

The Asymptotic Behavior of Tyler's M-Estimator of Scatter in High Dimension

Lutz Duembgen*

December 1994, revised May 1997

Note. This is an extended version of the paper “*On Tyler's M-functional of scatter in high dimension*” which has been tentatively accepted for publication in the *Annals of the Institute of Statistical Mathematics*. The present version contains some additional results and more detailed proofs.

Abstract. Let $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n \in \mathbf{R}^p$ be independent, identically distributed random vectors with nonsingular covariance matrix Σ , and let $S = S(\mathbf{y}_1, \dots, \mathbf{y}_n)$ be an estimator for Σ . A quantity of particular interest is the condition number of $\Sigma^{-1}S$. If the \mathbf{y}_i are Gaussian and S is the sample covariance matrix, the condition number of $\Sigma^{-1}S$, i.e. the ratio of its extreme eigenvalues, equals $1 + O_p((p/n)^{1/2})$ as $p \rightarrow \infty$ and $p/n \rightarrow 0$. The present paper shows that the same result can be achieved with estimators based on Tyler's (1987) M-functional of scatter, assuming only elliptical symmetry of $\mathcal{L}(\mathbf{y}_i)$ or less. The main tool is a linear expansion for this M-functional which holds uniformly in the dimension p . As a by-product we obtain continuous Fréchet-differentiability with respect to weak convergence.

Keywords and phrases: differentiability, dimensional asymptotics, elliptical symmetry, M-functional, scatter matrix, symmetrization

Author's address: Institut für Angewandte Mathematik, Universität Heidelberg, Im Neuenheimer Feld 294, D-69120 Heidelberg, Germany; lutz@statlab.uni-heidelberg.de

*Research supported in part by European Union Human Capital and Mobility Program ERB CHRX-CT 940693.

1 Introduction

It has been noted by numerous authors that asymptotic results, where the dimension of the underlying model is fixed while the number of observations tends to infinity, are often inappropriate for real applications; e.g. Portnoy (1988) or Girko (1995). In particular, the literature on M-estimation in linear regression models with increasing dimension is vast and still growing; see for instance Huber (1981), Portnoy (1984, 1985), Bai and Wu (1994 a-b), Mammen (1996) and the references cited therein. In the present paper we investigate the related problem of M-estimation of a high-dimensional covariance matrix.

Let \widehat{P}_n be the empirical distribution of independent random vectors $\mathbf{y}_n = \mathbf{y}_{n1}, \mathbf{y}_{n2}, \dots, \mathbf{y}_{nn}$ in \mathbf{R}^p with unknown distribution P_n , and let $S_n = S_n(\widehat{P}_n)$ be an estimator for the covariance matrix Σ_n of P_n , both assumed to be positive definite. Of particular interest is the condition number of $\Sigma_n^{-1}S_n$,

$$\gamma(\Sigma_n^{-1}S_n) := \frac{\lambda_1(\Sigma_n^{-1}S_n)}{\lambda_p(\Sigma_n^{-1}S_n)},$$

where $\lambda_1(A) \geq \lambda_2(A) \geq \lambda_3(A) \geq \dots$ denote the ordered real eigenvalues of $A \in \mathbf{R}^{p \times p}$. For there are explicit bounds for various scale-invariant functions of S_n and Σ_n such as correlations, partial and canonical correlations, regression coefficients or eigenspaces, all in terms of $\gamma(\Sigma_n^{-1}S_n)$ (cf. Dümbgen 1994). An example are the following sharp inequalities for correlations, where $(\cdot)'$ denotes transposition:

$$\left| \operatorname{arctanh}\left(\frac{x'\Sigma_n y}{\sqrt{x'\Sigma_n x y'\Sigma_n y}}\right) - \operatorname{arctanh}\left(\frac{x'S_n y}{\sqrt{x'S_n x y'S_n y}}\right) \right| \leq \frac{\log \gamma(\Sigma_n^{-1}S_n)}{2}$$

for arbitrary $x, y \in \mathbf{R}^p \setminus \{0\}$. Therefore it is of interest to study the probabilistic behavior of $\gamma(\Sigma_n^{-1}S_n)$. If P_n is multivariate normal and S_n is the sample covariance matrix, a modification of Silverstein's (1985) arguments reveals that

$$(1.1) \quad \gamma(\Sigma_n^{-1}S_n) = 1 + 4(p/n)^{1/2} + o_p((p/n)^{1/2});$$

see also the proof of Theorem 5.4. In connection with P_n, \widehat{P}_n we assume tacitly that the dimension $p = p_n$ may depend on the sample size n such that $p/n \rightarrow 0$. Asymptotic statements refer to $n \rightarrow \infty$. Expansions such as (1.1) hold under more general assumptions

on the distribution P_n , provided that it has sufficiently light tails (cf. Girko 1995). On the other hand, the distribution of the extremal eigenvalues of the sample covariance matrix is very sensitive to deviations from normality so that even the weaker assertion

$$(1.2) \quad \gamma(\Sigma_n^{-1}S_n) = 1 + O_p((p/n)^{1/2})$$

may be false, even under elliptical symmetry of P_n . It is thus desirable to have an estimator, whose distribution is less model-dependent, such that expansion (1.1) or at least (1.2) holds.

A possible alternative to the sample covariance matrix are M-estimators of scatter as proposed by Maronna (1976) and Tyler (1987). The present paper focuses on two estimators related to Tyler's (1987) M-functional. The latter is defined in Section 2 as a matrix-valued function $Q \mapsto \Sigma(Q)$ on the space of probability measures on $\mathbf{R}^p \setminus \{0\}$. Section 3 provides a basic linear expansion for $\Sigma(\cdot)$ with a rather explicit bound for the remainder term. As a by-product we obtain continuous Fréchet-differentiability of $\Sigma(\cdot)$ with respect to the weak topology on the space of probability measures on $\mathbf{R}^p \setminus \{0\}$.

Section 4 describes estimators based on $\Sigma(\cdot)$. One obvious choice is the M-estimator $\Sigma(\hat{P}_n)$, which is distribution-free if P is elliptically symmetric around zero. In addition we propose the estimator $\Sigma(\hat{P}_n^s)$, where \hat{P}_n^s is a symmetrization of \hat{P}_n . This is an intuitively appealing method to get rid of unknown location parameters. The linear expansion of Section 3 implies asymptotic normality of both estimators and consistency of certain bootstrap methods. Some of these results and conclusions are not entirely new but nevertheless stated explicitly for the reader's convenience. In connection with the bootstrap we use similar arguments as Bickel and Freedman (1981).

In order to prove Fréchet-differentiability for fixed dimension p , one could also apply general methods of Clarke (1983). An advantage of our explicit expansion is that it enables us to investigate the asymptotic behavior of $\Sigma(\hat{P}_n)$ and $\Sigma(\hat{P}_n^s)$ as $p = p_n \rightarrow \infty$. This is done in Section 5. Under certain regularity assumptions assertion (1.2) is valid for both estimators $\Sigma(\hat{P}_n)$ and $\Sigma(\hat{P}_n^s)$. In particular, if P is elliptically symmetric, $\Sigma(\hat{P}_n)$ is shown to have the same asymptotic behavior as the sample covariance matrix in the Gaussian

model, including expansion (1.1).

Another approach to the problem of unknown location, pursued by Tyler (1987), is to re-center \hat{P}_n around an estimator $\hat{\mu}_n = \hat{\mu}_n(\hat{P}_n)$ for P 's center. Section 6 contains some additional results on this method, also in view of dimensional asymptotics.

All proofs are deferred to Section 7.

2 Definition and basic properties of the M-functional $\Sigma(\cdot)$

Let us first introduce some notation. Throughout the set of symmetric matrices in $\mathbf{R}^{p \times p}$ is denoted by \mathbf{M} , while \mathbf{M}^+ denotes the set of positive definite $M \in \mathbf{M}$. For $M \in \mathbf{M}^+$ the unique matrix $N \in \mathbf{M}^+$ with $NN = M$ is denoted by $M^{1/2}$, and $M^{-1/2} := (M^{-1})^{1/2} = (M^{1/2})^{-1}$. Further we consider the following affine subspaces of \mathbf{M} , where I stands for the identity matrix in $\mathbf{R}^{p \times p}$:

$$\begin{aligned} \mathbf{M}(0) &:= \{M \in \mathbf{M} : \text{trace}(M) = 0\}, \\ \mathbf{M}(p) &:= \{M \in \mathbf{M} : \text{trace}(M) = p\} = I + \mathbf{M}(0). \end{aligned}$$

Let f be a real or vector-valued function on \mathbf{R}^p , and let Δ be a signed measure on \mathbf{R}^p . Then $f(\Delta)$ stands for $\int f(x) \Delta(dx)$. This convention will be particularly convenient for functions of several arguments. Further, for $A \in \mathbf{R}^{p \times p}$ we denote by $A\Delta$ the transformed signed measure $\Delta \circ A^{-1}$.

Throughout let P and Q be probability distributions on $\mathbf{R}^p \setminus \{0\}$. We regard Q as rotationally symmetric around zero in a weak sense if $G(Q) = \int G(x) Q(dx)$ is equal to I , where

$$G(x) := \begin{cases} p|x|^{-2} xx' \in \mathbf{M}(p) & \text{if } x \neq 0, \\ 0 & \text{else;} \end{cases}$$

here $|x|$ denotes the standard Euclidean norm $(x'x)^{1/2}$ of x . Note that $G(Q)$ equals p times the matrix of second moments of $|z|^{-1}z$, where $z \sim Q$. If Q is spherically symmetric around zero, one easily verifies that in fact $G(Q) = I$. More generally, this equality holds if the vectors $z = (z_i)_{1 \leq i \leq p}$ and $(\epsilon_i z_{\pi(i)})_{1 \leq i \leq p}$ have the same distribution for arbitrary $\epsilon \in \{-1, 1\}^p$ and permutations π of $\{1, 2, \dots, p\}$. In general one tries to find $M \in \mathbf{M}^+$

such that

$$G(M^{-1/2}Q) = p \int \frac{M^{-1/2} x x' M^{-1/2}}{x' M^{-1} x} Q(dx) = I.$$

Note that $G(M^{-1/2}Q) = G((sM)^{-1/2}Q)$ for all $s > 0$, so that $G(\cdot)$ is only useful in connection with scale-invariant functions on \mathbf{M}^+ such as correlations.

Definition. If the equality $G(M^{-1/2}Q) = I$ has a unique solution M in $\mathbf{M}^+(p) := \mathbf{M}^+ \cap \mathbf{M}(p)$, this matrix M is denoted by $\Sigma(Q)$. Otherwise we define arbitrarily $\Sigma(Q) := 0$.

An important property of $G(\cdot)$ and $\Sigma(\cdot)$ is linear equivariance. For nonsingular $A \in \mathbf{R}^{p \times p}$ and $M \in \mathbf{M}^+$ one easily verifies that

$$(2.1) \quad G\left((AMA')^{-1/2}AQ\right) = TG(M^{-1/2}Q)T'$$

where $T := (AMA')^{-1/2}AM^{1/2}$ is orthonormal.

Thus $G(M^{-1/2}Q) = I$ if, and only if, $G\left((AMA')^{-1/2}AQ\right) = I$. Hence

$$(2.2) \quad \Sigma(AQ) = r A\Sigma(Q)A' \quad \text{with } r := p / \text{trace}(A\Sigma(Q)A').$$

Necessary and sufficient conditions for $\Sigma(Q) \in \mathbf{M}^+$ are as follows.

Theorem 2.1 *Let \mathcal{V} be the set of proper linear subspaces \mathbf{V} of \mathbf{R}^p , i.e. $1 \leq \dim(\mathbf{V}) < p$.*

[a] *If $G(M^{-1/2}Q) = I$ for some $M \in \mathbf{M}^+$, then*

$$Q(\mathbf{V}) \leq \dim(\mathbf{V})/p \quad \text{for all } \mathbf{V} \in \mathcal{V}.$$

[b] *Suppose that*

$$(2.3) \quad Q(\mathbf{V}) < \dim(\mathbf{V})/p \quad \text{for all } \mathbf{V} \in \mathcal{V}.$$

Then there exists a unique $M \in \mathbf{M}^+(p)$ such that $G(M^{-1/2}Q) = I$.

[c] *Suppose that $G(Q) = I$ but $Q(\mathbf{V}) = \dim(\mathbf{V})/p$ for some $\mathbf{V} \in \mathcal{V}$. Then $Q(\mathbf{V} \cup \mathbf{V}^\perp) = 1$ and*

$$G\left((a\Pi + b(I - \Pi))^{-1/2}Q\right) = I \quad \text{for all } a, b > 0,$$

where $\Pi \in \mathbf{M}$ describes the orthogonal projection from \mathbf{R}^p onto \mathbf{V} , and \mathbf{V}^\perp stands for the orthogonal complement of \mathbf{V} .

Parts [a, b] are due to Tyler (1987) and Kent and Tyler (1988). Their proofs are formulated for empirical distributions Q , but extension to arbitrary distributions is mainly straightforward, requiring only notational changes. The only exception is the existence statement in part [b]. Two possible proofs are given in Section 7. Part [c], combined with (2.1), supplements part [b] in that condition (2.3) is even necessary for $\Sigma(Q) \in \mathbf{M}^+$. This will be needed in the proof of Theorem 3.2 below.

3 Differentiability of $\Sigma(\cdot)$

For $M \in \mathbf{M}$ we define $\|M\| := |\lambda_1(M)| \vee |\lambda_p(M)|$. Since the dimension p may vary, this particular choice of a norm is important. It is particularly useful in connection with eigenvalues, because $|\lambda_i(A) - \lambda_i(B)| \leq \|A - B\|$ for $A, B \in \mathbf{M}$ and $1 \leq i \leq p$. By way of contrast, for growing dimension p expansions involving the Euclidean norm $\|M\|_{\mathbf{E}} = \text{trace}(M^2)^{1/2}$ would be of little use. This is one reason why the results of Portnoy (1988) cannot be applied here without unnecessary restrictions on p . Generally, we always use the norm

$$\|L\| := \max_{y \in \mathbf{S}(\mathbf{B})} \|Ly\|$$

of a linear operator L from a normed vector space $(\mathbf{B}, \|\cdot\|)$ into another normed space, where $\mathbf{S}(\mathbf{B})$ denotes the unit sphere $\{y \in \mathbf{B} : \|y\| = 1\}$.

