefore,

$$f_0 = e^{-\nu_f(\mathbb{R}^b)}\sum_{j=0}^{\infty}\frac{h_c*\nu_f^{*j}}{j!} \geqslant e^{-\nu_f(\mathbb{R}^b)}\big(h_c*\nu_f\big), \tag{29}$$

where, with some abuse of notation, $\nu_f(\mathbb{R}^b)$ stands for the total mass of ν_f. Due to the compactness of the support of the Lévy measure corresponding to h_c, the density h_c admits a finite exponential moment [48, Theorem 26.1], that is there exists $\alpha > 0$ such that $\int_{\mathbb{R}^b} e^{\alpha|y|}h_c(y)\mathrm{d}y < \infty$.

For any $x \neq 0$, we have

$$\begin{aligned}|h_c*\nu_f(x) - \nu_f(x)| &= \Big|\int_{\mathbb{R}^b}\big(\nu_f(x-y)-\nu_f(x)\big)h_c(y)\mathrm{d}y\Big|\\ &\leqslant \int_{|y|\leqslant|x|/2}\big|\nu_f(x-y)-\nu_f(x)\big|h_c(y)\mathrm{d}y + \int_{|y|>|x|/2}\big|\nu_f(x-y)-\nu_f(x)\big|h_c(y)\mathrm{d}y\end{aligned}$$

Rewriting $\nu_f(x) = \mu(|x|)$ with $\mu(r) := \xi_b^{-1}(r^{-\beta-b}\wedge 1)$, we see that the expression in the last display does not exceed

$$\begin{aligned}&\sup_{|v|\leqslant|x|}|\mu'(|v|)|\int_{\mathbb{R}^b}|y|h_c(y)\mathrm{d}y + 2\|\nu_f\|_\infty e^{-\alpha|x|/2}\int_{|y|>|x|/2}e^{\alpha|y|}h_c(y)\mathrm{d}y\\ &\leqslant \Big(\sup_{|v|\leqslant|x|}|\mu'(|v|)| + 2\|\nu_f\|_\infty e^{-\alpha|x|/2}\Big)\int_{\mathbb{R}^b}(|y|\vee e^{\alpha|y|})h_c(y)\mathrm{d}y.\end{aligned} \tag{30}$$

By the polynomial decay of μ we have $|\mu'(|x|)| = o(\nu_f(x))$ as $|x|\to\infty$ implying that $|h_c*\nu_f(x) - \nu_f(x)| = \nu_f(x)o(1)$ as $|x|\to\infty$. Combining (29) and (30) yields

$$f_0(x) \geqslant e^{-\nu_f(\mathbb{R}^b)}\big(h_c*\nu_f\big)(x) = e^{-\nu_f(\mathbb{R}^b)}\nu_f(x)\big(1+o(1)\big) \geqslant \frac{C}{|x|^{\beta+b}}, \quad \forall |x|>r, \tag{31}$$

where $C > 0$ and $r > 0$ are constants depending only on b.

From the representation (28) we see that f_0 is strictly positive. By the decay of its characteristic function, f_0 is also continuous. Together with (31), we conclude that

$$f(x) \geqslant C_b(1+|x|^{\beta+b})^{-1}$$

for $\beta\in(0,1), R = 0, T = 1$.

In the case $\beta \in [1,2), R = 0, T = 1$ the proof is analogous with the only difference that the convolution exponential h_f is shifted by $a := (\int_{\mathbb{R}^b} x_1 \nu_f(x)\mathrm{d}x, \dots, \int_{\mathbb{R}^b} x_b \nu_f(x)\mathrm{d}x) \in \mathbb{R}^b$, i.e.,

$$h_f = e^{-\nu_f(\mathbb{R}^b)}\delta_a * \Big(\sum_{j=0}^{\infty} \frac{\nu_f^{*j}}{j!}\Big)$$

where δ_a is the Dirac distribution at a. Thus, we replace everywhere above $g_b * h_c$ by $g_b * h_c * \delta_a$. Clearly, the argument remains valid with such a modification.

Step 2. We now denote by f the density f_0 from Step 1 corresponding to $R = 0, T = 1$. Thus, f is a density with characteristic function $e^{-C|u|^\beta}$ for some $C > 0$. With this notation, for $R = 0, T > 0$ we have $f_0(x) = \mathcal{F}^{-1}[e^{-CT|u|^\beta}](x) = T^{-b/\beta} f(T^{-1/\beta}x)$. We now turn to the case $R > 0, T > 0$. Denoting the density of the normal distribution $\mathcal{N}(0, I_b)$ by g and using the lower bound from Step 1, we obtain

