Journal of Multivariate Analysis 196 (2023) 105184

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Tyler's and Maronna's M-estimators: Non-asymptotic concentration results

Elad Romanov [a,*], Gil Kur [b], Boaz Nadler [c]

[a] Department of Statistics, Stanford University, United States of America
[b] Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, United States of America
[c] Faculty of Mathematics and Computer Science, Weizmann Institute of Science, Israel

ARTICLE INFO

Article history:
Received 21 June 2022
Received in revised form 23 March 2023
Accepted 24 March 2023
Available online 28 March 2023

AMS 2020 subject classifications:
primary 62H12
secondary 15B52

Keywords:
Maronna's M-estimator
Robust covariance estimation
Tyler's M-estimator

ABSTRACT

Tyler's and Maronna's M-estimators, as well as their regularized variants, are popular robust methods to estimate the scatter or covariance matrix of a multivariate distribution. In this work, we study the non-asymptotic behavior of these estimators, for data sampled from a distribution that satisfies one of the following properties: (1) independent sub-Gaussian entries, up to a linear transformation; (2) log-concave distributions; (3) distributions satisfying a convex concentration property. Our main contribution is the derivation of tight non-asymptotic concentration bounds of these M-estimators around a suitably scaled version of the data sample covariance matrix. Prior to our work, non-asymptotic bounds were derived only for Elliptical and Gaussian distributions. Our proof uses a variety of tools from non asymptotic random matrix theory and high dimensional geometry. Finally, we illustrate the utility of our results on two examples of practical interest: sparse covariance and sparse precision matrix estimation.

1. Introduction

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be n i.i.d. samples from a p-dimensional random variable X. The $p \times p$ covariance matrix Σ of X is a central quantity of interest in multiple applications [5,57]. In the classical regime with $n \gg p$, if the random variable X is not heavily tailed and there are no outliers, the empirical covariance matrix yields a relatively accurate estimator for Σ.

To deal with heavy tails and potential outliers, several robust estimators were proposed and studied theoretically. Two popular procedures, applicable when $p < n$, include Maronna's and Tyler's M-estimators [50,72]. Regularized variants, applicable also when $p > n$, were also proposed and studied [1,22,59,61]. These estimators have found use in multiple applications, ranging from signal processing and radar detection to finance, see for example [24,59,61]. We remark that in addition to the above, many other robust covariance estimators have been proposed and analyzed, see for example [19,21,28,30,37,39,51,54,55,76] and references therein.

In this work we study the properties of Tyler's and Maronna's M-estimators under several families of multivariate distributions. Our analysis is non-asymptotic and generalizes previous results, which were either asymptotic or limited to elliptical distributions. Before presenting our results, we first briefly describe these estimators and related prior work. For simplicity, we describe the estimators assuming X has zero mean, and discuss how to relax this assumption later on.

[*] Corresponding author.
E-mail address: eromanov@stanford.edu (E. Romanov).

https://doi.org/10.1016/j.jmva.2023.105184
0047-259X/© 2023 Elsevier Inc. All rights reserved.

Maronna's M-estimator (ME). One of the first proposals for a robust covariance estimator, introduced by Maronna [50], is defined as follows. Let $u : [0, \infty) \to (0, \infty)$ be a function that is strictly positive, non-increasing, continuous, and such that the accompanying function $\phi(x) = xu(x)$ is non-decreasing and bounded. Then, for $n \geq p$, Maronna's M-estimator (if exists) is a solution to the non-linear matrix equation

$$\hat{\Sigma}_{\text{Mar}} = \frac{1}{n}\sum_{i=1}^{n} w_i^{\text{Mar}} \mathbf{x}_i \mathbf{x}_i^\top , \qquad w_i^{\text{Mar}} = u\left(\frac{1}{p}\mathbf{x}_i^\top \hat{\Sigma}_{\text{Mar}}^{-1} \mathbf{x}_i\right) . \tag{1}$$

Maronna [50] proved that under certain deterministic conditions on the samples, (1) has a unique solution. Couillet et al. [25] considered a high dimensional asymptotic framework, wherein $n, p \to \infty$ with their ratio converging to a constant. Assuming that X has i.i.d. entries with sufficiently many finite moments, and that $\phi_\infty := \sup_x \phi(x) > 1$, they proved that (1) has a unique solution with probability tending to one.

Maronna's regularized M-estimator (MRE). For $p > n$, Ollila and Tyler [59] proposed the following generalization of Maronna's M-estimator: For regularization parameter $\alpha > 0$,

$$\hat{\Sigma}_{\text{MRE}} = \frac{1}{1+\alpha}\frac{1}{n}\sum_{i=1}^{n} w_i^{\text{MRE}} \mathbf{x}_i \mathbf{x}_i^\top + \frac{\alpha}{1+\alpha} \mathrm{I}_{p\times p} , \qquad w_i^{\text{MRE}} = u\left(\frac{1}{p}\mathbf{x}_i^\top \hat{\Sigma}_{\text{MRE}}^{-1} \mathbf{x}_i\right) . \tag{2}$$

As proven in [59, Theorem 1], for any $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and α, (2) has a unique solution.

Tyler's M-estimator (TE). Introduced by Tyler in [72], TE is defined as the solution (if exists) of

$$\hat{\Sigma}_{\text{Tyl}} = \frac{1}{n}\sum_{i=1}^{n} w_i^{\text{Tyl}} \mathbf{x}_i \mathbf{x}_i^\top , \qquad w_i^{\text{Tyl}} = \left(\frac{1}{p}\mathbf{x}_i^\top \hat{\Sigma}_{\text{Tyl}}^{-1} \mathbf{x}_i\right)^{-1} , \qquad \mathrm{Tr}(\hat{\Sigma}_{\text{Tyl}}) = p \, . \tag{3}$$

While Tyler's estimator may seem like a special case of Maronna's with $u(x) = 1/x$, this is not so since $u(\cdot)$ is singular at $x = 0$. For $n > p$, Kent and Tyler [41, Theorems 1 and 2] proved existence and uniqueness of TE under the condition that any linear subspace of $\mathbb{R}^p$ of dimension $1 \leq d \leq p - 1$ contains less than nd/p samples. For i.i.d. samples from a random vector X with a proper density in $\mathbb{R}^p$, this condition is satisfied with probability one.

Tyler's regularized M-estimator (TRE). Similarly to ME, Tyler's M-estimator does not exist for $p > n$. In recent years, several regularized variants were proposed [1,22,59,61,70]. We focus on the estimator proposed in [61]. Given a regularization parameter $\alpha > 0$, it is defined by

$$\hat{\Sigma}_{\text{TRE}} = \frac{1}{1+\alpha}\frac{1}{n}\sum_{i=1}^{n} w_i^{\text{TRE}} \mathbf{x}_i \mathbf{x}_i^\top + \frac{\alpha}{1+\alpha} \mathrm{I}_{p\times p} , \qquad w_i^{\text{TRE}} = \left(\frac{1}{p}\mathbf{x}_i^\top \hat{\Sigma}_{\text{TRE}}^{-1} \mathbf{x}_i\right)^{-1} . \tag{4}$$

By [59, Theorem 2], when $\alpha > p - 1$, (4) always admits a unique solution. When $\alpha \leq p - 1$, [59, Theorem 3] gave a deterministic sufficient and almost necessary condition for existence and uniqueness; for $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in general position, the condition holds for $\alpha > \max\left\{0, \frac{p}{n} - 1\right\}$. When X has a density, this condition appeared earlier in [61].

Prior work. Maronna's and Tyler's M-estimators and their variants, have been studied extensively, with a particular focus under elliptical distributions; see [1,6,22–26,33,41,56,58–61,68,75,76,78]. The present paper extends several works that studied these estimators in a high-dimensional regime, where the number of samples n and the dimension p are both large and comparable. Couillet et al. [25] studied the asymptotic behavior of ME in the joint limit $p, n \to \infty$ with their ratio tending to a constant. Assuming that X has independent entries with zero mean, unit variance and sufficiently many finite moments, they proved that after a suitable scaling, ME converges asymptotically in spectral norm to the sample covariance matrix. In [26], this analysis was extended to X having an elliptical distribution with a general scatter matrix. A similar asymptotic analysis for MRE appeared in [6]. Two variants of TRE were studied in [24], assuming X has an elliptical distribution. A key result of their analysis is that asymptotically, these M-estimators behave similarly to regularized sample covariance matrices with Gaussian measurements. Zhang et al. [78], studied TE assuming X is Gaussian distributed with identity covariance. They showed that as $n, p \to \infty$, the limiting the spectral distribution of a properly scaled TE is a Marčenko–Pastur law. Moreover, similar to our own results in the present paper, they proved a non-asymptotic deviation bound for the weights (3), showing that they are concentrated around some particular value. Relying on their results, [33] extended the analysis to cover TRE, assuming X is elliptically distributed. We remark that the proofs in [33,78] rely on properties specific to the Gaussian and elliptical distributions, and do not extend easily to other distributions.

Another recent work is [47], which derived nonasymptotic concentration results for the Stieltjes transform of the spectral distribution of certain regularized M-estimators. In the context of this work, they derived results only for the regularized variants of Maronna's M-estimator. Their results apply under rather broad distributional assumptions, requiring only a concentration of measure property; see also their related papers [46,48]. In contrast, our analysis of Tyler's M-estimators and the unregularized Maronna's M-estimator requires an additional *anti-concentration* property (the small ball property) for the random vector X. It is an interesting question whether it is possible to derive similar results without the SBP assumption.

Our contributions. As mentioned above, most of the literature on Maronna's and Tyler's M-estimators has focused on asymptotic results, establishing convergence as $n, p \to \infty$ without specifying rates. Other works, that derived non-asymptotic finite-n bounds, mostly considered Gaussian or elliptical distributions. This paper extends and generalizes these works in several directions. We present a non-asymptotic analysis of both Tyler's and Marrona's M-estimators, and their regularized variants, under three broad families of multivariate data distributions: (1) independent sub-Gaussian entries, up to a linear transformation; (2) log-concave distributions; (3) distributions satisfying a convex concentration property (CCP). Our main results are given in terms of non-asymptotic concentration bounds for the weights of the M-estimators (1)–(4) around some particular deterministic value. They imply that for these three families of distributions, Maronna's and Tyler's M-estimators behave similarly to a rescaled sample covariance matrix. In Section 4 we illustrate the utility of these results for two concrete examples of practical interest: sparse covariance and sparse precision matrix estimation.

2. Main results

Let $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^p$ be n samples of the form

$$\mathbf{x}_i = \Sigma_p^{1/2} \mathbf{y}_i \,, \tag{5}$$

where Σ_p is a strictly positive p-by-p matrix, and $\mathbf{y}_1, \dots, \mathbf{y}_n \in \mathbb{R}^p$ are n i.i.d. realizations of a zero mean isotropic random vector Y, namely

$$\mathbb{E}[Y] = \mathbf{0}\,, \quad \mathbb{E}[YY^\top] = \mathrm{I}_{p\times p}\,. \tag{6}$$

We assume that $Y = (Y_1, \dots, Y_p)^\top$ is a continuous random vector, and satisfies one of the following distributional assumptions, with the precise definitions deferred to Section 3:

1. [SG-IND]: The coordinates of Y are independent, sub-Gaussian (Definition 1) and have a bounded density. The constant $K > 0$ denotes a uniform bound on sub-Gaussian constants, and $C_0 > 0$ a bound on the densities.
2. [CCP-SBP]: Y satisfies the convex concentration property (CCP) with constant $K > 0$ and also the small-ball property (SBP) with constant $C_0 > 0$ (Definition 3).
3. [LC]: Y has a log-concave distribution (Definition 4).

Remark 1. While the assumption that X has zero mean may not hold in practice, it has been used extensively in previous studies of Tyler's and Maronna's M-estimators, cf. [6,25,33,78]. In Section 6 we discuss how this restriction may be removed for some of our results.

Remark 2. We emphasize that the three families of distributions considered above are all distinct in the sense that neither one is contained in another. In particular, it is known that an i.i.d. sub-Gaussian vector does not necessarily satisfy the CCP, see for example [34,36]. Furthermore, below (after the statement of Theorem 2) we mention examples of two distributions that satisfy one of [CCP-SBP], [LC] but not the other.

We consider Maronna's and Tyler's M-estimators $\hat{\Sigma}_{\mathrm{Mar}}$ and $\hat{\Sigma}_{\mathrm{Tyl}}$, as well as their regularized variants $\hat{\Sigma}_{\mathrm{MRE}}$ and $\hat{\Sigma}_{\mathrm{TRE}}$, all computed from $\mathbf{x}_1, \dots, \mathbf{x}_n$. The latter two estimators depend on the regularization parameter $\alpha > 0$, which we omit to simplify notation. We study the nonasymptotic properties of these estimators in the high-dimensional regime, where the number of samples n and the dimension p are both large and comparable. Their ratio is denoted by

$$\gamma = \frac{p}{n} \in (0, \infty)\,. \tag{7}$$

Similarly to [33,78], while our results are nonasymptotic in n (in the form of finite-n deviation bounds), they involve constants that may depend on γ, often in a complicated manner that we do not keep track of explicitly.

2.1. Concentration for the weights of Maronna's and Tyler's estimators

Our first two results show concentration for the weights of Maronna's and Tyler's M-estimators for $\gamma = p/n < 1$, hence $n - p = \Omega(p)$. Below, Ψ_p denotes (up to a universal constant) the Cheeger constant of the family of p-dimensional log-concave distributions; it is known that $\Psi_p \geq 1/\mathrm{polylog}(p)$ and conjectured that $\Psi_p = \Theta(1)$; see Section 3, (19).

We start with Maronna's estimator. For $\varepsilon > 0$, let $\mathcal{E}_{ME}(\varepsilon)$ be the event that (1) has a unique solution whose weights satisfy

$$\max_{1\leq i\leq n} \left| w_i^{\mathrm{Mar}} - 1/\phi^{-1}(1) \right| \leq \varepsilon\,. \tag{8}$$

Theorem 1 (*The Weights of Maronna's Estimator*). *Assume $\gamma < 1$, and that the functions $u(x)$, $\phi(x) = xu(x)$ of Maronna's M-estimator further satisfy:*

(a) $\phi_\infty := \lim_{x\to\infty} \phi(x) > 1$.
(b) *There is a unique* d_0 *such that* $\phi(d_0) = 1$*, namely,* ϕ *is strictly increasing at* $\phi^{-1}(1)$*. Moreover, the inverse map* ϕ^{-1} *is locally Lipschitz at* 1.
(c) $u(\cdot)$ *is locally Lipschitz at* $\phi^{-1}(1)$.

Then, there are constants $c, C, \varepsilon_0 > 0$ *that depend on the distribution of* Y *and on* γ*, such that for any* $\varepsilon < \varepsilon_0$*, the following holds.*

(i) *Assume that* Y *satisfies [LC]. Then* $\Pr(\mathcal{E}_{ME}(\varepsilon)^c) \le Cn^2 e^{-c(\Psi_p\sqrt{n})\varepsilon}$.
(ii) *Assume that* Y *satisfies either [SG-IND] or [CCP-SBP]. Then* $\Pr(\mathcal{E}_{ME}(\varepsilon)^c) \le Cn^2 e^{-cn\varepsilon^2}$.

Remark 3. When saying that a constant C depends on the distribution of Y, we mean that it may depend on the parameters $K, C_0 > 0$ above, whenever Y is distributed according to [SG-IND] or [CCP-SBP]. Specifically, in the bound above in Theorem 1, the constants c, C can be taken to be monotonic in the parameters K, C_0. The reason is that as one increases K and decreases C_0, the class of distributions considered ([SG-IND] or [LC]) becomes larger. However, we do not keep track of this dependence explicitly.

It is interesting to compare Theorem 1 to the results of [25]. Assuming that the random vector Y has i.i.d. entries with sufficiently many finite moments, they proved that asymptotically, as $n, p \to \infty$ with their ratio tending to a constant, almost surely $\max_{1\le i\le n} \left| w_i^{\text{Mar}} - 1/\phi^{-1}(1) \right| \to 0$, see [25, Eq. (6)]. Our analysis extends theirs in two aspects: (i) It holds also for random vectors Y that do not have independent entries, but instead satisfy certain multivariate concentration properties; (ii) Our results are non-asymptotic, in the form of concentration inequalities for the weights.

Our next theorem regards Tyler's M-estimator. To this end, denote by $\tau_p = \frac{1}{p}\mathrm{Tr}\Sigma_p$ the normalized trace of the population covariance matrix. For any $\varepsilon > 0$, let $\mathcal{E}_{TE}(\varepsilon)$ be the event that Tyler's estimator (3) exists uniquely, with weights that satisfy

$$\max_{1\le i\le n} \left| \tau_p w_i^{\text{Tyl}} - 1 \right| \le \varepsilon. \tag{9}$$

Theorem 2 (*The Weights of Tyler's Estimator*)**.** *There are constants* $c, C, \varepsilon_0 > 0$*, that depend on the distribution of* Y *and on* $\gamma < 1$*, such that for any* $\varepsilon < \varepsilon_0$ *the following hold.*

(i) *Assume that* Y *satisfies [LC]. Then* $\Pr(\mathcal{E}_{TE}(\varepsilon)^c) \le Cn^2 e^{-c\Psi_p \min\left\{\sqrt{n}\varepsilon,\, n^{1/4}\right\}}$.
(ii) *Assume that* Y *satisfies either [SG-IND] or [CCP-SBP]. Then* $\Pr(\mathcal{E}_{TE}(\varepsilon)^c) \le Cn^2 e^{-c\min\left\{n\varepsilon^2,\, n^{1/2}\right\}}$.

We remark that Zhang et al. [78] provided a proof of Theorem 2 in the specific case where $Y \sim \mathcal{N}(\mathbf{0}, \mathrm{I}_{p\times p})$. They stated a non-asymptotic bound for the weights: $\Pr\left(\max_{1\le i\le n} |w_i - 1| \ge \varepsilon\right) \le Cne^{-cn\varepsilon^2}$. (Note that their definition of the weights differs from ours by a constant factor, and therefore does not depend on τ_p.) This bound should be compared to Theorem 2 under Assumption [SG-IND]. Importantly, their proof contains a non-trivial gap, which we correct in the present paper.

Next, we illustrate Theorems 1 and 2 by the following numerical experiment. We generated i.i.d. samples according to either of the following two distributions for Y:

1. *Laplace:* All coordinates Y_i are i.i.d. Laplace, with density $\mathrm{Lap}(y) = \frac{1}{\sqrt{2}}\exp(-\sqrt{2}|y|)$. Hence, Y is isotropic, log-concave, but not sub-Gaussian.
2. *Permuted smoothed*: $Y = \frac{1}{\sqrt{1+\sigma^2}}A + \frac{\sigma}{\sqrt{1+\sigma^2}}Z$, where $A \in \{\pm 1\}^p$ is uniform in the set $\left\{a \in \{\pm1\}^p : \sum_{i=1}^p a_i = 0\right\}$, $Z \sim \mathcal{N}(\mathbf{0}, \mathrm{I}_{p\times p})$ and $\sigma = 0.01$ is the smoothing level. The entries of A are clearly dependent; nonetheless, by a classical result of Maurey [52], it can be shown to satisfy the CCP. Consequently, Y satisfies the CCP and SBP.

We compute Tyler's and Maronna's M-estimators from $n = 2p$ samples, for the latter using $u(x) = 2/(1+x)$. Fig. 1 shows the deviation of the corresponding weights from their limiting value $w^* = 1$. We present on a log–log scale both the ℓ_∞ deviation $\max_{1\le i\le n} |\hat{w}_i - w^*|$, and the root mean squared error (RMSE) $\sqrt{\frac{1}{n}\sum_{i=1}^n (\hat{w}_i - w^*)^2}$, as a function of the dimension p. The slope of either line is approximately $\frac{1}{2}$, consistent with Theorems 1 and 2.

Remark 4. Observe that in Theorem 1, the weights of Maronna's estimator do not depend on the population covariance Σ_p. This is because the weights are preserved under an arbitrary linear full-rank transformation of the data. Let $\mathbf{x}_1, \dots, \mathbf{x}_n$ and consider the transformed measurements $\tilde{\mathbf{x}}_i = A\mathbf{x}_i$, where $A \in \mathbb{R}^{p\times p}$ is full-rank. Denote the corresponding Marrona's estimators (1) by $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n w_i \mathbf{x}_i \mathbf{x}_i^\top$ and $\tilde{\Sigma} = \frac{1}{n}\sum_{i=1}^n \tilde{w}_i \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^\top$. One may verify that $\tilde{\Sigma} = A\hat{\Sigma}A^\top$, hence $w_i = \tilde{w}_i$; setting $A = \Sigma_p^{-1/2}$, we deduce that the weights w_i^{Mar} do not depend on the covariance matrix of X.

In contrast, consistent with Theorem 2, the weights of Tyler's estimator do depend on Σ_p. The reason is that while the linearly transformed estimator $A\hat{\Sigma}A^\top$ does solve the unconstrained (3), corresponding to the transformed measurements (similarly to Maronna's estimator), one needs to rescale the weights due to the constraint $\mathrm{Tr}(\tilde{\Sigma}) = p$.

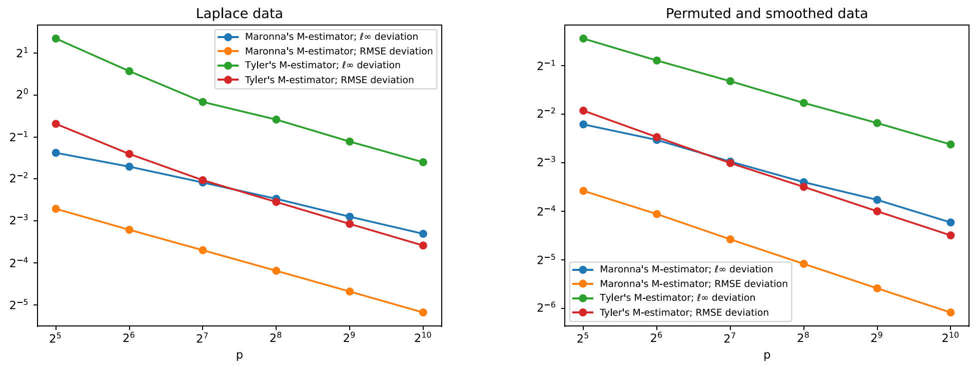

Fig. 1. The empirical deviation of the weights, plotted in log–log scale. Each point on the graph corresponds to the average of 50 random repetitions.