Now we investigate $\Sigma(Q)$ if Q is close to P in a certain sense and $G(P) = I$. By equivariance of $G(\cdot)$ and $\Sigma(\cdot)$ it suffices to consider the latter case.

The function $G(M^{-1/2}x)$ is differentiable with respect to $M \in \mathbf{M}^+$ with

$$\begin{aligned} D(x, B) &:= \left. \frac{\partial}{\partial t} \right|_{t=0} G\left((I + tB)^{-1/2}x\right) = F(x, B) - 2^{-1}\left(BG(x) + G(x)B\right), \\ F(x, B) &:= \begin{cases} |x|^{-2}x'Bx G(x) = p|x|^{-4}x'Bx xx' & \text{if } x \neq 0, \\ 0 & \text{else.} \end{cases} \end{aligned}$$

Note that $D(x, I) = 0$ and $\text{trace}(D(x, B)) = 0$ for all $B \in \mathbf{M}$. The next lemma shows that condition (2.3) is closely related to the operator $D(Q, \cdot)$.

Lemma 3.1 *The operator $D(Q, \cdot)$ is nonsingular on $\mathbf{M}(0)$ if, and only if, $Q(\mathbf{V} \cup \mathbf{V}^\perp) < 1$*

for arbitrary $\mathbf{V} \in \mathcal{V}$. In that case,

$$\text{trace}(D(Q, B)B) < 0 \quad \text{for all } B \in \mathbf{M}(0) \setminus \{0\}.$$

The inverse operator of $D(P, \cdot) : \mathbf{M}(0) \rightarrow \mathbf{M}(0)$, if existent, is denoted by $D^{-1}(P, \cdot)$. Here is our basic linear expansion for $\Sigma(\cdot)$.

Theorem 3.2 *For any $b < \infty$ there exist constants $\kappa(b) < \infty$ and $\epsilon(b) > 0$ (not depending on p or P) such that*

$$\left\| \Sigma(Q) - I + D^{-1}(P, G(Q - P)) \right\| \leq \kappa(b) \|F(Q - P, \cdot)\| \|G(Q - P)\|$$

whenever

$$\Sigma(P) = I, \quad \|D^{-1}(P, \cdot)\| \leq b \quad \text{and} \quad \|F(Q - P, \cdot)\| \leq \epsilon(b).$$

The latter two norms $\|\cdot\|$ refer to the linear operators $D^{-1}(P, \cdot) : \mathbf{M}(0) \rightarrow \mathbf{M}(0)$ and $F(Q - P, \cdot) : \mathbf{M} \rightarrow \mathbf{M}$. Note also that $\|G(Q - P)\| = \|F(Q - P, I)\| \leq \|F(Q - P, \cdot)\|$.

Theorem 3.2, Lemma 3.1 and (2.2) together imply that $\Sigma(\cdot)$ is Fréchet-differentiable with respect to the weak topology. The reason is that $x \mapsto F(x, \cdot)$ is a bounded, continuous mapping from $\mathbf{R}^p \setminus \{0\}$ into the finite-dimensional space of linear operators $L : \mathbf{M} \rightarrow \mathbf{M}$, so that $\|F(Q - P, \cdot)\| \rightarrow 0$ as $Q \rightarrow P$ weakly.

Corollary 3.3 *Suppose that $\Sigma(P) = I$. Then, as $Q \rightarrow P$ weakly,*

$$G(Q) \rightarrow I \quad \text{and} \quad \Sigma(Q) - I = -D^{-1}(P, G(Q) - I) + o(\|G(Q) - I\|). \quad \square$$

One can even show that $\Sigma(\cdot)$ is continuously Fréchet-differentiable. Instead of pursuing this issue, we shall prove a related statement about limiting distributions of $\Sigma(\hat{P}_n)$ and $\Sigma(\hat{P}_n^s)$ in the next section.

4 Related estimators and their properties in fixed dimension

At this point it is convenient to define $\Sigma(\tilde{Q}) := \Sigma(\tilde{Q}(\cdot | \mathbf{R}^p \setminus \{0\}))$ for any probability measure \tilde{Q} on \mathbf{R}^p with $\tilde{Q}\{0\} < 1$.

Suppose first that the distribution P_n has a known “center” $\mu_n \in \mathbf{R}^p$. Without loss of generality one may assume that $\mu_n = 0$. Then a straightforward estimator for $\Sigma(P_n)$ is given by $\Sigma(\hat{P}_n)$. An important example are elliptically symmetric distributions $P_n = \mathcal{L}(R_n \Sigma_n^{1/2} \mathbf{u})$, where $R_n > 0$ and \mathbf{u} are stochastically independent, \mathbf{u} is uniformly distributed on the unit sphere of \mathbf{R}^p , and $\Sigma_n \in \mathbf{M}^+(p)$. Clearly $\Sigma(P_n) = \Sigma_n$, and the empirical distribution \hat{P}_n satisfies condition (2.3) almost surely if $n > p$. Moreover, the distribution of $\gamma(\Sigma_n^{-1} \Sigma(\hat{P}_n))$ depends neither on Σ_n nor on $\mathcal{L}(R_n)$ (cf. Tyler 1987).

The center μ_n , no matter how it is defined, is rarely known in advance. In order to avoid definition and estimation of an unknown location parameter one can also consider the functional $Q \mapsto \Sigma(Q^s)$ with the symmetrized distribution

$$Q^s := \mathcal{L}(\mathbf{z}_1 - \mathbf{z}_2 \mid \mathbf{z}_1 \neq \mathbf{z}_2) \quad \text{where } (\mathbf{z}_1, \mathbf{z}_2) \sim Q \otimes Q.$$

Here $\Delta_1 \otimes \Delta_2$ denotes the product measure on $\mathbf{R}^p \times \mathbf{R}^p$ of (signed) measures Δ_1, Δ_2 on \mathbf{R}^p . One motivation for the functional $Q \mapsto \Sigma(Q^s)$ is the representation $2^{-1} \mathbb{E}((\mathbf{z}_1 - \mathbf{z}_2)(\mathbf{z}_1 - \mathbf{z}_2)')$ of the covariance matrix of Q . Moreover, if $\mathbf{z} \sim Q$ has independent, identically distributed components, then $G(Q^s) = I$, whereas $G(Q)$ may be different from I . Thus symmetrization partly corrects a possible deficiency of M-estimators.

One easily verifies that $Q \mapsto \Sigma(Q^s)$ is affinely invariant in that

$$(4.1) \quad A \Sigma(Q^s) A' = r \Sigma((\mu + A Q)^s) \quad \text{with } r := \text{trace}(A' A \Sigma(Q^s)) / p$$

for any nonsingular $A \in \mathbf{R}^{p \times p}$ and $\mu \in \mathbf{R}^p$, where $\mu + A Q := \mathcal{L}(\mu + A \mathbf{z})$, $\mathbf{z} \sim Q$. If Q is elliptically symmetric around μ with scatter matrix $\Sigma_o \in \mathbf{M}^+(p)$, then Q^s is elliptically symmetric around zero with the same scatter matrix Σ_o .

An application of Theorem 3.2 utilizing the explicit error bound is the following Central Limit Theorem for the distribution of $\Sigma(\hat{P}_n)$ and $\Sigma(\hat{P}_n^s)$.

Corollary 4.1 *Suppose that P_n converges weakly to some distribution P on \mathbf{R}^p .*

[a] *Let $P\{0\} = 0$ and $\Sigma(P) = I$. Let $L_n(\cdot \mid P_n)$ denote the distribution of*

$$n^{1/2} \left(\Sigma(\Sigma(P_n)^{-1/2} \hat{P}_n) - I \right)$$

(provided that $\Sigma(P_n) \in \mathbf{M}^+$). Then $\Sigma(P_n) \rightarrow I$ and

$$L_n(\cdot | P_n) \rightarrow_w \mathcal{L}(W),$$

where $W \in \mathbf{M}(0)$ is a random matrix with centered Gaussian distribution and the same covariance function as $D^{-1}(P, G(\mathbf{y}) - I)$, $\mathbf{y} \sim P$.

[b] Let $P\{\mu\} = 0$ for all $\mu \in \mathbf{R}^p$ and $\Sigma(P^s) = I$. Let $L_n^s(\cdot | P_n)$ denote the distribution of

$$n^{1/2} \left(\Sigma \left(\Sigma(P_n^s)^{-1/2} \hat{P}_n^s \right) - I \right)$$

(provided that $\Sigma(P_n^s) \in \mathbf{M}^+$). Then $\Sigma(P_n^s) \rightarrow I$ and

$$L_n^s(\cdot | P_n) \rightarrow_w \mathcal{L}(W^s),$$

where $W^s \in \mathbf{M}(0)$ is a random matrix with centered Gaussian distribution and the same covariance function as $2D^{-1}(P^s, \tilde{G}(\mathbf{y}, P) - I)$, $\mathbf{y} \sim P$. Here $\tilde{G}(x, y) := G(x - y)$.

Remark 4.2 The covariance function of a random matrix $W \in \mathbf{M}(0)$ is defined as the function $(A, B) \mapsto \text{Cov}(\text{trace}(WA), \text{trace}(WB))$ on $\mathbf{M}(0) \times \mathbf{M}(0)$.

Remark 4.3 In case of P being spherically symmetric around zero one can deduce from equations (7.11) and (7.12) in Lemma 7.5 that

$$\mathbb{E}(\text{trace}(WA) \text{trace}(WB)) = 2(1 + 2/p) \text{trace}(AB) \quad \text{for } A, B \in \mathbf{M}(0).$$

Remark 4.4 If $P_n \rightarrow P$ weakly, then the empirical distribution \hat{P}_n converges weakly to P in probability. More precisely, $d_w(\hat{P}_n, P)$ converges to zero in probability, where $d_w(\cdot, \cdot)$ metrizes weak convergence of probability measures on \mathbf{R}^p . Consequently, the bootstrap distributions $L_n(\cdot | \hat{P}_n)$ and $L_n^s(\cdot | \hat{P}_n^s)$ are consistent estimators of $L_n(\cdot | P_n)$ and $L_n^s(\cdot | P_n)$, respectively.

Remark 4.5 Utilizing the equivariance properties of $\Sigma(\cdot)$, (2.2) and (4.1), one can deduce from Corollary 4.1 that

$$\begin{aligned} n^{1/2} \left(\gamma \left(\Sigma(P_n)^{-1} \Sigma(\hat{P}_n) \right) - 1 \right) &\rightarrow_{\mathcal{L}} (\lambda_1 - \lambda_p)(W_o) \quad \text{in part [a]}, \\ n^{1/2} \left(\gamma \left(\Sigma(P_n^s)^{-1} \Sigma(\hat{P}_n^s) \right) - 1 \right) &\rightarrow_{\mathcal{L}} (\lambda_1 - \lambda_p)(W_o^s) \quad \text{in part [b]}. \end{aligned}$$

5 Asymptotic behavior of $\Sigma(\widehat{P}_n)$ and $\Sigma(\widehat{P}_n^s)$ in high dimension

Now we consider the case where

$$p = p_n \rightarrow \infty \quad \text{but} \quad p/n \rightarrow 0.$$

For the sake of simplicity it is assumed that P_n has no atoms.

Theorem 5.1 *Suppose that $\Sigma(P_n) = I$ for all n . Let*

$$\begin{aligned} \kappa_n^2 &:= \max_{u \in \mathbf{S}(\mathbf{R}^p)} \int (u'G(y)u)^2 P_n(dy) = O(1), \\ \sigma_n^2 &:= \max_{B \in \mathbf{S}(\mathbf{M}(0))} \int \left(\frac{y'By}{y'y} \right)^2 P_n(dy) = o(1). \end{aligned}$$

Further let $p = O(n^{1/2})$. Then

$$\mathbb{E} \|G(\widehat{P}_n) - I\| = o(1) \quad \text{and} \quad \mathbb{E} \|\Sigma(\widehat{P}_n) - G(\widehat{P}_n)\| = o(\mathbb{E} \|G(\widehat{P}_n) - I\|).$$

If in addition $p = O(n^{1/3})$, then

$$\mathbb{E} \|G(\widehat{P}_n) - I\| = O((p/n)^{1/2}).$$

Remark 5.2 Suppose that $\mathbf{y}_n = (y_{n,i})_{1 \leq i \leq p} \sim P_n$ has independent, identically distributed components with continuous, symmetric distribution such that $\mathbb{E}(y_{n,1}^2) = 1$ and $\mathbb{E}(y_{n,1}^4) = O(1)$. Then $\kappa_n^2 = O(1)$ and $\sigma_n^2 = O(p^{-1})$. For it follows from the one-sided version of Bennett's (1962) inequality that $\mathbb{P}\{|\mathbf{y}_n|^2/p \leq 1/2\} \leq \exp(-a_n p)$ for some number a_n depending on the fourth moment of $y_{n,1}$ and p such that $\liminf_{n \rightarrow \infty} a_n > 0$. Therefore, since $(u'G(y)u)^2 \leq p^2$ and $(y'By)^2/(y'y)^2 \leq 1$, one may replace these integrands of κ_n^2 and σ_n^2 with $4(u'y)^4$ and $4p^{-2}(y'By)^2$, respectively. Then the assertion follows from tedious but elementary moment calculations.

Remark 5.3 The conclusions of Theorem 5.1 and Remark 5.2 remain valid if (P_n, \widehat{P}_n) is replaced with (P_n^s, \widehat{P}_n^s) , where the symmetry condition in Remark 5.2 becomes superfluous. For the proof of Theorem 5.1 consists essentially of bounding $\mathbb{E}(\|F(\widehat{P}_n - P_n, \cdot)\|^2)$ and $\mathbb{E}(\|G(\widehat{P}_n - P_n)\|^2)$. But $F(\widehat{P}_n^s, B)$ can be written as a matrix-valued U-statistic

$$\binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} F(\mathbf{y}_{ni} - \mathbf{y}_{nj}, B).$$

Let $\check{P}_n^{\mathbf{s}}$ be the empirical distribution of $\mathbf{y}_{n1}^{\mathbf{s}}, \mathbf{y}_{n2}^{\mathbf{s}}, \dots, \mathbf{y}_{nm}^{\mathbf{s}}$, where $m = m_n := \lfloor n/2 \rfloor$ and $\mathbf{y}_{ni}^{\mathbf{s}} := \mathbf{y}_{n,2i-1} - \mathbf{y}_{n,2i}$. Then a simple convexity argument due to Hoeffding (1963) yields

$$(5.1) \quad \begin{aligned} \mathbb{E} \left(\|G(\hat{P}_n^{\mathbf{s}} - P_n^{\mathbf{s}})\|^2 \right) &\leq \mathbb{E} \left(\|G(\check{P}_n^{\mathbf{s}} - P_n^{\mathbf{s}})\|^2 \right), \\ \mathbb{E} \left(\|F(\hat{P}_n^{\mathbf{s}} - P_n^{\mathbf{s}}, \cdot)\|^2 \right) &\leq \mathbb{E} \left(\|F(\check{P}_n^{\mathbf{s}} - P_n^{\mathbf{s}}, \cdot)\|^2 \right); \end{aligned}$$

see also equation (7.20) in Section 7. Now the signed measure $\check{P}_n^{\mathbf{s}} - P_n^{\mathbf{s}}$ can be handled analogously as $\hat{P}_n - P_n$.