$$\begin{aligned}
f_0(x) &= \Big(\big(T^{-b/\beta} f(T^{-1/\beta}\cdot)\big) * \big(R^{-b/2} g(R^{-1/2}\cdot)\big)\Big)(x) \\
&\geqslant (2\pi R)^{-b/2} \int_{\mathbb{R}^b} \frac{C_b}{1 + |T^{-1/\beta}x - y|^{b+\beta}} e^{-\frac{T^{2/\beta}}{2R}|y|^2}\mathrm{d}y \\
&\geqslant C_b (2\pi R)^{-b/2} \frac{1}{(1 + 2^{b+\beta-1}T^{-1-b/\beta}|x|^{b+\beta})} \int_{\mathbb{R}^b} \frac{1}{1 + 2^{\beta+b-1}|y|^{b+\beta}} e^{-\frac{T^{2/\beta}}{2R}|y|^2}\mathrm{d}y \\
&\geqslant \frac{2\pi^{b/2} C_b}{4^{b+\beta}\Gamma(b/2)} (2\pi R)^{-b/2} \frac{1}{(1 + T^{-1-b/\beta}|x|^{b+\beta})} \int_{\mathbb{R}_+} \frac{r^{b-1}}{1 + r^{b+\beta}} e^{-\frac{T^{2/\beta}}{2R}r^2}\mathrm{d}r,
\end{aligned}$$

where in the third line we have used the fact that $b + \beta > 1$ and the convexity of $|x|^{b+\beta}$). Using the assumption $(\log n)^{-c} \leqslant C R^{\beta/2} T^{-1}$, we deduce that with some constants $\bar{C}_b, C_b' > 0$ depending only on b,

$$\begin{aligned}
f_0(x) &\geqslant \bar{C}_b R^{-b/2} \frac{1}{(1 + T^{-1-b/\beta}|x|^{b+\beta})} \int_{\mathbb{R}_+} \frac{r^{b-1}}{1 + r^{b+\beta}} e^{-(C(\log n)^c)^{2/\beta} r^2/2}\mathrm{d}r \\
&\geqslant \bar{C}_b R^{-b/2} (C(\log n)^c)^{-b/\beta} \frac{1}{(1 + T^{-1-b/\beta}|x|^{b+\beta})} \int_{\mathbb{R}_+} \frac{r^{b-1}}{1 + (C(\log n)^c)^{-1-b/\beta} r^{b+\beta}} e^{-r^2/2}\mathrm{d}r \\
&\geqslant C_b' R^{-b/2} (\log n)^{-cb/\beta} \frac{1}{(1 + T^{-1-b/\beta}|x|^{b+\beta})} \int_{\mathbb{R}_+} \frac{r^{b-1} e^{-r^2/2}}{1 + r^{b+\beta}}\mathrm{d}r.
\end{aligned}$$

Since the last integral is finite and positive we obtain the result of the lemma for $R > 0, T > 0$. □

Due to Lemma 15 and the assumption $T(\log n)^{-1+\beta/2} \leqslant C_3 R^{\beta/2}$, the χ^2-divergence (27) satisfies

$$\begin{aligned}
\chi^2(\mathbb{P}_{j,l}, \mathbb{P}_0) \leqslant & C R^{b/2} (\log n)^{(1-\beta/2)b/\beta} \Big(\int_{\mathbb{R}^b} \big(f_{j,l}(x) - f_0(x)\big)^2 \mathrm{d}x \\
& + \frac{1}{T^{1+b/\beta}} \int_{\mathbb{R}^b} |x|^{\beta+b} \big(f_{j,l}(x) - f_0(x)\big)^2 \mathrm{d}x \Big).
\end{aligned} \tag{32}$$

We now bound separately the first and the second integral in (32). Using Plancherel's identity, we rewrite the first integral as

$$\int_{\mathbb{R}^b} \big(f_{j,l}(x) - f_0(x)\big)^2 \mathrm{d}x = \frac{1}{(2\pi)^b} \big\|\varphi_{j,l} - \varphi_0\big\|_{L^2}^2, \tag{33}$$

where $\varphi_{j,l}$ and φ_0 are the characteristic functions corresponding to $f_{j,l}$ and f_0, respectively. We now consider the difference of the characteristic exponents

$$\eta_j(u) := \log \varphi_{j,l}(u) - \log \varphi_0(u) = -\frac{1}{2}\langle u, \bar{A}_{j,l} u\rangle \big(1 - \mathcal{F}K(\delta_{n,p} u)\big),$$

where $\bar{A}_{j,l} = (\rho T \delta_{n,p}^{2-\beta}/b) A_{j,l}$. A first order Taylor expansion yields

$$\begin{aligned}\varphi_{j,l}(u) - \varphi_0(u) &= \eta_j(u)\varphi_0(u) \int_0^1 e^{t\eta_j(u)} \mathrm{d}t \\ &= \eta_j(u) e^{-R|u|^2/2} \int_0^1 \psi_{0,l}^{1-t}(u) \psi_{j,l}^t(u) \exp\big(-\frac{t}{2}\langle u, \bar{A}_{j,l} u\rangle\big) \mathrm{d}t.\end{aligned}$$

