Author(s): Dümbgen, Lutz; Nordhausen, Klaus

Title: Approximating symmetrized estimators of scatter via balanced incomplete U-statistics

Year: 2024

Version: Accepted version (Final draft)

Please cite the original version:

Dümbgen, L., & Nordhausen, K. (2024). Approximating symmetrized estimators of scatter via balanced incomplete U-statistics. Annals of the Institute of Statistical Mathematics, 76(2), 185-207. https://doi.org/10.1007/s10463-023-00879-1

Approximating Symmetrized Estimators of Scatter via Balanced Incomplete U-Statistics

Lutz Dümbgen* and Klaus Nordhausen
University of Bern and University of Jyväskylä

Abstract

We derive limiting distributions of symmetrized estimators of scatter, where instead of all $n(n-1)/2$ pairs of the n observations we only consider nd suitably chosen pairs, $1 \le d < \lfloor n/2 \rfloor$. It turns out that the resulting estimators are asymptotically equivalent to the original one whenever $d = d(n) \to \infty$ at arbitrarily slow speed. We also investigate the asymptotic properties for arbitrary fixed d. These considerations and numerical examples indicate that for practical purposes, moderate fixed values of d between 10 and 20 yield already estimators which are computationally feasible and rather close to the original ones.

*Work supported by Swiss National Science Foundation.

AMS subject classifications: 62H12, 65C60.

Key words: Asymptotic normality, Incomplete U-statistic, independent component analysis, linear expansion, U-statistic.

Corresponding author: Lutz Dümbgen, e-mail: `duembgen@stat.unibe.ch`

1 Introduction

Robust estimation of multivariate scatter for a distribution P on $\mathbb{R}^q$, $q \geq 1$, is a recurring topic in statistics. For instance, different estimators of multivariate scatter are an important ingredient for independent component analysis (ICA) or invariant coordinate selection (ICS), see Nordhausen et al. (2008), Tyler et al. (2009) and the references therein. Other potential applications are classification methods and multivariate regression, see for instance Nordhausen and Tyler (2015). Of particular interest are symmetrized estimators of scatter which are defined in Section 2. Throughout this paper we consider independent random vectors $X_1, X_2, \ldots, X_n$ with distribution P. The symmetrized estimators are just standard functionals of scatter (with given center $0 \in \mathbb{R}^q$) applied to the empirical distribution

$$\hat{Q}_n \ := \ \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} \delta^{\mathrm{s}}_{X_j - X_i},$$

where $\delta^{\mathrm{s}}_z := 2^{-1}(\delta_z + \delta_{-z})$, and δ_z denotes Dirac measure at $z \in \mathbb{R}^q$. Thus, $\hat{Q}_n$ is the empirical distribution of all $n(n-1)$ differences of two different observations, and it may be viewed as a measure-valued version of a U-statistic as introduced by Hoeffding (1948). It is an unbiased estimator of the symmetrized distribution

$$Q = Q(P) \ := \ \mathcal{L}(X_1 - X_2). \tag{1}$$

Here and throughout, $\mathcal{L}(\cdot)$ stands for 'distribution of'. The computation of symmetrized M-estimators of scatter is rather time-consuming, whence some people refrain from using them. However, the symmetrized estimators have two desirable properties: one avoids the estimation of a location nuisance parameter, and the underlying scatter functional has the so-called block independence property as explained in Section 2; see also Dümbgen (1998) and Sirkiä et al. (2007).