Under spherical symmetry of P_n , restrictions on p beyond $p = o(n)$ are superfluous, and one can obtain rather precise expansions.

Theorem 5.4 *Suppose that P_n is spherically symmetric around zero for all n .*

[a] *Then*

$$\begin{aligned} \mathbb{E} \|G(\hat{P}_n) - I\| &= O((p/n)^{1/2}), \\ \mathbb{E} \left\| \Sigma(\hat{P}_n) - I - (1 + 2/p)(G(\hat{P}_n) - I) \right\| &= O(\log(n/p)p/n). \end{aligned}$$

Moreover, one can couple $\Sigma(\hat{P}_n)$ with a standard Wishart matrix $M_n \in \mathbf{M}$ with n degrees of freedom such that

$$\mathbb{E} \|\Sigma(\hat{P}_n) - n^{-1}M_n\| = o((p/n)^{1/2}).$$

In particular, $\gamma(\Sigma(\hat{P}_n)) = 1 + 4(p/n)^{1/2} + o_p((p/n)^{1/2})$.

[b] *As for $\hat{P}_n^{\mathbf{s}}$,*

$$\begin{aligned} \mathbb{E} \|\Sigma(\hat{P}_n^{\mathbf{s}}) - I\| &= O((p/n)^{1/2}), \\ \mathbb{E} \left\| \Sigma(\hat{P}_n^{\mathbf{s}}) - I - n^{-1} \sum_{i=1}^n H_n(|\mathbf{y}_{ni}|)(G(\mathbf{y}_{ni}) - I) \right\| &= o((p/n)^{1/2}), \end{aligned}$$

where H_n is an increasing function from $[0, \infty[$ into $[0, 2[$. If in addition $|\mathbf{y}_n|^2/p$ converges in probability to a constant $\kappa_o > 0$, then

$$\mathbb{E} \|\Sigma(\hat{P}_n^{\mathbf{s}}) - \Sigma(\hat{P}_n)\| = o((p/n)^{1/2}).$$

6 The impact of plugging in estimates of location

For $\mu \in \mathbf{R}^p$ let

$$Q^{(\mu)} := \mathcal{L}(\mathbf{z} - \mu \mid \mathbf{z} \neq \mu) \quad \text{where } \mathbf{z} \sim Q.$$

If $P\{0\} = 0$, one can easily show that $Q^{(\mu)}$ converges weakly to P as $Q \rightarrow_w P$ and $\mu \rightarrow 0$.

Thus Corollary 3.3 implies that

$$(6.1) \quad \Sigma(Q^{(\mu)}) \rightarrow \Sigma(P) \text{ as } Q \rightarrow_w P \text{ and } \mu \rightarrow 0$$

whenever $\Sigma(P) \in \mathbf{M}^+$. The following two results show that under moderate moment assumptions on P the difference $\Sigma(Q^{(\mu)}) - \Sigma(Q)$ can be expanded explicitly, extending results of Tyler (1987, Section 4).

Theorem 6.1 *Suppose that $P\{0\} = 0$, $\Sigma(P) = I$ and $\int |x|^{-1} P(dx) < \infty$. Define*

$$H(x, \mu) := p|x|^{-2}(\mu x' + x \mu') - 2|x|^{-2} x' \mu G(x).$$

Then

$$\Sigma(Q^{(\mu)}) - I = -D^{-1}(P, G(Q) - I + H(P, \mu)) + o(\|G(P - Q)\| + |\mu|)$$

as

$$Q \rightarrow_w P, \quad \int |x|^{-1} Q(dx) \rightarrow \int |x|^{-1} P(dx), \quad \mu \rightarrow 0.$$

Note that $H(\cdot, \mu)$ is an odd function. Thus the bias term $H(P, \mu)$ equals zero if P is symmetric in that $P(S) = P(-S)$ for all Borel sets $S \subset \mathbf{R}^p$. As for the moment condition, note that $\int |x|^{-r} P(dx) < \infty$ if $r < p$ and P has a bounded density with respect to Lebesgue measure.

Theorem 6.2 *Suppose that $p = p_n \rightarrow \infty$ and $p/n \rightarrow 0$. Let $\Sigma(P_n) = I$ and $P_n(\{\mu\}) = 0$ for arbitrary $\mu \in \mathbf{R}^p$ and all n . Moreover, let*

$$p \int |x|^{-2} P_n(dx) = O(1),$$

and suppose that either

$$\kappa_n^2 = O(1), \quad \sigma_n^2 = o(1) \quad \text{and} \quad p = O(n^{1/3})$$

(cf. Theorem 5.1), or,

P_n is spherically symmetric around zero for all n .

Then $\|H(P_n, \cdot)\| = O(1)$ and for any sequence of positive numbers $\epsilon_n = o((p/n)^{1/4})$,

$$\sup_{|\mu| \leq \epsilon_n} \left\| \Sigma(\hat{P}_n^{(\mu)}) - I + D^{-1} \left(P_n, G(\hat{P}_n) - I + H(P_n, \mu) \right) \right\| = o_p((p/n)^{1/2}).$$

Since $\int |x|^{-2} \mathcal{N}_p(0, I)(dx) = (p-2)^{-1}$, the first moment condition is satisfied for mixtures

$$P_n = \int \mathcal{N}_p(0, \sigma^2 I) \pi_n(d\sigma),$$

provided that $\int \sigma^{-2} \pi_n(d\sigma) = O(1)$. If $\hat{\mu}_n = \hat{\mu}_n(\hat{P}_n)$ is an estimator such that

$$(6.2) \quad \hat{\mu}_n = O_p((p/n)^{1/2}),$$

then under the assumptions of Theorem 6.2,

$$\begin{aligned} \left\| \Sigma(\hat{P}_n^{(\hat{\mu}_n)}) - I \right\| &= O_p((p/n)^{1/2}), \\ \left\| \Sigma(\hat{P}_n^{(\hat{\mu}_n)}) - \Sigma(\hat{P}_n) + D^{-1}(P_n, H(P_n, \hat{\mu}_n)) \right\| &= o_p((p/n)^{1/2}). \end{aligned}$$

In case of $p^{-1} \int |x|^2 P_n(dx) = O(1)$, the sample mean $\hat{\mu}_n = \int x \hat{P}_n(dx)$ satisfies condition (6.2). Alternatively consider Tukey's median

$$\hat{\mu}_n = \arg \max_{\mu \in \mathbf{R}^p} \inf_{u \in \mathbf{S}(\mathbf{R}^p)} \hat{P}_n \{ x \in \mathbf{R}^p : x'u \leq \mu'u \}.$$

Here $\hat{\mu}_n = O_p((p/n)^{1/2})$, provided that

$$(6.3) \quad \liminf_{n \rightarrow \infty} \inf_{u \in \mathbf{S}(\mathbf{R}^p)} \epsilon_n^{-1} \left(P_n \{ x \in \mathbf{R}^p : u'x \leq \epsilon_n \} - 1/2 \right) > 0 \quad \text{whenever } \epsilon_n \downarrow 0.$$

This follows straightforwardly from the fact that

$$\mathbb{E} \sup_{\text{halfspaces } H \subset \mathbf{R}^p} (\hat{P}_n(H) - P_n(H))^2 \leq cp/n$$

for some universal constant c . This is a consequence of Alexander (1984, Corollary 2.9); see also Pollard (1990, Sections 1-4) for techniques to prove it. If P_n is a mixture of normal distributions as above, condition (6.3) is satisfied if

$$\liminf_{n \rightarrow \infty} \pi_n([0, r]) > 0 \quad \text{for some } r < \infty.$$

7 Proofs

7.1 Proofs for Section 2

Proof of Theorem 2.1 [a, c]: Let $G(Q) = I$, and let $\mathbf{V} \in \mathcal{V}$ with corresponding projection matrix $\Pi \in \mathbf{M}$. Then

$$\dim(\mathbf{V}) = \text{trace}(\Pi) = p \int |x|^{-2} x' \Pi x Q(dx) \geq pQ(\mathbf{V})$$

with equality if, and only if, $Q(\mathbf{V} \cup \mathbf{V}^\perp) = 1$. In this case $G\left((a\Pi + b(I - \Pi))^{-1/2}Q\right)$ equals $G(Q) = I$ for all $a, b > 0$, because $G\left((a\Pi + b(I - \Pi))^{-1/2}x\right) = G(x)$ for any nonzero $x \in \mathbf{V} \cup \mathbf{V}^\perp$. Note that $(a\Pi + b(I - \Pi))^\lambda = a^\lambda\Pi + b^\lambda(I - \Pi)$ for any real λ . \square

First proof of the existence statement in Theorem 2.1 [b]. The arguments of Kent and Tyler (1988) can be modified as follows. Without loss of generality let Q be supported by the unit sphere $\mathbf{S}(\mathbf{R}^p)$. Any local maximum $A \in \mathbf{M}^+(p)$ of the functional

$$\ell(A) := \log \det A - p \int \log(x'Ax) Q(dx)$$

satisfies $G(A^{1/2}Q) = I$, because

$$\begin{aligned} \left. \frac{\partial}{\partial t} \right|_{t=0} \ell(A + t\Delta) &= \text{trace}(A^{-1/2}\Delta A^{-1/2}) - p \int \frac{x'\Delta x}{x'Ax} Q(dx) \\ &= \text{trace}\left(A^{-1/2}\Delta A^{-1/2}(I - G(A^{1/2}Q))\right) \end{aligned}$$

for arbitrary $\Delta \in \mathbf{M}$. Existence of such a local maximum is guaranteed if we can show that $\lim_{k \rightarrow \infty} \ell(A_k) = -\infty$ for any sequence $(A_k)_k$ in $\mathbf{M}^+(p)$ with limit $A \in \mathbf{M}(p) \setminus \mathbf{M}^+$.

For that purpose assume without loss of generality that $A_k = \sum_{i=1}^p \lambda_i(A_k) \tau_{ki} \tau'_{ki}$ with an orthonormal matrix $(\tau_{k1}, \tau_{k2}, \dots, \tau_{kp})$ converging to $(\tau_1, \tau_2, \dots, \tau_p)$ as $k \rightarrow \infty$. For fixed $\epsilon > 0$ and $1 \leq j \leq p$ define

$$S_j := \left\{ x \in \mathbf{S}(\mathbf{R}^p) : \sum_{i=j}^p (\tau'_i x)^2 > 1 - \epsilon^2 \right\} \quad \text{and} \quad D_j := S_j \setminus S_{j+1},$$

where $S_{p+1} := \emptyset$. Note that S_j is just the intersection of the unit sphere $\mathbf{S}(\mathbf{R}^p)$ with the open ϵ -neighborhood of the space $\text{span}(\tau_j, \dots, \tau_p)$. Since

$$\liminf_{k \rightarrow \infty} \min_{x \in \mathbf{S}(\mathbf{R}^p) \setminus S_{j+1}} \frac{x'A_k x}{\lambda_j(A_k)} \geq \liminf_{k \rightarrow \infty} \min_{x \in \mathbf{S}(\mathbf{R}^p) \setminus S_{j+1}} \sum_{i=1}^j (\tau'_{ki} x)^2$$

$$\begin{aligned}
&= \min_{x \in \mathbf{S}(\mathbf{R}^p) \setminus S_{j+1}} \sum_{i=1}^j (\tau'_i x)^2 \\
&\geq \epsilon^2,
\end{aligned}$$

it follows that

$$\begin{aligned}
\ell(A_k) &= \sum_{j=1}^p \left(\log \lambda_j(A_k) - p \int_{D_j} \log(x' A_k x) Q(dx) \right) \\
&\leq \sum_{j=1}^p \left(\log \lambda_j(A_k) - pQ(D_j) \left(\log \lambda_j(A_k) + O(1) \right) \right) \\
&= \sum_{j=1}^p \log \lambda_j(A_k) (1 - pQ(D_j)) + O(1) \\
&= \sum_{j=J}^p \log \lambda_j(A_k) (1 - pQ(D_j)) + O(1) \quad \text{as } k \rightarrow \infty,
\end{aligned}$$

where $J := \min\{j : \lambda_j(A) = 0\} > 1$. If ϵ is sufficiently small, condition (2.3) entails that

$$pQ(S_j) < p - j + 1 \quad \text{for } 2 \leq j \leq p.$$

This will be shown to imply that

$$\begin{aligned}
(7.1) \quad \sum_{j=t}^p \log \lambda_j(A_k) (1 - pQ(D_j)) &\leq \log \lambda_t(A_k) (p - t + 1 - pQ(S_t)) \\
&\text{whenever } \lambda_t(A_k) \leq 1.
\end{aligned}$$

In particular,

$$\ell(A_k) \leq \log \lambda_J(A_k) (p - J + 1 - pQ(S_J)) + O(1) \rightarrow -\infty \quad (k \rightarrow \infty).$$

Inequality (7.1) is proved by reverse induction on t . If $t = p$, then equality holds in (7.1). Now suppose that $\lambda_s(A_k) \leq 1$ and (7.1) is true for $t = s + 1$. Then

$$\begin{aligned}
&\sum_{j=s}^p \log \lambda_j(A_k) (1 - pQ(D_j)) \\
&\leq \log \lambda_s(A_k) (1 - pQ(D_s)) + \log \lambda_t(A_k) (p - t + 1 - pQ(S_t)) \\
&= \log \lambda_s(A_k) (1 - pQ(S_s) + pQ(S_t)) + \log \lambda_t(A_k) (p - t + 1 - pQ(S_t)) \\
&\leq \log \lambda_s(A_k) (1 - pQ(S_s) + pQ(S_t)) + \log \lambda_s(A_k) (p - t + 1 - pQ(S_t)) \\
&= \log \lambda_s(A_k) (p - s + 1 - pQ(S_s)),
\end{aligned}$$

because $\log \lambda_t(A_k) \leq \log \lambda_s(A_k) \leq 0$ and $p - t + 1 - pQ(S_t) > 0$. \square

Second proof of the existence statement in Theorem 2.1 [b]. This proof may be of independent interest and is based on a well-known result from topology (cf. Deimling 1985, Chapter 1 and problem 3.3), which is closely related to Brouwer's fixed-point theorem.