2.2. *Regularized estimators*

We next consider the regularized variants of Maronna's and Tyler's M-estimators. As mentioned in Section 1, MRE exists uniquely for all $\gamma, \alpha > 0$, whereas for any $\gamma > 0$, TRE is only guaranteed to exist uniquely when $\alpha = \alpha(\gamma) > 0$ is sufficiently large.

The regularization term $\frac{\alpha}{1+\alpha}\mathrm{I}_{p\times p}$ in (2) and (4) shrinks the solution towards the identity matrix. As a result, the weights of MRE and TRE, and our deviation bounds for them, depend on the underlying population matrix Σ_p. Accordingly, throughout this section we operate under the following additional constraints on Σ_p, so to ensure that it has the same scale as the identity matrix. Specifically, for constants $s_{\max} \geq \tau > 0$, we assume:

- Bounded operator norm:

$$\|\Sigma_p\| \leq s_{\max}\,. \tag{10}$$

- Lower bound on total energy:

$$\tau_p = \frac{1}{p}\mathrm{Tr}\,\Sigma_p \geq \tau\,. \tag{11}$$

Towards stating our results, given a function $\phi(x)$, we define the following two functions $Q, F : \mathbb{R}_+ \to \mathbb{R}_+$:

$$Q(d) = \frac{1}{p}\mathbb{E}\mathrm{Tr}\,\Sigma_p\left(\phi(d)\sum_{i=1}^{n-1}\mathbf{x}_i\mathbf{x}_i^\top + \alpha d\mathrm{I}_{p\times p}\right)^{-1}, \quad F(d) = (1+\alpha)\frac{Q(d)}{1+\gamma\phi(d)Q(d)}\,. \tag{12}$$

As we shall see below, the functions Q and F play a decisive role in determining the weights of MRE and TRE. A key quantity is the solution $d^* > 0$ to the following "master equation":

$$F(d^*) = 1\,. \tag{13}$$

Later, we show that if $u(\cdot)$ is non-increasing and $\phi(x) = xu(x)$ is non-decreasing, then (13) admits a unique solution.

The theorem below regards Maronna's regularized M-estimator. For $\varepsilon > 0$, let $\mathcal{E}_{MRE}(\varepsilon)$ be the event that MRE exists uniquely with weights that satisfy

$$\max_{1\leq i\leq n}|w_i^{\mathrm{MRE}} - u(d^*)| \leq \varepsilon\,. \tag{14}$$

Theorem 3 (*The Weights of MRE*). *Let $\alpha > 0$ and u be a Lipschitz continuous on any compact sub-interval of $[0, \infty)$. Then (13) has a unique solution which satisfies $d^* \in [\underline{d}, \overline{d}]$, where the constants $0 < \underline{d} \leq \overline{d}$ depend on the distribution of Y, γ, α, $s_{\max}$ and τ.*

Furthermore, there are constants $c, C, \varepsilon_0 > 0$, that depend on the distribution of Y, γ, α, $s_{\max}$ and τ, such that for all $\varepsilon < \varepsilon_0$,

(i) *Assume that Y satisfies [LC]. Then $\Pr(\mathcal{E}_{MRE}(\varepsilon)^c) \leq Cn^2e^{-c(\Psi_p\sqrt{n})\varepsilon}$.*
(ii) *Assume that Y satisfies either [SG-IND] or [CCP-SBP]. Then $\Pr(\mathcal{E}_{MRE}(\varepsilon)^c) \leq Cn^2e^{-cn\varepsilon^2}$.*

We note that the weights of MRE were previously studied in [6], assuming Y has i.i.d. entries. They proved that asymptotically, as $n, p \to \infty$ at a fixed ratio $p/n = \gamma$, one has $\max_{1\leq i\leq n}\|w_i^{\mathrm{MRE}} - u(d^*)\| \to 0$ with probability 1.

Lastly, we consider Tyler's regularized M-estimator (TRE). TRE is superficially a special case of MRE, corresponding to $u(x) = 1/x$ and $\phi(x) = 1$; a crucial difference, however, is that $u(\cdot)$ is singular at $x = 0$. Accordingly, to carry out our analysis, we require an additional constraint on Σ_p: there exists a constant $\underline{\tau} > 0$ such that

$$\frac{1}{p}\mathrm{Tr}\Sigma_p^{-1} \le \underline{\tau}\,. \tag{15}$$

For $\varepsilon > 0$, let $\mathcal{E}_{TRE}(\varepsilon)$ be the event that TRE exists uniquely, and that furthermore its weights satisfy

$$\max_{1\le i\le n} |w_i^{\mathrm{TRE}} - 1/d^*| \le \varepsilon\,. \tag{16}$$

Theorem 4 (*The Weights of TRE*)**.** *Let $u(x) = 1/x$, $\phi(x) = 1$ and $\alpha > \max\{0, \gamma - 1\}$. Then (13) has a unique solution which satisfies $d^* \in [\underline{d}, \overline{d}]$, where the constants $0 < \underline{d} \le \overline{d}$ depend on the distribution of Y, γ, α, $s_{\max}$, τ and $\underline{\tau}$.*

Furthermore, there are constant $c, C, \varepsilon_0 > 0$, that depend on the distribution of Y, γ, α, $s_{\max}$, τ and $\underline{\tau}$, such that for all $\varepsilon < \varepsilon_0$,

(i) *Assume that Y satisfies [LC]. Then $\Pr(\mathcal{E}_{TRE}(\varepsilon)^c) \le Cn^2 e^{-c(\Psi_p\sqrt{n})\varepsilon}$.*

(ii) *Assume that Y satisfies either [SG-IND] or [CCP-SBP]. Then $\Pr(\mathcal{E}_{TRE}(\varepsilon)^c) \le Cn^2 e^{-cn\varepsilon^2}$.*

For the special case where X follows a Gaussian or an elliptical distribution, similar non-asymptotic concentration bounds for the weights of TRE were provided in [33]. Lastly, we remark that as noted in [33], the dependence on $\underline{\tau}$ can be removed when α is sufficiently large, i.e $\alpha \ge Cs_{\max}$ where C is a suitable constant (see Remark 5 in the appendix for further details).

Paper outline. The rest of the manuscript is organized as follows. In Section 3 we provide definitions and technical background related to the distributional assumptions from Section 2. In Section 4 we describe some applications of our results to robust covariance and precision matrix estimation. Section 5 is devoted to the proofs of Theorems 1–4, with some technical details deferred to the Appendix. Finally, we offer some concluding remarks in Section 6.

3. Preliminaries and technical background

We provide a brief background and definitions for the distributions considered in Section 2. Recall that $Y \in \mathbb{R}^p$ is an isotropic random vector. Throughout this text, $\|\cdot\|$ denotes the Euclidean norm in the suitable dimension p.

Definition 1. Y is a sub-Gaussian vector with constant $K > 0$ if for any $\|u\| = 1$,

$$\Pr\left(\left|u^\top Y - \mathbb{E}[u^\top Y]\right| \ge t\right) \le 2\exp\left[-(t/2K)^2\right]\,.$$

Sub-Gaussianity implies that linear functions of Y concentrate. For our analysis we shall also need concentration of some sufficiently well-behaved non-linear functions. Following [3,53], we consider the following property:

Definition 2. Y satisfies the convex concentration property (CCP) with constant $K > 0$ if for every 1-Lipschitz convex $f : \mathbb{R}^p \to \mathbb{R}$, the random variable $f(Y)$ is sub-Gaussian with constant K:

$$\Pr\left(|f(Y) - \mathbb{E}[f(Y)]| \ge t\right) \le 2\exp\left[-(t/2K)^2\right]\,. \tag{17}$$

The CCP was introduced by Talagrand [71], who proved that if Y has independent, uniformly bounded entries then it satisfies the CCP. Subsequent works established weaker conditions under which the CCP holds, see [2,73] and references therein. Another family of distributions satisfying the CCP, which has received attention in machine learning and statistics (e.g., [9,13,62]), are the distributions satisfying a Log-Sobolev Inequality (such distributions in fact satisfy a stronger Lipschitz concentration property: the function $f(\cdot)$ in Definition 2 does not need to be convex).

For parts of our analysis, we shall also need the following:

Definition 3. Y satisfies the small ball property (SBP) with constant C_0 if for any $\|u\| = 1$ and $a \in \mathbb{R}$,

$$\Pr\left(\left|u^\top Y - a\right| \le t\right) \le C_0 t\,, \quad t \ge 0\,. \tag{18}$$

The SBP is an anti-concentration property: it states that the law of $u^\top Y$ cannot put large mass around any particular value $a \in \mathbb{R}$. For our purposes, the SBP will be especially important for bounding the smallest eigenvalue of the sample covariance matrix, which is a key step in several of our proofs. Note that if the density of $u^\top Y$ is bounded, then (18) holds for some appropriate C_0. The following remarkable result, due to Rudelson and Vershynin, states that if Y has independent entries, each with bounded density, then it satisfies the SBP [66, Theorem 1.2]:

Lemma 1. *Let $Y = (Y_1, \dots, Y_p)^\top$ have independent entries, with univariate densities all uniformly bounded by C. Then Y satisfies the SBP with constant $2\sqrt{2}C$.*

Log-concave distributions. We next consider the family of log-concave distributions on $\mathbb{R}^p$. This rich family has found multiple applications, for example, in statistics [7], pure mathematics [14,69], computer science [10,49] and economics [4]. Particular members in this family include the Gaussian, exponential, uniform over convex bodies, logistic, Gamma, Laplace, Weibull, Chi and Chi-Squared, Beta distributions and more.

Definition 4. Y is log-concave if it has a density $p_Y(y) = e^{-V(y)}$ for $V : \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$ convex.

It is known that log-concave random vectors are sub-exponential with a universal constant, see e.g., [14]. The following is implied by a recent breakthrough result of Klartag and Lehec [42, Theorem 1.1]:

Lemma 2. *There exists a universal $c > 0$ and some Ψ_p satisfying*

$$(\log p)^{-5} \le \Psi_p \le 1 \tag{19}$$

such that for every isotropic log-concave $Y \in \mathbb{R}^p$ and 1-Lipschitz function $f(\cdot)$,

$$\Pr\left(|f(Y) - \mathbb{E}[f(Y)]| \ge t\right) \le \exp\left[-c\Psi_p t\right] . \tag{20}$$

The quantity Ψ_p is (up to a universal constant) the Cheeger constant corresponding to the family of log-concave distributions on $\mathbb{R}^p$. It is conjectured [38] that in fact $\Psi_p = \Theta(1)$ (the KLS conjecture). For background on the KLS conjecture and its consequences, see [20,45]. Lastly, [49, Lemma 5.5] implies the following:

Lemma 3. *There is a universal $0 < C_0 \le 2$ so that every isotropic log-concave Y satisfies the SBP with constant C_0.*

4. Applications to robust sparse covariance and inverse covariance estimation

As mentioned in the introduction, robust estimation of the covariance and inverse covariance matrices, given possibly heavy tailed data are important tasks in statistics. In high dimensional settings, these matrices are often assumed to be sparse, allowing their estimation from a limited number of samples. A common model for multivariate heavy tailed data, under which various estimators were derived and analyzed is to assume that the samples are elliptically distributed [18,31,32,40]. Specifically, a random vector $\tilde{X}$ follows an elliptical distribution with mean $\boldsymbol{\mu}$ and shape matrix $\Sigma_p \in S^p_{++}$ if it has the form

$$\tilde{X} = \boldsymbol{\mu} + z\Sigma_p^{1/2} Y , \tag{21}$$

where $Y \sim \text{Unif}(\mathbb{S}^{p-1})$, and z is a strictly positive random variable which is independent of Y (but otherwise arbitrary). To make the model (21) identifiable, the scaling $p^{-1}\text{Tr}\,\Sigma_p = 1$ is assumed.

The authors of [33] considered the problem of shape matrix estimation under a sparsity constraint. They showed that given (possibly heavy tailed) elliptical samples, a sparse Σ_p may nonetheless be estimated at the same rate as if one had sub-Gaussian samples with covariance Σ_p. Their estimator is remarkably simple: compute Tyler's M-estimator corresponding to the n samples, and then threshold its entries as proposed by [12].

In this section, building upon the theorems of Section 2, we show that the above approach yields accurate estimates for heavy tailed distributions beyond the elliptical model (21). Specifically, the vector Y in (21) may be replaced by any isotropic random vector satisfying the assumptions of Section 2. Furthermore, we show a similar result for estimating the inverse shape matrix Σ_p^{-1} assuming it is sparse. For simplicity, we assume that $\tilde{X}$ is zero mean with $\boldsymbol{\mu} = \mathbf{0}$ whereas in Section 6 we discuss how this restriction may be overcome.

4.1. Sparse shape matrix estimation

Observe that Tyler's M-estimator, Eq. (3), is invariant to an arbitrary scaling of the samples. Consequently, Tyler's estimator computed from the elliptically-distributed samples $\tilde{\mathbf{x}}_i = z_i\Sigma_p^{1/2}\mathbf{y}_i$, call it $\hat{\Sigma}_{\text{Tyl}}$, is exactly the same as the estimator computed from the rescaled samples $\mathbf{x}_i = \Sigma_p^{1/2}\mathbf{y}_i$. We emphasize that the rescaled vectors $\mathbf{x}_1, \dots, \mathbf{x}_n$ are *not* available to the estimator. Denote $S = n^{-1}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top$, and $\hat{\Sigma}_{\text{Tyl}} = n^{-1}\sum_{i=1}^n w_i^{\text{Tyl}}\mathbf{x}_i\mathbf{x}_i^\top$. For a matrix M, denote the norms,

$$\|M\|_{\max} = \max_{i,j}|M_{ij}|, \quad \|M\|_1 = \sum_{i,j}|M_{ij}| .$$

We start with the following important lemma, which asserts that $\hat{\Sigma}_{\text{Tyl}}$ is close to Σ_p entrywise:

Lemma 4. *Assume the setting of Theorem 2, recalling the normalization $\tau_p = 1$. There is $C > 0$, that depends on the distribution of Y and on γ, such that with probability $1 - o(1)$, the following holds.*

(i) *Assume that Y satisfies [LC]. Then* $\|\hat{\Sigma}_{Tyl} - \Sigma_p\|_{\max} \le C\|\Sigma_p\| \frac{\log p}{\Psi_p \sqrt{n}}$.

(ii) *Assume that Y satisfies either [SG-IND] or [CCP-SBP]. Then* $\|\hat{\Sigma}_{Tyl} - \Sigma_p\|_{\max} \le C\|\Sigma_p\| \sqrt{\frac{\log p}{n}}$.

Proof. By the triangle inequality, $\|\hat{\Sigma}_{\mathrm{Tyl}} - \Sigma_p\|_{\max} \le \|\hat{\Sigma}_{\mathrm{Tyl}} - S\|_{\max} + \|S - \Sigma_p\|_{\max}$. We bound the second term, by Lemma 23 in the Appendix, which bounds the deviations of the entries of S about their expectation Σ_p. For the first term,

$$\|\hat{\Sigma}_{\mathrm{Tyl}} - S\|_{\max} \le \|\hat{\Sigma}_{\mathrm{Tyl}} - S\| = \left\| \frac{1}{n}\sum_{i=1}^{n} (w_i^{\mathrm{Tyl}} - 1)\mathbf{x}_i\mathbf{x}_i^\top \right\| \le \max_{1\le i\le n} |w_i^{\mathrm{Tyl}} - 1| \cdot \|S\| \,.$$

By Lemma 20 in the Appendix, $\|S\| \lesssim \|\Sigma_p\|$ with high probability, whereas $\max_{1\le i\le n} |w_i^{\mathrm{Tyl}} - 1|$ may be bounded by Theorem 2. □

Following Bickel and Levina [12], consider the class of approximately sparse covariance matrices, $q \in [0, 1)$,

$$\mathcal{U}_p(q, s_q(p)) = \left\{ \Sigma \in \mathcal{S}_{++}^p \ : \ \sum_{j=1}^{p} |\Sigma_{ij}|^q \le s_q(p), \ 1 \le i \le p \right\} . \tag{22}$$

Let $\mathcal{T}_t$ be the entry-wise hard-thresholding operator $\mathcal{T}_t(M)_{ij} = M_{ij}\mathbb{1}_{\{|M_{ij}|\ge t\}}$. The authors of [12] showed that to accurately estimate a matrix $\Sigma_p \in \mathcal{U}_p(q, s_q(p))$ with respect to operator norm, it suffices to construct a matrix A which is close to Σ_p entrywise, and then threshold it. We cite the following form of their result, as stated in [33, Lemma 6]:

Lemma 5. *Let $\Sigma_p \in \mathcal{U}_p(q, s_q(p))$ and A such that $\|A - \Sigma_p\|_{\max} \le \varepsilon$. There is a threshold $t = c_1\varepsilon$ so that for some $c_2 = c_2(q)$, $\|\mathcal{T}_t(A) - \Sigma_p\| \le c_2 s_q(p)\varepsilon^{1-q}$.*

Combining Lemmas 4 and 5 yields the following generalization of [33, Theorem 1], which proved a similar result under the more restrictive assumption that Y has an elliptical distribution:

Corollary 1. *There are $c, C > 0$, that may depend on the distribution of Y and on γ, such that the following holds. For an appropriately chosen threshold t, if $\Sigma_p \in \mathcal{U}_p(q, s_q(p))$, then with probability $1 - o(1)$,*

(i) *Assume that Y satisfies [LC]. Then* $\|\Sigma_p - \mathcal{T}_t(\hat{\Sigma}_{Tyl})\| \le Cs_q(p)\|\Sigma_p\|^{1-q} \left(\frac{(\log p)^2}{\Psi_p^2 n}\right)^{(1-q)/2}$.

(ii) *Assume that Y satisfies either [SG-IND] or [CCP-SBP]. Then* $\|\Sigma_p - \mathcal{T}_t(\hat{\Sigma}_{Tyl})\| \le Cs_q(p)\|\Sigma_p\|^{1-q} \left(\frac{\log p}{n}\right)^{(1-q)/2}$.

Under [SG-IND] and [CCP-SBP], the attained rate is minimax optimal in n, p [17].

4.2. Sparse inverse shape matrix estimation

We now consider the problem of estimating Σ_p^{-1}, assuming that it is sparse: $\Sigma_p^{-1} \in \mathcal{U}_p(q, s_q(p))$. Cai et al. proposed the CLIME estimator [15], which solves a linear program of the form:

$$\min_{\Omega} \|\Omega\|_1 \quad \text{subject to} \quad \|\hat{S}\Omega - \mathrm{I}_{p\times p}\|_{\max} \le \lambda \,, \tag{23}$$

where λ is a tuning parameter and $\hat{S}$ is a proxy for Σ_p ([15] proposes to use the data sample covariance matrix). Having solved (23), a symmetrization step is applied to get the final estimator. [15, Theorem 6] states that if $\hat{S}$ is close entrywise to Σ_p, then the estimator $\hat{\Omega} = \hat{\Omega}(\hat{S})$ is close in operator norm to Σ_p^{-1}:

Lemma 6. *Suppose that $\Sigma_p^{-1} \in \mathcal{U}_p(q, s_q(p))$ and $\hat{S}$ satisfies $\|\Sigma_p - \hat{S}\|_{\max} \le \varepsilon$. For any $\lambda \ge \|\Sigma_p^{-1}\|_1\varepsilon$, the estimator obtained by solving (23) and applying symmetrization satisfies $\|\hat{\Omega} - \Sigma_p^{-1}\| \le Cs_q(p)\|\Sigma_p^{-1}\|_1^{1-q}\lambda^{1-q}$.*

Setting $\hat{S} = \hat{\Sigma}_{\mathrm{Tyl}}$ in (23) and using Lemma 4, yields:

Corollary 2. *There are $c, C > 0$, that may depend on the distribution of Y and on γ, such that the following holds. For an appropriately chosen λ, if $\Sigma_p^{-1} \in \mathcal{U}_p(q, s_q(p))$, then with probability $1 - o(1)$,*

(i) *Assume that Y satisfies [LC]. Then* $\|\Sigma_p^{-1} - \hat{\Omega}(\hat{\Sigma}_{Tyl})\| \le Cs_q(p)\|\Sigma_p^{-1}\|_1^{2-q}\|\Sigma_p\|^{1-q} \left(\frac{(\log p)^2}{\Psi_p^2 n}\right)^{(1-q)/2}$.

(ii) *Assume that Y satisfies either [SG-IND] or [CCP-SBP]. Then* $\|\Sigma_p^{-1} - \hat{\Omega}(\hat{\Sigma}_{Tyl})\| \le Cs_q(p)\|\Sigma_p^{-1}\|_1^{2-q}\|\Sigma_p\|^{1-q} \left(\frac{\log p}{n}\right)^{(1-q)/2}$.

Lastly, we remark that CLIME is known to have a sub-optimal rate for i.i.d. sub-Gaussian data. In [16], the minimax rate is computed and a rate-optimal adaptive estimator ACLIME is proposed. While beyond the scope of the present paper, we conjecture that a similar result, as Corollary 2, may be derived for ACLIME as well.

5. Proofs

Recall that the input data consists of n samples $\mathbf{x}_i = \Sigma_p^{1/2}\mathbf{y}_i$ with $\mathbf{y}_i$ isotropic. Denote by S (respectively T) the sample covariance corresponding to the $\mathbf{x}_i$-s (resp. $\mathbf{y}_i$-s),

$$S = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^\top = \Sigma_p^{1/2}T\Sigma_p^{1/2}\,, \quad T = \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i\mathbf{y}_i^\top .$$

For $1 \le j \le n$, denote by S_{-j} (similarly T_{-j}) the sample covariance of $(n-1)$ samples excluding $\mathbf{x}_j$,

$$S_{-j} = S - \frac{1}{n}\mathbf{x}_i\mathbf{x}_i^\top = \frac{1}{n}\sum_{i=1, i\neq j}^{n}\mathbf{x}_i\mathbf{x}_i^\top .$$

The following lemma, whose proof is deferred to the appendix, is key to the analysis of Maronna's and Tyler's estimators.

Lemma 7. *Assume $\gamma < 1$. There are $c, C, \varepsilon_0 > 0$, that depend on the distribution of Y and on γ, so that for all $\varepsilon < \varepsilon_0$:*

(i) *Assume [LC]. Then* $\Pr(\max_{1\le i\le n}|p^{-1}\mathbf{x}_i^\top S^{-1}\mathbf{x}_i - 1| \ge \varepsilon) \le Cn^2e^{-c(\Psi_p\sqrt{n})\varepsilon}$.
(ii) *Assume [SG-IND] or [CCP-SBP]. Then* $\Pr(\max_{1\le i\le n}|p^{-1}\mathbf{x}_i^\top S^{-1}\mathbf{x}_i - 1| \ge \varepsilon) \le Cn^2e^{-cn\varepsilon^2}$.