Due to the property $1 - \mathcal{F}K(\delta_{n,p} u) = 0$ for $|u| \leqslant \delta_{n,p}^{-1}$ and the elementary inequality $x^2 e^{|x|} \leqslant \exp(3|x|)$, $\forall x \in \mathbb{R}$, we obtain

$$\begin{aligned}\|\varphi_{j,l} - \varphi_0\|_{L^2}^2 &\leqslant \int_0^1 \Big\| \eta_j(u) \exp\Big(-\frac{R}{2}|u|^2 - \frac{t}{2}\langle u, \bar{A}_{j,l} u\rangle\Big)\Big\|_{L^2}^2 \mathrm{d}t \\ &\leqslant \frac{1}{4} \int_0^1 \int_{|u|>1/\delta_{n,p}} |\langle u, \bar{A}_{j,l} u\rangle|^2 \exp\Big(-R|u|^2 - t\langle u, \bar{A}_{j,l} u\rangle\Big) \mathrm{d}u \, \mathrm{d}t \\ &\leqslant \frac{1}{4} \int_{|u|>1/\delta_{n,p}} \exp\Big(-R|u|^2 + 3|\langle u, \bar{A}_{j,l} u\rangle|\Big) \mathrm{d}u \\ &\leqslant \frac{1}{4} \int_{|u|>1/\delta_{n,p}} \exp\Big(-(R - 3\rho T \delta_{n,p}^{2-\beta})|u|^2\Big) \mathrm{d}u,\end{aligned}$$

where we have used the bound $\|A_{j,l}\|_\infty \leqslant b$. Finally, we choose $\rho > 0$ sufficiently small to satisfy $3\rho T \delta_{n,p}^{2-\beta} \leqslant R/4$. Then,

$$\|\varphi_{j,l} - \varphi_0\|_{L^2}^2 \leqslant \frac{1}{4} e^{-R/(4\delta_{n,p}^2)} \int_{|u|>\delta_{n,p}^{-1}} e^{-R|u|^2/2} \mathrm{d}u \leqslant \frac{1}{4}\Big(\frac{2\pi}{R}\Big)^{b/2} e^{-R/(4\delta_{n,p}^2)}. \tag{34}$$

To take into account the first factor in (32), we note that the definition of $\delta_{n,p}$ in (22) imply

$$R^{b/2} (\log n)^{(1-\beta/2)b/\beta} R^{-b/2} e^{-R/(12\delta_{n,p}^2)} \leqslant (\log n)^{(1-\beta/2)b/\beta} \Big(\frac{\rho' \log(1 + c_1' bp/k)}{n}\Big)^{1/2}, \tag{35}$$

where the last expression is uniformly bounded by a constant. Combining this remark with (33) and (34), we finally get that there is a constant $C > 0$ such that

$$R^{b/2} (\log n)^{(1-\beta/2)b/\beta} \int_{\mathbb{R}^b} \big(f_{j,l}(x) - f_0(x)\big)^2 \mathrm{d}x \leqslant C \exp\Big(-\frac{R}{6\delta_{n,p}^2}\Big). \tag{36}$$

To bound the second integral in (32) we use the following proposition proved in Supplement A.

Proposition 16. *There is a constant $C > 0$ depending only on the kernel K and on b such that, for all $\beta \in (0, 2)$ and $j = 1, \dots, M$, $l = 1, \dots, L$,*

$$\int_{\mathbb{R}^b} \xi_b |x|^{\beta+b} \big(f_{j,l}(x) - f_0(x)\big)^2 \mathrm{d}x \leqslant C(1 \vee R^{\beta/2}) \exp\Big(-\frac{R}{5\delta_{n,p}^2}\Big). \tag{37}$$

This proposition and the assumption $T(\log n)^{c_*} \geqslant 1 \vee R^{\beta/2}$ yield, via an argument analogous to (35), that

$$\begin{aligned}&(\log n)^{(1-\beta/2)b/\beta} \frac{R^{b/2}}{T^{1+b/\beta}} \int_{\mathbb{R}^b} |x|^{\beta+b} \big(f_{j,l}(x) - f_0(x)\big)^2 \mathrm{d}x \\ &\quad \leqslant C' (\log n)^{(1-\beta/2)b/\beta} \frac{R^{b/2} \vee R^{(b+\beta)/2}}{T^{(b+\beta)/\beta}} \exp\Big(-\frac{R}{5\delta_{n,p}^2}\Big) \leqslant C \exp\Big(-\frac{R}{6\delta_{n,p}^2}\Big)\end{aligned} \tag{38}$$

where $C, C' > 0$ are constants. Combining (32), (36) and (38) and using the definition of $\delta_{n,p}$ in (22), we find