Lemma 7.1 *Let Ω be a compact, convex subset of \mathbf{R}^m with 0 in its interior, and let $f : \Omega \rightarrow \mathbf{R}^m$ be continuous such that*

$$f(x) \notin \{-rx : r > 0\} \quad \text{for all } x \in \partial\Omega.$$

Then $f(x) = 0$ for some $x \in \Omega$. \square

For $0 < \epsilon < 1$ let Ω_ϵ be the set of all $A \in \mathbf{M}(p)$ such that $\lambda_p(A) \geq \epsilon$. Then Ω_ϵ is a compact, convex subset of $\mathbf{M}(p)$ with I in its (relative) interior, and $A \mapsto G(A^{1/2}Q)$ defines a continuous map from Ω_ϵ into $\mathbf{M}(p)$. Now suppose that

$$(7.2) \quad G(A^{1/2}Q) - I \notin \{-r(A - I) : r > 0\} \quad \text{for all } A \in \partial\Omega_\epsilon.$$

Then Lemma 7.1 implies that $G(A^{1/2}Q) = I$ for some $A \in \Omega_\epsilon$.

It remains to be shown that condition (7.2) holds for sufficiently small ϵ . Assume the contrary. Then there exists a sequence $(A_k)_{k \geq 1}$ in $\mathbf{M}^+(p)$ such that $A_k \rightarrow A \in \mathbf{M}(p) \setminus \mathbf{M}^+$ ($k \rightarrow \infty$) and $G(A_k^{1/2}Q) = I - r_k(A_k - I)$ for suitable $r_k > 0$ for all k . Let $\Pi \in \mathbf{M}$ be the projection matrix corresponding to $\mathbf{V} := \{x \in \mathbf{R}^p : Ax = 0\} \in \mathcal{V}$. Then Fatou's Lemma entails that

$$\text{trace}(\Pi G(A_k^{1/2}Q)) = p \int (x' A_k x)^{-1} x' A_k^{1/2} \Pi A_k^{1/2} x Q(dx) \leq pQ(\mathbf{V}) + o(1)$$

as $k \rightarrow \infty$, because $x' A_k^{1/2} \Pi A_k^{1/2} x / x' A_k x \rightarrow 0$ for all $x \in \mathbf{R}^p \setminus \mathbf{V}$. On the other hand,

$$\text{trace}(\Pi G(A_k^{1/2}Q)) = \dim(\mathbf{V}) + r_k \text{trace}(\Pi - \Pi A_k) = \dim(\mathbf{V}) + r_k(\dim(\mathbf{V}) + o(1)),$$

whence $Q(\mathbf{V}) \geq \dim(\mathbf{V})/p$, in contradiction to (2.3). \square

7.2 Proofs for Section 3

Proof of Lemma 3.1: For any $B \in \mathbf{M}(0)$,

$$\text{trace}(D(Q, B)B) = p \int \left(|x|^{-4} (x' B x)^2 - |x|^{-2} x' B^2 x \right) Q(dx).$$

By the Cauchy-Schwarz inequality, $(x' B x)^2 \leq |x|^2 (x' B^2 x)$ with equality if, and only if, x is an eigenvector of B . Hence, if $\lambda_{(1)} > \dots > \lambda_{(m)}$ are the distinct eigenvalues of B , and if $\mathbf{V}_i := \{x \in \mathbf{R}^p : Bx = \lambda_{(i)}x\}$, then $\text{trace}(D(Q, B)B) \leq 0$ with equality if, and only if,

$$Q(\mathbf{V}_1 \cup \dots \cup \mathbf{V}_m) = 1.$$

Now the assertion follows from the fact that $m > 1$ whenever $B \neq 0$. \square

In order to prove Theorem 3.2 one needs explicit bounds for the norm of the remainder term $G((I + B)^{-1/2}Q) - G(Q) - D(Q, B)$.

Lemma 7.2 *There is a universal constant $\kappa_o \in \mathbf{R}^+$ (not depending on Q or p) such that*

$$\left\| G((I + B)^{-1/2}Q) - G(Q) - D(Q, B) \right\| \leq \kappa_o \|G(Q)\| \|B\|^2$$

for arbitrary $B \in \mathbf{M}$ with $\|B\| \leq 1/2$.

Proof of Lemma 7.2: For $A \in \mathbf{M}$ with $\lambda_1(A) < 1$ define

$$K(x, A) := G((I - A)x) - G(x) - 2D(x, A).$$

Then for $y := |x|^{-1}x \in \mathbf{S}(\mathbf{R}^p)$,

$$\begin{aligned} K(x, A) &= \left(\frac{(I - A)G(y)(I - A)}{y'(I - A)^2 y} - G(y) \right) - 2D(y, A) \\ &= \frac{(I - A)G(y)(I - A) - (1 - 2y'Ay + y'A^2y)G(y)}{y'(I - A)^2 y} - 2D(y, A) \\ &= \frac{AG(y)A - y'A^2y G(y) + 2D(y, A)}{y'(I - A)^2 y} - 2D(y, A) \\ &= \frac{AG(y)A - y'A^2y G(y) + 2(2y'Ay - y'A^2y)D(y, A)}{y'(I - A)^2 y}. \end{aligned}$$

The denominator $y'(I - A)^2y$ is not smaller than $\lambda_p((I - A)^2)$. As for the numerator, given any unit vector u , pick $v \in \mathbf{S}(\mathbf{R}^p)$ such that $Au = |Au|v$. Then

$$\begin{aligned} \left| u'(AG(y)A - y'A^2yG(y))u \right| &\leq \|A\|^2 \left(v'G(y)v + u'G(y)u \right), \\ \left| u'(2y'Ay - y'A^2y)D(y, A)u \right| &\leq (2\|A\| + \|A\|^2) |u'D(y, A)u| \\ &\leq (2\|A\|^2 + \|A\|^3) \left(u'G(y)u + |u'G(y)v| \right). \end{aligned}$$

Further there are orthonormal vectors \tilde{u}, \tilde{v} such that

$$\begin{aligned} u &= ((1 + u'v)/2)^{1/2}\tilde{u} + ((1 - u'v)/2)^{1/2}\tilde{v}, \\ v &= ((1 + u'v)/2)^{1/2}\tilde{u} - ((1 - u'v)/2)^{1/2}\tilde{v}, \end{aligned}$$

so that

$$\begin{aligned} |u'G(y)v| &= 2^{-1} \left| (1 + u'v)\tilde{u}'G(y)\tilde{u} - (1 - u'v)\tilde{v}'G(y)\tilde{v} \right| \\ &\leq 2^{-1} (1 + u'v)\tilde{u}'G(y)\tilde{u} + 2^{-1} (1 - u'v)\tilde{v}'G(y)\tilde{v}. \end{aligned}$$

Hence

$$\begin{aligned} \|K(Q, A)\| &\leq \max_{u \in \mathbf{S}(\mathbf{R}^p)} \int |u'K(x, A)u| Q(dx) \\ &\leq \lambda_1((I - A)^{-2}) \left(10\|A\|^2 + 4\|A\|^3 \right) \max_{u \in \mathbf{S}(\mathbf{R}^p)} u'G(Q)u \\ (7.3) \quad &= \lambda_1((I - A)^{-2}) \left(10\|A\|^2 + 4\|A\|^3 \right) \|G(Q)\|. \end{aligned}$$

Moreover, since $\|F(Q, \cdot)\| = \|G(Q)\|$,

$$(7.4) \quad \|D(Q, \cdot)\| \leq 2\|G(Q)\|$$

Now let $B \in \mathbf{M}$ with $\|B\| \leq 1/2$ and define $A := I - (I + B)^{-1/2}$, i.e. $I + B = (I - A)^{-2}$. Then it follows from the spectral representation of B and A , together with a Taylor expansion of the function $t \mapsto 1 - (1 + t)^{-1/2}$, that $\lambda_1((I - A)^{-2}) \leq 1 + \|B\|$, $\|2A - B\| \leq (3/4)\|B\|^2 + \kappa'\|B\|^3$ and $\|A\| \leq \|B\|/2 + \kappa''\|B\|^2$ for universal constants $\kappa', \kappa'' \in \mathbf{R}^+$. Hence (7.3) and (7.4) imply that

$$\left\| G((I + B)^{-1/2}Q) - G(Q) - D(Q, B) \right\|$$

$$\begin{aligned}
&\leq \|K(Q, A)\| + \|D(Q, 2A - B)\| \\
&\leq (1 + \|B\|)(10\|A\|^2 + 4\|A\|^3)\|G(Q)\| + 2\|G(Q)\|\|2A - B\| \\
&\leq \|G(Q)\|(4\|B\|^2 + \kappa'''\|B\|^3)
\end{aligned}$$

for suitable $\kappa''' = \kappa'''(\kappa', \kappa'')$. □

Proof of Theorem 3.2: For notational convenience let $L := D^{-1}(P, \cdot)$. Suppose that $\|L\| \leq b < \infty$ and $\|F(Q - P, \cdot)\| \leq \epsilon \in]0, 1]$. Now $f(B) := L\left(G((I + B)^{-1/2}Q) - I\right)$ defines a continuous mapping from $\Omega := \{B \in \mathbf{M}(0) : \|B\| \leq \rho\}$ into $\mathbf{M}(0)$, where $\rho \in]0, 1/2]$ is some constant. One can write

$$\begin{aligned}
f(B) &= LG(Q - P) + L\left(G((I + B)^{-1/2}Q) - G(Q)\right) \\
&= LG(Q - P) + B + LD(Q - P, B) + L\left(G((I + B)^{-1/2}Q) - G(Q) - D(Q, B)\right) \\
&= LG(Q - P) + B + R(B),
\end{aligned}$$

where

$$\begin{aligned}
\|R(B)\| &\leq b\|D(Q - P, \cdot)\|\|B\| + b\kappa_o\|G(Q)\|\|B\|^2 \\
(7.5) \quad &\leq 2b\|F(Q - P, \cdot)\|\|B\| + 2b\kappa_o\|B\|^2 \\
&\leq 2b(\epsilon + \kappa_o\rho)\|B\|,
\end{aligned}$$

according to Lemma 7.2. Since $\|LG(Q - P)\| \leq b\epsilon$, this implies that

$$\|R(B)\| \leq \|B\|/2 \quad \text{and} \quad \|B - f(B)\| \leq \rho \quad \text{for all } B \in \Omega,$$

provided that $b\epsilon$, $b\rho$ and ϵ/ρ are sufficiently small. Then Brouwer's Fixed Point theorem shows that $f(B_o) = 0$ for some $B_o \in \Omega$. If $f(B_1) = 0$ for some point $B_1 \in \Omega$, which is equivalent to $G((I + B_1)^{-1/2}Q) = I$, then $\|B_1\| \leq \|LG(Q - P)\| + \|R(B_1)\| \leq b\|G(Q - P)\| + \|B_1\|/2$, whence

$$(7.6) \quad \|B_1\| \leq 2b\|G(Q - P)\| \leq 2b\epsilon.$$

Combined with inequality (7.5) this yields

$$\begin{aligned}
\|B_o + LG(Q - P)\| &= \|R(B_o)\| \\
&\leq 4b^2\|F(Q - P, \cdot)\|\|G(Q - P)\| + 8b^3\kappa_o\|G(Q - P)\|^2 \\
&\leq 4b^2(1 + 2b\kappa_o)\|F(Q - P, \cdot)\|\|G(Q - P)\|.
\end{aligned}$$

It remains to be shown that $\Sigma(Q) = I + B_o$, i.e. that Q satisfies condition (2.3). Suppose the contrary. Then, by Theorem 2.1 [c] and (2.2), there exists a proper projection matrix $\Pi \in \mathbf{M}$ such that $G(M^{-1/2}Q) = I$ with $M = (I + B_o)^{1/2}(a\Pi + b(I - \Pi))(I + B_o)^{1/2}$ for arbitrary $a, b > 0$. But then one easily verifies that for suitable $a, b > 0$ the matrix $B_1 := M - I$ belongs to $\partial\Omega$, i.e. $\|B_1\| = \rho$. For sufficiently small ϵ/ρ this is in contradiction to (7.6). \square

7.3 Proof of Corollary 4.1

As for part [a], $P\{0\} = 0$ implies that $P_n\{0\} \rightarrow 0$ and $P_{no} := \mathcal{L}(\mathbf{y}_n | \mathbf{y}_n \neq 0) \rightarrow_w P$. Consequently, by Corollary 3.3, $\Sigma(P_n) = \Sigma(P_{no}) \rightarrow I$, and thus $\Sigma(P_n)^{-1/2}P_n \rightarrow_w P$, according to Rubin's extended Continuous Mapping Theorem (cf. Billingsley 1968, Theorem 5.5). Thus we may assume without loss of generality that $\Sigma(P_n) = I$ for all n . Note further that $\|F(P_{no} - P, \cdot)\| \rightarrow 0$ and thus $\|D^{-1}(P_{no}, \cdot) - D^{-1}(P, \cdot)\| \rightarrow 0$. Furthermore, $\|F(\hat{P}_n - P_n, \cdot)\| = O_p(n^{-1/2})$ and $\hat{P}_n\{0\} - P_n\{0\} = o_p(n^{-1/2})$. Defining $\hat{P}_{no} := \hat{P}_n(\cdot | \mathbf{R}^p \setminus \{0\})$ we thus conclude that

$$\|F(\hat{P}_{no} - P_{no}, \cdot)\| = O_p(n^{-1/2}) \quad \text{and} \quad G(\hat{P}_{no} - P_{no}) = G(\hat{P}_n - P_n) + o_p(n^{-1/2}).$$

Hence Theorem 3.2 yields

$$\begin{aligned} \Sigma(\hat{P}_n) - I &= \Sigma(\hat{P}_{no}) - I = -D^{-1}(P_{no}, G(\hat{P}_{no} - P_{no})) + o_p(n^{-1/2}) \\ &= -D^{-1}(P, G(\hat{P}_n - P_n)) + o_p(n^{-1/2}). \end{aligned}$$

But Lindeberg's multivariate Central Limit Theorem entails that $\mathcal{L}(n^{1/2}G(\hat{P}_n - P_n))$ converges weakly to a centered Gaussian distribution on $\mathbf{M}(0)$ with the same covariances as $G(\mathbf{y})$, $\mathbf{y} \sim P$.