Note that $\mathbf{x}_i^\top S^{-1}\mathbf{x}_i = \mathbf{y}_i^\top\Sigma_p^{1/2}(\Sigma_p^{1/2}T\Sigma_p^{1/2})^{-1}\Sigma_p^{1/2}\mathbf{y}_i = \mathbf{y}_i^\top T^{-1}\mathbf{y}_i$, so the quadratic form $p^{-1}\mathbf{x}_i^\top S^{-1}\mathbf{x}_i$ does not depend on Σ_p. Assuming that $\mathbf{x}_i$-s are Gaussian, a similar result was derived in [78]. Their proof relies on the orthogonal invariance of the isotropic Gaussian distribution, and does not generalize to other distributions. Lemma 7 implies that $\max_{1\le i\le n}|\mathbf{x}_i^\top S^{-1}\mathbf{x}_i - 1| \to 1$ with probability 1 in the asymptotic limit $n, p \to \infty$ with $\frac{p}{n} \to \gamma$. Assuming Y has independent entries with finite fourth moment, this asymptotic result was proven in [25].

With these preparations, we now present the proof of Theorem 1 (Maronna's M-Estimator). The proof has two parts, one regarding existence and the second regarding uniqueness and concentration.

Existence. We first prove that a solution to the fixed point equation (1) indeed exists. This proof follows that of [25]. We briefly describe it, for the sake of completeness. Denote $\mathbf{d} = (d_1, \dots, d_n)$. Consider the function $h : \mathbb{R}_+^n \to \mathbb{R}_+^n$,

$$h_j(\mathbf{d}) = \frac{1}{p}\mathbf{x}_j^\top\left(\frac{1}{n}\sum_{i=1}^{n}u(d_i)\mathbf{x}_i\mathbf{x}_i^\top\right)^{-1}\mathbf{x}_j, \quad 1 \le j \le n\,. \tag{24}$$

Since by assumption X has a density and $n > p$, h is well-defined with probability 1. Moreover, a vector $\boldsymbol{d}$ yields a Maronna's estimator with weights $w_j = u(d_j)$ if and only if $d_j = h_j(\mathbf{d})$ for all j. Hence, to establish the existence of Maronna's estimator, we need to show that h has a fixed point. To this end, as proven in [25], the function $h(\cdot)$ satisfies the following three properties. (1) Positivity: $h(\mathbf{d}) > \mathbf{0}$ for any vector $\mathbf{d} \ge \mathbf{0}$, namely with $d_j \ge 0$ for all j; (2) Monotonicity: if $\mathbf{d} \ge \mathbf{d}' \ge \mathbf{0}$, then $h(\mathbf{d}) \ge h(\mathbf{d}')$; (3) Scalability: for any $\alpha > 1$, $\alpha h(\mathbf{d}) \ge h(\alpha\mathbf{d})$. A function $h(\cdot)$ satisfying these properties is (almost) a standard interference function, in the sense of [77]. By [77, Theorem 1], if there exists some $\mathbf{d}$ with $\mathbf{d} \ge h(\mathbf{d})$, then $h(\cdot)$ has a fixed point. The following lemma shows that with high probability, such a vector $\mathbf{d}$ exists, which in turn implies the existence of Maronna's estimator.

Lemma 8. *There are $c, C > 0$, that depend on the distribution of Y and on γ, such that*

(i) *Assume [LC].* $\Pr(\exists\boldsymbol{d} > \mathbf{0} \;:\; \boldsymbol{d} \ge h(\boldsymbol{d})) \ge 1 - Cn^2e^{-c(\Psi_p\sqrt{n})}$.
(ii) *Assume [SG-IND] or [CCP-SBP]. Then* $\Pr(\exists\boldsymbol{d} > \mathbf{0} \;:\; \boldsymbol{d} \ge h(\boldsymbol{d})) \ge 1 - Cn^2e^{-cn}$.

Proof. By assumption (a) of Theorem 1, $\phi_\infty > 1$. Hence, we may take some $d_0 > 0$ such that $\phi(d_0) > 1$. Setting $\boldsymbol{d}_0 = d_0\mathbf{1} = (d_0, \dots, d_0)^\top$ in (24) and using $u(d) = \phi(d)/d$, one may verify that $h_j(\boldsymbol{d}_0) = \frac{d_0}{\phi(d_0)}\frac{1}{p}\mathbf{x}_j^\top S^{-1}\mathbf{x}_j$. By Lemma 7, with high probability, $\max_{1\le j\le n}\frac{1}{p}\mathbf{x}_j^\top S^{-1}\mathbf{x}_j \le \phi(d_0)$, and so $h(\boldsymbol{d}_0) \le \boldsymbol{d}_0$. □

Uniqueness and concentration. Let $\mathbf{d} = (d_1, \dots, d_n)^\top$ be a vector satisfying $h(\mathbf{d}) = \mathbf{d}$, which yields a valid Maronna's estimator. Next, we prove that only one such $\boldsymbol{d}$ exists. To this end, denote $j_{\min} = \operatorname{argmin}_{1\le j\le n} d_j$, $j_{\max} = \operatorname{argmax}_{1\le j\le n} d_j$. Since $h(\cdot)$ is non-decreasing, $d_j = h_j(\boldsymbol{d}) \le h_j(d_{j_{\max}}\mathbf{1}) = (u(d_{j_{\max}}))^{-1}p^{-1}\mathbf{x}_j^\top S^{-1}\mathbf{x}_j$. Setting $j = j_{\max}$ and multiplying by $u(d_{j_{\max}})$ gives $\phi(d_{j_{\max}}) \le p^{-1}\mathbf{x}_{j_{\max}}^\top S^{-1}\mathbf{x}_{j_{\max}}$. Similarly, $h(\boldsymbol{d}) \ge h(d_{j_{\min}}\mathbf{1})$, and thus $\phi(d_{j_{\min}}) \ge p^{-1}\mathbf{x}_{j_{\min}}^\top S^{-1}\mathbf{x}_{j_{\min}}$. Finally, since ϕ is non-decreasing, we deduce that for all j, $p^{-1}\mathbf{x}_{j_{\min}}^\top S^{-1}\mathbf{x}_{j_{\min}} \le \phi(d_j) \le p^{-1}\mathbf{x}_{j_{\max}}^\top S^{-1}\mathbf{x}_{j_{\max}}$. Consequently,

$$\max_{1\le j\le n}|\phi(d_j) - 1| \le \max_{1\le j\le n}|p^{-1}\mathbf{x}_j^\top S^{-1}\mathbf{x}_j - 1|\,. \tag{25}$$

The next lemma shows that with high probability, $h(\cdot)$ cannot have any fixed points whose entries are far from the constant $\phi^{-1}(1)$:

Lemma 9. *For $\varepsilon > 0$, let $\mathcal{F}_{\text{bad}}(\varepsilon)$ the set of "bad" fixed points of $h(\cdot)$, namely, such that $\|\boldsymbol{d} - \phi^{-1}(1)\mathbf{1}\|_\infty \geq \varepsilon$. There $c, C, \varepsilon_0 > 0$, that depend on the distribution of Y and on γ, such that for all $\varepsilon < \varepsilon_0$,*

(i) *Assume [LC]. Then* $\Pr\left(\mathcal{F}_{\text{bad}}(\varepsilon) \neq \emptyset\right) \leq Cn^2 e^{-c(\Psi_p\sqrt{n})\varepsilon}$.
(ii) *Assume [SG-IND] or [CCP-SBP]. Then* $\Pr\left(\mathcal{F}_{\text{bad}}(\varepsilon) \neq \emptyset\right) \leq Cn^2 e^{-cn\varepsilon^2}$.

Proof. Follows immediately by Lemma 7, Eq. (25) and Assumption (b) of Theorem 1, which states that ϕ is invertible in a neighborhood of $\phi^{-1}(1)$ and that the inverse ϕ^{-1} is locally Lipschitz at 1. □

Lastly, we prove that Maronna's estimator exists uniquely with high probability:

Lemma 10. *There are $c, C > 0$, that depend on the distribution of Y and on γ, such that*

(i) *Assume [LC]. Then* $\Pr\left(\textit{ME exists uniquely}\right) \geq 1 - Cn^2 e^{-c(\Psi_p\sqrt{n})}$.
(ii) *Assume [SG-IND] or [CCP-SBP]. Then* $\Pr\left(\textit{ME exists uniquely}\right) \geq 1 - Cn^2 e^{-cn}$.

Proof. Existence, with high probability, is guaranteed by Lemma 8. Assume by contradiction that h has two different fixed points, $\mathbf{d} \neq \mathbf{d}'$. Take $\alpha = \max_j(d_j/d_j')$, and let l be a coordinate where the maximum is attained. Assume without loss of generality that $\alpha > 1$ (otherwise replace $\mathbf{d}$ with $\mathbf{d}'$), and note that by its definition, $\mathbf{d} \leq \alpha\mathbf{d}'$.

Let $\eta > 0$ be so that $\phi(\cdot)$ is strictly increasing in $[\phi^{-1}(1) - \eta, \phi^{-1}(1) + \eta]$. By Lemma 9, with high probability both $\mathbf{d}, \mathbf{d}' \in [\phi^{-1}(1) - \eta/2, \phi^{-1}(1) + \eta/2]^n$. In addition, since $\alpha > 1$, then $\phi(\alpha d_j') > \phi(d_j')$ for all j. This, in turn, implies that

$$u(\alpha d_j') = \frac{\phi(\alpha d_j')}{\alpha d_j'} > \frac{\phi(d_j')}{\alpha d_j'} = \frac{1}{\alpha} u(d_j').$$

Plugging this into (24) gives $h(\alpha\boldsymbol{d}') < \alpha h(\boldsymbol{d}')$. Since $\mathbf{d} \leq \alpha\mathbf{d}'$ and h is non-decreasing, we deduce $h(\boldsymbol{d}) < \alpha h(\boldsymbol{d}')$. But, since $\mathbf{d}, \mathbf{d}'$ are fixed points of $h(\cdot)$, this yields a contradiction: $d_l = h_l(\mathbf{d}) < \alpha h_l(\mathbf{d}') = \alpha d_l' = d_l$. □

Proof of Theorem 1. Existence and uniqueness follow from Lemma 10. By Lemma 9, the coordinates of the fixed point $\mathbf{d}$ concentrate around $\phi^{-1}(1)$. Hence, the weights of Maronna's estimator, $w_i^{\text{Mar}} = u(d_i)$, concentrate around $u(\phi^{-1}(1)) = \phi(\phi^{-1}(1))/\phi^{-1}(1) = 1/\phi^{-1}(1)$. □

Next, we describe the proof of Theorem 2 (Tyler's M-Estimator). Our proof combines the strategy of Zhang et al. [78] with Lemma 7. Since X has a density, by [41, Theorems 1 and 2] Tyler's estimator exists uniquely with probability 1. By [78, Lemma 2.1], its weights are

$$w_i^{\text{Tyl}} = \frac{p \cdot \widehat{w}_i}{\operatorname{Tr}\left(\frac{1}{n}\sum_{i=1}^n \widehat{w}_i \mathbf{x}_i\mathbf{x}_i^\top\right)}, \tag{26}$$

where $\widehat{\mathbf{w}} = (\widehat{w}_1, \ldots, \widehat{w}_n)^\top$ is the unique minimizer of

$$F(\mathbf{w}) = -\sum_{i=1}^n \log w_i + \frac{n}{p}\log\det\left(\sum_{i=1}^n w_i\mathbf{x}_i\mathbf{x}_i^\top\right), \qquad \mathbf{w} > \mathbf{0}, \quad \sum_{i=1}^n \mathbf{w}_i = n. \tag{27}$$

As in [78], the proof proceeds in two steps: (I) Show that $\widehat{w}_1, \ldots, \widehat{w}_n$ all concentrate around 1; (II) Using (26), deduce concentration for the weights $w_1^{\text{Tyl}}, \ldots, w_n^{\text{Tyl}}$.

We start with the weights $\widehat{\mathbf{w}}$. The argument of [78, Section 3.2] starts with the following observation:

Lemma 11. *Let $\beta > 0$ be arbitrary. Then $\widehat{\mathbf{w}}$ is the unique stationary point of the following function $G_\beta : \mathbb{R}^n \to \mathbb{R}$:*

$$G_\beta(\mathbf{w}) = F(\mathbf{w}) + \frac{\beta}{2}\left(\sum_{i=1}^n w_i - n\right)^2 = -\sum_{i=1}^n \log w_i + \frac{n}{p}\log\det\left(\sum_{i=1}^n w_i\mathbf{x}_i\mathbf{x}_i^\top\right) + \frac{\beta}{2}\left(\sum_{i=1}^n w_i - n\right)^2. \tag{28}$$

Next, note that $\mathbf{w} = \mathbf{1} = (1, \ldots, 1)^\top$ is "almost" a zero of $g_\beta(\cdot) = \nabla G_\beta(\cdot)$ (for any β), in the sense that with high probability $g_\beta(\mathbf{1})$ is small. Indeed, a straightforward calculation, Appendix Eq. (B.9), gives

$$\left(g_\beta(\mathbf{1})\right)_\ell = -1 + \frac{1}{p}\mathbf{x}_\ell^\top\left(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top\right)^{-1}\mathbf{x}_\ell = -1 + \frac{1}{p}\mathbf{x}_\ell^\top S^{-1}\mathbf{x}_\ell, \qquad 1 \leq \ell \leq n. \tag{29}$$

By Lemma 7, the right-hand-side of (29) concentrates tightly around 0. That is, the following deviation bound holds:

Lemma 12. *There are $c, C, \varepsilon_0 > 0$, that depend on the distribution of Y and on γ, so that for all $\varepsilon < \varepsilon_0$ and all β*

(i) *Assume [LC]. Then* $\Pr\left(\|g_\beta(\mathbf{1})\|_\infty \geq \varepsilon\right) \leq Cn^2 e^{-c(\Psi_p\sqrt{n})\varepsilon}$.
(ii) *Assume either [SG-IND] or [CCP-SBP]. Then* $\Pr\left(\|g_\beta(\mathbf{1})\|_\infty \geq \varepsilon\right) \leq Cn^2 e^{-cn\varepsilon^2}$.

Next, we carry out a perturbation argument: we show that $\|g_\beta(\mathbf{1})\|_\infty$ being small implies that the unique root $g_\beta(\widehat{\mathbf{w}}) = \mathbf{0}$ is close to $\mathbf{1}$. To this end, we use the following result, [78, Lemma 3.1]. Below, for a matrix $M \in \mathbb{R}^{n\times n}$, we denote by $\|M\|_{\infty,\infty} = \max_{1\le j\le n}\sum_{\ell=1}^n |M_{j\ell}|$ its ℓ_∞-to-ℓ_∞ operator norm.

Lemma 13. *Let $h : \mathbb{R}^p \to \mathbb{R}^p$ be differentiable, $\mathbf{w}_0 \in \mathbb{R}^p$. Suppose that for some $L, R > 0$ and $0 < \varepsilon \le \min\{R, L^{-1}\}$:*

(i) *($\mathbf{w}_0$ is "almost" a zero.) $\|h(\mathbf{w}_0)\|_\infty \le \frac{1}{2}\varepsilon$.*
(ii) *(Non-degeneracy at $\mathbf{w}_0$.) $\nabla h(\mathbf{w}_0) = \mathrm{I}_{p\times p}$, namely $(\nabla h)_{j,\ell} = \frac{\partial h_j}{\partial w_\ell} = \delta_{j,\ell}$.*
(iii) *(Smoothness around $\mathbf{w}_0$.) $\|\nabla h(\mathbf{w}) - \nabla h(\mathbf{w}_0)\|_{\infty,\infty} \le L\|\mathbf{w} - \mathbf{w}_0\|_\infty$ for all $\mathbf{w}$ in $\|\mathbf{w} - \mathbf{w}_0\|_\infty \le R$.*

Then $h(\cdot)$ has a zero $\widehat{\mathbf{w}}$ close to $\mathbf{w}_0$ such that $\|\widehat{\mathbf{w}} - \mathbf{w}_0\|_\infty \le \varepsilon$.

To prove that g_β has a zero near $\mathbf{w}_0 = \mathbf{1}$ via Lemma 13, we consider $h_\beta(\mathbf{w}) = \left(\nabla g_\beta(\mathbf{1})\right)^{-1} g_\beta(\mathbf{w})$. Clearly, g_β and h_β have the same zeros. Also, by its definition, h_β satisfies condition (II) above. We next show that with high probability h_β satisfies condition (I). Since $\|h_\beta(\mathbf{1})\|_\infty \le \left\|\left(\nabla g_\beta(\mathbf{1})\right)^{-1}\right\|_{\infty,\infty}\|g_\beta(\mathbf{1})\|_\infty$, it suffices to bound the matrix norm $\left\|\left(\nabla g_\beta(\mathbf{1})\right)^{-1}\right\|_{\infty,\infty}$:

Lemma 14. *There are $c, C, B, \beta_0 > 0$, that depend the distribution of Y and on γ, so that setting $\beta = \beta_0/n$,*

(i) *Assume [LC]. Then $\Pr\left(\|\left(\nabla g_\beta(\mathbf{1})\right)^{-1}\|_{\infty,\infty} \ge B\right) \le Cn^2e^{-c\Psi_p n^{1/4}}$.*
(ii) *Assume [SG-IND] or [CCP-SBP]. Then $\Pr\left(\|\left(\nabla g_\beta(\mathbf{1})\right)^{-1}\|_{\infty,\infty} \ge B\right) \le Cn^2e^{-cn^{1/2}}$.*

The proof of Lemma 14 is deferred to the appendix. Our proof follows Zhang et al. [78, Lemma 3.3]. Their argument, however, contains a mathematical error that we correct.

Lastly, we address condition (iii) of Lemma 13. We prove the following in the appendix:

Lemma 15. *Let L_g be the ℓ_∞ Lipschitz constant of ∇g_β on an ℓ_∞ ball of radius $\frac{1}{2}$ around $\mathbf{1}$:*

$$L_g = \max_{\mathbf{w}:\,\|\mathbf{w}-\mathbf{1}\|_\infty\le\frac{1}{2}} \frac{\|\nabla g_\beta(\mathbf{w}) - \nabla g_\beta(\mathbf{1})\|_{\infty,\infty}}{\|\mathbf{w}-\mathbf{1}\|_\infty}.$$

There are $c, C, L > 0$, that depend on the distribution of Y and on γ, so that

(i) *Assume [LC]. Then $\Pr(L_g \ge L) \le Cn^2e^{-c\Psi_p\sqrt{n}}$.*
(ii) *Assume [SG-IND] or [CCP-SBP]. Then $\Pr(L_g \ge L) \le Cn^2e^{-cn}$.*

Equipped with the preceding lemmas, we are ready to prove our concentration result for $\widehat{\mathbf{w}}$:

Lemma 16. *Let $\widehat{\mathbf{w}} = (\widehat{w}_1,\ldots,\widehat{w}_n) > \mathbf{0}$ be the unique minimizer of (27). There are $c, C, \varepsilon_0 > 0$, that depend on the distribution of Y and on γ, such that for all $\varepsilon < \varepsilon_0$,*

(i) *Assume [LC]. Then $\Pr\left(\max_{1\le i\le n}|\widehat{w}_i - 1| \ge \varepsilon\right) \le Cn^2e^{-c\Psi_p\min\left\{\sqrt{n}\varepsilon,\, n^{1/4}\right\}}$.*
(ii) *Assume [SG-IND] or [CCP-SBP]. Then $\Pr\left(\max_{1\le i\le n}|\widehat{w}_i - 1| \ge \varepsilon\right) \le Cn^2e^{-c\min\left\{n\varepsilon^2,\, n^{1/2}\right\}}$.*

Proof. Choose $\beta = \beta_0/n$ per Lemma 14, such that $\|\left(\nabla g_\beta(\mathbf{1})\right)^{-1}\|_{\infty,\infty} \le B$ holds with high probability. By Lemma 15, for some $L > 0$, with high probability $\|\nabla g_\beta(\mathbf{w}) - \nabla g_\beta(\mathbf{1})\|_{\infty,\infty} \le \frac{L}{B}\|\mathbf{w}-\mathbf{1}\|_\infty$ holds uniformly inside the ℓ_∞ ball $\|\mathbf{w}-\mathbf{1}\|_\infty \le \frac{1}{2}$. By Lemma 12, $\|g_\beta(\mathbf{1})\|_\infty < \frac{\varepsilon}{2B}$ holds with high probability. Under the intersection of these events, $h_\beta(\mathbf{w}) = \left(\nabla g_\beta(\mathbf{1})\right)^{-1} g_\beta(\mathbf{w})$ satisfies the conditions of Lemma 13, with constants $R = \frac{1}{2}$ and L. Assuming $\varepsilon \le \varepsilon_0 := \min\{\frac{1}{2}, L^{-1}\}$, by Lemma 13 $h_\beta(\cdot)$ has a zero $\mathbf{w}^*$, equivalently a stationary point of $G_\beta(\cdot)$, with $\|\mathbf{w}^* - \mathbf{1}\|_\infty \le \varepsilon$. By Lemma 11, $\mathbf{w}^* = \widehat{\mathbf{w}}$. □

Proof of Theorem 2. Recall that the weights w_i^{Tyl} are related to $\widehat{w}_i$ via (26). Denote $\tau_p = p^{-1}\mathrm{Tr}(\Sigma_p)$. Under the high-probability event $\max_{1\le i\le n}|\widehat{w}_i - 1| \le \varepsilon$, we have

$$\frac{1-\varepsilon}{(1+\varepsilon)\cdot p^{-1}\mathrm{Tr}(S)/\tau_p} \le \tau_p w_i^{\mathrm{Tyl}} = \frac{\widehat{w}_i}{\tau_p^{-1}p^{-1}\mathrm{Tr}\left(\frac{1}{n}\sum_{i=1}^n \widehat{w}_i\mathbf{x}_i\mathbf{x}_i^\top\right)} \le \frac{1+\varepsilon}{(1-\varepsilon)\cdot p^{-1}\mathrm{Tr}(S)/\tau_p}. \qquad (30)$$

We next show that the denominator of (30) concentrates tightly around 1, namely, that with high probability $|p^{-1}\mathrm{Tr}(S) - \tau_p| \le \varepsilon\tau_p$. To this end, let $u_1,\ldots,u_p$ be an orthonormal basis of eigenvectors of Σ_p, so that $\Sigma_p u_i = \lambda_i u_i$. Since $S = \Sigma_p^{1/2}T\Sigma_p^{1/2}$,

$$|p^{-1}\mathrm{Tr}(S) - \tau_p| = \left|p^{-1}\sum_{i=1}^p \lambda_i(u_i^\top Tu_i - 1)\right| \le \tau_p\max_{1\le i\le p}|u_i^\top Tu_i - 1|.$$