$$\chi^2(\mathbb{P}_{j,l}, \mathbb{P}_0) \leqslant C \exp\Big(- \frac{R}{6\delta_{n,p}^2}\Big) \leqslant C \frac{\rho' \log(1 + c_1' bp/k)}{n} \leqslant \frac{C\rho' \log M}{C_1' kn} := \frac{\rho'' \log M}{kn}, \tag{39}$$

where the last inequality follows from (21). Taking into account (26) and (39) we get

$$\chi^2(\mathbb{P}_{\Sigma_j,\psi_j}^{\otimes n}, \mathbb{P}_{\Sigma_0,\psi_0}^{\otimes n}) \leqslant \exp\Big(Ln \max_l \chi^2(\mathbb{P}_{j,l}, \mathbb{P}_0)\Big) - 1 \leqslant \exp\Big(\frac{p\rho'' \log M}{bk}\Big) - 1 \leqslant M^{2\rho''/b} - 1, \tag{40}$$

where we have used that, by construction, $L = p/b$ and $p \leqslant 2k$. Finally, choose ρ' (and thus ρ'') small enough to guarantee that $M^{2\rho''/b} - 1 \leqslant M/3$. Hence, condition (ii) is verified, which concludes the proof of the theorem. □

Appendix A: Proof of Theorem 9

For the later reference we set $\xi_U := \inf_{|u| \leqslant U/2} |\varphi(u)|$ and introduce the events

$$\Omega(u) := \big\{ |\varphi_n(u) - \varphi(u)| \leqslant |\varphi(u)|/2 \big\}, \qquad u \in \mathbb{R}^p.$$

Due to the decomposition (2), we obtain the bound

$$\begin{aligned}
\|\mathcal{R}_n\|_\infty &\leqslant S_n^{(1)} + S_n^{(2)} + D_n, \qquad (41)\\
\text{where } S_n^{(1)} &:= \int_{\mathbb{R}^p} \big| \operatorname{Re}\big(\log \varphi_n(u) \mathbb{1}_{\{|\varphi_n(u)| \geqslant \iota\}} - \log \varphi(u) \big) \big| \mathbb{1}_{\Omega(u)} \frac{\|\Theta(u)\|_\infty}{|u|^2} w_U(u) \mathrm{d}u,\\
S_n^{(2)} &:= \int_{\mathbb{R}^p} \big| \log |\varphi_n(u)| \mathbb{1}_{\{|\varphi_n(u)| \geqslant \iota\}} - \log |\varphi(u)| \big| \mathbb{1}_{\Omega(u)^c} \frac{\|\Theta(u)\|_\infty}{|u|^2} w_U(u) \mathrm{d}u,\\
D_n &:= \int_{\mathbb{R}^p} \big| \log |\psi(u)| \big| \frac{\|\Theta(u)\|_\infty}{|u|^2} w_U(u) \mathrm{d}u.
\end{aligned}$$

Here, $S_n^{(i)}$ are stochastic error terms and D_n is a deterministic error term. Using the decay of $\psi \in \mathcal{H}_\beta'(T)$, the form of the support of w_U and the fact that $\|\Theta(u)\|_\infty = 1$, we obtain

$$D_n \leqslant 16 U^{-2} \sup_{|u| \leqslant U/2} \big| \log |\psi(u)| \big| \int_{\mathbb{R}^p} \frac{w(v)}{|v|^2} \mathrm{d}v \leqslant C(w) T U^{-(2-\beta)} \tag{42}$$

for a constant $C(w) > 0$ depending only on w.

To bound $S_n^{(1)}$ in (41), we first note that we have on $\Omega(u)$ under the assumption $\xi_U \geqslant 1/\sqrt{n}$

$$|\varphi_n(u)| \geqslant |\varphi(u)| - |\varphi_n(n) - \varphi(u)| \geqslant |\varphi(u)|/2 \geqslant \frac{1}{2\sqrt{n}} = \iota,$$

for all u in the support of w_U. Thus, the indicator function $\mathbb{1}_{\{|\varphi_n(u)| \geqslant \iota\}}$ in $S_n^{(1)}$ can be omitted. Linearizing the logarithm yields

$$\log \varphi_n(u) - \log \varphi(u) = \log\Big(\frac{\varphi_n(u) - \varphi(u)}{\varphi(u)} + 1 \Big) = \frac{\varphi_n(u) - \varphi(u)}{\varphi(u)} + r_n(u)$$

with a residual r_n satisfying on $\Omega(u)$

$$|r_n(u)| \leqslant \bar{c} \Big| \frac{\varphi_n(u) - \varphi(u)}{\varphi(u)} \Big|^2 \tag{43}$$

where $\bar{c} > 0$ is a constant. Hence, we have

$$S_n^{(1)} \leqslant L_n + T_n \tag{44}$$

where

$$L_n := \int_{\mathbb{R}^p} \frac{|\varphi_n(u)-\varphi(u)|}{|u|^2|\varphi(u)|} w_U(u)\mathrm{d}u, \quad T_n := \int_{\mathbb{R}^p} \frac{|r_n(u)|}{|u|^2}\mathbb{1}_{\Omega(u)} w_U(u)\mathrm{d}u.$$