To diminish the computational burden, one could replace the empirical distribution $\hat{Q}_n$ with the empirical distribution

$$\hat{Q}_{n,d} \ := \ (nd)^{-1} \sum_{i=1}^{n} \sum_{j=i+1}^{i+d} \delta^{\mathrm{s}}_{X_j - X_i}$$

for some integer $1 \leq d \leq (n-1)/2$, where $X_{n+s} := X_s$ for $1 \leq s \leq n$. This is a measure-valued version of a reduced U-statistic as introduced by Blom (1976) and Brown and Kildea (1978). Other authors, e.g. Lee (1990), call this a balanced incomplete U-statistic. In the context of estimation of scatter, Miettinen et al. (2016) illustrate the potential benefits of $\hat{Q}_{n,d}$ compared to $\hat{Q}_n$ in simulations. As a preliminary proof of concept, they present the asymptotic properties of the estimator $2^{-1} \int_{\mathbb{R}^q} yy^\top \, \hat{Q}_{n,d}(dy)$ in comparison to the usual sample covariance matrix $2^{-1} \int_{\mathbb{R}^q} yy^\top \hat{Q}_n(dy)$. Their findings are encouraging, but the latter estimator can be computed rather easily in $O(n)$ steps and is non-robust of course.

The purpose of the present paper is to provide an in-depth analysis of robust and smooth symmetrized scatter estimators based on $\hat{Q}_n$ and $\hat{Q}_{n,d}$, where the computation time with $\hat{Q}_n$ can definitely become a limiting factor. It turns out that these two scatter estimators are asymptotically equivalent whenever $d = d(n) \to \infty$. Here and throughout the sequel, asymptotic statements are meant as $n \to \infty$. More precisely, if $\mathbf{\Sigma}(\cdot)$ is our functional of scatter, then it will be shown that the following statements are true: There exist two stochastically independent and centered Gaussian random matrices $\boldsymbol{G}_1, \boldsymbol{G}_2$ whose distribution depends only on P and $\mathbf{\Sigma}(\cdot)$ such that

$$\sqrt{n}\big(\mathbf{\Sigma}(\hat{Q}_{n,d(n)}) - \mathbf{\Sigma}(Q)\big) \;=\; \sqrt{n}\big(\mathbf{\Sigma}(\hat{Q}_n) - \mathbf{\Sigma}(Q)\big) + o_p(1) \;\to_{\mathcal{L}}\; \boldsymbol{G}_1$$

provided that $d(n) \to \infty$. Here '$o_p(1)$' denotes a random term converging to zero in probability, and '$\to_{\mathcal{L}}$' denotes convergence in distribution. For any fixed integer $d \geq 1$,

$$\sqrt{n}\big(\mathbf{\Sigma}(\hat{Q}_{n,d}) - \mathbf{\Sigma}(Q)\big) \;\to_{\mathcal{L}}\; \boldsymbol{G}_1 + d^{-1/2}\boldsymbol{G}_2.$$

This explains why for sufficiently large but fixed d, the estimator $\mathbf{\Sigma}(\hat{Q}_{n,d})$ is a good surrogate for $\mathbf{\Sigma}(\hat{Q}_n)$.

An easy way to compute $\mathbf{\Sigma}(\hat{Q}_{n,d})$ is to generate a data matrix containing the nd differences $X_j - X_i$, where $i \in \{1, \ldots, n\}$ and $j \in \{i+1, \ldots, i+d\}$, and to apply $\mathbf{\Sigma}(\cdot)$ to the empirical distribution $(nd)^{-1} \sum_{k=1}^{nd} \delta^{\mathrm{s}}_{Y_k}$ of these nd vectors Y_k. But for large values of nd, this may be too cumbersome. A possible alternative is to compute the average $d^{-1} \sum_{\ell=1}^{d} \mathbf{\Sigma}(\hat{Q}_{n,1}^{(\ell)})$ with the same $d \geq 1$, where $\hat{Q}_{n,1}^{(1)}, \ldots, \hat{Q}_{n,1}^{(d)}$ are defined as $\hat{Q}_{n,1}$, but with d random permutations of the observations $X_1, X_2, \ldots, X_n$. It turns out that for fixed d, this average has the same asymptotic distribution as $\mathbf{\Sigma}(\hat{Q}_{n,d})$.