By Lemma 23 (and a union bound over $1 \leq i \leq p$), under [LC], $\Pr\left(\left|p^{-1}\mathrm{Tr}(S) - \tau_p\right| \geq \varepsilon\, \tau_p\right) \leq Cpe^{-c(\Psi_p\sqrt{n})\varepsilon}$, whereas under [SG-IND] or [CCP-SBP], $\Pr\left(\left|p^{-1}\mathrm{Tr}(S) - \tau_p\right| \geq \varepsilon\, \tau_p\right) \leq Cpe^{-cn\varepsilon^2}$. Combining with (30) yields that with high probability $\max_{1\leq i\leq p} |\tau_p w_i^{\mathrm{Tyl}} - 1| \leq C\varepsilon$, and the theorem follows. □

Proof of Theorem 3 (MRE). By [59, Theorem 1], MRE exists uniquely with probability 1. We proceed similarly to the proof of Theorem 1. Define

$$\bar{h} : \mathbb{R}^n_+ \to \mathbb{R}^n_+, :\quad \bar{h}_j(\boldsymbol{d}) = \frac{1+\alpha}{p}\mathbf{x}_j^\top \left(\frac{1}{n}\sum_{i=1}^{n} u(d_i)\mathbf{x}_i\mathbf{x}_i^\top + \alpha \mathrm{I}_{p\times p}\right)^{-1} \mathbf{x}_j\,, \quad 1 \leq j \leq n\,. \tag{31}$$

By the definition of MRE, (2), its weights are $w_i^{\mathrm{MRE}} = u(\hat{d}_i)$ where $\bar{h}(\hat{\boldsymbol{d}}) = \hat{\boldsymbol{d}}$ is a fixed point. Accordingly, we study the fixed points of $\bar{h}$. Let $\hat{\boldsymbol{d}} > \mathbf{0}$ be a fixed point, and $j_{\min} = \mathrm{argmin}_{1\leq j\leq n}\, \hat{d}_j$, $j_{\max} = \mathrm{argmax}_{1\leq j\leq n}\, \hat{d}_j$. Since $u(\cdot)$ is non-increasing, $\bar{h}(\cdot)$ is non-decreasing, and so $\bar{h}(\hat{d}_{j_{\min}}\mathbf{1}) \leq \bar{h}(\hat{\boldsymbol{d}}) \leq \bar{h}(\hat{d}_{j_{\max}}\mathbf{1})$. Considering coordinates $j \in \{j_{\min}, j_{\max}\}$, and bearing in mind that $d_j = \bar{h}_j(\hat{\boldsymbol{d}})$,

$$\hat{d}_{j_{\min}} \geq (1+\alpha)p^{-1}\mathbf{x}_{j_{\min}}^\top(u(\hat{d}_{j_{\min}})S + \alpha\mathrm{I}_{p\times p})^{-1}\mathbf{x}_{j_{\min}}^\top, \quad \hat{d}_{j_{\max}} \leq (1+\alpha)p^{-1}\mathbf{x}_{j_{\max}}^\top(u(\hat{d}_{j_{\max}})S + \alpha\mathrm{I}_{p\times p})^{-1}\mathbf{x}_{j_{\max}}^\top\,. \tag{32}$$

Define the following n functions $\hat{F}_j : \mathbb{R}_+ \to \mathbb{R}_+$, $1 \leq j \leq n$, by

$$\hat{F}_j(d) = (1+\alpha)d^{-1}p^{-1}\mathbf{x}_j^\top(u(d)S + \alpha\mathrm{I}_{p\times p})^{-1}\mathbf{x}_j = (1+\alpha)p^{-1}\mathbf{x}_j^\top(\phi(d)S + \alpha d\mathrm{I}_{p\times p})^{-1}\mathbf{x}_j\,. \tag{33}$$

By assumption, $\phi(d) = du(d)$ is non-decreasing hence $\hat{F}_j$ is decreasing. Dividing the left and right inequalities in (32) by $\hat{d}_{j_{\min}}$ and $\hat{d}_{j_{\max}}$ respectively, gives $\hat{F}_{j_{\min}}(\hat{d}_{j_{\min}}) \leq 1$, $\hat{F}_{j_{\max}}(\hat{d}_{j_{\max}}) \geq 1$. Since $\hat{F}_j$ is decreasing, we deduce that $\hat{d}_{j_{\min}} \geq \hat{F}_{j_{\min}}^{-1}(1)$, $\hat{d}_{j_{\max}} \leq \hat{F}_{j_{\max}}^{-1}(1)$ provided that 1 is indeed in the range of $\hat{F}_{j_{\min}}(\cdot)$, $\hat{F}_{j_{\max}}(\cdot)$. Since $\hat{d}_{j_{\min}}, \hat{d}_{j_{\max}}$ are the smallest and largest coordinates of $\hat{\boldsymbol{d}}$ respectively, we deduce that for all $1 \leq j \leq n$,

$$\min_{1\leq i\leq n} \hat{F}_i^{-1}(1) \leq \hat{d}_j \leq \max_{1\leq i\leq n} \hat{F}_i^{-1}(1)\,, \tag{34}$$

provided that 1 is in the range of all $\hat{F}_i$-s. We shall soon see that this is indeed the case, and moreover, that the (data-dependent) quantities $\hat{F}_i^{-1}(1)$ all concentrate around a particular deterministic quantity.

We now analyze $\hat{F}_i$, defined in (33). Decomposing $S = S_{-i} + n^{-1}\mathbf{x}_i\mathbf{x}_i^\top$, by the Sherman–Morrison formula,

$$\hat{F}_i(d) = (1+\alpha)\frac{\hat{Q}_i(d)}{1+\gamma\phi(d)\hat{Q}_i(d)}, \qquad \hat{Q}_i(d) = p^{-1}\mathbf{x}_i^\top(\phi(d)S_{-i} + \alpha d\mathrm{I}_{p\times p})^{-1}\mathbf{x}_i\,. \tag{35}$$

Next we consider a deterministic analog of $\hat{F}_i$, where $\hat{Q}_i$ is replaced by its expectation Q. Define

$$Q(d) = \mathbb{E}Q_i(d) = p^{-1}\mathbb{E}\mathrm{Tr}\,\Sigma_p(\phi(d)S_{-i} + \alpha d\mathrm{I}_{p\times p})^{-1}\,, \quad F(d) = (1+\alpha)\frac{Q(d)}{1+\gamma\phi(d)Q(d)}\,. \tag{36}$$

Lemma 17. *Let $d_0 > 0$ be given. There are $c, C, \varepsilon_0 > 0$, that depend on the distribution of Y, γ, $s_{\max}$, α and d_0, such that the following holds. For all $d \geq d_0$ and $\varepsilon \leq \varepsilon_0$,*

(i) *Assume [LC]. Then $\Pr(\max_{1\leq i\leq n} |\hat{Q}_i(d) - Q(d)| \geq \varepsilon) \leq Cne^{-c\Psi_p\sqrt{n}\varepsilon}$.*
(ii) *Assume [SG-IND] or [CCP-SBP]. Then $\Pr(\max_{1\leq i\leq n} |\hat{Q}_i(d) - Q(d)| \geq \varepsilon) \leq Cne^{-cn\varepsilon^2}$.*

The proof of Lemma 17 appears in the Appendix. By Lemma 17, the functions $\hat{F}_i$ concentrate pointwise around the deterministic function F. To proceed, we show that $1 \in \mathrm{Range}(F)$ and study the local behavior of F around this point. Before stating our next result, we remark that up to this point, the analysis in this section applies both to MRE and TRE, the latter corresponding to $u(x) = x^{-1}$. The proof of the next lemma, however, relies on the boundedness of the function u, which is always assumed for MRE, but does not hold for TRE.

Lemma 18. *There is a unique root $F(d^*) = 1$. Moreover, there exist constants $0 < \underline{d} < \overline{d}$, and $\eta > 0$, depending on the distributions of Y, γ, $s_{\max}$, τ and α, so that: (1) $d^* \in (\underline{d}, \overline{d})$; (2) For every $d_1, d_2 \in (\underline{d}, \overline{d})$, $|F(d_1) - F(d_2)| \geq \eta|d_1 - d_2|$.*

The proof of Lemma 18 is deferred to the appendix. We are ready to conclude the proof of Theorem 3. Fix a small enough $\varepsilon > 0$ so that $d_1 = d^* - \varepsilon$, $d_2 = d^* + \varepsilon$ satisfy $[d_1, d_2] \subseteq (\underline{d}, \overline{d})$. Let $\eta > 0$ be the constant from Lemma 18. By Lemma 17, with high probability $|\hat{F}_i(d_\ell) - F(d_\ell)| \leq \eta\varepsilon/2$ for all $1 \leq i \leq n$ and $\ell = 1, 2$. Under this event, in particular, $\hat{F}_i(d_2) \leq F(d_2) + \eta\varepsilon/2 \leq F(d^*) - \eta\varepsilon + \eta\varepsilon/2 = 1 - \eta\varepsilon/2$. Similarly, $\hat{F}_i(d_1) \geq 1 + \eta\varepsilon/2$. Since the functions $\hat{F}_i$ are decreasing and continuous, it follows that $\hat{F}_i^{-1}(1) \in (d_1, d_2) = (d^* - \varepsilon, d^* + \varepsilon)$ for all i. Thus, by (34), $\hat{d}_j \in (d^* - \varepsilon, d^* + \varepsilon)$ for all

$1 \le j \le n$. Finally, recalling that the weights of MRE are $w_j^{\text{MRE}} = u(\hat{d}_j)$, we conclude that $|w^{\text{MRE}} - u(d^*)| \le L\varepsilon$ where L is the Lipschitz constant of u. □

Proof of Theorem 4 (TRE). Since the samples $\mathbf{x}_i$ are assumed to have a density, they are in general position with probability 1. Consequently, since by assumption $\alpha > \max\{0, \gamma - 1\}$, [59, Theorem 3] implies that TRE exists uniquely with probability 1.

Recall that TRE has the same form as MRE, with a crucial difference that the function $u(x) = 1/x$ is not bounded. We follow along the proof of Theorem 3. The argument carries over, verbatim, with the exception of Lemma 18. Thus, Theorem 4 follows from Lemma 19, stated below and proven in the Appendix. □

Lemma 19. *Let F, Q be as in (36) with $u(x) = 1/x$, and suppose that $\alpha > \max\{0, p/n - 1\}$. There is a unique root $F(d^*) = 1$. Moreover, there exist constants $0 < \underline{d} < \overline{d}$, and $\eta > 0$, depending on the distributions of Y, γ, $s_{\max}$, τ, $\underline{\tau}$ and α, so that: (1) $d^* \in (\underline{d}, \overline{d})$; (2) For every $d_1, d_2 \in (\underline{d}, \overline{d})$, $|F(d_1) - F(d_2)| \ge \eta |d_1 - d_2|$.*

6. Conclusion and further discussion

This paper presented a non-asymptotic analysis of Tyler's and Maronna's M-estimators, as well as their regularized variants, under a substantially broader class of distributions than those considered in previous works. Specifically, we assumed a data distribution of the form $X = \Sigma_p^{1/2} Y$, where Y is isotropic and satisfies one of several abstract concentration properties. Some of these distributions allow for the coordinates of Y to be statistically dependent.

Results for non-centered distributions. In our analysis, we assumed that X has zero mean. This is often not the case in real-world applications. A more reasonable model is $X = \boldsymbol{\mu} + \Sigma_p^{1/2} Y$, where $\mathbb{E}[X] = \boldsymbol{\mu}$ is in general not zero. Note that the standard method of coping with a non-zero mean, namely subtracting the sample mean, creates a statistical inter-dependency between the modified samples, so that the results of Section 2 do not immediately apply.

To overcome this difficulty, [29] suggested the following "symmetrization" procedure. Given a data set of $2n$ samples $\mathbf{x}_1, \dots, \mathbf{x}_{2n}$, construct a symmetrized set of n samples:

$$\mathbf{x}_i^{\text{sym}} = 2^{-1/2}(\mathbf{x}_i - \mathbf{x}_{i+n}) = 2^{-1/2} \Sigma_p^{1/2}(\mathbf{y}_i - \mathbf{y}_{i+n}), \quad 1 \le i \le n.$$

Clearly, $\mathbb{E}[\mathbf{x}_i^{\text{sym}}] = \mathbf{0}$ and $\text{Cov}(\mathbf{x}_i^{\text{sym}}) = \Sigma_p$. To apply our main results, the isotropic random vector $Y^{\text{sym}} = 2^{-1/2}(Y - Y')$, where Y' is an independent copy of Y, has to satisfy the same properties as Y. Indeed:

(i) under [SG-IND], Y^{sym} has independent entries, with sub-Gaussian constants $\lesssim K$. In addition, by Lemma 1, the density of each entry is bounded by $\lesssim C_0$;
(ii) under [LC], Y^{sym} is a log-concave random vector, since the log-concave family is closed under convolution of the densities (e.g., [67]);
(iii) under [CCP-SBP]: By separately conditioning on Y, Y', one may verify that Y^{sym} satisfies both the CCP and the SBP, possibly with a larger constant.

In Appendix C we discuss some further details regarding the symmetrization procedure under the model (21).

Projection pursuit in high dimension. Both [11] and recently [79] (who extended the results of [11]) studied some fundamental limitations on the ability to detect structure by projection pursuit in the high-dimensional setting, assuming multivariate Gaussian observations. Specifically, asymptotically as $p/n \to \gamma \in (1, \infty)$, they proved that with high probability, for any i.i.d. observations $X_1, \dots, X_n \sim \mathcal{N}(0, I_{p\times p})$ and a given distribution G with mean zero and variance bounded by $\gamma - 1$, one can find a sequence of (data-dependent) projections $u_{p,G} \in \mathbb{S}^{p-1}$ such that the sequence of empirical distributions $F_{n,u} := \sum_{i=1}^n \mathbb{1}_{X_i^\top \cdot u_{p,G}}(t)$ converges to G in Kolmogorov–Smirnov distance: $\lim_{n\to\infty} \|F_{n,u} - G\|_\infty = 0$. Informally speaking, this result implies that in the high dimensional regime, one can find structure in the data that does not exist in its underlying distribution, as all its marginals are $\mathcal{N}(0, 1)$. This is in contrast to the results of [27], whereby in the classical regime where $p/n \to 0$, all the empirical marginals converge to their population counterparts $\mathcal{N}(0, 1)$, namely $\lim_{n,p\to\infty} \sup_{u \in \mathbb{S}^{p-1}} \|F_{n,u} - \mathcal{N}(0, 1)\|_\infty = 0$.

The mathematical analysis in the present paper can be used to generalize the results of [11]. Specifically, most of its results on projection pursuit continue to hold for any $X = (X_1, \dots, X_p)$ where each entry is a zero mean and variance one independent sub-Gaussian with a uniform constant K (these entries need not be identically distributed).

CRediT authorship contribution statement

Elad Romanov: Conceptualization, Formal analysis, Writing. **Gil Kur:** Conceptualization, Formal analysis, Writing. **Boaz Nadler:** Conceptualization, Formal analysis, Writing.

Acknowledgments

We are grateful to Mark Rudelson for help regarding the convex concentration property; and to the anonymous reviewers, whose comments helped improve this manuscript considerably.

Appendix A. Auxiliary technical lemmas

Let $\mathbf{y}_1,\dots,\mathbf{y}_n$ be i.i.d. realizations of an isotropic random vector $Y\in\mathbb{R}^p$, with sample covariance $T=\frac{1}{n}\sum_{i=1}^n \mathbf{y}_i\mathbf{y}_i^\top$. Recall that $\gamma=\frac{p}{n}$.

The next two lemmas present well-known bounds on the largest and smallest eigenvalues of T:

Lemma 20. *There are $c_1,c_2>0$, that depend on the distribution of Y and on γ, such that:*

(i) *Assume [LC]. Then* $\Pr\left(\|T\|\ge c_1\right)\le e^{-c_2\sqrt{n}}$.
(ii) *Assume [SG-IND] or [CCP-SBP]Then* $\Pr\left(\|T\|\ge c_1\right)\le e^{-c_2 n}$.

Under [SG-IND] and [CCP-SBP], Lemma 20 follows from [74, Theorem 5.39]. Under [LC], it follows from [3, Theorem 1].

Lemma 21. *Suppose that $\gamma<1$. There are $c_1,c_2>0$, that depend on the distribution of Y and on γ, such that:*

(i) *Assume [LC]. Then* $\Pr\left(\lambda_{\min}(T)\le c_1\right)\le e^{-c_2\sqrt{n}}$.
(ii) *Assume [SG-IND] or [CCP-SBP]. Then* $\Pr\left(\lambda_{\min}(T)\le c_1\right)\le e^{-c_2 n}$.

For Y with i.i.d. sub-Gaussian entries, Lemma 21 follows from the work of Rudelson and Vershynin [64], see also [74, Theorem 5.38]. Their non-asymptotic bound remarkably captures the exact "true" location of $\lambda_{\min}(T)$; to wit, c_1 may be taken up to the Marčenko–Pastur lower edge $(1-\sqrt{\gamma})^2$. To our knowledge, for $Y\in\mathbb{R}^p$ with dependent (but uncorrelated) entries, similarly strong results are not currently available. Moreover, existing results which bound the two-sided deviation $\|T-\mathrm{I}_{p\times p}\|$ typically fail (barring the i.i.d. case) to produce a positive bound on $\lambda_{\min}(T)$ in the entire range $\gamma\in(0,1)$. As observed by [43], if Y satisfies the SBP then this difficulty can be overcome, albeit with non-sharp c_1; this will suffice for our purposes. For completeness, we give a proof of Lemma 21 in Appendix B.

Next, we state a concentration inequality for quadratic forms. Note that for any fixed matrix $A\in\mathbb{R}^{p\times p}$, $\mathbb{E}[Y^\top AY]=\mathrm{Tr}(A)$. The next lemma bounds the deviation $|Y^\top AY-\mathrm{Tr}(A)|$:

Lemma 22. *There is $c>0$ that depends on the distribution of Y, and universal $C>0$, such that for any fixed matrix $A\in\mathbb{R}^{p\times p}$ and $\varepsilon\in(0,1)$:*

(i) *Assume [SG-IND] or [CCP-SBP]. Then* $\Pr\left(\left|p^{-1}Y^\top AY-p^{-1}Tr(A)\right|\ge\varepsilon\|A\|\right)\le Ce^{-cp\varepsilon^2}$.
(ii) *Assume [LC]. Then* $\Pr\left(\left|p^{-1}Y^\top AY-p^{-1}Tr(A)\right|\ge\varepsilon\|A\|\right)\le Ce^{-c(\Psi_p\sqrt{p})\varepsilon}$.

Under [SG-IND], Lemma 22 follows from [65]. Under [CCP-SBP], it follows from [2]. Both of these are extensions of the Hanson–Wright bound [35]. A proof under [LC], appears in Appendix B. We remark that for large ε, a tighter tail bound can be derived in the log-concave case (without Ψ_p), see for example [44]. However, to the best of our knowledge, for small ε no sharper bound is currently known.

The next lemma is an entrywise concentration bound for the sample covariance matrix:

Lemma 23. *There are $C,c>0$, that depend on the distribution of Y, such that for all unit vectors u,v and $\varepsilon\in(0,1)$,*

(i) *Assume [SG-IND] or [CCP-SBP]. Then* $\Pr\left(\left|u^\top Tv-u^\top v\right|\ge\varepsilon\right)\le Ce^{-cn\varepsilon^2}$.
(ii) *Assume [LC]. Then* $\Pr\left(\left|u^\top Tv-u^\top v\right|\ge\varepsilon\right)\le Ce^{-c(\Psi_n\sqrt{n})\varepsilon}$.

Proof. Since $u^\top Tv=\frac{1}{4}[(u+v)^\top T(u+v)-(u-v)^\top T(u-v)]$, it suffices to prove the lemma for $u=v$. Consider the random vector $W\in\mathbb{R}^n$ with entries $W_i=\mathbf{y}_i^\top u$. Observe that W is centered, isotropic, and: (1) under [SG-IND] or [CCP-SBP], W has i.i.d. sub-Gaussian entries; (2) under [LC], W is log-concave. Since $u^\top Tu=\frac{1}{n}W^\top W$, the result follows from Lemma 22 with $A=I$. □

The following is well-known, see for example [25, Lemma 4]:

Lemma 24. *Let $A,B,C\succeq 0$ be non-negative matrices and $z>0$ a positive number. Then*

$$\left|TrC\,(zI+A)^{-1}-TrC\,(zI+B)^{-1}\right|\le\mathrm{rank}(A-B)\frac{\|C\|}{z}\,. \tag{A.1}$$

The following is a standard concentration inequality for "resolvent-like" expressions:

Lemma 25. *Let $C\succeq 0$ and $A\succ 0$ be fixed matrices, and let $X_1,\dots,X_n$ be independent, non-negative random matrices, such that for all i, $\mathrm{rank}(X_i)=1$ with probability 1. Denote $S_n=\sum_{i=1}^n X_i$ and $R_n=TrC\,(A+S_n)^{-1}$. There is universal $c>0$ such that for all $t\ge 0$,*

$$\Pr\left(|R_n-\mathbb{E}(R_n)|>t\right)\le 2\exp\left(-c\frac{t^2}{n\left\|CA^{-1}\right\|^2}\right)\,. \tag{A.2}$$

Proof. Consider the filtration $\mathcal{F}_k = \sigma(X_1, \dots, X_k)$ and the martingale $M_k = \mathbb{E}[R_n|\mathcal{F}_k]$. We have $M_n = R_n$, $M_0 = \mathbb{E}R_n$, and by Lemma 24, $|M_{k+1} - M_k| \leq 2\|CA^{-1}\|$. The lemma follows from the Azuma–Hoeffding inequality for martingales with bounded increments. □

Appendix B. Deferred proofs

We start with the proof of Lemma 7. Proving this lemma by a direct analysis of the quadratic form $\mathbf{x}_i^\top S^{-1}\mathbf{x}_i$ is difficult, since the sample covariance S depends on $\mathbf{x}_i$. To disentangle this dependency, as in [25], we apply the Sherman–Morrison formula,

$$S^{-1} = \left(\frac{1}{n}\sum_{j\neq i}^{n} \mathbf{x}_j\mathbf{x}_j^\top + \frac{1}{n}\mathbf{x}_i\mathbf{x}_i^\top \right)^{-1} = S_{-i}^{-1} - \frac{\frac{1}{n}S_{-i}^{-1}\mathbf{x}_i\mathbf{x}_i^\top S_{-i}^{-1}}{1 + \frac{1}{n}\mathbf{x}_i^\top S_{-i}^{-1}\mathbf{x}_i} \,.$$

Importantly, $\mathbf{x}_i$ and S_{-i} are statistically independent. Furthermore,

$$\frac{1}{p}\mathbf{x}_i^\top S^{-1}\mathbf{x}_i = \frac{1}{p}\mathbf{x}_i^\top \left[S_{-i}^{-1} - \frac{\frac{1}{n}S_{-i}^{-1}\mathbf{x}_i\mathbf{x}_i^\top S_{-i}^{-1}}{1 + \frac{1}{n}\mathbf{x}_i^\top S_{-i}^{-1}\mathbf{x}_i} \right]\mathbf{x}_i = \frac{\frac{1}{p}\mathbf{x}_i^\top S_{-i}^{-1}\mathbf{x}_i}{1 + \gamma\cdot\frac{1}{p}\mathbf{x}_i^\top S_{-i}^{-1}\mathbf{x}_i} \,, \tag{B.1}$$

and so,

$$\left|\frac{1}{p}\mathbf{x}_i^\top S^{-1}\mathbf{x}_i - 1\right| = \frac{\left|(1-\gamma)\cdot\frac{1}{p}\mathbf{x}_i^\top S_{-i}^{-1}\mathbf{x}_i - 1\right|}{1 + \gamma\cdot\frac{1}{p}\mathbf{x}_i^\top S_{-i}^{-1}\mathbf{x}_i} \leq (1-\gamma)\left|\frac{1}{p}\mathbf{x}_i^\top S_{-i}^{-1}\mathbf{x}_i - \frac{1}{1-\gamma}\right| \,. \tag{B.2}$$

The following lemma, proven below, shows that the right-hand side of (B.2) is small with high probability.