Here, L_n is the linearized stochastic error and T_n is a remainder. By the Cauchy-Schwarz inequality

$$L_n \leqslant \frac{16}{U^2\xi_U}\int_{\mathbb{R}^p}|\varphi_n(u)-\varphi(u)|w_U(u)\mathrm{d}u \leqslant \frac{16\bar{\varkappa}_w^{1/2}}{U^2\xi_U} Z \tag{45}$$

with $\bar{\varkappa}_w = \|w\|_{L^1}$ and

$$Z = Z(Y_1,\dots,Y_n) = \Big(\int_{\mathbb{R}^p}|\varphi_n(u)-\varphi(u)|^2 w_U(u)\mathrm{d}u\Big)^{1/2}.$$

Similarly, we deduce from (43)

$$T_n \leqslant \frac{16\bar{c}}{U^2}\int_{\mathbb{R}^p}\frac{|\varphi_n(u)-\varphi(u)|^2}{|\varphi(u)|^2} w_U(u)\mathrm{d}u \leqslant \frac{16\bar{c}}{U^2\xi_U^2} Z^2. \tag{46}$$

Note that Z satisfies the bounded difference condition

$$\forall Y_i, Y_i' \in \mathbb{R}^p: \quad |Z(Y_1,\dots,Y_{i-1},Y_i',Y_{i+1},\dots,Y_n) - Z(Y_1,\dots,Y_n)| \leqslant 2\bar{\varkappa}_w^{1/2}/n$$

By the bounded difference inequality [29, Theorem 3.3.14] we get $\mathbb{P}(Z \geqslant \mathbb{E}(Z)+t) \leqslant \exp(-\frac{nt^2}{4\bar{\varkappa}_w})$, for all $t>0$. Since $\mathbb{E}(Z) \leqslant (\bar{\varkappa}_w/n)^{1/2}$ this implies

$$\mathbb{P}\Big(Z \geqslant \frac{\bar{\varkappa}_w^{1/2}}{\sqrt{n}}(2\gamma+1)\Big) \leqslant e^{-\gamma^2}, \quad \forall \gamma>0.$$

Using (45),(46), and the assumption that $\gamma \geqslant 1$ we find that there exists a numerical constant $c_1^*>0$ such that

$$\mathbb{P}\Big(S_n^{(1)} \geqslant \frac{c_1^*\bar{\varkappa}_w\gamma}{U^2\xi_U\sqrt{n}}\Big(1+\frac{\gamma}{\xi_U\sqrt{n}}\Big)\Big) \leqslant 2e^{-\gamma^2}.$$

Since $\xi_U \leqslant 1$, this implies

$$\mathbb{P}\Big(S_n^{(1)} \geqslant \frac{2c_1^*\bar{\varkappa}_w\gamma^2}{U^2\xi_U^2\sqrt{n}}\Big) \leqslant 2e^{-\gamma^2}. \tag{47}$$

Using lower bounds ι and ξ_U for $|\varphi_n(u)|$ and $|\varphi(u)|$, respectively, and applying the elementary bound $\mathbb{1}_{\{a>1\}} < a$ for any $a>0$, the term $S_n^{(2)}$ in (41) is bounded as follows:

$$\begin{aligned}
S_n^{(2)} &\leqslant \int_{\mathbb{R}^p}\big(\log\iota^{-1}+\log\xi_U^{-1}\big)\mathbb{1}_{\Omega(u)^c}\frac{w_U(u)}{|u|^2}\mathrm{d}u\\
&\leqslant 2\int_{\mathbb{R}^p}\big(\log\iota^{-1}+\log\xi_U^{-1}\big)\frac{|\varphi_n(u)-\varphi(u)|}{|\varphi(u)|}\frac{w_U(u)}{|u|^2}\mathrm{d}u\\
&\leqslant \frac{32(2\log\xi_U^{-1}+\log 2)}{U^2\xi_U}\int_{\mathbb{R}^p}|\varphi_n(u)-\varphi(u)|w_U(u)\mathrm{d}u \leqslant \frac{32\sqrt{\bar{\varkappa}_w}}{U^2\xi_U^2}Z.
\end{aligned}$$

Hence, for some numerical constant $c_2^*>0$ we have

$$\mathbb{P}\Big(S_n^{(2)} \geqslant \frac{c_2^*\bar{\varkappa}_w}{U^2\xi_U^2\sqrt{n}}\gamma\Big) \leqslant e^{-\gamma^2}, \quad \forall\gamma>0. \tag{48}$$

Combining (41), (42), (47) and (48) and using the fact that $\gamma \geqslant 1$ we obtain

$$\mathbb{P}\Big(\|\mathcal{R}_n\|_\infty \geqslant \frac{(2c_1^*+c_2^*)\bar{\varkappa}_w\gamma^2}{U^2\xi_U^2\sqrt{n}} + C(w)TU^{-2+\beta}\Big) \leqslant 3e^{-\gamma^2}.$$

Finally, we use the bound $\xi_U \geqslant \exp(-\|\Sigma\|_\infty U^2/8 - 2TU^\beta)$ that is shown similarly to the analogous bound in the proof of Theorem 1.