The remainder of this paper is organized as follows: In Section 2, we recall some basic facts about scatter functionals and symmetrized scatter functionals as presented by Dümbgen et al. (2015). In Section 3, the asymptotic results mentioned before are stated in detail. The theory is illustrated with numerical examples in Section 4. All proofs are deferred to Section 5 and Appendix A. The starting point is standard theory for complete and incomplete U-statistics as presented, for instance, by Serfling (1980) and Lee (1990). Suitable modifications of these results, combined with linear expansions for functionals of scatter yield the asymptotic distributions of $\mathbf{\Sigma}(\hat{Q}_n)$ and $\mathbf{\Sigma}(\hat{Q}_{n,d})$. For the averaging estimator $d^{-1} \sum_{\ell=1}^{d} \mathbf{\Sigma}(\hat{Q}_{n,1}^{(\ell)})$, we derive and use a variation of the combinatorial central limit theorem of Hoeffding (1951). This result is potentially of independent interest, for instance, in the context of kernel mean embeddings as used in machine learning (Muandet et al., 2017).

2 Functionals of Scatter

The material in this section is adapted from the survey of Dümbgen et al. (2015), where the latter builds on previous work of Tyler (1987), Kent and Tyler (1991) and Dudley et al. (2009).

The space of symmetric matrices in $\mathbb{R}^{q \times q}$ is denoted by $\mathbb{R}^{q \times q}_{\mathrm{sym}}$, and $\mathbb{R}^{q \times q}_{\mathrm{sym},+}$ stands for its subset of positive definite matrices. The identity matrix in $\mathbb{R}^{q \times q}$ is written as I_q. The Euclidean norm of a vector $v \in \mathbb{R}^q$ is denoted by $\|v\| = \sqrt{v^\top v}$. For matrices M, N with identical dimensions we write

$$\langle M, N\rangle \ := \ \mathrm{tr}(M^\top N) \quad \text{and} \quad \|M\| \ := \ \sqrt{\langle M, M\rangle},$$

so $\|M\|$ is the Frobenius norm of M.

2.1 Functionals of scatter for centered distributions

Let $\mathcal{Q}$ be a given family of probability distributions on $\mathbb{R}^q$ which are viewed as centered around 0. In our specific applications, this is plausible, because $\mathcal{Q}$ consists of symmetrized distributions. We consider a function $\mathbf{\Sigma} : \mathcal{Q} \to \mathbb{R}^{q\times q}_{\mathrm{sym},+}$, called a functional of scatter, and $\mathbf{\Sigma}(Q)$ is the scatter matrix of $Q \in \mathcal{Q}$. For the general theory presented in the next section, we assume that $\mathcal{Q}$ and $\mathbf{\Sigma}$ have two important properties.

Linear equivariance. We assume that for any nonsingular matrix $B \in \mathbb{R}^{q\times q}$ and any distribution $Q \in \mathcal{Q}$, the distribution $Q^B := \mathcal{L}(BY)$ with $Y \sim Q$ belongs to $\mathcal{Q}$ too, and that

$$\mathbf{\Sigma}(Q^B) \ = \ B\mathbf{\Sigma}(Q)B^\top. \tag{2}$$

Linear equivariance has some interesting implications. For instance, if $Q \in \mathcal{Q}$ is spherically symmetric in the sense that $Q^B = Q$ for all orthogonal matrices $B \in \mathbb{R}^{q\times q}$, then $\mathbf{\Sigma} = cI_q$ for some $c > 0$. Furthermore, if $Q^B = Q$ for some matrix $B = \mathrm{diag}(\xi_1, \ldots, \xi_q)$ with $\xi \in \{-1, 1\}^q$, then for arbitrary different indices $i, j \in \{1, \ldots, q\}$, the (i, j)-th component of $\mathbf{\Sigma}(Q)$ satisfies

$$\mathbf{\Sigma}(Q)_{ij} \ = \ 0 \quad \text{whenever } \xi_i \neq \xi_j. \tag{3}$$