As for part [b], note first that $P_n \otimes P_n \rightarrow_w P \otimes P$. Since P has no atoms, this implies that $P_n^s \rightarrow_w P^s$. Again one may assume without loss of generality that $\Sigma(P_n^s) = I$. The operators $F(P_n^s, \cdot)$ and $F(\hat{P}_n^s, \cdot)$ can be written as

$$\tilde{h}(P_n \otimes P_n)^{-1}F(P_n \otimes P_n, \cdot) \quad \text{and} \quad \tilde{h}(\hat{U}_n)^{-1}\tilde{F}(\hat{U}_n, \cdot),$$

respectively, where $\tilde{h}(x, y) := 1\{x \neq y\}$, $\tilde{F}(x, y, B) := F(x - y, B)$, and \hat{U}_n stands for the random measure

$$\hat{U}_n := \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} \delta_{\mathbf{y}_{ni}} \otimes \delta_{\mathbf{y}_{nj}}$$

on $(\mathbf{R}^p)^2$. But $0 \leq \tilde{h} \leq 1$ and $\tilde{h}(P_n \otimes P_n) \rightarrow 1$, while $\|\tilde{F}(x, y, \cdot)\| \leq p$ for all $x, y \in \mathbf{R}^p$.

Thus standard theory for U-statistics implies that

$$\begin{aligned} \|\tilde{F}(\hat{U}_n - P_n \otimes P_n, \cdot)\| &= O_p(n^{-1/2}), \\ \tilde{h}(\hat{U}_n - P_n \otimes P_n) &= o_p(n^{-1/2}), \\ \tilde{G}(\hat{U}_n - P_n \otimes P_n) &= 2\tilde{G}((\hat{P}_n - P_n) \otimes P_n) + o_p(n^{-1/2}) \end{aligned}$$

(cf. Hoeffding 1948). Consequently,

$$\|F(\hat{P}_n^s - P_n^s, \cdot)\| = O_p(n^{-1/2}) \quad \text{and} \quad G(\hat{P}_n^s - P_n^s) = 2\tilde{G}((\hat{P}_n - P_n) \otimes P_n) + o_p(n^{-1/2}).$$

Since $\|D^{-1}(P_n^s, \cdot) - D^{-1}(P^s, \cdot)\| \rightarrow 0$, Theorem 3.2 entails that

$$\Sigma(\hat{P}_n^s) - I = -D^{-1}\left(P^s, 2\tilde{G}((\hat{P}_n - P_n) \otimes P_n)\right) + o_p(n^{-1/2}).$$

But it follows from Rubin's Theorem that $\tilde{G}(\cdot, P_n) \rightarrow \tilde{G}(\cdot, P)$ uniformly on compact subsets of \mathbf{R}^p . Consequently, one can deduce from the Central Limit Theorem that $n^{1/2}\tilde{G}((\hat{P}_n - P_n) \otimes P_n)$ converges in distribution to a random matrix in $\mathbf{M}(0)$ with centered Gaussian distribution and the same covariance function as $\tilde{G}(\mathbf{y}, P)$, $\mathbf{y} \sim P$. \square

7.4 Proofs for Section 5

The proofs of Theorem 5.1 and Theorem 5.4 utilize the following two lemmas.

Lemma 7.3 *For any normed vector space $(\mathbf{B}, \|\cdot\|)$ let $\mathbf{F}(\mathbf{B})$ be a maximal subset of the sphere $\mathbf{S}(\mathbf{B})$ such that $\|x - y\| > 1/3$ for different $x, y \in \mathbf{F}(\mathbf{B})$. Then*

$$\#\mathbf{F}(\mathbf{B}) \leq \exp(2 \dim(\mathbf{B})) \quad \text{and} \quad \|L\| \leq (3/2) \max_{x \in \mathbf{F}(\mathbf{B})} \|Lx\|$$

for any linear function L from \mathbf{B} into another normed space. In particular,

$$\|M\| \leq 3 \max_{v \in \mathbf{F}(\mathbf{R}^p)} |v'Mv| \quad \text{for all } M \in \mathbf{M}. \quad \square$$

Lemma 7.4 For any finite collection of functions $g_1, g_2, \dots, g_m \in \mathcal{L}^1(P_n)$ and arbitrary numbers $t > 0$,

$$\left(\mathbb{E} \max_{1 \leq j \leq m} g_j(\widehat{\Delta}_n)^2 \right)^{1/2} \leq 2 \left(\frac{\log(2m)}{n} + \max_{1 \leq j \leq m} \log \mathbb{E} \cosh(tg_j(\mathbf{y}_n)) \right) / t,$$

where $\widehat{\Delta}_n := \widehat{P}_n - P_n$.

In Lemma 7.3 the bound $\exp(2 \dim(\mathbf{B}))$ for the cardinality of $\mathbf{F}(\mathbf{B})$ is standard and follows from considering balls of radius $1/6$ with center in $\mathbf{F}(\mathbf{B})$ (cf. Pollard 1990, Section 4). The bounds for $\|L\|$ and $\|M\|$ are elementary.

Proof of Lemma 7.4: This inequality is a modification of Pisier's (1983) Lemma 1.6, which is tailored for our purposes. It follows from Jensen's inequality and convexity of $\exp(\cdot)$ that

$$\begin{aligned} \left(\mathbb{E} \exp(\pm ntg_j(\widehat{\Delta}_n)/2) \right)^{1/n} &= \mathbb{E} \exp(\pm t(g_j(\mathbf{y}_n) - g_j(P_n))/2) \\ &= \mathbb{E} \exp\left(\mathbb{E}(\pm t(g_j(\mathbf{y}_{n1}) - g_j(\mathbf{y}_{n2}))/2 \mid \mathbf{y}_{n1}) \right) \\ &\leq \mathbb{E} \exp(\pm t(g_j(\mathbf{y}_{n1}) - g_j(\mathbf{y}_{n2}))/2) \\ &\leq \left(\mathbb{E} \exp(\pm tg_j(\mathbf{y}_{n1})) + \mathbb{E} \exp(\mp tg_j(\mathbf{y}_{n2})) \right) / 2 \\ &= \mathbb{E} \cosh(tg_j(\mathbf{y}_n)). \end{aligned}$$

Thus $\mathbb{E} \psi(g_j(\widehat{\Delta}_n)^2) \leq \left(\mathbb{E} \cosh(tg_j(\mathbf{y}_n)) \right)^n$, where $\psi(x) := \cosh(ntx^{1/2}/2)$ is convex and increasing in $x \geq 0$. Since $\psi^{-1}(y) \leq \left(2 \log(2y)/(nt) \right)^2$ for $y \geq 1$, a second application of Jensen's inequality yields

$$\begin{aligned} \mathbb{E} \max_j g_j(\widehat{\Delta}_n)^2 &\leq \psi^{-1}\left(\mathbb{E} \max_j \psi(g_j(\widehat{\Delta}_n)^2) \right) \\ &\leq \psi^{-1}\left(\sum_j \mathbb{E} \psi(g_j(\widehat{\Delta}_n)^2) \right) \\ &\leq \psi^{-1}\left(m \max_j \mathbb{E} \psi(g_j(\widehat{\Delta}_n)^2) \right) \\ &\leq \left((2/nt) \log(2m) + (2/t) \max_j \log \mathbb{E} \cosh(tg_j(\mathbf{y}_n)) \right)^2. \quad \square \end{aligned}$$

Proof of Theorem 5.1: Note first that $\|D(P_n, B) + B\| = \|F(P_n, B)\| \leq \sigma_n \kappa_n \|B\|$ for all $B \in \mathbf{M}(0)$, by the Cauchy-Schwarz inequality. Thus $\sup_{B \in \mathbf{S}(\mathbf{M}(0))} \|D^{-1}(P_n, B) + B\|$

converges to zero. Therefore, according to Theorem 3.2, it suffices to show that

$$\begin{aligned}\mathbb{E} \|F(\widehat{\Delta}_n, \cdot)\|^2 &= o(1), \\ \mathbb{E} \|G(\widehat{\Delta}_n)\|^2 &= O(p/n) \quad \text{if } p = O(n^{1/3}),\end{aligned}$$

where $\widehat{\Delta}_n = \widehat{P}_n - P_n$. Lemma 7.3 yields

$$\begin{aligned}\|G(\widehat{\Delta}_n)\| &\leq 3 \max_{v \in \mathbf{F}(\mathbf{R}^p)} |v'G(\widehat{\Delta}_n)v|, \\ \|F(\widehat{\Delta}_n, \cdot)\| &\leq \|G(\widehat{\Delta}_n)\| + \left\| F(\widehat{\Delta}_n, \cdot) \Big|_{\mathbf{M}(0)} \right\|, \\ \left\| F(\widehat{\Delta}_n, \cdot) \Big|_{\mathbf{M}(0)} \right\| &\leq (9/2) \max_{v \in \mathbf{F}(\mathbf{R}^p), B \in \mathbf{F}(\mathbf{M}(0))} |v'F(\widehat{\Delta}_n, B)v|.\end{aligned}$$

In order to bound the latter maximum we use a truncation argument. For any constant $K \geq 0$,

$$\begin{aligned}|v'F(\widehat{\Delta}_n, B)v| &\leq \left| \int |x|^{-2} x' B x \mathbf{1}\{v'G(x)v < K\} v'G(x)v \widehat{\Delta}_n(dx) \right| \\ &\quad + \int \mathbf{1}\{v'G(x)v \geq K\} v'G(x)v (\widehat{P}_n + P_n)(dx) \\ &= K \left| f(\widehat{\Delta}_n | v, B, K) \right| + g(\widehat{\Delta}_n | v, K) + 2g(P_n | v, K),\end{aligned}$$

where

$$\begin{aligned}g(x | v, K) &:= \mathbf{1}\{v'G(x)v \geq K\} v'G(x)v, \\ f(x | v, B, K) &:= |x|^{-2} x' B x \mathbf{1}\{v'G(x)v < K\} v'G(x)v / K\end{aligned}$$

(and $f(x | v, B, 0) := 0$). Note that $v'G(\widehat{\Delta}_n)v = g(\widehat{\Delta}_n | v, 0)$ and $g(P_n | v, K) \leq \kappa_n^2 / K$ for all $K > 0$. Thus it suffices to show that for arbitrary fixed $K \geq 0$,

$$(7.7) \quad \mathbb{E} \max_{v \in \mathbf{F}(\mathbf{R}^p)} g(\widehat{\Delta}_n | v, K)^2 = \begin{cases} o(1), \\ O(p/n) \quad \text{if } p = O(n^{1/3}), \end{cases}$$

$$(7.8) \quad \mathbb{E} \max_{v \in \mathbf{F}(\mathbf{R}^p), B \in \mathbf{F}(\mathbf{M}(0))} f(\widehat{\Delta}_n | v, B, K)^2 = o(1).$$

Since $|g(\cdot)| \leq p$, $\mathbb{E}(g(\mathbf{y}_n | v, K)^2) \leq \kappa_n^2$, $|f(\cdot)| \leq 1$ and $\mathbb{E}(f(\mathbf{y}_n | v, B, K)^2) \leq \sigma_n^2$, one obtains

$$\mathbb{E} \cosh\left(tg(\mathbf{y}_n | v, K)\right) \leq 1 + \sum_{k=1}^{\infty} t^{2k} \kappa_n^2 p^{2k-2} / (2k)!$$

$$\begin{aligned}
&= 1 + (\kappa_n/p)^2(\cosh(pt) - 1) \\
&\leq \exp\left((\kappa_n/p)^2(\cosh(pt) - 1)\right), \\
\mathbb{E} \cosh\left(tf(\mathbf{y}_n | v, B, K)\right) &\leq \exp\left(\sigma_n^2(\cosh(t) - 1)\right)
\end{aligned}$$

for all $t > 0$. Combined this with Lemma 7.3 and 7.4 we get

$$\begin{aligned}
&\left(\mathbb{E} \max_{v \in \mathbf{F}(\mathbf{R}^p)} g(\widehat{\Delta}_n | v, K)^2\right)^{1/2} \\
&\leq 2 \min_{t>0} \left((2p+1)/n + (\kappa_n/p)^2(\cosh(pt) - 1) \right) / t \\
&\leq 2\kappa_n^2/p \min_{r>0} \left((3/\kappa_n^2)p^3/n + \cosh(r) - 1 \right) / r \\
(7.9) \quad &= 2\kappa_n^2/p h\left((3/\kappa_n^2)p^3/n\right),
\end{aligned}$$

$$\begin{aligned}
&\left(\mathbb{E} \max_{v \in \mathbf{F}(\mathbf{R}^p), B \in \mathbf{F}(\mathbf{M}(0))} f(\widehat{\Delta}_n | v, B, K)^2\right)^{1/2} \\
&\leq 2 \min_{t>0} \left((2p+p(p+1)+1)/n + \sigma_n^2(\cosh(t) - 1) \right) / t \\
(7.10) \quad &\leq 2\sigma_n^2 h\left(5p^2/(n\sigma_n^2)\right).
\end{aligned}$$

Here $h(a) := \min_{r>0} (a + \cosh(r) - 1)/r$ is increasing in $a > 0$ with

$$h(a) = \begin{cases} (2a)^{1/2} (1 + o(1)) & \text{as } a \rightarrow 0, \\ a / \log a (1 + o(1)) & \text{as } a \rightarrow \infty. \end{cases}$$

Consequently, (7.9) and (7.10) imply (7.7) and (7.8). \square

The next lemma summarizes some (in)equalities for spherically symmetric distributions.