Appendix B: Proofs for Section 5.3

Proof of Lemma 10. By Taylor's formula we have for some $\xi \in [0,1]$ that

$$\begin{aligned} R(u) :=& \eta^{-1}\big(-\log|\varphi|(u)\big) - \langle u, \Sigma u\rangle - \frac{\log|\psi|(u)}{\eta'(\langle u,\Sigma u\rangle)} \\ =& \frac{(\log|\psi|(u))^2}{2}(\eta^{-1})''\big(\eta\big(\langle u,\Sigma u\rangle\big) - \xi\log|\psi|(u)\big) \end{aligned}$$

Since $(\eta^{-1})''(x) = -\eta''(\eta^{-1}(x))/\eta'(\eta^{-1}(x))^3$ and thus $|(\eta^{-1})''(x)| \leqslant T|\eta^{-1}(x)|^{-1}|\eta'(\eta^{-1}(x))|^{-2}$ we have

$$|R(u)| \leqslant \frac{T|\log|\psi|(u)|^2}{2|g(u,\xi)||\eta'(\eta^{-1}(\eta(\langle u,\Sigma u\rangle) - \xi\log|\psi|(u)))|^2} \tag{49}$$

with $g(u,\xi) := \eta^{-1}(\eta(\langle u,\Sigma u\rangle) - \xi\log|\psi|(u))$. Since η^{-1} is non-negative and monotone increasing and $\log|\psi(u)| \leqslant 0$, we have

$$|g(u,\xi)| = g(u,\xi) \geqslant \eta^{-1}\big(\eta(\langle u,\Sigma u\rangle)\big) = \langle u,\Sigma u\rangle.$$

Another Taylor expansion for the second term in the denominator in (49) yields for some $\xi' \in [0,1]$

$$\begin{aligned} &\eta'(\eta^{-1}(\eta\big(\langle u,\Sigma u\rangle\big) - \xi\log|\psi|(u)) \\ &\quad= \eta'(\langle u,\Sigma u\rangle) - \xi(\log|\psi(u)|)\,(\eta'\circ\eta^{-1})'(\eta(\langle u,\Sigma u\rangle) - \xi'\log|\psi|(u)) \\ &\quad= \eta'(\langle u,\Sigma u\rangle) + \xi(\log|\psi(u)|)\Big(\frac{\eta''\circ\eta^{-1}}{\eta'\circ\eta^{-1}}\Big)(\eta(\langle u,\Sigma u\rangle) - \xi'\log|\psi|(u)) \\ &\quad\geqslant \eta'(\langle u,\Sigma u\rangle) - T|\log|\psi|(u)|/g(u,\xi') \\ &\quad\geqslant \eta'(\langle u,\Sigma u\rangle)\big(1 - T^2(1+|u|)^\beta/\langle u,\Sigma u\rangle\big). \end{aligned} \tag{50}$$

If $|u| \geqslant (2^{\beta+1}T^2/\lambda_{min})^{1/(2-\beta)}$, then we conclude

$$|R(u)| \leqslant \frac{2|\log|\psi|(u)|^2}{\langle u,\Sigma u\rangle\eta'(\langle u,\Sigma u\rangle)^2} \leqslant \frac{4T^2}{\lambda_{min}}|u|^{2\beta-2}. \qquad \square$$

Proof of Proposition 11. Due to Lemma 10 and the mean value theorem, the estimation error can be bounded for any $U \geqslant (2^{\beta+1}T^2/\lambda_{min})^{1/(2-\beta)}$ by

$$\begin{aligned} |\hat\sigma^\Phi_{i,i} - \sigma_{i,i}| &\leqslant U^{-2}\Big|\eta^{-1}\big(-\log|\varphi_n(Uu^{(i)})|\big) - \eta^{-1}\big(-\log|\varphi(Uu^{(i)})|\big)\Big| + \big(2T + \frac{4T^2}{\lambda_{min}}U^{-2+\beta}\big)U^{-2+\beta} \\ &= \frac{|S(Uu^{(i)})|}{U^2\eta'(\eta^{-1}(-\log|\varphi(Uu^{(i)})| + \xi S(Uu^{(i)})))} + \big(2T+2\big)U^{-2+\beta} \end{aligned}$$

for some $\xi \in [0,1]$ and $S(u) = \log|\varphi_n(u)| - \log|\varphi(u)|$ from Lemma 12. As in (50) (for any u with $g(u,\xi) \geqslant \langle u,\Sigma u\rangle \geqslant 1$), we deduce

$$\eta'\big(\eta^{-1}(-\log|\varphi(u)| + \xi S(u))\big) \geqslant \eta'(\langle u,\Sigma u\rangle) - T(|\log|\psi|(u)| + S(u)|.$$