Lemma 26. *Assume $\gamma < 1$. There are $c, C, \varepsilon_0 > 0$, that depend on the distribution of Y and on γ, so that for all $\varepsilon < \varepsilon_0$:*

(i) *Assume [LC]. Then* $\Pr(\max_{1\leq i\leq n} |p^{-1}\mathbf{x}_i^\top S_{-i}^{-1}\mathbf{x}_i - \frac{1}{1-\gamma}| \geq \varepsilon) \leq Cn^2e^{-c(\Psi_p\sqrt{n})\varepsilon}$.

(ii) *Assume [SG-IND] or [CCP-SBP]. Then* $\Pr(\max_{1\leq i\leq n} |p^{-1}\mathbf{x}_i^\top S_{-i}^{-1}\mathbf{x}_i - \frac{1}{1-\gamma}| \geq \varepsilon) \leq Cn^2e^{-cn\varepsilon^2}$.

Proof of Lemma 7. Immediate from Lemma 26 combined with Eq. (B.2). □

Next, we consider the proof of Lemma 26. As previously mentioned, $\mathbf{x}_i^\top S_{-i}^{-1}\mathbf{x}_i = \mathbf{y}_i^\top T_{-i}^{-1}\mathbf{y}_i$ does not depend on the population covariance Σ_p. Hence we bound the deviations of $\mathbf{y}_i^\top T_{-i}^{-1}\mathbf{y}_i$ from $1/(1-\gamma)$. Lemma 26 then follows by a union bound over $1 \leq i \leq n$. The analysis proceeds as follows. First, we show that $p^{-1}\mathbf{y}_i^\top T_{-i}^{-1}\mathbf{y}_i$ concentrates around $\mathbb{E}[p^{-1}\mathbf{y}_i^\top T_{-i}^{-1}\mathbf{y}_i|T_{-i}] = p^{-1}\mathrm{Tr}(T_{-i}^{-1})$. Next, we prove that $p^{-1}\mathrm{Tr}(T_{-i}^{-1})$ concentrates around $1/(1-\gamma)$. Proving this directly is difficult, since the smallest eigenvalue of T_{-i} can take very small value, though with overwhelming small probability. To circumvent this, we consider a regularized variant $\hat{m}(\varepsilon) = p^{-1}\mathrm{Tr}(T_{-i} + \varepsilon \mathrm{I}_{p\times p})^{-1}$. We then show that $|\hat{m}(\varepsilon) - p^{-1}\mathrm{Tr}(T_{-i}^{-1})| \lesssim \varepsilon$, and finally that $\hat{m}(\varepsilon)$ concentrates around $1/(1-\gamma)$.

Note that $\hat{m}(\varepsilon)$ is the Stieltjes transform of the empirical spectral distribution (ESD) of T_{-i}, evaluated at $-\varepsilon$. Since T_{-i} is a sample covariance matrix of i.i.d. isotropic samples, its ESD converges to a Marčenko–Pastur law with shape parameter γ. This reveals the reason for the value $(1-\gamma)^{-1}$: it is the value of the Stieltjes transform of the Marčenko–Pastur law, evaluated at 0, cf. [8,25].

In light of the above roadmap, write $p^{-1}\mathbf{y}_i T_{-i}^{-1}\mathbf{y}_i - (1-\gamma)^{-1} = \Delta_1 + \Delta_2 + \Delta_3$, where

$$\begin{gathered} \Delta_1 = p^{-1}\mathbf{y}_i^\top T_{-i}\mathbf{y}_i - p^{-1}\mathrm{Tr}(T_{-i}^{-1}), \quad \Delta_2 = p^{-1}\mathrm{Tr}(T_{-i}^{-1}) - \widehat{m}(\varepsilon), \quad \Delta_3 = \widehat{m}(\varepsilon) - (1-\gamma)^{-1}, \\ \widehat{m}(\varepsilon) = p^{-1}\mathrm{Tr}(T_{-i} + \varepsilon \mathrm{I}_{p\times p})^{-1} \,. \end{gathered} \tag{B.3}$$

It suffices to show that with high probability, $|\Delta_\ell| \lesssim \varepsilon$ for $\ell = 1, 2, 3$. We start with a high-probability bound on Δ_1, Δ_2:

Lemma 27. *Assume the conditions of Lemma 26. There are $c_1, c_2, C_1, \varepsilon_0 > 0$, that may depend on the distribution of Y and on γ, such that for all $\varepsilon \in (0, \varepsilon_0)$,*

(i) *Assume [LC]. Then* $\Pr(|\Delta_1| \geq \varepsilon) \leq c_1e^{-c_2(\Psi_p\sqrt{p})\varepsilon}$ *and* $\Pr(|\Delta_2| \geq C_1\varepsilon) \leq c_1e^{-c_2(\Psi_p\sqrt{p})\varepsilon}$.

(ii) *Assume [SG-IND] or [CCP-SBP]. Then* $\Pr(|\Delta_1| \geq \varepsilon) \leq c_1e^{-c_2p\varepsilon^2}$ *and* $\Pr(|\Delta_2| \geq C_1\varepsilon) \leq c_1e^{-c_2p\varepsilon^2}$.

Proof. Let $C > 0$ be such that the event $\mathcal{E} = \{\lambda_{\min}(T_{-i}) \geq C\}$ holds with high probability, per Lemma 21. Of course, $\Pr(|\Delta_\ell| \geq \varepsilon) \leq \Pr(|\Delta_\ell| \geq \varepsilon \mid \mathcal{E}) + \Pr(\mathcal{E}^c)$. Starting with $\ell = 1$, since $\mathbf{y}_i$ and T_{-i}^{-1} are independent, $\Pr(|\Delta_1| \geq \varepsilon \mid \mathcal{E})$ may be bounded using Lemma 22, a concentration inequality for the quadratic form $\mathbf{y}_i \mapsto p^{-1}\mathbf{y}_i T_{-i}^{-1}\mathbf{y}_i$, applied conditionally on T_{-i}^{-1}. Importantly, note that under $\mathcal{E}$, we have $\|T_{-i}^{-1}\| = 1/\lambda_{\min}(T_{-i}) \leq 1/C$. As for $\ell = 2$, under $\mathcal{E}$,

$$|\Delta_2| = p^{-1}\mathrm{Tr}\left[T_{-i}^{-1} - (T_{-i}+\varepsilon \mathrm{I}_{p\times p})^{-1}\right] = \varepsilon \cdot p^{-1}\mathrm{Tr}\left[(T_{-i}+\varepsilon \mathrm{I}_{p\times p})^{-1}T_{-i}^{-1}\right] \le \frac{\varepsilon}{(\lambda_{\min}(T_{-i}))^2} \le C^{-2}\varepsilon,$$

and so the claimed result follows. □

Finally, Lemma 28, proven below, provides a high-probability bound for $|\Delta_3|$:

Lemma 28. *Assume the conditions of* Lemma 26. *There are* $c_1, c_2, C_1, \varepsilon_0 > 0$, *that may depend on the distribution of* Y *and on* γ, *such that for all* $\varepsilon \in (0, \varepsilon_0)$,

(i) *Assume [LC]. Then* $\Pr(|\Delta_3| \ge C_1\varepsilon) \le c_1 n e^{-c_2(\Psi_p\sqrt{p})\varepsilon}$.
(ii) *Assume [SG-IND] or [CCP-SBP]. Then* $\Pr(|\Delta_3| \ge C_1\varepsilon) \le c_1 n e^{-c_2 p\varepsilon^2}$.

Proof of Lemma 26. Recall that $|p^{-1}\mathbf{y}_i T_{-i}^{-1}\mathbf{y}_i - (1-\gamma)^{-1}| \le |\Delta_1|+|\Delta_2|+|\Delta_3|$, and so the result follows from Lemma 27, Lemma 28, and a union bound over $1 \le i \le n$. □

Proof of Lemma 28. To simplify notation, assume without loss of generality that $i = n$. Set $\bar{n} = n-1$, and let $\mathbf{K} \in \mathbb{R}^{\bar{n}\times p}$ be the matrix whose rows are $\mathbf{w}_1^\top, \dots, \mathbf{w}_{\bar{n}}^\top$, with $\mathbf{w}_j = n^{-1/2}\mathbf{y}_j$. Note that $T_{-n} = \mathbf{K}^\top\mathbf{K}$ and so $\widehat{m}(\varepsilon) = p^{-1}\mathrm{Tr}(\mathbf{K}^\top\mathbf{K}+\varepsilon\mathrm{I}_{p\times p})^{-1}$. For $1 \le j \le \bar{n}$, let $\mathbf{K}_j \in \mathbb{R}^{(\bar{n}-1)\times p}$ be obtained by removing the jth row from $\mathbf{K}$.

The proof relies on several algebraic "tricks" which are classical in random matrix theory, see [8]. Recall that for any matrix $A \in \mathbb{R}^{d_1\times d_2}$, the spectra of $AA^\top$ and $A^\top A$ are identical up to $|d_1-d_2|$ zeros. Thus,

$$\widehat{m}(\varepsilon) = p^{-1}\mathrm{Tr}(\mathbf{K}^\top\mathbf{K}+\varepsilon\mathrm{I}_{p\times p})^{-1} = p^{-1}\mathrm{Tr}(\mathbf{K}\mathbf{K}^\top+\varepsilon\mathrm{I}_{\bar{n}\times\bar{n}})^{-1} + \varepsilon^{-1}(1-p^{-1}\bar{n})\,. \tag{B.4}$$

Note that $\mathbf{K}\mathbf{K}^\top$ is the Gram matrix of the vectors $\mathbf{w}_i$. We write the jth diagonal entry of $(\mathbf{K}\mathbf{K}^\top+\varepsilon\mathrm{I}_{\bar{n}\times\bar{n}})^{-1}$ as:

$$\begin{aligned}(\mathbf{K}\mathbf{K}^\top+\varepsilon\mathrm{I}_{\bar{n}\times\bar{n}})^{-1}_{jj} &\overset{(i)}{=} \left(\varepsilon + \|\mathbf{w}_j\|^2 - \mathbf{w}_j^\top\mathbf{K}_j^\top\left(\mathbf{K}_j\mathbf{K}_j^\top + \varepsilon\mathrm{I}_{(\bar{n}-1)\times(\bar{n}-1)}\right)^{-1}\mathbf{K}_j\mathbf{w}_j\right)^{-1} \\ &\overset{(ii)}{=} \left(\varepsilon + \mathbf{w}_j^\top\left[\mathrm{I}_{p\times p} - \left(\mathbf{K}_j^\top\mathbf{K}_j+\varepsilon\mathrm{I}_{p\times p}\right)^{-1}\mathbf{K}_j^\top\mathbf{K}_j\right]\mathbf{w}_j\right)^{-1} \overset{(iii)}{=} \left(\varepsilon + \varepsilon\mathbf{w}_j^\top\left(\mathbf{K}_j^\top\mathbf{K}_j+\varepsilon\mathrm{I}_{p\times p}\right)^{-1}\mathbf{w}_j\right)^{-1},\end{aligned} \tag{B.5}$$

where: (i) follows from the block matrix inversion formula; (ii) follows since for any $f(\cdot)$ and matrix A, $Af(A^\top A)A^\top = f(AA^\top)AA^\top$, where for a symmetric matrix P, $f(P)$ is the matrix obtained by applying $f(\cdot)$ on the eigenvalues of P (the spectral calculus for symmetric matrices); this may be verified readily by considering the SVD of A; and (iii) follows by straightforward algebraic manipulation.

Consider the quadratic form $\mathbf{w}_j^\top(\mathbf{K}_j^\top\mathbf{K}_j+\varepsilon\mathrm{I}_{p\times p})^{-1}\mathbf{w}_j = \gamma p^{-1}\mathbf{y}_j^\top(\mathbf{K}_j^\top\mathbf{K}_j+\varepsilon\mathrm{I}_{p\times p})^{-1}\mathbf{y}_j$. We claim that $p^{-1}\mathbf{y}_j^\top(\mathbf{K}_j^\top\mathbf{K}_j+\varepsilon\mathrm{I}_{p\times p})^{-1}\mathbf{y}_j$ is very close (with high probability) to $\widehat{m}(\varepsilon)$, the quantity of interest. To wit, define the residual

$$\eta_j = p^{-1}\mathbf{y}_j^\top(\mathbf{K}_j^\top\mathbf{K}_j+\varepsilon\mathrm{I}_{p\times p})^{-1}\mathbf{y}_j - \widehat{m}(\varepsilon)\,, \tag{B.6}$$

so that (B.5) reads

$$(\mathbf{K}\mathbf{K}^\top+\varepsilon\mathrm{I}_{\bar{n}\times\bar{n}})^{-1}_{jj} = \varepsilon^{-1}\left(1+\gamma\widehat{m}(\varepsilon)+\gamma\eta_j\right)^{-1}\,. \tag{B.7}$$

Lemma 29. *Assume the conditions of* Lemma 26. *There are* $c_1, c_2, C_1, \varepsilon_0 > 0$, *that may depend on the distribution of* Y *and on* γ, *such that for all* $\varepsilon \in (0, \varepsilon_0)$,

(i) *Assume [LC]. Then* $\Pr(\max_{1\le j\le\bar{n}}|\eta_j| \ge C_1\varepsilon + (p\varepsilon)^{-1}) \le c_1 n e^{-c_2(\Psi_p\sqrt{p})\varepsilon}$.
(ii) *Assume [SG-IND] or [CCP-SBP]. Then* $\Pr(\max_{1\le j\le\bar{n}}|\eta_j| \ge C_1\varepsilon + (p\varepsilon)^{-1}) \le c_1 n e^{-c_2 p\varepsilon^2}$.

Proof. Decompose $\eta_j = \eta_{j,1} + \eta_{j,2}$ where

$$\eta_{j,1} = p^{-1}\mathbf{y}_j^\top(\mathbf{K}_j^\top\mathbf{K}_j+\varepsilon\mathrm{I}_{p\times p})^{-1}\mathbf{y}_j - p^{-1}\mathrm{Tr}(\mathbf{K}_j^\top\mathbf{K}_j+\varepsilon\mathrm{I}_{p\times p})^{-1}, \quad \eta_{j,2} = p^{-1}\mathrm{Tr}(\mathbf{K}_j^\top\mathbf{K}_j+\varepsilon\mathrm{I}_{p\times p})^{-1} - p^{-1}\mathrm{Tr}(\mathbf{K}^\top\mathbf{K}+\varepsilon\mathrm{I}_{p\times p})^{-1}.$$

Since $\mathbf{K}_j$ and $\mathbf{y}_j$ are independent, we can bound $|\eta_{j,1}|$ similarly to Lemma 27; we omit the details. As for $\eta_{j,2}$, note that $\mathbf{K}^\top\mathbf{K} - \mathbf{K}_j^\top\mathbf{K}_j = \mathbf{w}_j\mathbf{w}_j^\top$ is rank 1. Thus, by Lemma 24, $|\eta_{j,2}| \le (p\varepsilon)^{-1}$ with probability 1. □

Now, considering (B.7), $\left|\left(1+\gamma\widehat{m}(\varepsilon)+\gamma\eta_j\right)^{-1} - (1+\gamma\widehat{m}(\varepsilon))^{-1}\right| \le 2\gamma|\eta_j|$ holds whenever $\gamma|\eta_j| \le \frac{1}{2}$. Accordingly, $\left|\frac{1}{\bar{n}}\sum_{j=1}^{\bar{n}}\left(1+\gamma\widehat{m}(\varepsilon)+\gamma\eta_j\right)^{-1} - (1+\gamma\widehat{m}(\varepsilon))^{-1}\right| \le \max_{1\le j\le\bar{n}} 2\gamma|\eta_j|$ holds whenever $\max_{1\le j\le\bar{n}}\gamma|\eta_j| \le \frac{1}{2}$. Using Eqs. (B.4), (B.7), also $p^{-1}\bar{n} = p^{-1}n - p^{-1} = \gamma^{-1} - p^{-1}$, write

$$\widehat{m}(\varepsilon) = p^{-1}\sum_{j=1}^{\bar{n}}\varepsilon^{-1}\left(1+\gamma\widehat{m}(\varepsilon)+\gamma\eta_j\right)^{-1} + \varepsilon^{-1}(1-p^{-1}\bar{n}) = \varepsilon^{-1}(p^{-1}\bar{n})\frac{1}{\bar{n}}\sum_{j=1}^{\bar{n}}\left(1+\gamma\widehat{m}(\varepsilon)+\gamma\eta_j\right)^{-1} + \varepsilon^{-1}(1-p^{-1}\bar{n})$$

$$= \varepsilon^{-1}\left[\gamma^{-1}\left(1+\gamma\widehat{m}(\varepsilon)\right)^{-1} + 1 - \gamma^{-1} + \xi\right],$$

where $|\xi| \le 2p^{-1} + \max_{1\le j\le \bar{n}} 2\gamma|\eta_j|$ whenever $\max_{1\le j\le \bar{n}} \gamma|\eta_j| \le 1/2$. Rearranging terms, we deduce that $\widehat{m}(\varepsilon)$ satisfies the quadratic equation

$$\gamma\varepsilon\widehat{m}(\varepsilon) + (\varepsilon + 1 - \gamma - \gamma\xi)\widehat{m}(\varepsilon) - (1+\xi) = 0\,. \tag{B.8}$$

Now, assume without loss of generality that $\varepsilon \ge p^{-1/2}$; we can do this since the bound of Lemma 28 is vacuous for $\varepsilon < p^{-1/2}$, provided that c_1, c_2 are chosen appropriately. Under the high-probability event of Lemma 29, $|\xi| \le C_1\varepsilon$. Thus, under this event, $\widehat{m}(\varepsilon)$ satisfies a quadratic equation, whose coefficients are $O(\varepsilon)$-close to the coefficients of the linear equation $(1-\gamma)m - 1 = 0$; note that the unique root of the linear equation is $m = (1+\gamma)^{-1}$. Let $m_1 \le m_2$ be the two roots of (B.8). One may readily verify the following: there are $C_2, \varepsilon_0 > 0$, that depend on C_1, such that for all $\varepsilon \le \varepsilon_0$, and ξ satisfying $|\xi| \le C_1\varepsilon$, we have $|m_1 - (1-\gamma)^{-1}| \le C_2\varepsilon$ whereas $m_2 > 1/(C_2\varepsilon)$. It remains to argue that, with high probability, necessarily $\widehat{m}(\varepsilon) = m_1$. Indeed, recall that $\widehat{m}(\varepsilon) \le 1/\lambda_{\min}(T_{-i})$. By Lemma 21, we can find some $C_3 > 0$ such that $\widehat{m}(\varepsilon) \le C_3$ holds with high probability. Consequently, for all $\varepsilon < \min\{\varepsilon_0, 1/(C_2C_3)\}$, the high-probability event $\{|\xi| \le C_1\varepsilon\} \cap \{\widehat{m}(\varepsilon) \le C_3\}$ implies that necessarily $\widehat{m}(\varepsilon) = m_1$, and so $|\widehat{m}(\varepsilon) - (1-\gamma)^{-1}| \le C_2\varepsilon$. Thus, the proof of Lemma 28 is concluded. □

Proof of Lemma 14. Starting from the definition of G_β, (28), one may readily calculate,

$$(g_\beta(\mathbf{w}))_\ell = \left(\nabla G_\beta(\mathbf{w})\right)_\ell = \nabla F(\mathbf{w})_\ell + \beta\left(\sum_{i=1}^n w_i - n\right) = -\frac{1}{w_\ell} + \frac{n}{p}\mathbf{x}_\ell^\top\left(\sum_{i=1}^n w_i\mathbf{x}_i\mathbf{x}_i^\top\right)^{-1}\mathbf{x}_\ell + \beta\left(\sum_{i=1}^n w_i - n\right). \tag{B.9}$$

Taking the second derivative,

$$\left(\nabla g_\beta(\mathbf{w})\right)_{j,\ell} = \frac{1}{w_\ell^2}\mathbb{1}_{\{j=\ell\}} - \gamma\left(\frac{1}{p}\mathbf{x}_j^\top\left(\frac{1}{n}\sum_{i=1}^n w_i\mathbf{x}_i\mathbf{x}_i^\top\right)^{-1}\mathbf{x}_\ell\right)^2 + \beta\,. \tag{B.10}$$

In particular, setting $\mathbf{w} = \mathbf{1}$,

$$\nabla g_\beta(\mathbf{1}) = \mathrm{I}_{p\times p} - (A - \beta\mathbf{1}\mathbf{1}^\top), \qquad A_{j,\ell} = \gamma\left(p^{-1}\mathbf{y}_j^\top T^{-1}\mathbf{y}_\ell\right)^2. \tag{B.11}$$

Following [78, Lemma 3.3], $\left(\nabla g_\beta(\mathbf{1})\right)^{-1} = \left(\mathrm{I}_{p\times p} - (A-\beta\mathbf{1}\mathbf{1}^\top)\right)^{-1} = \sum_{k=0}^\infty \left(A - \beta\mathbf{1}\mathbf{1}^\top\right)^k$ provided that the sum converges. Since $\|\left(A-\beta\mathbf{1}\mathbf{1}^\top\right)^k\|_{\infty,\infty} \le \|A - \beta\mathbf{1}\mathbf{1}^\top\|_{\infty,\infty}^k$, the sum clearly converges when $\|A-\beta\mathbf{1}\mathbf{1}^\top\|_{\infty,\infty} < 1$, and then

$$\left\|\left(\nabla g_\beta(\mathbf{1})\right)^{-1}\right\|_{\infty,\infty} \le \sum_{k=0}^\infty \|A-\beta\mathbf{1}\mathbf{1}^\top\|_{\infty,\infty}^k \le \frac{1}{1-\|A-\beta\mathbf{1}\mathbf{1}^\top\|_{\infty,\infty}}.$$

Thus, to prove Lemma 14, it suffices to show that with high probability, $\|A-\beta\mathbf{1}\mathbf{1}^\top\|_{\infty,\infty} \le 1-c$ for some constant $c > 0$.