Lemma 7.5 *Let $\mathbf{y} \sim \mathcal{N}_p(0, I)$. Then $\mathbb{E} G(\mathbf{y}) = I$ and*

$$(7.11) \quad \mathbb{E} \left(\text{trace}(G(\mathbf{y})A) \text{trace}(G(\mathbf{y})B) \right) = \frac{\text{trace}(A) \text{trace}(B) + 2 \text{trace}(AB)}{1 + 2/p}$$

for all $A, B \in \mathbf{M}$. In particular, for spherically symmetric distributions P on $\mathbf{R}^p \setminus \{0\}$,

$$(7.12) \quad D(P, B) = -(1 + 2/p)^{-1} B \quad \text{for all } B \in \mathbf{M}(0).$$

Moreover,

$$(7.13) \quad \mathbb{E} \left((v'G(\mathbf{y})v)^k \right) \leq \mathbb{E} \left((v'\mathbf{y})^{2k} \right) \leq 2^k k!$$

for arbitrary $v \in \mathbf{S}(\mathbf{R}^p)$ and integers $k \geq 1$,

$$(7.14) \quad \mathbb{E} \exp(\text{trace}(G(\mathbf{y})B)) \leq \mathbb{E} \exp(\mathbf{y}'B\mathbf{y}) \leq \exp\left(\text{trace}(B) + \frac{\text{trace}(B^2)}{(1 - 2\|B\|)^+}\right)$$

for arbitrary $B \in \mathbf{M}$.

Proof of Lemma 7.5: The key point is that $G(\mathbf{y}) = G(\mathbf{u})$ and $\mathbf{y}\mathbf{y}' = p^{-1}|\mathbf{y}|^2G(\mathbf{u})$, where $\mathbf{u} := |\mathbf{y}|^{-1}\mathbf{y}$ and $|\mathbf{y}|^2$ are stochastically independent with $\mathcal{L}(|\mathbf{y}|^2) = \chi_p^2$. Consequently, for $A, B \in \mathbf{M}$,

$$\begin{aligned} & \mathbb{E}\left(\text{trace}(G(\mathbf{y})A) \text{trace}(G(\mathbf{y})B)\right) \\ &= \mathbb{E}\left(\text{trace}(\mathbf{y}\mathbf{y}'A) \text{trace}(\mathbf{y}\mathbf{y}'B)\right) / \mathbb{E}(p^{-2}|\mathbf{y}|^4) \\ &= (1 + 2/p)^{-1} \mathbb{E}\left(\text{trace}(\mathbf{y}\mathbf{y}'A) \text{trace}(\mathbf{y}\mathbf{y}'B)\right) \\ &= (1 + 2/p)^{-1} \mathbb{E}(\mathbf{y}'A\mathbf{y} \mathbf{y}'B\mathbf{y}) \\ &= (1 + 2/p)^{-1} \sum_{i,j,k,\ell=1}^p A_{ij}B_{k\ell} \mathbb{E}(y_i y_j y_k y_\ell) \\ &= (1 + 2/p)^{-1} \sum_{i,j=1}^p \left(3\{i=j\}A_{ii}B_{ii} + 1\{i \neq j\}(A_{ii}B_{jj} + A_{ij}B_{ij} + A_{ij}B_{ji})\right) \\ &= (1 + 2/p)^{-1} \sum_{i,j=1}^p (A_{ii}B_{jj} + 2A_{ij}B_{ji}) \\ &= (1 + 2/p)^{-1} \left(\text{trace}(A) \text{trace}(B) + 2 \text{trace}(AB)\right). \end{aligned}$$

In particular, if $A, B \in \mathcal{M}(0)$, then

$$\text{trace}(F(P, B)A) = p^{-1} \mathbb{E}\left(\text{trace}(G(\mathbf{y})A) \text{trace}(G(\mathbf{y})B)\right) = 2(p+2)^{-1} \text{trace}(AB).$$

Hence $F(P, B) = 2(2+p)^{-1}B$ and $D(P, B) = -(1+2/p)^{-1}B$.

Generally, for any convex function $\psi : \mathbf{M} \rightarrow \mathbf{R}$, Jensen's inequality yields

$$\mathbb{E} \psi(G(\mathbf{y})) = \mathbb{E} \psi\left(\mathbb{E}(\mathbf{y}\mathbf{y}' \mid \mathbf{u})\right) \leq \mathbb{E} \mathbb{E}\left(\psi(\mathbf{y}\mathbf{y}') \mid \mathbf{u}\right) = \mathbb{E} \psi(\mathbf{y}\mathbf{y}').$$

In particular, for $v \in \mathbf{S}(\mathbf{R}^p)$ and integers $k \geq 1$,

$$\mathbb{E}\left(|v'G(\mathbf{y})v|^k\right) \leq \mathbb{E}((v'\mathbf{y})^{2k}) = \mathbb{E}(y_1^{2k}) = \prod_{j=0}^{k-1} (1+2j) \leq 2^k k!,$$

while for any $B \in \mathbf{M}$,

$$\begin{aligned}
\mathbb{E} \exp(\text{trace}(G(\mathbf{y})B)) &\leq \mathbb{E} \exp(\mathbf{y}'B\mathbf{y}) \\
&= \mathbb{E} \exp\left(\sum_{i=1}^p \lambda_i(B) g_i^2\right) \\
&= \mathbb{E} \exp\left(-2^{-1} \sum_{i=1}^p \log((1 - 2\lambda_i(B))^+)\right) \\
&= \mathbb{E} \exp\left(\sum_{i=1}^p \sum_{k=1}^{\infty} 2^{k-1} \lambda_i(B)^k / k\right) \\
&\leq \mathbb{E} \exp\left(\sum_{i=1}^p \left(\lambda_i(B) + \lambda_i(B)^2 \sum_{k=0}^{\infty} (2\|B\|)^k\right)\right) \\
&= \exp\left(\text{trace}(B) + \text{trace}(B^2)/(1 - 2\|B\|)^+\right). \quad \square
\end{aligned}$$

Proof of Theorem 5.4 [a]: Since $M \mapsto G(M^{-1/2}\widehat{P}_n)$ depends only on the directions $\mathbf{u}_{ni} := |\mathbf{y}_{ni}|^{-1}\mathbf{y}_{ni}$, which are uniformly distributed on $\mathbf{S}(\mathbf{R}^p)$, one may assume without loss of generality that P_n is a standard normal distribution on \mathbf{R}^p . With the same notation as in the proof of Theorem 5.1, the first two assertions in part [a] follow from Theorem 3.2 and (7.12), provided that the following two claims are true:

$$(7.15) \quad \mathbb{E} \max_{v \in \mathbf{F}(\mathbf{R}^p)} g(\widehat{\Delta}_n | v, K)^2 = O(p/n) \quad \text{uniformly in } K \geq 0,$$

$$\begin{aligned}
&\max_{v \in \mathbf{F}(\mathbf{R}^p)} g(P_n | v, K_n)^2 + K_n^2 \mathbb{E} \max_{v \in \mathbf{F}(\mathbf{R}^p), B \in \mathbf{F}(\mathbf{M}(0))} f(\widehat{\Delta}_n | v, B, K_n)^2 \\
(7.16) \quad &= O\left(\log(n/p)^2 p/n\right)
\end{aligned}$$

for suitable numbers K_n in \mathbf{R}^+ .

Note first that

$$\mathbb{E} \cosh\left(tg(\mathbf{y}_n | v, K)\right) \leq \mathbb{E} \cosh\left(tv'G(\mathbf{y}_n)v\right) \leq \sum_{k=0}^{\infty} (2t)^{2k} = (1 - 4t^2)^{-1},$$

according to (7.13). Thus Lemma 7.3 and 7.4 yield

$$\left(\mathbb{E} \max_{v \in \mathbf{F}(\mathbf{R}^p)} g(\widehat{\Delta}_n | v, K)^2\right)^{1/2} \leq (2/t_n) \left(2p/n - \log(1 - 4t_n^2)\right) = O((p/n)^{1/2})$$

if $t_n := (p/n)^{1/2} \wedge 1/2$. Moreover, it follows from (7.14) that

$$\begin{aligned} g(P_n | v, K_n) &\leq K_n \exp(-K_n/3) \mathbb{E} \exp\left(v' G(\mathbf{y}_n) v/3\right) \\ &\leq K_n \exp(-K_n/3) \exp(2/3) \quad \text{for all } v \in \mathbf{S}(\mathbf{R}^p) \end{aligned}$$

whenever $K_n \geq 3$, because $x \exp(-x/3)$ is decreasing in $x \geq 3$. Setting K_n equal to $(3/2) \log(n/p)$ shows that a sufficient condition for (7.16) is given by

$$(7.17) \quad \mathbb{E} \max_{v \in \mathbf{F}(\mathbf{R}^p), B \in \mathbf{F}(\mathbf{M}(0))} f(\widehat{\Delta}_n | v, B, K_n)^2 = O(p/n).$$

But for any $B \in \mathbf{S}(\mathbf{M}(0))$ and $0 < t \leq p/2$,

$$\begin{aligned} \mathbb{E} \cosh\left(tf(\mathbf{y}_n | v, B, K_n)\right) &\leq \mathbb{E} \cosh\left((t/p) \text{trace}(G(\mathbf{y}_n) B)\right) \\ &\leq \exp\left((t/p)^2 \text{trace}(B^2)/(1 - 2t/p)\right) \\ &\leq \exp\left((t^2/p)/(1 - 2t/p)\right), \end{aligned}$$

according to (7.14). Now (7.17) follows from Lemma 7.3 and 7.4 if $t = t_n = p\left((p/n)^{1/2} \wedge 1/2\right)$.

As for the assertion about coupling with a Wishart matrix, the preceding results show that $\Sigma(\widehat{P}_n)$ may be replaced with $G(\widehat{P}_n)$. The matrix $M_n := n \int x x' \widehat{P}_n(dx)$ has the desired Wishart distribution. Further,

$$\begin{aligned} \mathbb{E}(G(\mathbf{y}_n) - \mathbf{y}_n \mathbf{y}_n') &= 0, \\ \mathbb{E}\left(\left|v'(G(\mathbf{y}_n) - \mathbf{y}_n \mathbf{y}_n')v\right|^2\right) &= \mathbb{E}\left(\left(v'G(\mathbf{y}_n)v\right)^2\right) \text{Var}\left(|\mathbf{y}_n|^2/p\right) = 6/(p+2), \\ \mathbb{E}\left(\left|v'(G(\mathbf{y}_n) - \mathbf{y}_n \mathbf{y}_n')v\right|^k\right) &\leq \left(\mathbb{E}\left(\left(2v'G(\mathbf{y}_n)v\right)^k\right) + \mathbb{E}\left(\left(2v'\mathbf{y}_n \mathbf{y}_n'v\right)^k\right)\right)/2 \leq 4^k k!, \end{aligned}$$

see (7.11) and (7.13). Consequently,

$$\mathbb{E} \cosh\left(tv'(G(\mathbf{y}_n) - \mathbf{y}_n \mathbf{y}_n')v\right) \leq 1 + 3t^2/p + (4t)^4/(1 - 16t^2) \quad \text{for } 0 < t \leq 1/4.$$

If we take $t = t_n = (\epsilon^{-1}(p/n)^{1/2}) \wedge 1/4$ for arbitrarily small $\epsilon > 0$, it follows from Lemma 7.3 and 7.4 that $\mathbb{E}\left(\|G(\widehat{P}_n) - n^{-1}M_n\|^2\right) = o(p/n)$.

As for the assertion about the eigenvalues of $\Sigma(\widehat{P}_n)$, one can modify Silverstein's (1985) arguments in order to show that

$$\|n^{-1}M_n - T_p D_n T_p'\| = O_p\left((\log(p)/n)^{1/2}\right),$$

where T_p is Haar-distributed on the group of orthonormal matrices in $\mathbf{R}^{p \times p}$, while D_n denotes the non-random tridiagonal matrix

$$D_n := n^{-1} \begin{pmatrix} n & (n(p-1))^{1/2} & 0 & \cdots & 0 \\ & n+p-2 & ((n-1)(p-2))^{1/2} & \ddots & \vdots \\ & & n+p-4 & \ddots & 0 \\ & & & \ddots & (n-p+2)^{1/2} \\ \text{(symm.)} & & & & n-p+2 \end{pmatrix}.$$

Silverstein (1985) derived from Geršgorin's theorem that

$$\lambda_1(D_n) \leq (1 + (p/n)^{1/2})^2 \quad \text{and} \quad \lambda_p(D_n) \geq (1 - (p/n)^{1/2})^2.$$

On the other hand consider unit vectors

$$u_{n,+} := k^{-1/2}(0, 1, 1, \dots, 1, 0, \dots, 0)', \quad u_{n,-} := k^{-1/2}(0, -1, 1, -1, \dots, (-1)^k, 0, \dots, 0)'$$

in \mathbf{R}^p with $k = k_n = p^{1/2} + O(1)$ nonzero coefficients. Then $\lambda_1(D_n) \geq u'_{n,+} D_n u_{n,+}$ and $\lambda_p(D_n) \leq u'_{n,-} D_n u_{n,-}$ with

$$\begin{aligned} u'_{n,\pm} D_n u_{n,\pm} &= (kn)^{-1} \left(\sum_{i=2}^{k+1} (n+p-2i) \pm 2 \sum_{i=2}^k ((n-i)(p-i-1))^{1/2} \right) \\ &= (1 \pm (p/n)^{1/2})^2 + O(n^{-1/2}). \quad \square \end{aligned}$$

Proof of Theorem 5.4 [b]: Because of (5.1) and the proof of part [a], we know that $\|\Sigma(\hat{P}_n^s) - G(\hat{P}_n^s)\|$ has expectation $o((p/n)^{1/2})$. Thus it suffices to analyze $G(\hat{P}_n^s)$ in more detail. This is just a matrix-valued U-statistic with Hoeffding-decomposition

$$G(\hat{P}_n^s) = \tilde{G}(\hat{U}_n) = I + 2\tilde{G}(\hat{\Delta}_n \otimes P_n) + \tilde{G}(\hat{R}_n),$$

where $\tilde{G}(x, y) = G(x - y)$ and

$$\hat{U}_n := \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} \delta_{\mathbf{y}_{ni}} \otimes \delta_{\mathbf{y}_{nj}} \quad \text{and} \quad \hat{R}_n := \hat{U}_n - 2\hat{P}_n \otimes P_n + P_n \otimes P_n.$$

Now we show that

$$\mathbb{E} \|\tilde{G}(\hat{R}_n)\| \leq 3 \mathbb{E} \max_{v \in \mathbf{F}(\mathbf{R}^p)} |v' \tilde{G}(\hat{R}_n) v| = o((p/n)^{1/2}).$$