On the event $\{|\log|\psi|(u)| + S(u)| \leqslant \eta'(U^2\sigma_{ii})/2\}$, we thus obtain

$$|\hat\sigma^\Phi_{i,i} - \sigma_{i,i}| \leqslant \frac{2|S(Uu^{(i)})|}{U^2\eta'(U^2\sigma_{ii})} + \big(2T+2\big)U^{-2+\beta}.$$

From this line, the argument is analogous to the proof of Theorem 1. $\square$

Acknowledgements

D. Belomestny acknowledges the financial support from the Russian Academic Excellence Project "5-100" and from Deutsche Forschungsgemeinschaft (DFG) through the SFB 823 "Statistical modelling of nonlinear dynamic processes". M. Trabs gratefully acknowledges the financial support by the DFG research fellowship TR 1349/1-1. The work of A.B. Tsybakov was supported by GENES and by the French National Research Agency (ANR) under the grants IPANEMA (ANR-13-BSH1-0004-02) and Labex Ecodec (ANR-11-LABEX-0047). This work has been started while M.T. was affiliated to the Université Paris-Dauphine.

Supplementary Material

Supplement A: A bound for the L^2-distance of certain densities and their derivatives (doi: COMPLETED BY THE TYPESETTER). We prove Proposition 16 which is needed to show the lower bound for the covariance estimation in the deconvolution model. More precisely, a bound for the L^2-distance of the certain densities and their derivatives is proven.

References

[1] Belomestny, D. and Reiß, M. (2006). Spectral calibration of exponential Lévy models. *Finance Stoch.*, 10(4):449–474.

[2] Belomestny, D. and Trabs, M. (2017). Low-rank diffusion matrix estimation for high-dimensional time-changed Lévy processes. *Ann. Inst. Henri Poincaré Probab. Stat.* To appear. ArXiv preprint arXiv:1510.04638.

[3] Bickel, P. J. and Levina, E. (2008a). Covariance regularization by thresholding. *Ann. Statist.*, 36(6):2577–2604.

[4] Bickel, P. J. and Levina, E. (2008b). Regularized estimation of large covariance matrices. *Ann. Statist.*, 36(1):199–227.

[5] Butucea, C. and Matias, C. (2005). Minimax estimation of the noise level and of the deconvolution density in a semiparametric convolution model. *Bernoulli*, 11:309–340.

[6] Butucea, C., Matias, C., and Pouet, C. (2008). Adaptivity in convolution models with partially known noise distribution. *Electron. J. Stat.*, 2:897–915.

[7] Butucea, C. and Tsybakov, A. B. (2008). Sharp optimality in density deconvolution with dominating bias. i. *Theory Probab. Appl.*, 52(1):24–39.

[8] Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc.*, 106(494):672–684.

[9] Cai, T. T., Ren, Z., and Zhou, H. H. (2016). Estimating structured high-dimensional covariance and precision matrices: optimal rates and adaptive estimation. *Electron. J. Stat.*, 10(1):1–59.

[10] Cai, T. T. and Zhang, A. (2016). Minimax rate-optimal estimation of high-dimensional covariance matrices with incomplete data. *arXiv preprint arXiv:1605.04358*.

[11] Cai, T. T., Zhang, C.-H., and Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.*, 38(4):2118–2144.

[12] Cai, T. T. and Zhou, H. H. (2012). Minimax estimation of large covariance matrices under ℓ_1-norm. *Statist. Sinica*, pages 1319–1349.

[13] Carroll, R. J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.*, 83(404):1184–1186.

[14] Comte, F. and Lacour, C. (2011). Data-driven density estimation in the presence of additive noise with unknown distribution. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 73(4):601–627.

[15] Cressie, N. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. John Wiley & Sons.

[16] Dattner, I., Reiß, M., and Trabs, M. (2016). Adaptive quantile estimation in deconvolution with unknown error distribution. *Bernoulli*, 22(1):143–192.

[17] Delaigle, A. and Hall, P. (2016). Methodology for non-parametric deconvolution when the error distribution is unknown. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 78(1):231–252.