By definition,

$$\|A-\beta\mathbf{1}\mathbf{1}^\top\|_{\infty,\infty} = \max_{1\le j\le n}\sum_{\ell=1}^n |A_{j,\ell} - \beta|\,. \tag{B.12}$$

It is instructive to consider (B.12) with $\beta = 0$. Observe that $\sum_{\ell=1}^n |A_{j,\ell}| = \frac{1}{np}\sum_{\ell=1}^n \mathbf{y}_j^\top T^{-1}\mathbf{y}_\ell\mathbf{y}_\ell^\top T^{-1}\mathbf{y}_j = \frac{1}{p}\mathbf{y}_j^\top T^{-1}\mathbf{y}_j$, and recall that by Lemma 7, this quantity concentrates tightly around 1. Accordingly, when $\beta = 0$, the norm (B.12) concentrates around 1. Our goal, then, is to find some β so to consistently bias (B.12) away from 1. To this end, we proceed along the argument of Zhang et al. [78, Lemma 3.3]. Parameterize $\beta = \beta_0/n$, for constant β_0, so that β has the same scale as the off-diagonal entries $A_{j,\ell}$ in (B.11). Let $N_j[\beta_0]$ be the number of entries $A_{j,\ell}$, in the jth row ($1 \le \ell \le n$), such that $A_{j,\ell} \ge \beta_0/n$. For $A_{j,\ell} \ge \beta_0/n$ we have $|A_{j,\ell} - \beta_0/n| = A_{j,\ell} - \beta_0/n$ whereas if $A_{j,\ell} < \beta_0/n$, we may bound $|A_{j,\ell} - \beta_0/n| \le A_{j,\ell} + \beta_0/n$. Hence,

$$\|A-\beta\mathbf{1}\mathbf{1}^\top\|_{\infty,\infty} \le \max_{1\le j\le n}\left\{\sum_{\ell=1}^n A_{j,\ell} - \frac{\beta_0}{n}N_j[\beta_0] + (n - N_j[\beta_0])\frac{\beta_0}{n}\right\} \le \max_{1\le j\le n} p^{-1}\mathbf{y}_j^\top T^{-1}\mathbf{y}_j - 2\beta_0\min_{1\le j\le n}\left(\frac{N_j[\beta_0]}{n} - \frac{1}{2}\right). \tag{B.13}$$

Thus, to conclude the proof of the lemma, we need to find some constant β_0 so that with high probability, $\min_{1\le j\le n} N_j[\beta_0]/n \ge \frac{1}{2} + c$ (say, $c = 0.1$); that is, such that at least $(\frac{1}{2}+c)n$ of the entries of A in every row are consistently larger than β_0. Observe that for any row $1 \le j \le n$, the off-diagonal entries $A_{j,\ell}$, $\ell \ne j$, are identically distributed. A source of difficulty is that they are not independent, and this dependence is manifested in two ways: (1) They all depend on $\mathbf{y}_j$; (2) The sample covariance T depends on all $\mathbf{y}_\ell$-s. The first dependence is easy to overcome (by conditioning on $\mathbf{y}_j$), but the second one is more involved. To deal with the latter, we shall lower bound $A_{j,\ell}$ by a different set of random variables, which are easier to analyze.

The mathematical error in [78]. In their attempt to lower bound $A_{j,\ell}$, in the proof of their Lemma 3.4, [78] used the following inequality (top of page 122 in their paper),

$$A_{j,\ell} = \frac{1}{np}(\mathbf{y}_j^\top T^{-1}\mathbf{y}_\ell)^2 \geq \frac{1}{np}\cdot\frac{(\mathbf{y}_j^\top\mathbf{y}_\ell)^2}{\lambda_{\max}(T)^2}\,. \tag{B.14}$$

They next analyzed the simpler expressions for the numerator and denominator above. Unfortunately, (B.14) is false, as $|u^\top B^{-1}v| \geq |u^\top v|/\lambda_{\max}(B)$ is not generally true for positive matrices B.

A corrected argument. Let $k = o(n)$ be an integer, to be chosen later. Partition $[n] = \{1,\dots,n\}$ into $M = \lceil n/k\rceil$ subsets $\mathcal{I}_1,\dots,\mathcal{I}_M$ of size $k/2 \leq |\mathcal{I}_i| \leq k$ each. The idea is to approximate $\{\mathbf{y}_j^\top T^{-1}\mathbf{y}_\ell\}_{\ell\neq j}$ by a different set of random variables, so that variables within the same class $\ell\in\mathcal{I}_i$ are independent of one another (conditioned on $\mathbf{y}_j$). For an index ℓ, denote by $\mathcal{I}(\ell)$ the unique class such that $\ell\in\mathcal{I}(\ell)$.

For a set $\mathcal{I}\subseteq[n]$, denote by $\mathbf{Y}_{\mathcal{I}}\in\mathbb{R}^{|\mathcal{I}|\times p}$ the matrix whose rows are $\{\mathbf{y}_\ell : \ell\in\mathcal{I}\}$. Given indices $j\neq\ell$, we decompose $T = T_{-\ell} + n^{-1}\mathbf{y}_\ell\mathbf{y}_\ell^\top = T_{-\ell,-j} + n^{-1}\mathbf{y}_\ell\mathbf{y}_\ell^\top + n^{-1}\mathbf{y}_j\mathbf{y}_j^\top$. By the Sherman–Morrison formula,

$$\frac{1}{p}\mathbf{y}_j^\top T^{-1}\mathbf{y}_\ell = \frac{\frac{1}{p}\mathbf{y}_j^\top T_{-\ell}^{-1}\mathbf{y}_\ell}{1+\gamma\cdot\frac{1}{p}\mathbf{y}_\ell^\top T_{-\ell}^{-1}\mathbf{y}_\ell} = \frac{\frac{1}{p}\mathbf{y}_j^\top T_{-\ell,-j}^{-1}\mathbf{y}_\ell}{\left(1+\gamma\cdot\frac{1}{p}\mathbf{y}_\ell^\top T_{-\ell}^{-1}\mathbf{y}_\ell\right)\left(1+\gamma\cdot\frac{1}{p}\mathbf{y}_j^\top T_{-\ell,-j}^{-1}\mathbf{y}_j\right)}\,. \tag{B.15}$$

We next simplify $T_{-\ell,-j}^{-1}$. Decompose $T_{-\ell,-j} = T_{-j,-\mathcal{I}(\ell)} + n^{-1}\mathbf{Y}_{\mathcal{I}(\ell)\setminus\{j,\ell\}}^\top\mathbf{Y}_{\mathcal{I}(\ell)\setminus\{j,\ell\}}$. Recall that by the Woodbury formula, for invertible matrices A, C, one has $(A+UCV)^{-1} = A^{-1} - A^{-1}U\left(C^{-1}+VA^{-1}U\right)^{-1}VA^{-1}$. Applying this with $A = T_{-\ell,-j}^{-1}$, $C = I$, $U = V^\top = n^{-1/2}\mathbf{Y}_{\mathcal{I}(\ell)\setminus\{j,\ell\}}^\top$, gives

$$T_{-\ell,-j}^{-1} = T_{-j,-\mathcal{I}(\ell)}^{-1} - n^{-1}T_{-j,-\mathcal{I}(\ell)}^{-1}\mathbf{Y}_{\mathcal{I}(\ell)\setminus\{j,\ell\}}^\top\left(I + n^{-1}\mathbf{Y}_{\mathcal{I}(\ell)\setminus\{j,\ell\}}T_{-j,-\mathcal{I}(\ell)}^{-1}\mathbf{Y}_{\mathcal{I}(\ell)\setminus\{j,\ell\}}^\top\right)^{-1}\mathbf{Y}_{\mathcal{I}(\ell)\setminus\{j,\ell\}}T_{-j,-\mathcal{I}(\ell)}^{-1}\,. \tag{B.16}$$

Let $\underline{\Omega}$ be the following upper bound on the denominator of (B.15):

$$\underline{\Omega} = \max\left\{\left(1+\gamma\cdot\frac{1}{p}\mathbf{y}_\ell^\top T_{-\ell}^{-1}\mathbf{y}_\ell\right)\left(1+\gamma\cdot\frac{1}{p}\mathbf{y}_j^\top T_{-\ell,-j}^{-1}\mathbf{y}_j\right) \;:\; j\in[n],\,\ell\in[n]\setminus\{j\}\right\}\,. \tag{B.17}$$

Similarly, define

$$\overline{\Omega} = \max\left\{(np)^{-1}\left\|\mathbf{Y}_{\mathcal{I}(\ell)\setminus\{j,\ell\}}T_{-j,-\mathcal{I}(\ell)}^{-1}\mathbf{y}_\ell\right\|\left\|\mathbf{Y}_{\mathcal{I}(\ell)\setminus\{j,\ell\}}T_{-j,-\mathcal{I}(\ell)}^{-1}\mathbf{y}_j\right\| \;:\; j\in[n],\,\ell\in[n]\setminus\{j\}\right\}\,. \tag{B.18}$$

Observe that per (B.16), $|\frac{1}{p}\mathbf{y}_j^\top T_{-\ell,-j}^{-1}\mathbf{y}_\ell| \geq |\frac{1}{p}\mathbf{y}_j^\top T_{-j,-\mathcal{I}(\ell)}^{-1}\mathbf{y}_\ell| - \overline{\Omega}$. Finally, denote

$$\xi_{j,\ell} = p^{-1/2}\left\langle\frac{T_{-j,-\mathcal{I}(\ell)}^{-1}\mathbf{y}_j}{\left\|T_{-j,-\mathcal{I}(\ell)}^{-1}\mathbf{y}_j\right\|},\mathbf{y}_\ell\right\rangle\,,\qquad \nu = \min_{j,\ell}\left\|T_{-j,-\mathcal{I}(\ell)}^{-1}\mathbf{y}_j\right\|\,. \tag{B.19}$$

Importantly, observe that the random variables $\xi_{j,\ell}$ within the same class $\ell\in\mathcal{I}_i$ are statistically independent of one another, conditioned on $\mathbf{y}_j$. Combining (B.15)–(B.20) yields the following lower bound,

$$\left|\frac{1}{p}\mathbf{y}_j^\top T^{-1}\mathbf{y}_\ell\right| \geq p^{-1/2}|\xi_{j,\ell}|\frac{\nu}{\underline{\Omega}} - \frac{\overline{\Omega}}{\underline{\Omega}}\,. \tag{B.20}$$

Next we derive high-probability bounds on $\overline{\Omega}$, $\underline{\Omega}$, ν.

Lemma 30. *For a number C_1, let $\mathcal{E}_{\text{Lem.30}}$ be the event that: (1) $\underline{\Omega}\leq C_1$; (2) $\nu\geq 1/C_1$; (3) $\overline{\Omega}\leq C_1\frac{k}{n}$.*

Assume $k\leq\frac{1-\gamma}{1+\gamma}n-1$. There are $c, C, C_1>0$, that depend on the distribution of Y and on γ, so that

(i) *Assume [LC]. Then $\Pr(\mathcal{E}_{\text{Lem.30}}^c)\leq Cn^2e^{-c\Psi_p\sqrt{k}}$.*
(ii) *Assume [SG-IND] or [CCP-SBP]. Then $\Pr(\mathcal{E}_{\text{Lem.30}}^c)\leq Cn^2e^{-ck}$.*

Proof. We start by bounding $\underline{\Omega}$ and ν. Considering their definitions, in (B.17) and (B.19), it suffices to show that the following are all high-probability events (for some constant $C>0$): (I) $\max_{j,\ell}\lambda_{\max}(T_{-j,\mathcal{I}(\ell)})\leq C$; (II) $\min_{j,\ell}\lambda_{\max}(T_{-j,\mathcal{I}(\ell)})\geq 1/C$; (III) $\max_\ell|p^{-1}\|\mathbf{y}_\ell\|^2-1|\leq\frac{1}{2}$. Observe that $T_{-j,\mathcal{I}(\ell)}$ is,

up to the normalization, the sample covariance of at least $n-k-1\geq\frac{2\gamma}{1+\gamma}n = \frac{1}{1/2+\gamma/2}p$ i.i.d. measurements. Thus, conditions (I) and (II) can be verified using Lemmas 20 and 21 respectively. Condition (III) can be verified using Lemma 22. We omit the details.

We next consider $\overline{\Omega}$, defined in (B.18). Let us bound, $\omega_{j,\ell} := (np)^{-1} \left\| \mathbf{Y}_{\mathcal{I}(\ell)\setminus\{j,\ell\}} T^{-1}_{-j,-\mathcal{I}(\ell)} \mathbf{y}_\ell \right\|^2$, so that a high-probability bound on $\overline{\Omega}$ may be attained by union bound over ℓ, j. Denote $u_{j,\ell} = T^{-1}_{-j,-\mathcal{I}(\ell)}\mathbf{y}_\ell$, $\hat{u}_{j,\ell} = u_{j,\ell}/\|u_{j,\ell}\|$ so $\omega_{j,\ell} = \frac{k}{n} \cdot p^{-1}\|u_{j,\ell}\|^2 \cdot \hat{u}_{j,\ell}^\top \left(\frac{1}{k} \mathbf{Y}^\top_{\mathcal{I}(\ell)\setminus\{j,\ell\}} \mathbf{Y}_{\mathcal{I}(\ell)\setminus\{j,\ell\}} \right) \hat{u}_{j,\ell}$. The terms $p^{-1}\|u_{j,\ell}\|^2$ may be treated upper-bounded similarly to the previous paragraph. As for the term $\hat{u}_{j,\ell}^\top \left(\frac{1}{k} \mathbf{Y}^\top_{\mathcal{I}(\ell)\setminus\{j,\ell\}} \mathbf{Y}_{\mathcal{I}(\ell)\setminus\{j,\ell\}} \right) \hat{u}_{j,\ell}$, observe that $\frac{1}{k} \mathbf{Y}^\top_{\mathcal{I}(\ell)\setminus\{j,\ell\}} \mathbf{Y}_{\mathcal{I}(\ell)\setminus\{j,\ell\}}$ is a sample covariance matrix consisting of (at most) k samples, and $\hat{u}_{j,\ell}$ is a unit vector which is statistically independent of $\mathbf{Y}_{\mathcal{I}(\ell)\setminus\{j,\ell\}}$. By Lemma 23, this quadratic form is bounded by a constant with high probability. □

Next, we show that with high probability, there are many large $|\xi_{j,\ell}|$-s. For a number α, let $\tilde{N}_j[\alpha]$ be the number of variables $\xi_{j,\ell}$ $(1 \le \ell \le n)$ in row j, such that $|\xi_{j,\ell}| \ge \alpha$.

Lemma 31. *There are $c, C, \alpha_* > 0$, that depend on the distribution of Y, so that* $\Pr(\min_{1\le j\le n} \tilde{N}_j[\alpha_*] \le 0.6n) \le Cn^2 e^{-ck}$.

Proof. For $1 \le i \le M$, let $\tilde{N}_j^i[\alpha] = \sum_{\ell\in\mathcal{I}_i} \mathbb{1}_{\{|\xi_{j,\ell}|\ge\alpha\}}$, so that $\tilde{N}_j[\alpha] = \sum_{i=1}^M \tilde{N}_j^i[\alpha]$. Let $\mathcal{F}_{j,i}$ be the σ-algebra generated by $\{\mathbf{y}_j\} \cup \{\mathbf{y}_\ell\}_{\ell\notin\mathcal{I}_i}$. Conditioned on $\mathcal{F}_{j,i}$, $\{\xi_{j,\ell}\}_{\ell\in\mathcal{I}_i\setminus\{j\}}$ are i.i.d. Since Y satisfies the SBP, Definition 3, there is some α_* such that $\Pr(|\xi_{j,\ell}| \ge \alpha_* | \mathcal{F}_{j,i}) \ge 0.8$. By Hoeffding's inequality, $\Pr(\tilde{N}_j^i[\alpha_*] \le 0.7|\mathcal{I}_i \setminus \{j\}| \, |\mathcal{F}_{j,i}) \le 2e^{-c_1|\mathcal{I}_i\setminus\{j\}|} \le 2e^{-c_2 k}$. Taking a union bound over $i \in [M]$, with probability $\ge 1 - 2ne^{-c_2 k}$ it holds that $\tilde{N}_j^i[\alpha_*] \ge 0.7|\mathcal{I}_i \setminus \{j\}|$ simultaneously for all i, and in particular $\sum_{i=1}^M \tilde{N}_j^i[\alpha_*] \ge 0.7(n-1)$. To finish the proof of the Lemma, take a union bound over all $1 \le j \le n$. □

We are ready to conclude the proof of Lemma 14. Set $k = c_K n^{1/2}$, for a small constant $c_K > 0$, to be chosen momentarily. By Lemma 30 and Eq. (B.20), there are C_1, C_2, α_* such that with high probability,

$$p^{-1}|\mathbf{y}_j T^{-1}\mathbf{y}_\ell| \ge n^{-1/2}(C_1|\xi_{j,\ell}| - C_2 c_K) \quad \text{for all } j, l \in [n],\ \ell \ne j$$

and

$$\tilde{N}_j[\alpha_*] \ge 0.6n \quad \text{for all } j \in [n].$$

Accordingly, choose c_K so that $C_2 c_K \le 0.5 C_1 \alpha_*$. Recall, by Eq. (B.11), that $A_{j,\ell} = \gamma \left(p^{-1}|\mathbf{y}_j T^{-1}\mathbf{y}_\ell|\right)^2$. Taking $\beta_0 = \gamma(0.5C_1\alpha_*)^2$, observe that the high-probability event above implies that $\min_{1\le j\le n} N_j[\beta_0] \ge 0.6n$, that is, each row j of A contains at least $0.6n$ entries $A_{j,\ell}$ satisfying $A_{j,\ell} \ge \beta_0/n$. As explained in the beginning of this section, this establishes the proof of the Lemma. □

Proof of Lemma 15. Denote $S(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n w_i \mathbf{x}_i \mathbf{x}_i$, $T(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n w_i \mathbf{y}_i \mathbf{y}_i^\top$, so that $\mathbf{x}_j^\top S^{-1}(\mathbf{w})\mathbf{x}_\ell = \mathbf{y}_j^\top T^{-1}(\mathbf{w})\mathbf{y}_\ell$. By (B.10),

$$\left| \left(\nabla g_\beta(\mathbf{w}) - \nabla g_\beta(\mathbf{1}) \right)_{j,\ell} \right| = \left| \frac{1}{w_\ell^2} - 1 \right| \mathbb{1}_{\{j=\ell\}} + \frac{1}{np} \left| (\mathbf{y}_j^\top T^{-1}(\mathbf{1})\mathbf{x}_\ell)^2 - (\mathbf{y}_j^\top T^{-1}(\mathbf{w})\mathbf{y}_\ell)^2 \right| .$$

If $\|\mathbf{w} - \mathbf{1}\|_\infty \le 1/2$ then $\left| \frac{1}{w_\ell^2} - 1 \right| = \left| \frac{1}{w_\ell^2} + \frac{1}{w_\ell} \right| |1 - w_\ell| \le 6\|\mathbf{w} - \mathbf{1}\|_\infty$. And so, for all $1 \le j \le n$,

$$\sum_{\ell=1}^n \left| \left(\nabla g_\beta(\mathbf{w}) - \nabla g_\beta(\mathbf{1}) \right)_{j,\ell} \right| \le 6\|\mathbf{w} - \mathbf{1}\|_\infty + \frac{1}{np} \sum_{\ell=1}^n \left| (\mathbf{y}_j^\top T^{-1}(\mathbf{1})\mathbf{y}_\ell)^2 - (\mathbf{y}_j^\top T^{-1}(\mathbf{w})\mathbf{y}_\ell)^2 \right| .$$

Write

$$\begin{aligned}(\mathbf{y}_j^\top T^{-1}(\mathbf{1})\mathbf{y}_\ell)^2 - (\mathbf{y}_j^\top T^{-1}(\mathbf{w})\mathbf{y}_\ell)^2 &= \left((\mathbf{y}_j^\top T^{-1}(\mathbf{1})\mathbf{y}_\ell) + (\mathbf{y}_j^\top T^{-1}(\mathbf{w})\mathbf{y}_\ell) \right) \left((\mathbf{y}_j^\top T^{-1}(\mathbf{1})\mathbf{y}_\ell) - (\mathbf{y}_j^\top T^{-1}(\mathbf{w})\mathbf{y}_\ell) \right) \\ &= \mathbf{y}_j^\top \left(T^{-1} + T^{-1}(\mathbf{w}) \right) \mathbf{y}_l \cdot \mathbf{y}_j^\top \left(T^{-1} - T^{-1}(\mathbf{w}) \right) \mathbf{y}_l \,.\end{aligned}$$