For that purpose we use once more a truncation argument. Let $g(x|v, K)$ be defined as in the proof of Theorem 5.1 and $h(x|v, K) := 1\{v'G(x)v < K\}v'G(x)v = v'G(x)v - g(x|v, K)$. Further let $\tilde{g}(x, y|v, K) := g(x - y|v, K)$ and $\tilde{h}(x, y|v, K) := h(x - y|v, K)$. Then it suffices to show that for suitable positive numbers K_n ,

$$(7.18) \quad \mathbb{E} \max_{v \in \mathbf{F}(\mathbf{R}^p)} \left| \tilde{g}(\hat{R}_n | v, K_n) \right| = o_p((p/n)^{1/2}),$$

$$(7.19) \quad \mathbb{E} \max_{v \in \mathbf{F}(\mathbf{R}^p)} \left| \tilde{h}(\hat{R}_n | v, K_n) \right| = o_p((p/n)^{1/2}).$$

In order to prove (7.18), let π be uniformly distributed on the set of permutations of $\{1, 2, \dots, n\}$ and independent from $(\mathbf{y}_{ni})_{1 \leq i \leq n}$. Then

$$(7.20) \quad \begin{aligned} \hat{U}_n &= \mathbb{E} \left(m^{-1} \sum_{i=1}^n \delta_{\mathbf{y}_{n, \pi(2i-1)}} \otimes \delta_{\mathbf{y}_{n, \pi(2i)}} \mid \mathbf{y}_{n1}, \mathbf{y}_{n2}, \dots, \mathbf{y}_{nn} \right), \\ \hat{P}_n \otimes P_n &= \mathbb{E} \left(m^{-1} \sum_{i=1}^n \delta_{\mathbf{y}_{n, \pi(2i-1)}} \otimes P_n \mid \mathbf{y}_{n1}, \mathbf{y}_{n2}, \dots, \mathbf{y}_{nn} \right), \\ m^{-1} \sum_{i=1}^n \delta_{\mathbf{y}_{n, \pi(2i-1)}} \otimes P_n &= \mathbb{E} \left(m^{-1} \sum_{i=1}^n \delta_{\mathbf{y}_{n, 2i-1}} \otimes \delta_{\mathbf{y}_{n, 2i}} \mid \mathbf{y}_{n2}, \mathbf{y}_{n4}, \mathbf{y}_{n6}, \dots \right), \end{aligned}$$

where $m = m_n = \lfloor n/2 \rfloor$. Consequently, applying Jensen's inequality three times while using the fact that $\mathcal{L}((\mathbf{y}_{ni})_i) = \mathcal{L}((\mathbf{y}_{n, \pi(i)})_i)$ gives us

$$(7.21) \quad \begin{aligned} &\mathbb{E} \max_{v \in \mathbf{F}(\mathbf{R}^p)} \left| \tilde{g}(\hat{R}_n | v, K_n) \right| \\ &\leq \mathbb{E} \max_{v \in \mathbf{F}(\mathbf{R}^p)} \left| \tilde{g}(\hat{U}_n - P_n \otimes P_n | v, K_n) \right| \\ &\quad + 2 \mathbb{E} \max_{v \in \mathbf{F}(\mathbf{R}^p)} \left| \tilde{g}(\hat{P}_n \otimes P_n - P_n \otimes P_n | v, K_n) \right| \\ &\leq 3 \mathbb{E} \max_{v \in \mathbf{F}(\mathbf{R}^p)} \left| \tilde{g} \left(m^{-1} \sum_{i=1}^n \delta_{\mathbf{y}_{n, 2i-1}} \otimes \delta_{\mathbf{y}_{n, 2i}} - P_n \otimes P_n \mid v, K_n \right) \right| \\ &= 3 \mathbb{E} \max_{v \in \mathbf{F}(\mathbf{R}^p)} \left| g(\check{P}_n^s - P_n^s | v, K_n) \right| \\ &\leq 6 \left((2p+1)/m + \max_{v \in \mathbf{F}(\mathbf{R}^p)} \log \mathbb{E} \cosh \left(tg(\mathbf{y}_{n1} - \mathbf{y}_{n2} | v, K_n) \right) \right) / t \\ &= 6 \left((2p+1)/m + \max_{v \in \mathbf{F}(\mathbf{R}^p)} \log \mathbb{E} \cosh \left(tg(\mathbf{y}_n | v, K_n) \right) \right) / t \end{aligned}$$

for arbitrary $t > 0$, where \check{P}_n^s was defined in Remark 5.3. The last inequality follows from Lemma 7.3 and 7.4, applied to $(m, \check{P}_n^s, P_n^s)$ in place of (n, \hat{P}_n, P_n) . The last equality is

due to $G(\mathbf{y}_{n1} - \mathbf{y}_{n2})$ being distributed as $G(\mathbf{y}_n)$. Now we deduce from (7.13) that

$$\begin{aligned}\mathbb{E}\left(g(\mathbf{y}_n | v, K_n)^k\right) &\leq \mathbb{E}\left((v'G(\mathbf{y}_n)v)^k\right) \leq 2^k k!, \\ \mathbb{E}\left(g(\mathbf{y}_n | v, K_n)^2\right) &\leq \mathbb{E}\left((v'G(\mathbf{y}_n)v)^4\right)/K_n^2 \leq C/K_n^2,\end{aligned}$$

whence $\log \mathbb{E} \cosh\left(tg(\mathbf{y}_n | v, K_n)\right) \leq Ct^2/K_n^2 + 16t^4/(1 - 4t^2)_+$. Consequently, if $K_n \rightarrow \infty$ but $K_n^4 \leq n/p$, then (7.18) follows by setting $t = K_n(p/n)^{1/2}$ in (7.21).

As for (7.19), an exponential inequality for degenerate U-statistics yields

$$\mathbb{E} \cosh\left(cn\tilde{h}(\hat{R}_n | v, K_n)/K_n\right) \leq e$$

for some universal constant $c > 0$. This follows from Nolan and Pollard (1987, Section 2) or Arcones and Giné (1994, Proposition 2.3(d)). Hence, with $\psi(x) := \cosh(cx^{1/2}/K_n)$ for $x \geq 0$, one can conclude from Lemma 7.3 and Pisier's (1983) Lemma 1.6 that

$$\begin{aligned}\left(\mathbb{E} \max_{v \in \mathbf{F}(\mathbf{R}^p)} \tilde{h}(\hat{R}_n | v, K_n)^2\right)^{1/2} &\leq \left(\psi^{-1}\left(\exp(2p) \max_{v \in \mathbf{F}(\mathbf{R}^p)} \mathbb{E} \psi\left(\tilde{h}(\hat{R}_n | v, K_n)^2\right)\right)\right)^{1/2} \\ &\leq 2K_n(p+1)/(nc) \\ &= O((p/n)^{3/4}),\end{aligned}$$

because $\psi^{-1}(y) \leq \left(K_n \log(2y)/(nc)\right)^2$.

Now we consider the random matrix $\tilde{G}(\hat{\Delta}_n \otimes P_n)$ in more detail. For fixed $x \in \mathbf{R}^p$ let v_1, v_2, \dots, v_p be an orthonormal basis of \mathbf{R}^p such that $x = |x|v_1$. Then with $\mathbf{y}_n = (y_{n,i})_{1 \leq i \leq n}$ and $h_n(r) := \mathbb{E}\left((r - y_{n,1})^2 / \left((r - y_{n,1})^2 + |\mathbf{y}_n|^2 - y_{n,1}^2\right)\right)$,

$$\begin{aligned}v_1' \tilde{G}(x, P_n) v_1 &= p \mathbb{E}\left((|x| - v_1' \mathbf{y}_n)^2 / \left((|x| - v_1' \mathbf{y}_n)^2 + |\mathbf{y}_n|^2 - (v_1' \mathbf{y}_n)^2\right)\right) \\ &= ph_n(|x|), \\ v_i' \tilde{G}(x, P_n) v_i &= p \mathbb{E}\left((v_i' \mathbf{y}_n)^2 / \left((|x| - v_1' \mathbf{y}_n)^2 + |\mathbf{y}_n|^2 - (v_1' \mathbf{y}_n)^2\right)\right) \\ &= p(p-1)^{-1} \mathbb{E}\left((|\mathbf{y}_n|^2 - y_{n,1}^2) / \left((|x| - y_{n,1})^2 + |\mathbf{y}_n|^2 - y_{n,1}^2\right)\right) \\ &= p(p-1)^{-1}(1 - h_n(|x|)) \quad \text{for } 2 \leq i \leq p, \\ v_i' \tilde{G}(x, P_n) v_j &= 0 \quad \text{for } 1 \leq i < j \leq p.\end{aligned}$$

Hence $\tilde{G}(x, P_n) - I$ can be written as

$$ph_n(|x|)v_1v_1' + p(p-1)^{-1}(1 - h_n(|x|))(I - v_1v_1') - I = 2^{-1}H_n(|x|)(G(x) - I),$$

where

$$H_n(r) := 2(ph_n(r) - 1)/(p - 1) \in [0, 2].$$

This leads to the representation $n^{-1} \sum_{i=1}^n H_n(|\mathbf{y}_{ni}|)(G(\mathbf{y}_{ni}) - I)$ of $2\tilde{G}(\hat{\Delta}_n \otimes P_n)$.

Finally, suppose that $|\mathbf{y}_n|^2/p \rightarrow_p \kappa_o > 0$. Then one easily verifies that $h_n(|\mathbf{y}_n|) \rightarrow_p 1/2$ and thus $H_n(|\mathbf{y}_n|) \rightarrow_p 1$. Since $0 \leq H_n < 2$, this implies that

$$\begin{aligned} \mathbb{E} \left(\left| (H_n(|\mathbf{y}_n|) - 1)v'(G(\mathbf{u}_n) - I)v \right|^k \right) &= \mathbb{E} \left(|H_n(|\mathbf{y}_n|) - 1|^k \right) \mathbb{E} \left(|v'(G(\mathbf{y}_n) - I)v|^k \right) \\ &\leq \epsilon_n 4^k k! \end{aligned}$$

for $k \geq 2$, where $\epsilon_n \rightarrow 0$. Thus $\log \mathbb{E} \cosh \left(t(H_n(|\mathbf{y}_n|) - 1)v'(G(\mathbf{y}_n) - I)v \right)$ is not greater than $16\epsilon_n t^2/(1 - 16t^2)^+$, and a final application of Lemmas 7.3 and 7.4 gives us

$$\mathbb{E} \left\| G(\hat{P}_n) - I - n^{-1} \sum_{i=1}^n H_n(|\mathbf{y}_{ni}|)(G(\mathbf{y}_{ni}) - I) \right\| = o((p/n)^{1/2}).$$

But part [a] provides the expansion $\mathbb{E} \|\Sigma(\hat{P}_n) - G(\hat{P}_n)\| = o_p((p/n)^{1/2})$. Consequently, $\mathbb{E} \|\Sigma(\hat{P}_n^s) - \Sigma(\hat{P}_n)\| = o_p((p/n)^{1/2})$. \square

7.5 Proofs for Section 6

Here is a preliminary result for proving Theorems 6.1 and 6.2.

Lemma 7.6 For $x \in \mathbf{R}^p \setminus \{0\}, \mu \in \mathbf{R}^p, B \in \mathbf{M}$ let

$$H(x, \mu, B) := p|x|^{-4}x'Bx(x\mu' + \mu x') + 2|x|^{-2}x'B\mu G(x) - 4|x|^{-2}x'\mu F(x, B),$$

i.e. $H(x, \mu) = H(x, \mu, I)$. Then there is a universal constant $c \in \mathbf{R}^+$ such that for arbitrary $\mu \in \mathbf{R}^p$,

$$\begin{aligned} \left\| F(Q - Q^{(\mu)}, \cdot) - H(Q, \mu, \cdot) \right\| &\leq \frac{2pQ\{\mu\}}{1 - Q\{\mu\}} + cp|\mu| \int |x|^{-1} \min\{1, |\mu|/|x|\} Q(dx), \\ \|H(Q, \mu, \cdot)\| &\leq 4|\mu| \left(p \int |x|^{-2} Q(dx) \right)^{1/2} \|G(Q)\|^{1/2}. \end{aligned}$$

Proof of Lemma 7.6: For any fixed $B \in \mathbf{S}(\mathbf{M})$,

$$F(Q - Q^{(\mu)}, B) = \int \left(F(x, B) - (1 - Q\{\mu\})^{-1} \{x \neq \mu\} F(x - \mu, B) \right) Q(dx).$$

Since $\|F(x, B)\| \leq p$,

$$(7.22) \quad \begin{aligned} & \left\| F(Q - Q^{(\mu)}, B) - \int \{x \neq \mu\} (F(x, B) - F(x - \mu, B)) Q(dx) \right\| \\ & \leq 2p(1 - Q\{\mu\})^{-1} Q\{\mu\}. \end{aligned}$$

Now define

$$\tilde{H}(x, \gamma, B) := p|x|^{-3}(x'Bx(x\gamma' + \gamma x') + 2x'B\gamma xx').$$

For $u, v \in \mathbf{S}(\mathbf{R}^p)$ and $\gamma := v - u$,

$$\begin{aligned} |v'Bv - u'Bu| & \leq 2|\gamma| \quad \text{and} \quad v'Bv - u'Bu = 2u'B\gamma + \gamma'B\gamma, \\ \|vv' - uu'\| & \leq 2|\gamma| \quad \text{and} \quad vv' - uu' = u\gamma' + \gamma u' + \gamma\gamma'. \end{aligned}$$

Thus

$$\begin{aligned} & \left\| F(v, B) - F(u, B) - \tilde{H}(u, \gamma, B) \right\| \\ & \leq p|v'Bv - u'Bu| \|vv' - uu'\| \\ & \quad + \left\| pu'Bu(vv' - uu') + p(v'Bv - u'Bu)uu' - \tilde{H}(u, \gamma, B) \right\| \\ & \leq 4p|\gamma|^2 + p\|u'Bu\gamma\gamma' + \gamma'B\gamma uu'\| \\ & \leq 6p|\gamma|^2. \end{aligned}$$

In particular, let $u := |x|^{-1}x$ and $v := |x - \mu|^{-1}(x - \mu)$, where $|\mu| < |x|$. Then elementary calculations show that for $\delta := -|x|^{-1}\mu$,

$$|\gamma|^2 \leq c'|\delta|^2 \quad \text{and} \quad |\gamma - \delta + u'\delta u| \leq c''|\delta|^2,$$

where c', c'' are universal constants. Since

$$H(x, \mu, B) = -\tilde{H}(x, \delta - u'\delta u, B)$$

(for arbitrary x, μ), these considerations yield

$$(7.23) \quad \left\| \{x \neq \mu\} (F(x, B) - F(x - \mu, B)) - H(x, \mu, B) \right\| \leq p + 4p|x|^{-1}|\mu|$$

in general, while

$$\begin{aligned}
& \left\| F(x, B) - F(x - \mu, B) - H(x, \mu, B) \right\| \\
& \leq \left\| F(v, B) - F(u, B) - \tilde{H}(u, \gamma, B) \right\| + \left\| \tilde{H}(u, \gamma - \delta + u'\delta u, B) \right\| \\
(7.24) \quad & \leq (6c' + 4c'')p|x|^{-2}|\mu|^2 \text{ if } |\mu| < |x|.
\end{aligned}$$

Now (7.22, 7.23, 7.24) together yield the first asserted inequality.