[18] Delaigle, A., Hall, P., and Meister, A. (2008). On deconvolution with repeated measurements. *Ann. Statist.*, 36(2):665–685.
[19] Delaigle, A. and Meister, A. (2011). Nonparametric function estimation under Fourier-oscillating noise. *Statist. Sinica*, 21(3):1065–1092.
[20] Eckle, K., Bissantz, N., and Dette, H. (2016). Multiscale inference for multivariate deconvolution. *arXiv preprint arXiv:1611.05201*.
[21] El Karoui, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.*, 36(6):2717–2756.
[22] Fan, J. (1991). On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems. *Ann. Statist.*, 19(3):1257–1272.
[23] Fan, J., Li, Y., and Yu, K. (2012). Vast volatility matrix estimation using high-frequency data for portfolio selection. *J. Amer. Statist. Assoc.*, 107(497):412–428.
[24] Fan, J., Liao, Y., and Liu, H. (2016). An overview of the estimation of large covariance and precision matrices. *Econom. J.*, 19(1):C1–C32.
[25] Fan, J., Liao, Y., and Mincheva, M. (2011). High-dimensional covariance matrix estimation in approximate factor models. *Ann. Statist.*, 39(6):3320–3356.
[26] Fan, J., Liao, Y., and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 75(4):603–680. With 33 discussions by 57 authors and a reply by Fan, Liao and Mincheva.
[27] Fang, K., Kotz, S., and Ng, K. W. (1990). *Symmetric multivariate and related distributions*, volume 36. Chapman & Hall/CRC.
[28] Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.
[29] Gine, E. and Nickl, R. (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge Univ. Press, Cambridge.
[30] Jacod, J. and Reiß, M. (2014). A remark on the rates of convergence for integrated volatility estimation in the presence of jumps. *Ann. Statist.*, 42(3):1131–1144.
[31] Johannes, J. (2009). Deconvolution with unknown error distribution. *Ann. Statist.*, 37(5A):2301–2323.
[32] Kappus, J. and Mabon, G. (2014). Adaptive density estimation in deconvolution problems with unknown error distribution. *Electron. J. Stat.*, 8(2):2879–2904.
[33] Koltchinskii, V., Lounici, K., and Tsybakov, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, 39(5):2302–2329.
[34] Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.*, 37(6B):4254–4278.
[35] Lepski, O. and Willer, T. (2017a). Estimation in the convolution structure density model. Part I: oracle inequalities. *arXiv preprint arXiv:1704.04418*.
[36] Lepski, O. and Willer, T. (2017b). Estimation in the convolution structure density model. Part II: adaptation over the scale of anisotropic classes. *arXiv preprint arXiv:1704.04420*.
[37] Lounici, K. (2014). High-dimensional covariance matrix estimation with missing observations. *Bernoulli*, 20(3):1029–1058.
[38] Low, M. G. (1997). On nonparametric confidence intervals. *Ann. Statist.*, 25(6):2547–2554.
[39] Masry, E. (1993). Strong consistency and rates for deconvolution of multivariate densities of stationary processes. *Stochastic Process. Appl.*, 47(1):53–74.
[40] Matias, C. (2002). Semiparametric deconvolution with unknown noise variance. *ESAIM Probab. Stat*, 6:271–292.
[41] Meister, A. (2008). Deconvolution from Fourier-oscillating error densities under decay and smoothness restrictions. *Inverse Problems*, 24(1):015003, 14.
[42] Neumann, M. H. (1997). On the effect of estimating the error density in nonparametric deconvolution. *J. Nonparametr. Statist.*, 7(4):307–330.
[43] Rigollet, P. and Tsybakov, A. (2011). Exponential screening and optimal rates of sparse estimation. *Ann. Statist.*, 39(2):731–771.
[44] Rigollet, P. and Tsybakov, A. B. (2012). Comment:" minimax estimation of large covariance matrices under ℓ_1-norm". *Statist. Sinica*, 22(4):1358–1367.

[45] Rothman, A. J. (2012). Positive definite estimators of large covariance matrices. *Biometrika*, 99(3):733–740.

[46] Rothman, A. J., Levina, E., and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *J. Amer. Statist. Assoc.*, 104(485):177–186.

[47] Sanandaji, B. M., Tascikaraoglu, A., Poolla, K., and Varaiya, P. (2015). Low-dimensional models in spatio-temporal wind speed forecasting. In *American Control Conference (ACC), 2015*, pages 4485–4490. IEEE.

[48] Sato, K.-i. (2013). *Lévy processes and infinitely divisible distributions*, volume 68 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge. Translated from the 1990 Japanese original, Revised edition of the 1999 English translation.

[49] Tao, M., Wang, Y., and Zhou, H. H. (2013). Optimal sparse volatility matrix estimation for high-dimensional Itô processes with measurement errors. *Ann. Statist.*, 41(4):1816–1864.

[50] Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

[51] Tsybakov, A. B. (2013). Aggregation and high-dimensional statistics. Saint Flour Lecture notes.