By Cauchy–Schwarz, $\frac{1}{np}\sum_{\ell=1}^n \left| (\mathbf{y}_j^\top T^{-1}(\mathbf{1})\mathbf{y}_\ell)^2 - (\mathbf{y}_j^\top T^{-1}(\mathbf{w})\mathbf{y}_\ell)^2 \right| \le (I_1 I_2)^{1/2}$, where

$$I_1 = \frac{1}{np} \sum_{\ell=1}^n \left(\mathbf{y}_j^\top \left(T^{-1} + T^{-1}(\mathbf{w}) \right) \mathbf{y}_l \right)^2 = p^{-1} \mathbf{y}_j^\top \left((T^{-1} + T^{-1}(\mathbf{w})) T (T^{-1} + T^{-1}(\mathbf{w})) \right) \mathbf{y}_j \,,$$

$$I_2 = \frac{1}{np} \sum_{\ell=1}^n \left(\mathbf{y}_j^\top \left(T^{-1} - T^{-1}(\mathbf{w}) \right) \mathbf{y}_l \right)^2 = p^{-1} \mathbf{y}_j^\top \left((T^{-1} - T^{-1}(\mathbf{w})) T (T^{-1} - T^{-1}(\mathbf{w})) \right) \mathbf{y}_j \,.$$

$\|\mathbf{w} - \mathbf{1}\|_\infty \le 1/2$ implies $\|T^{-1} + T^{-1}(\mathbf{w})\| \le 3\|T^{-1}\|$, and $\|T^{-1} - T^{-1}(\mathbf{w})\| = \|T^{-1}(T(\mathbf{w}) - T)(T^{-1}(\mathbf{w}))\| \le (3/2)\|T^{-1}\|^2\|T\|\|\mathbf{w} - \mathbf{1}\|_\infty$. Thus, for numerical $c > 0$, $(I_1 I_2)^{1/2} \le cp^{-1}\|\mathbf{y}_j\|^2\|T^{-1}\|^3\|T\|^2\|\mathbf{w} - \mathbf{1}\|_\infty$. We get

$$\|\nabla g_\beta(\mathbf{w}) - \nabla g_\beta(\mathbf{1})\|_{\infty,\infty} = \max_{1\le j\le n} \sum_{\ell=1}^n \left| \left(\nabla g_\beta(\mathbf{w}) - \nabla g_\beta(\mathbf{1}) \right)_{j,\ell} \right| \le \left(3 + c\|T^{-1}\|^3\|T\|^2 \max_{1\le j\le n} p^{-1}\|\mathbf{y}_j\|^2 \right) \|\mathbf{w} - \mathbf{1}\|_\infty .$$

For conclude the proof, recall that with high probability: (1) $\|T\| \le C_1$ (by Lemma 20); (2) $\|T^{-1}\| \le C_2$ (by Lemma 21); and ((3) $\max_{1\le j\le p} p^{-1}\|\mathbf{y}_j\|^2 \le C_3$ (by Lemma 22 and a union bound over $1 \le j \le n$). □

Proof of Lemma 17. Write $\hat{Q}_i(d) - Q(d) = \Delta_1 + \Delta_2$ where

$$\begin{aligned}\Delta_1 &= p^{-1}\mathbf{x}_i^\top(\phi(d)S_{-i} + \alpha d\mathrm{I}_{p\times p})^{-1}\mathbf{x}_i - p^{-1}\mathrm{Tr}\,\Sigma_p(\phi(d)S_{-i} + \alpha d\mathrm{I}_{p\times p})^{-1}\,,\\ \Delta_2 &= p^{-1}\mathrm{Tr}\,\Sigma_p(\phi(d)S_{-i} + \alpha d\mathrm{I}_{p\times p})^{-1} - p^{-1}\mathbb{E}\mathrm{Tr}\,\Sigma_p(\phi(d)S_{-i} + \alpha d\mathrm{I}_{p\times p})^{-1}\,.\end{aligned}$$

To bound $|\Delta_1|$, use Lemma 22, applied to the quadratic form $p^{-1}\mathbf{y}_i^\top B\mathbf{y}_i$, where $B = \Sigma_p^{1/2}(\phi(d)S_{-i}+\alpha d\mathrm{I}_{p\times p})^{-1}\Sigma_p^{1/2}$. Since $\|\Sigma_p\| \le s_{\max}$ and $d \ge d_0$ then $\|B\| \le s_{\max}(\alpha d_0)^{-1}$. To bound $|\Delta_2|$, use Lemma 25 with $S_n = S_{-i}$, $C = \Sigma_p$, $A = \alpha d\mathrm{I}_{p\times p}$. □

Proof of Lemma 18. We first show that d^* indeed exists. By (36), the functions Q, F are continuous and strictly decreasing. Since $\phi(d) = du(d)$ and u is bounded, then $\lim_{d\to 0}\phi(d) = 0$. This, in turn, implies $\lim_{d\to 0} Q(d) = \infty$ and $\lim_{d\to 0} F(d) = \infty$. Moreover, $\lim_{d\to\infty} Q(d) = 0$, and so $\lim_{d\to\infty} F(d) = 0$. Consequently, $d^* = F^{-1}(1)$ exists uniquely.

Next, we bound d^*. To this end, we first upper bound F. By Eq. (36), $Q(d) \le p^{-1}\mathbb{E}\mathrm{Tr}\,\Sigma_p(\alpha d\mathrm{I}_{p\times p})^{-1} = \tau_p/(\alpha d)$, where $\tau_p = p^{-1}\mathrm{Tr}\,\Sigma_p \le s_{\max}$. Since $Q(d) \ge 0$, clearly $F(d) \le (1+\alpha)Q(d) \le (s_{\max}\frac{1+\alpha}{\alpha})d^{-1}$. Setting $d = d^*$, $F(d^*) = 1$, we conclude $d^* \le \bar{d} = \frac{1+\alpha}{\alpha}s_{\max}$.

To lower bound d^*, we need a lower bound on Q. Clearly, $S_{-i} \preceq \|S_{-i}\|\mathrm{I}_{p\times p}$, so

$$Q(d) \ge p^{-1}\mathrm{Tr}\,\Sigma_p\mathbb{E}\left(\phi(d)\|S_{-i}\|\mathrm{I}_{p\times p} + \alpha d\mathrm{I}_{p\times p}\right)^{-1} \ge \tau\cdot\mathbb{E}\frac{1}{\phi(d)\|S_{-i}\| + \alpha d}\,. \tag{B.21}$$

By Lemma 20, there is some $C_0 > 0$ such that $\|S_{-i}\| \le s_{\max}C_0$ holds with probability $\ge \frac{1}{2}$. Moreover, $\phi(d) = u(d)d \le u(0)d$, since u is decreasing. Combining this with (B.21) yields $Q(d) \ge \frac{1}{2}\frac{\tau}{u(d)dC_0s_{\max}+\alpha d} = C_1/d$. Plugging this bound and the previously derived upper bound $Q(d) \le C_2/d$ into (36) yields $F(d) \ge (1+\alpha)\frac{C_1/d}{1+\gamma\phi(d)(C_2/d)} \ge (1+\alpha)\frac{C_1/d}{1+\gamma u(0)C_2}$, from which a lower bound on d^* follows.

Finally, it remains to show that for some η, $|F(d_1) - F(d_2)| \ge \eta|d_1 - d_2|$ inside the interval $[\underline{d}, \bar{d}]$. Let $d_1 \le d_2$. Then,

$$\begin{aligned}(1+\alpha)^{-1}(F(d_1) - F(d_2)) &= \frac{Q(d_1)}{1+\gamma\phi(d_1)Q(d_1)} - \frac{Q(d_2)}{1+\gamma\phi(d_2)Q(d_2)} = \frac{Q(d_1) - Q(d_2) + \gamma(\phi(d_2) - \phi(d_1))Q(d_1)Q(d_2)}{(1+\gamma\phi(d_1)Q(d_1))(1+\gamma\phi(d_2)Q(d_2))}\\ &\overset{(\star)}{\ge} \frac{Q(d_1) - Q(d_2)}{(1+\gamma\phi(d_1)Q(d_1))(1+\gamma\phi(d_2)Q(d_2))}\,,\end{aligned}$$

where in $(\star)$ we used $\phi(d_2) - \phi(d_1) \ge 0$, since ϕ is non-decreasing. The denominator is upper bounded by a constant inside the interval, and the numerator is non-negative. Thus, it suffices to show that $Q(d_1) - Q(d_2) \ge \eta_0(d_1 - d_2)$ for some $\eta_0 > 0$. Since ϕ is non-decreasing, $Q(d_1) = p^{-1}\mathbb{E}\mathrm{Tr}\,\Sigma_p(\phi(d_1)S_{-i} + \alpha d_1\mathrm{I}_{p\times p})^{-1} \ge p^{-1}\mathbb{E}\mathrm{Tr}\,\Sigma_p(\phi(d_2)S_{-i} + \alpha d_1\mathrm{I}_{p\times p})^{-1}$, and so

$$Q(d_1) - Q(d_2) \ge p^{-1}\mathbb{E}\mathrm{Tr}\,\Sigma_p(\phi(d_2)S_{-i} + \alpha d_2\mathrm{I}_{p\times p})^{-1}((d_1 - d_2)\mathrm{I}_{p\times p})(\phi(d_2)S_{-i} + \alpha d_1\mathrm{I}_{p\times p})^{-1}.$$

Consider the matrix $A = (\phi(d_2)S_{-i} + \alpha d_2\mathrm{I}_{p\times p})^{-1}(\phi(d_2)S_{-i} + \alpha d_1\mathrm{I}_{p\times p})^{-1}$. It is positive, being the product of commuting positive matrices. Consequently, $Q(d_1) - Q(d_2) \ge (d_1 - d_2)p^{-1}\mathbb{E}\mathrm{Tr}\,\Sigma_pA \ge (d_1 - d_2)\tau_p\mathbb{E}\lambda_{\min}(A)$. Lastly, $\mathbb{E}\lambda_{\min}(A) \ge \eta_0$ for some η_0 since, by Lemma 20, for some C, $\|S_{-i}\| \le Cs_{\max}$ holds with high probability. □

Proof of Lemma 19. We focus on Item (1), namely showing the existence and boundedness of d^*. The proof of Item (2) is identical to the corresponding part in Lemma 18 (MRE). Using $\phi(d) = 1$ in (36), $F(d^*) = 1$ is equivalent to

$$Q(d^*) = \frac{1}{1+\alpha-\gamma}\,,\qquad Q(d) = p^{-1}\mathbb{E}\mathrm{Tr}\,\Sigma_p(S_{-i} + \alpha d\mathrm{I}_{p\times p})^{-1}\,. \tag{B.22}$$

Since $\alpha > \max\{0, \gamma - 1\}$ then $\frac{1}{1+\alpha-\gamma} > 0$. The function Q is positive, continuous and strictly decreasing, with $\lim_{d\to\infty} Q(d) = 0$. We have $Q(d) \le s_{\max}/(\alpha d)$, and therefor if a solution $Q(d^*) = (1+\alpha-\gamma)^{-1}$ exists, then necessarily $d^* \le \bar{d} \le \frac{s_{\max}}{\alpha}(1+\alpha-\gamma)$. As for establishing existence, by continuity it suffices to show that $\lim_{d\to 0} Q(d) > (1+\alpha-\gamma)^{-1}$; in other words, we need to study the behavior of $Q(d)$ near $d = 0$. To this end, we consider separately the regimes $\gamma < 1$ and $\gamma \ge 1$, noting that S_{-i} is only invertible (with probability 1) in the regime $p/n = \gamma < 1$.

The case $\gamma < 1$. Note that $(1+x)^{-1} \ge 1 - x$ for all $x \ge 0$. Thus, for any non-negative matrix P, $(I+P)^{-1} \succeq I - P$. Write $\Sigma_p^{1/2}(S_{-i}+\alpha d\mathrm{I}_{p\times p})^{-1}\Sigma_p^{1/2} = (T_{-i}+\alpha d\Sigma_p^{-1})^{-1} = T_{-i}^{-1/2}(\mathrm{I}_{p\times p}+\alpha dT_{-i}^{-1/2}\Sigma_p^{-1}T_{-i}^{-1/2})T_{-i}^{-1/2}$, and so $\Sigma_p^{1/2}(S_{-i}+\alpha d\mathrm{I}_{p\times p})^{-1}\Sigma_p^{1/2} \succeq T_{-i}^{-1} - \alpha dT_{-i}^{-1}\Sigma_p^{-1}T_{-i}^{-1} \succeq T_{-i}^{-1} - \alpha d\Sigma_p^{-1}/\lambda_{\min}(T_{-i})^2$. For an event $\mathcal{E}$, denote for brevity $\mathbb{E}^{\mathcal{E}}[\cdot] = \mathbb{E}[\cdot\mathbb{1}_{\{\mathcal{E}\}}]$. By Lemmas 26 and 21, assuming large enough n, there is $C_* = C_*(\gamma, Y, \alpha)$ such that for the event $\mathcal{E} = \{\lambda_{\min}(T_{-i}) \ge C_*\}$, we have $\mathbb{E}^{\mathcal{E}}[p^{-1}\mathrm{Tr}T_{-i}^{-1}] \ge \frac{1}{2}(\frac{1}{1-\gamma} + \frac{1}{1-\gamma+\alpha})$. Thus,

$$Q(d) \ge p^{-1}\mathbb{E}^{\mathcal{E}}\mathrm{Tr}\,\Sigma_p^{1/2}(S_{-i}+\alpha d\mathrm{I}_{p\times p})^{-1}\Sigma_p^{1/2} \ge p^{-1}\mathbb{E}^{\mathcal{E}}T_{-i}^{-1} - \alpha dp^{-1}\mathbb{E}^{\mathcal{E}}\mathrm{Tr}\,\Sigma_p^{-1}/\lambda_{\min}(T_{-i})^2 \ge \frac{1}{2}\left(\frac{1}{1-\gamma} + \frac{1}{1-\gamma+\alpha}\right) - \frac{\alpha}{C_*^2}\underline{\tau}\cdot d.$$

This implies that $\lim_{d\to 0} Q(d) > (1-\gamma+\alpha)^{-1}$; moreover, setting $d = d^*$, yields an explicit lower bound on d^*.

Remark 5. When α is sufficiently large, we can obtain a lower bound $\underline{d}$ which does not depend on $p^{-1}\mathrm{Tr}\Sigma_p^{-1}$, similarly to the analysis of [33]. They use (B.21), recalling that by Lemma 20, there is $\overline{C} = \overline{C}(\gamma, Y)$ such that $\Pr(\|S_{-i}\| \le \overline{C}s_{\max}) = 1 - o(1)$. Thus, $Q(d) \ge (1-o(1))\tau_p(\overline{C}s_{\max} + \alpha d)^{-1}$, which yields a positive lower bound on d^* whenever $\frac{1}{1+\alpha-\gamma} < \frac{\tau}{\overline{C}s_{\max}}$, that is, $\alpha > \gamma - 1 + \overline{C}s_{\max}/\tau$. The resulting lower bound may be arbitrarily better (larger) than the previously derived lower bound (which depends on $\underline{\tau}$), since $p^{-1}\mathrm{Tr}\Sigma_p^{-1}$ may be very large when Σ_p has only one eigenvalue close to 0.

The case $\gamma \ge 1$. We analyze this case essentially by reduction to the case $\gamma < 1$. Making explicit the dependence of Q on n, denote $Q_n(d) = p^{-1}\mathbb{E}\mathrm{Tr}\Sigma_p(S_{-i}^{(n)} + \alpha d \mathrm{I}_{p\times p})^{-1}$, where $S^{(n)} = n^{-1}\sum_{j=1}^n \mathbf{x}_j\mathbf{x}_j^\top$ is the sample covariance of n i.i.d. measurements. For an integer $m \ge 0$, let $\mathbf{x}_{n+1}, \dots, \mathbf{x}_{n+m}$ be m new i.i.d. samples from X. With probability 1,

$$\mathrm{Tr}\Sigma_p(S_{-i}^{(n)} + \alpha d\mathrm{I}_{p\times p})^{-1} \ge p^{-1}\mathrm{Tr}\Sigma_p\left(\frac{1}{n}\sum_{j=1}^{n+m-1}\mathbf{x}_j\mathbf{x}_j^\top + \alpha d\mathrm{I}_{p\times p}\right)^{-1} = \frac{n}{n+m}p^{-1}\mathrm{Tr}\Sigma_p\left(S_{-i}^{(n+m)} + \alpha\frac{n}{n+m}d\mathrm{I}_{p\times p}\right)^{-1}.$$

Consequently, $Q_n(d) \ge \frac{n}{n+m}Q_{n+m}\left(\frac{n}{n+m}d\right)$. For $n+m-p = \Omega(p)$, Lemma 26 implies, assuming n is large, that $Q_{n+m}(0) \ge 0.99\frac{1}{1-\frac{p}{n+m}}$, hence $\frac{n}{n+m}Q_{n+m}(0) \ge 0.99\frac{n}{n+m-p} = 0.99\frac{1}{1+m/n-\gamma}$. Clearly, we may set $m = c_0 n$ for some $c_0 = c_0(\gamma,\alpha)$ such that $0.99\frac{1}{1+c_0-\gamma}$ is larger than $\frac{1}{1-\gamma+\alpha}$ by a constant. We can then lower bound $Q_{n+m}(d)$ following the same argument as in the case $\gamma < 1$, noting that $S^{(n+m)}$ is invertible with probability 1; we omit the details. □

Proof of Lemma 21. Recall that under either assumption on Y, [SG-IND], [CCP-SBP] or [LC], it satisfies the SBP with some constant C_0 (for [SG-IND] and [LC], see Lemmas 1 and 3 respectively). Let $\mathbf{Y} \in \mathbb{R}^{n\times p}$ be the matrix whose rows are $\mathbf{y}_1^\top, \dots, \mathbf{y}_n^\top$. Since $T = n^{-1}\mathbf{Y}^\top\mathbf{Y}$, our goal is to show that with high probability, $\sigma_{\min}(\mathbf{Y}) = \min_{v\in\mathbb{S}^{p-1}}\|\mathbf{Y}v\| = \Omega(\sqrt{n})$.

The following proof is based on [63, Corollary 4.6]. We first show that for any fixed $v \in \mathbb{S}^{p-1}$, $\|\mathbf{Y}v\|$ is large with high probability; the desired result will then follow by a standard net argument. Let $t > 0$ and $s \in (0,1)$; observe that whenever $\|\mathbf{Y}v\|^2 \le tn$, then there are at most sn entries of $\mathbf{Y}v$ for which $|(\mathbf{Y}v)_i|^2 > t/s$; equivalently, there are (at least) $(1-s)n$ entries for which $|(\mathbf{Y}v)_i|^2 \le t/s$. Note that all n rows of $\mathbf{Y}v$ are i.i.d., with the same law as $v^\top Y$. Taking a union bound over all possible subsets S of $[n]$ with size $(1-s)n$, corresponding to "small" coordinates in $\mathbf{Y}v$,

$$\Pr\left(\|\mathbf{Y}v\|^2 \le tn\right) \le \binom{n}{(1-s)n}\left(\Pr(|v^\top Y|^2 \le t/s)\right)^{(1-s)n} \le \left(C_0\sqrt{\frac{t}{s}}\frac{e}{(1-s)}\right)^{(1-s)n}. \tag{B.23}$$

Above, we used $\binom{n}{k} \le (en/k)^k$ and the small-ball property for Y, Definition 3.

Let $\varepsilon_0 \in (0,1)$, to be chosen later, and let $\mathcal{N}$ be an ε_0-net of $\mathbb{S}^{p-1}$ of minimal size. By a standard packing argument [74, Lemma 5.2], $|\mathcal{N}| \le (1+\frac{2}{\varepsilon_0})^p \le (3/\varepsilon_0)^p = (3/\varepsilon_0)^{\gamma n}$. Now,

$$\sigma_{\min}(\mathbf{Y}) = \min_{v\in\mathbb{S}^{p-1}}\|\mathbf{Y}v\| \ge \min_{v\in\mathbb{S}^{p-1}}\min_{v^*\in\mathcal{N}}\left\{\|\mathbf{Y}v^*\| - \|\mathbf{Y}(v-v^*)\|\right\} \ge \min_{v^*\in\mathcal{N}}\|\mathbf{Y}v^*\| - \sigma_{\max}(\mathbf{Y})\varepsilon_0\,. \tag{B.24}$$

By Lemma 20, there is C_1 such that with probability $\ge 1 - e^{-c\sqrt{n}}$, $\sigma_{\max}(\mathbf{Y}) \le C_1\sqrt{n}$. Thus, it suffices to show that for some fixed ε_0, with high probability, $\min_{v^*\in\mathcal{N}}\|\mathbf{Y}v^*\| > 2C_1\varepsilon_0\sqrt{n}$. Using (B.23) with $t = (2C_1\varepsilon_0)^2$,

$$\Pr\left(\min_{v^*\in\mathcal{N}}\|\mathbf{Y}v^*\| \le 2C_1\varepsilon_0\sqrt{n}\right) \le |\mathcal{N}|\left(\frac{2eC_0C_1}{\sqrt{s}(1-s)}\varepsilon_0\right)^{(1-s)n} \le \left(C_3(\sqrt{s}(1-s))^{s-1}\varepsilon_0^{1-s-\gamma}\right)^n. \tag{B.25}$$

Recall that $\gamma < 1$, and fix any $s \in (0, 1-\gamma)$. As $\varepsilon_0 \to 0$, the RHS of (B.25) tends to zero. Thus, for all small enough (but constant) ε_0, the RHS of (B.25) is $\le e^{-C_4 n}$. As discussed above, this concludes the proof of the Lemma. □