As for the second inequality, note that for arbitrary $w \in \mathbf{S}(\mathbf{R}^p)$,

$$|w'H(x, \mu, B)w| = |w'\tilde{H}(x, \delta - u'\delta u, B)w| \leq 4p|x|^{-1}|\mu||x|^{-1}|x'w|,$$

Consequently, by the Cauchy-Schwarz inequality,

$$\begin{aligned}
|w'H(Q, \mu, B)w| & \leq 4|\mu| \int p^{1/2}|x|^{-1} (w'G(x)w)^{1/2} Q(dx) \\
& \leq 4|\mu| \left(p \int |x|^{-2} Q(dx) \right)^{1/2} \|G(Q)\|^{1/2}. \quad \square
\end{aligned}$$

Proof of Theorem 6.1: According to Corollary 3.3 it suffices to show that

$$G(Q - Q^{(\mu)}) = H(P, \mu) + o(|\mu|)$$

as

$$(7.25) \quad Q \xrightarrow{w} P, \quad \int |x|^{-1} Q(dx) \rightarrow \int |x|^{-1} P(dx), \quad \mu \rightarrow 0.$$

Lemma 7.6 implies that

$$\begin{aligned}
\left\| G(Q - Q^{(\mu)}) - H(P, \mu) \right\| & \leq |\mu| \|H(Q - P, \cdot)\| + 2p(1 - Q\{\mu\})^{-1} Q\{\mu\} \\
& \quad + cp|\mu| \int |x|^{-1} \min\{1, |\mu|/|x|\} Q(dx).
\end{aligned}$$

Under (7.25) the right hand side is of order $o(|\mu|)$. For

$$|\mu|^{-1} Q\{\mu\} \leq \int |x|^{-1} \min\{1, |\mu|/|x|\} Q(dx) \rightarrow 0,$$

and $\|H(Q - P, \cdot)\| \rightarrow 0$, because $\mathbf{R}^p \setminus \{0\} \ni x \mapsto H(x, \cdot)$ is a continuous mapping into the space of linear mappings $L : \mathbf{R}^p \rightarrow \mathbf{M}$ such that $\|H(x, \cdot)\| \leq 4p|x|^{-1}$. \square

Proof of Theorem 6.2: It follows from Theorems 5.1 and 5.4 that $\mathbb{E} \|F(\widehat{\Delta}_n, \cdot)\| = o(1)$ and $\mathbb{E} \|G(\widehat{\Delta}_n)\| = O((p/n)^{1/2})$. Thus, by Theorem 3.2, it suffices to show that

$$(7.26) \quad \mathbb{E} \sup_{|\mu| \leq \epsilon_n} \|F(\widehat{P}_n^{(\mu)} - \widehat{P}_n, \cdot)\| = o(1),$$

$$(7.27) \quad \|H(P_n, \cdot)\| = O(1),$$

$$(7.28) \quad \mathbb{E} \sup_{|\mu| \leq \epsilon_n} \|G(\widehat{P}_n^{(\mu)} - \widehat{P}_n) - H(P_n, \mu)\| = o((p/n)^{1/2}).$$

Since $\widehat{P}_n\{\mu\} \leq 1/n$ for all $\mu \in \mathbf{R}^p$ almost surely, it follows from Lemma 7.6 and the Cauchy-Schwarz inequality that

$$(7.29) \quad \begin{aligned} & \mathbb{E} \sup_{|\mu| \leq \epsilon_n} \|F(\widehat{P}_n - \widehat{P}_n^{(\mu)}, \cdot) - H(\widehat{P}_n, \mu, \cdot)\| \\ & \leq 2p/(n-1) + cp\epsilon_n^2 \int |x|^{-2} P_n(dx) = O(p/n + \epsilon_n^2), \end{aligned}$$

$$(7.30) \quad \begin{aligned} & \mathbb{E} \sup_{v \in \mathbf{S}(\mathbf{R}^p)} \|H(\widehat{P}_n, v, \cdot)\| + \sup_{v \in \mathbf{S}(\mathbf{R}^p)} \|H(P_n, v, \cdot)\| \\ & \leq 8 \left(p \int |x|^{-2} P_n(dx) \right)^{1/2} \left(\mathbb{E} \|G(\widehat{P}_n)\| \right)^{1/2} = O(1). \end{aligned}$$

These two inequalities imply (7.26) and (7.27). Claim (7.28) follows from (7.29), together with

$$\mathbb{E} \|H(\widehat{\Delta}_n, \cdot)\| = O((p/n)^{1/4}).$$

The latter assertion can be verified as follows:

$$\begin{aligned} \|H(\widehat{\Delta}_n, \cdot)\| & \leq 2p \left| \int |x|^{-2} x \widehat{\Delta}_n(dx) \right| \\ & \quad + 2 \sup_{v \in \mathbf{S}(\mathbf{R}^p)} \left\| \int |x|^{-2} x' v G(x) \widehat{\Delta}_n(dx) \right\|. \end{aligned}$$

On the one hand,

$$p \mathbb{E} \left| \int |x|^{-2} x \widehat{\Delta}_n(dx) \right| \leq pn^{-1/2} \left(\int |x|^{-2} P_n(dx) \right)^{1/2} = O((p/n)^{1/2}).$$

On the other hand, with $\delta_n := (p/n)^{1/4}$ Lemma 7.3 gives

$$\begin{aligned} & \mathbb{E} \sup_{v \in \mathbf{S}(\mathbf{R}^p)} \left\| \int |x|^{-2} x' v G(x) \widehat{\Delta}_n(dx) \right\| \\ & \leq 2p \int \{|x| \leq \delta_n\} |x|^{-1} P_n(dx) \end{aligned}$$

$$\begin{aligned}
& + \mathbb{E} \max_{u,v \in \mathbf{F}(\mathbf{R}^p)} \left\| \int \{|x| > \delta_n\} |x|^{-2} x' v u' G(x) u \widehat{\Delta}_n(dx) \right\| \\
\leq & 2\delta_n p \int |x|^{-2} P_n(dx) \\
& + \delta_n^{-1} \mathbb{E} \max_{u,v \in \mathbf{F}(\mathbf{R}^p)} \left\| f_n(\widehat{\Delta}_n | u, v) \right\|,
\end{aligned}$$

where $f_n(x | u, v) := \{|x| > \delta_n\} \delta_n |x|^{-2} x' v u' G(x) u$, and it suffices to show that

$$(7.31) \quad \mathbb{E} \max_{u,v \in \mathbf{F}(\mathbf{R}^p)} \left\| f_n(\widehat{\Delta}_n | u, v) \right\| = O((p/n)^{1/2}).$$

As in the proof of Theorem 5.1 or Theorem 5.4 one can show that

$$\max_{u,v \in \mathbf{F}(\mathbf{R}^p)} \log \mathbb{E} \cosh(t f_n(\mathbf{y}_n | u, v)) = O(t^2) \quad (t \rightarrow 0, n \rightarrow \infty).$$

Since $\#(\mathbf{F}(\mathbf{R}^p) \times \mathbf{F}(\mathbf{R}^p)) \leq \exp(4p)$, assertion (7.31) follows from Lemma 7.4. \square

8 Some final remarks

In principle different M-estimators such as in Maronna (1976) could be treated similarly. But this would require stronger regularity assumptions (not to mention more complicated notation) without giving substantially better results.

An interesting special case is the maximum likelihood estimator for the multivariate Cauchy distribution on \mathbf{R}^{p-1} . Suppose that $\mathbf{y}_{ni} = (\tilde{\mathbf{y}}_{ni}', 1)'$ with random vectors $\tilde{\mathbf{y}}_{ni} \in \mathbf{R}^{p-1}$ having distribution \tilde{P}_n . If we write

$$\Sigma(\hat{P}_n) = (\Sigma(\hat{P}_n))_{pp} \begin{pmatrix} \check{\Sigma}_n - \check{\mu}_n \check{\mu}_n' & \check{\mu}_n \\ \check{\mu}_n' & 1 \end{pmatrix}$$

with $\check{\mu}_n = \check{\mu}_n(\hat{P}_n) \in \mathbf{R}^{p-1}$ and $\check{\Sigma}_n = \check{\Sigma}_n(\hat{P}_n) \in \mathbf{R}^{(p-1) \times (p-1)}$, then $(\check{\mu}_n, \check{\Sigma}_n)$ is the maximum likelihood estimator for $(\tilde{\mu}_n, \tilde{\Sigma}_n)$ under the model assumption that

$$(8.1) \quad \tilde{P}_n(d\tilde{\mathbf{y}}) = \text{const.} (p-1) \det(\tilde{\Sigma}_n)^{-1/2} \left(1 + (\tilde{\mathbf{y}} - \tilde{\mu}_n)' \tilde{\Sigma}_n^{-1} (\tilde{\mathbf{y}} - \tilde{\mu}_n) \right)^{-p/2} d\tilde{\mathbf{y}};$$

see Kent and Tyler (1991). The results of the present paper can be used directly to derive asymptotic properties of $(\check{\mu}_n, \check{\Sigma}_n)$, where (8.1) is replaced with general regularity conditions on \tilde{P}_n .

A possible objection to M-estimators of scatter is their breakdown-point of order $O(1/p)$ only (cf. Stahel 1981, Tyler 1986). However one should keep in mind that computation of M-estimators is much easier than computation of the high-breakdown estimators considered by Maronna *et al.* (1992), so that resampling methods become feasible. The breakdown properties of the estimator $\Sigma(\hat{P}_n^s)$ are currently investigated.

References

- ALEXANDER, K.S. (1984). Probability inequalities for empirical processes and a law of the iterated logarithm. *Ann. Prob.* **12**, 1041-1067
- ARCONES, M.A. AND E. GINE (1991). Limit Theorems for U -processes. *Ann. Prob.* **21**, 1494-1542
- BAI, Z.D. AND Y. WU (1994a). Limiting behavior of M -estimators of regression coefficients in high dimensional linear models. I. Scale-dependent case. *J. Multivariate Anal.* **51**, 211-239
- BAI, Z.D. AND Y. WU (1994b). Limiting behavior of M -estimators of regression coefficients in high dimensional linear models. II. Scale-invariant case. *J. Multivariate Anal.* **51**, 240-251
- BENNETT, G. (1962). Probability inequalities for the sum of independent random variables. *J. Amer. Statist. Assoc.* **57**, 33-45
- BICKEL, P. AND D. FREEDMAN (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9**, 1196-1217
- BILLINGSLEY, P. (1968). *Convergence of probability measures*. Wiley, New York
- CLARKE, B.R. (1983). Uniqueness and Fréchet differentiability of functional solutions to maximum likelihood type equations. *Ann. Statist.* **11**, 1196-1205
- DEIMLING, K. (1985). *Nonlinear Functional Analysis*. Springer, Berlin Heidelberg

- DÜMBGEN, L. (1994). Perturbation inequalities and confidence sets for functions of a scatter matrix. tentatively accepted for publication in *J. Multivariate Anal.*
- GIRKO, V.L. (1995). *Statistical Analysis of Observations of Increasing Dimension*. Kluwer, Dordrecht
- HOEFFDING, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.* **19**, 293-325.
- HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58**, 13-30
- HUBER, P.J. (1981). *Robust Statistics*. Wiley, New York
- KENT, J.T. AND D.E. TYLER (1988). Maximum likelihood estimation for the wrapped Cauchy distribution. *J. Appl. Statist.* **15**, 247-254
- KENT, J.T. AND D.E. TYLER (1991). Redescending M-estimates of multivariate location and scatter. *Ann. Statist.* **19**, 2102-2119
- MAMMEN, E. (1996). Empirical process of residuals for high-dimensional linear models. *Ann. Statist.* **24**, 307-335
- MARONNA, R.A. (1976). Robust M-estimators of multivariate location and scatter. *Ann. Statist.* **4**, 51-67
- MARONNA, R.A., W.A. STAHEL AND V.J. YOHAI (1992). Bias-robust estimators of multivariate scatter based on projections. *J. Multivariate Anal.* **42**, 141-161
- NOLAN, D. AND D. POLLARD (1987). U-processes: rates of convergence. *Ann. Statist.* **15**, 780-799
- PISIER, G. (1983). Some applications of the metric entropy condition to harmonic analysis. In: *Banach Spaces, Harmonic Analysis, and Probability Theory* (R.C. Blei and S.J. Sidney, eds.), pp. 123-154, Lect. Notes Math. 995, Springer, Berlin Heidelberg

- POLLARD, D. (1990). *Empirical Processes: Theory and Applications*. NSF-CBMS Regional Conf. Series Prob. Statist. **2**, IMS, Hayward CA
- PORTNOY, S. (1984). Asymptotic behavior M -estimators of p regression parameters when p^2/n is large, I. Consistency. *Ann. Statist.* **12**, 1298-1309
- PORTNOY, S. (1985). Asymptotic behavior M -estimators of p regression parameters when p^2/n is large, II. Asymptotic normality. *Ann. Statist.* **13**, 1403-1417
- PORTNOY, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Statist.* **16**, 356-366
- SILVERSTEIN, J.W. (1985). The smallest eigenvalue of a large dimensional Wishart matrix. *Ann. Prob.* **13**, 1364-1368
- Stahel, W.A. (1981): Breakdown of covariance estimators. Research report 31, Fachgruppe Statistik, ETH Zürich
- TYLER, D.E. (1986). Breakdown properties of the M -estimators of multivariate scatter. Technical report, Rutgers University
- TYLER, D.E. (1987). A distribution-free M -estimator of multivariate scatter. *Ann. Statist.* **15**, 234-251