Proof of Lemma 22. We prove Lemma 22 assuming Y is an isotropic log-concave random vector. Denote the ball $\mathcal{B} = \left\{y \in \mathbb{R}^p \,:\, \|y\| \le 2\sqrt{p}\right\}$, and let $Y_{\mathcal{B}}$ be a random vector distributed according to the law of Y, conditioned on $Y \in \mathcal{B}$. Clearly,

$$\begin{aligned}\Pr\left(\left|p^{-1}Y^\top AY - p^{-1}\mathrm{Tr}(A)\right| \ge \varepsilon\|A\|\right) &\le \Pr(Y \notin \mathcal{B}) + \Pr\left(\left|p^{-1}Y_{\mathcal{B}}^\top AY_{\mathcal{B}} - p^{-1}\mathrm{Tr}(A)\right| \ge \varepsilon\|A\|\right)\\ &\le \Pr(Y \notin \mathcal{B}) + \Pr\left(\left|p^{-1}Y_{\mathcal{B}}^\top AY_{\mathcal{B}} - \mathbb{E}\left[p^{-1}Y_{\mathcal{B}}^\top AY_{\mathcal{B}}\right]\right| \ge (\varepsilon-\varepsilon_0)\|A\|\right),\end{aligned}$$

where $\varepsilon_0 = \|A\|^{-1}p^{-1}\left|\mathbb{E}Y_{\mathcal{B}}^\top AY_{\mathcal{B}} - \mathbb{E}Y^\top AY\right|$. Since $\mathbb{E}\|Y\| \le \sqrt{p}$, Lemma 2 implies that $\Pr(Y \notin \mathcal{B}) \le e^{-c_1\Psi_p\sqrt{p}}$. Observe that $Y_{\mathcal{B}}$ is a log-concave random vector, being the restriction of Y onto a convex set. Moreover, for any $u \in \mathbb{S}^{p-1}$,

$$\mathrm{Var}(u^\top Y_{\mathcal{B}}) \le \mathbb{E}(u^\top Y_{\mathcal{B}})^2 \le \frac{\mathbb{E}(u^\top Y)^2}{\Pr(Y \in \mathcal{B})} = 1 + O(e^{-c\Psi_p\sqrt{p}}) = O(1)\,,$$

hence $\|\mathrm{Cov}(Y_{\mathcal{B}})\| = O(1)$. Since the function $y \mapsto p^{-1}y^\top Ay$ is $L = O(\|A\|p^{-1/2})$-Lipschitz on $\mathcal{B}$, Lemma 2 implies

$$\Pr\left(\left|p^{-1}Y_{\mathcal{B}}^\top AY_{\mathcal{B}} - \mathbb{E}\left[p^{-1}Y_{\mathcal{B}}^\top AY_{\mathcal{B}}\right]\right| \ge (\varepsilon-\varepsilon_0)\|A\|\right) \le e^{-c\Psi_p\frac{(\varepsilon-\varepsilon_0)\|A\|}{L}} \le e^{-c_2\Psi_p\sqrt{p}(\varepsilon-\varepsilon_0)},$$

so that

$$\Pr\left(\left|p^{-1}Y^\top AY - p^{-1}\mathrm{Tr}(A)\right| \geq \varepsilon\|A\|\right) \leq e^{-c_1\Psi_p\sqrt{p}} + e^{-c_2\Psi_p\sqrt{p}(\varepsilon-\varepsilon_0)} \leq C_3 e^{-c_3\Psi_p\sqrt{p}(\varepsilon-\varepsilon_0)}.$$

It remains to show that $\varepsilon_0 = O((\Psi_p\sqrt{p})^{-1})$. Decompose $\mathbb{E}\left[Y^\top AY\right] = \mathbb{E}\left[Y_{\mathcal{B}}^\top AY_{\mathcal{B}}\right]\Pr(Y \in \mathcal{B}) + \mathbb{E}\left[Y^\top AY \cdot \mathbb{1}_{\{Y\notin\mathcal{B}\}}\right]$, so

$$\varepsilon_0 \leq \|A\|^{-1}p^{-1}\left|\mathbb{E}\left[Y_{\mathcal{B}}^\top AY_{\mathcal{B}}\right]\right|\Pr(Y\notin\mathcal{B}) + \|A\|^{-1}p^{-1}\left|\mathbb{E}\left[Y^\top AY\mathbb{1}_{\{Y\notin\mathcal{B}\}}\right]\right| \leq O(e^{-c_1\Psi_p\sqrt{p}}) + p^{-1}\mathbb{E}\left[\|Y\|^2\mathbb{1}_{\{\|Y\|^2>4p\}}\right].$$

It remains to bound the second term above. Use

$$\begin{aligned}\mathbb{E}\left[\|Y\|^2\mathbb{1}_{\{\|Y\|^2>4p\}}\right] &= \int_0^\infty \Pr\left(\|Y\|^2\mathbb{1}_{\{\|Y\|^2>4p\}} \geq t\right)dt \\ &= 4p\Pr(\|Y\|^2 \geq 4p) + \int_{4p}^\infty \Pr(\|Y\|^2 \geq t)dt = O(pe^{-c_1\Psi_p\sqrt{p}}) + \int_{4p}^\infty \Pr(\|Y\| \geq \sqrt{t})dt\,.\end{aligned}$$

By Lemma 2, $\Pr(\|Y\| \geq \sqrt{t}) \leq e^{-c_4\Psi_p(\sqrt{t}-\sqrt{p})}$. Moreover, when $\sqrt{t} \geq 2\sqrt{p}$, we have $\sqrt{t}-\sqrt{p} \geq \frac{1}{2}\sqrt{t}$. Thus,

$$\int_{4p}^\infty \Pr(\|Y\| \geq \sqrt{t})dt \leq \int_{4p}^\infty e^{-(c_4/2)\Psi_p\sqrt{t}}dt \leq e^{-(c_4/4)\Psi_p\sqrt{4p}}\int_{4p}^\infty e^{-(c_4/4)\Psi_p\sqrt{t}}dt = O(e^{-c_5\Psi_p\sqrt{p}}),$$

and we are done. □

Appendix C. Relaxing the zero mean assumption

As described in Section 6, we considered the symmetrization procedure of [29] to relax the zero mean assumption. We note that under the elliptical model, with Y uniform on the sphere, this procedure is especially appealing, as the scaled difference $(zY - z'Y')/R$ with $R = \sqrt{z^2+z'^2}$ is also uniformly distributed on the sphere.

Here we show that our main results continue to hold under a data distribution of the form $X = \Sigma_p^{1/2}Y^\circ$, where $Y^\circ = \zeta Y + \zeta' Y'$ and $\zeta = z/R$, $\zeta' = -z'/R$. By construction, the random vector Y° is isotropic; however, since z, z' are arbitrary, Y° in general does not inherit the favorable distributional properties of Y. Fortunately, our analysis does not require these properties in their full detail. In fact, to carry out the proofs, it suffices to verify that Y° satisfies the following:

- *Small-ball property:* Y° satisfies the SBP. To see this, observe that with probability 1, either $|\zeta| \geq 1/\sqrt{2}$ or $|\zeta'| \geq 1/\sqrt{2}$ (because $\zeta^2 + \zeta'^2 = 1$). Condition on ζ, ζ' and assume without loss of generality that $|\zeta| \geq 1/\sqrt{2}$. Then $|Y^\circ - a| \leq t$ implies that $|Y - (a - \zeta'Y')/\zeta| \leq t/|\zeta| \leq \sqrt{2}t$. Since $(a - \zeta'Y')/\zeta$ is independent of Y, $\Pr(|Y^\circ - a| \leq t|\zeta,\zeta') \leq \Pr(|Y - (\zeta'Y' + a)/\zeta| \leq \sqrt{2}t|\zeta,\zeta') \leq \sqrt{2}C_0 t$, where C_0 is the small-ball constant of Y.
- *Eigenvalue bounds for the sample covariance:* Let $S^\circ = n^{-1}\sum_{i=1}^n \mathbf{y}_i^\circ\mathbf{y}_i^{\circ\top}$ be the sample covariance matrix of n Y°-distributed measurements. Also denote $S = n^{-1}\sum_{i=1}^n \mathbf{y}_i\mathbf{y}_i^\top$, $S' = n^{-1}\sum_{i=1}^n \mathbf{y}_i'\mathbf{y}_i'^\top$.
First, we need a high-probability bound on $\lambda_{\max}(S^\circ)$. Observe that for any u, by Cauchy–Schwarz, $(u^\top Y^\circ)^2 = (\zeta u^\top Y + \zeta' u^\top Y)^2 \leq (\zeta^2 + \zeta'^2)((u^\top Y)^2 + (u^\top Y')^2) = (u^\top Y)^2 + (u^\top Y')^2$. Consequently, $\lambda_{\max}(S^\circ) \leq \lambda_{\max}(S) + \lambda_{\max}(S')$, which may be bounded with high probability using Lemma 20. Next, when $\gamma < 1$ we need a high-probability lower bound on $\lambda_{\min}(S^\circ)$. To this end, one can follow the proof of Lemma 21. To carry it out, we needed two components: the SBP, and a high-probability upper bound on $\lambda_{\max}(S^\circ)$; as explained, both hold.
- *Concentration for quadratic forms:* While complicated functions of Y° should not be expected to concentrate, since ζ, ζ' are arbitrary, concentration of quadratic forms is maintained due to their bilinear nature. We need to prove an analog of Lemma 22. Note that for *fixed* ζ, ζ', the random vector $\zeta Y + \zeta' Y$ inherits the favorable concentration properties of Y. Since the conditional expectation of a quadratic form does not depend on ζ, ζ', $\mathbb{E}[Y^{\circ\top}AY^\circ|\zeta,\zeta'] = \mathrm{Tr}(A)$, we may simply apply Lemma 22 pointwise conditioned on ζ, ζ'.
- *Entrywise concentration for the sample covariance:* We need an analog of Lemma 23. We may carry out the proof of Lemma 23, essentially verbatim, conditioned on $\{\zeta_i, \zeta_i'\}_{1\leq i\leq n}$ and noting that $\mathbb{E}[u^\top S^\circ v|\{\zeta_i,\zeta_i'\}_{1\leq i\leq n}] = u^\top v$ does not depend on $\{\zeta_i,\zeta_i'\}_{1\leq i\leq n}$.

References

[1] Y.I. Abramovich, N.K. Spencer, Diagonally loaded normalised sample matrix inversion (LNSMI) for outlier-resistant adaptive filtering, in: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, Vol. 3, 2007, pp. III–1105–III–1108, http://dx.doi.org/10.1109/ICASSP.2007.366877.

[2] R. Adamczak, A note on the Hanson-Wright inequality for random vectors with dependencies, Electron. Commun. Probab. 20 (72) (2015) 1–13.

[3] R. Adamczak, A.E. Litvak, A. Pajor, N. Tomczak-Jaegermann, Sharp bounds on the rate of convergence of the empirical covariance matrix, C. R. Math. 349 (3–4) (2011) 195–200.

[4] M.Y. An, Log-concave probability distributions: Theory and statistical testing, Technical Report 95–03, Duke University Dept of Economics Working Paper, 1997.

[5] T.W. Anderson, An Introduction to Multivariate Statistical Analysis, in: Wiley Series in Probability and Statistics, Wiley, New York, 2003.

[6] N. Auguin, D. Morales-Jimenez, M.R. McKay, R. Couillet, Large-dimensional behavior of regularized Maronna's M-estimators of covariance matrices, IEEE Trans. Signal Process. 66 (13) (2018) 3529–3542.
[7] M. Bagnoli, T. Bergstrom, Log-concave probability and its applications, Econom. Theory 26 (2) (2005) 445–469.
[8] Z. Bai, J. Silverstein, Spectral Analysis of Large Dimensional Random Matrices, in: Springer Series in Statistics, Springer, New York, 2009.
[9] D. Bakry, I. Gentil, M. Ledoux, Analysis and Geometry of Markov Diffusion Operators, Vol. 348, Springer Science & Business Media, 2013.
[10] M.-F. Balcan, P. Long, Active and passive learning of linear separators under log-concave distributions, in: Conference on Learning Theory, 2013, pp. 288–316.
[11] P.J. Bickel, G. Kur, B. Nadler, Projection pursuit in high dimensions, Proc. Natl. Acad. Sci. 115 (37) (2018) 9151–9156.
[12] P.J. Bickel, E. Levina, Covariance regularization by thresholding, Ann. Statist. 36 (6) (2008) 2577–2604.
[13] A. Block, Y. Mroueh, A. Rakhlin, Generative modeling with denoising auto-encoders and Langevin sampling, 2020, arXiv preprint arXiv:2002.00107.
[14] S. Brazitikos, A. Giannopoulos, P. Valettas, B.-H. Vritsiou, Geometry of Isotropic Convex Bodies, Vol. 196, American Mathematical Soc., 2014.
[15] T. Cai, W. Liu, X. Luo, A constrained $\ell 1$ minimization approach to sparse precision matrix estimation, J. Amer. Statist. Assoc. 106 (494) (2011) 594–607.
[16] T.T. Cai, W. Liu, H.H. Zhou, Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation, Ann. Statist. 44 (2) (2016) 455–488.
[17] T.T. Cai, H.H. Zhou, Optimal rates of convergence for sparse covariance matrix estimation, Ann. Statist. 40 (5) (2012) 2389–2420.
[18] S. Cambanis, S. Huang, G. Simons, On the theory of elliptically contoured distributions, J. Multivariate Anal. 11 (3) (1981) 368–385.
[19] O. Catoni, PAC-Bayesian bounds for the Gram matrix and least squares regression with a random design, 2016, arXiv preprint arXiv:1603.05229.
[20] Y. Chen, An almost constant lower bound of the isoperimetric coefficient in the KLS conjecture, Geom. Funct. Anal. 31 (1) (2021) 34–61.
[21] M. Chen, C. Gao, Z. Ren, Robust covariance and scatter matrix estimation under Huber's contamination model, Ann. Statist. 46 (5) (2018) 1932–1960.
[22] Y. Chen, A. Wiesel, A. Hero III, Robust shrinkage estimation of high-dimensional covariance matrices, IEEE Trans. Signal Process. 59 (9) (2011) 4097–4107.
[23] R. Couillet, A. Kammoun, F. Pascal, Second order statistics of robust estimators of scatter. Application to GLRT detection for elliptical signals, J. Multivariate Anal. 143 (2016) 249–274.
[24] R. Couillet, M. McKay, Large dimensional analysis and optimization of robust shrinkage covariance matrix estimators, J. Multivariate Anal. 131 (2014) 99–120.
[25] R. Couillet, F. Pascal, J.W. Silverstein, Robust estimates of covariance matrices in the large dimensional regime, IEEE Trans. Inform. Theory 60 (11) (2014) 7269–7278.
[26] R. Couillet, F. Pascal, J.W. Silverstein, The random matrix regime of Maronna's M-estimator with elliptically distributed samples, J. Multivariate Anal. 139 (2015) 56–78.
[27] P. Diaconis, D. Freedman, Asymptotics of graphical projection pursuit, Ann. Statist. 12 (1984) 793–815.
[28] I. Diakonikolas, D.M. Kane, Recent advances in algorithmic high-dimensional robust statistics, 2019, arXiv preprint arXiv:1911.05911.
[29] L. Dümbgen, On Tyler's M-functional of scatter in high dimension, Ann. Inst. Statist. Math. 50 (3) (1998) 471–491.
[30] L. Dümbgen, K. Nordhausen, H. Schuhmacher, New algorithms for M-estimation of multivariate scatter and location, J. Multivariate Anal. 144 (2016) 200–217.
[31] K. Fang, S. Kotz, K. Ng, Symmetric Multivariate and Related Distributions, in: Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Taylor & Francis, New York, 1990.
[32] G. Frahm, Generalized Elliptical Distributions: Theory and Applications (Ph.D. thesis), University of Cologne, 2004.
[33] J. Goes, G. Lerman, B. Nadler, Robust sparse covariance estimation by thresholding Tyler's M-estimator, Ann. Statist. 48 (1) (2020) 86–110.
[34] N. Gozlan, C. Roberto, P.-M. Samson, Y. Shu, P. Tetali, Characterization of a class of weak transport-entropy inequalities on the line, in: Annales de l'Institut Henri Poincaré, Probabilités et Statistiques, Vol. 54, (3) Institut Henri Poincaré, 2018, pp. 1667–1693.
[35] D.L. Hanson, F.T. Wright, A bound on tail probabilities for quadratic forms in independent random variables, Ann. Math. Stat. 42 (3) (1971) 1079–1083.
[36] H. Huang, K. Tikhomirov, On dimension-dependent concentration for convex Lipschitz functions in product spaces, 2021, arXiv preprint arXiv:2106.06121.
[37] M. Hubert, P.J. Rousseeuw, S. Van Aelst, High-breakdown robust multivariate methods, Statist. Sci. 23 (1) (2008) 92–119.
[38] R. Kannan, L. Lovász, M. Simonovits, Isoperimetric problems for convex bodies and a localization lemma, Discrete Comput. Geom. 13 (3) (1995) 541–559.
[39] Y. Ke, S. Minsker, Z. Ren, Q. Sun, W.-X. Zhou, User-friendly covariance estimation for heavy-tailed distributions, Statist. Sci. 34 (3) (2019) 454–471.
[40] D. Kelker, Distribution theory of spherical distributions and a location-scale parameter generalization, Sankhyā (1970) 419–430.
[41] J.T. Kent, D.E. Tyler, Maximum likelihood estimation for the wrapped Cauchy distribution, J. Appl. Stat. 15 (2) (1988) 247–254.
[42] B. Klartag, J. Lehec, Bourgain's slicing problem and KLS isoperimetry up to polylog, 2022, arXiv preprint arXiv:2203.15551.
[43] V. Koltchinskii, S. Mendelson, Bounding the smallest singular value of a random matrix without concentration, Int. Math. Res. Not. 2015 (23) (2015) 12991–13008.
[44] Y.T. Lee, S.S. Vempala, Eldan's stochastic localization and the KLS hyperplane conjecture: an improved lower bound for expansion, in: 2017 IEEE 58th Annual Symposium on Foundations of Computer Science, FOCS, IEEE, 2017, pp. 998–1007.
[45] Y.T. Lee, S.S. Vempala, Stochastic localization + Stieltjes barrier=tight bound for log-Sobolev, in: Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, ACM, 2018, pp. 1122–1129.
[46] C. Louart, R. Couillet, Concentration of measure and large random matrices with an application to sample covariance matrices, 2018, arXiv preprint arXiv:1805.08295.
[47] C. Louart, R. Couillet, A concentration of measure and random matrix approach to large dimensional robust statistics, 2020a, arXiv preprint arXiv:2006.09728.
[48] C. Louart, R. Couillet, Concentration of solutions to random equations with concentration of measure hypotheses, 2020b, arXiv preprint arXiv:2010.09877.
[49] L. Lovász, S. Vempala, The geometry of logconcave functions and sampling algorithms, Random Struct. Algorithms 30 (3) (2007) 307–358.
[50] R.A. Maronna, Robust M-estimators of multivariate location and scatter, Ann. Statist. 4 (1) (1976) 51–67.
[51] R.A. Maronna, R.D. Martin, V.J. Yohai, M. Salibián-Barrera, Robust Statistics: Theory and Methods (With R), John Wiley & Sons, New York, 2019.
[52] B. Maurey, Construction de suites symétriques, CR Acad. Sci. Paris SÉR. AB 288 (14) (1979) A679–A681.
[53] M. Meckes, S. Szarek, Concentration for noncommutative polynomials in random matrices, Proc. Amer. Math. Soc. 140 (5) (2012) 1803–1813.
[54] S. Mendelson, N. Zhivotovskiy, Robust covariance estimation under $L\_\{4\}-L\_\{2\}$ norm equivalence, Ann. Statist. 48 (3) (2020) 1648–1664.
[55] S. Minsker, L. Wang, Robust estimation of covariance matrices: Adversarial contamination and beyond, 2022, arXiv preprint arXiv:2203.02880.

[56] D. Morales-Jimenez, R. Couillet, M.R. McKay, Large dimensional analysis of robust M-estimators of covariance with outliers, IEEE Trans. Signal Process. 63 (21) (2015) 5784–5797.
[57] R. Muirhead, Aspects of Multivariate Statistical Theory, in: Wiley Series in Probability and Statistics, Wiley, New York, 2009.
[58] E. Ollila, D.P. Palomar, F. Pascal, Shrinking the eigenvalues of M-estimators of covariance matrix, IEEE Trans. Signal Process. 69 (2020) 256–269.
[59] E. Ollila, D.E. Tyler, Regularized M-estimators of scatter matrix, IEEE Trans. Signal Process. 62 (22) (2014) 6059–6070.
[60] E. Ollila, D.E. Tyler, V. Koivunen, H.V. Poor, Complex elliptically symmetric distributions: Survey, new results and applications, IEEE Trans. Signal Process. 60 (11) (2012) 5597–5625.
[61] F. Pascal, Y. Chitour, Y. Quek, Generalized robust shrinkage estimator and its application to STAP detection problem, IEEE Trans. Signal Process. 62 (21) (2014) 5640–5651.
[62] M. Raginsky, A. Rakhlin, M. Telgarsky, Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis, in: Conference on Learning Theory, PMLR, 2017, pp. 1674–1703.
[63] M. Rudelson, Lecture notes on non-aymptotic random matrix theory, 8, 2013, arXiv preprint arXiv:1301.2382.
[64] M. Rudelson, R. Vershynin, Smallest singular value of a random rectangular matrix, Comm. Pure Appl. Math. 62 (12) (2009) 1707–1739.
[65] M. Rudelson, R. Vershynin, Hanson-Wright inequality and sub-gaussian concentration, Electron. Commun. Probab. 18 (82) (2013) 1–9.
[66] M. Rudelson, R. Vershynin, Small ball probabilities for linear images of high-dimensional distributions, Int. Math. Res. Not. 2015 (19) (2015) 9594–9617.
[67] A. Saumard, J.A. Wellner, Log-concavity and strong log-concavity: a review, Stat. Surv. 8 (2014) 45.
[68] I. Soloveychik, A. Wiesel, Performance analysis of Tyler's covariance estimator, IEEE Trans. Signal Process. 63 (2) (2015) 418–426.
[69] R.P. Stanley, Log-concave and unimodal sequences in algebra, combinatorics, and geometry, Ann. New York Acad. Sci. 576 (1) (1989) 500–535.
[70] Y. Sun, P. Babu, D. Palomar, Regularized tyler's scatter estimator: existence, uniqueness, and algorithms, IEEE Trans. Signal Process. 62 (19) (2014) 5143–5156.
[71] M. Talagrand, Concentration of measure and isoperimetric inequalities in product spaces, Publications Mathématiques de l'Institut des Hautes Etudes Scientifiques 81 (1) (1995) 73–205.
[72] D.E. Tyler, A distribution-free M-estimator of multivariate scatter, Ann. Statist. (1987) 234–251.
[73] R. van Handel, Probability in high dimension, Technical Report, Princeton University, New Jersey, 2014, URL http://www.princeton.edu/~rvan/APC550.pdf.
[74] R. Vershynin, Introduction to the Non-Asymptotic Analysis of Random Matrices, in: Compressed Sensing, Cambridge Univ. Press, Cambridge, 2012, pp. 210–268.
[75] A. Wiesel, Geodesic convexity and covariance estimation, IEEE Trans. Signal Process. 60 (12) (2012) 6182–6189.
[76] A. Wiesel, T. Zhang, Structured robust covariance estimation, Found. Trends Signal Process. 8 (3) (2015) 127–216.
[77] R.D. Yates, et al., A framework for uplink power control in cellular radio systems, IEEE J. Sel. Areas Commun. 13 (7) (1995) 1341–1347.
[78] T. Zhang, X. Cheng, A. Singer, Marčenko–Pastur law for Tyler's M-estimator, J. Multivariate Anal. 149 (2016) 114–123.
[79] K. Zhou, A. Montanari, High-dimensional projection pursuit: Outer bounds and applications to interpolation in neural networks, in: Conference on Learning Theory, PMLR, 2022, pp. 5525–5527.