Differentiability. We assume that $\mathcal{Q}$ is an open subset of the family of all probability distributions on $\mathbb{R}^q$ in the topology of weak convergence. Moreover, for any distribution $Q \in \mathcal{Q}$, there exists a bounded, measurable and even function $J = J_Q : \mathbb{R}^q \to \mathbb{R}^{q\times q}_{\mathrm{sym}}$ such that $\int_{\mathbb{R}^q} J\, dQ = 0$, and for other distributions $\check{Q} \in \mathcal{Q}$,

$$\mathbf{\Sigma}(\check{Q}) \ = \ \mathbf{\Sigma}(Q) + \int_{\mathbb{R}^q} J\, d\check{Q} + o\Big(\Big\|\int_{\mathbb{R}^q} J\, d\check{Q}\Big\|\Big)$$

as $\check{Q} \to Q$ weakly. Note that this differentiability property of $\mathbf{\Sigma}(\cdot)$ implies its robustness in the sense that $\mathbf{\Sigma}(\check{Q}) \to \mathbf{\Sigma}(Q)$ as $\check{Q} \to Q$ weakly, because then $\int_{\mathbb{R}^q} J\, d\check{Q} \to \int_{\mathbb{R}^q} J\, dQ = 0$.

***M*-functionals of scatter.** An important example for $\mathbf{\Sigma}$ are M-functionals of scatter, driven by a function $\rho : [0, \infty) \to \mathbb{R}$ with the following properties: ρ is twice continuously differentiable such that $\psi(s) := s\rho'(s)$ satisfies the inequalities $\psi'(s) > 0$ for $s > 0$ and $q < \psi(\infty) := \lim_{s\to\infty} \psi(s) < \infty$. For any distribution Q on $\mathbb{R}^q$ and $\Sigma \in \mathbb{R}^{q\times q}_{\mathrm{sym},+}$, let

$$L_\rho(\Sigma, Q) \ := \ \int_{\mathbb{R}^q} \big[\rho(y^\top\Sigma^{-1}y) - \rho(y^\top y)\big]\, Q(dy) + \log\det(\Sigma). \tag{4}$$

The function $L_\rho(\cdot, Q)$ has a unique minimizer $\mathbf{\Sigma}(Q)$ on $\mathbb{R}^{q\times q}_{\mathrm{sym},+}$ if and only if

$$Q(\mathbb{W}) \ < \ \frac{\psi(\infty) - q + \dim(\mathbb{W})}{\psi(\infty)}$$

for any linear subspace $\mathbb{W}$ of $\mathbb{R}^q$ with $0 \le \dim(\mathbb{W}) < q$. The set $\mathcal{Q}$ of distributions which satisfy the latter constraints is open with respect to weak convergence.

A particular example for a function ρ with the stated properties is given by $\rho(s) = \rho_\nu(s) := (\nu + q)\log(s + \nu)$, where $\nu > 0$.

The function $J = J_Q$ is rather complicated in general. But in case of a spherically symmetric distribution Q with $\mathbf{\Sigma}(Q) = I_q$,

$$J(y) \ = \ \frac{q+2}{q+2+2\kappa}\, \rho'(\|y\|^2)\Big(yy^\top - \frac{\|y\|^2}{q} I_q\Big) + \frac{1}{1+\kappa}\Big(\rho'(\|y\|^2)\frac{\|y\|^2}{q} - 1\Big) I_q$$

for $y \in \mathbb{R}^q$, where $\kappa := q^{-1}\int_{\mathbb{R}^q} \rho''(\|y\|^2)\|y\|^4\, Q(dy) \in (-1, \infty)$.

2.2 Tyler's (1987) functional of scatter

For any distribution Q on $\mathbb{R}^q$ such that $Q(\{0\}) = 0$ and $\Sigma \in \mathbb{R}^{q\times q}_{\mathrm{sym},+}$, let

$$L_0(\Sigma, Q) \ := \ q \int_{\mathbb{R}^q} \log\Big(\frac{y^\top \Sigma^{-1} y}{y^\top y}\Big)\, Q(dy) + \log\det(\Sigma).$$

Note that $L_0(t\Sigma, Q) = L_0(\Sigma, Q)$ for all $t > 0$. The function $L_0(\cdot, Q)$ has a unique minimizer $\mathbf{\Sigma}_0(Q)$ on the set $\big\{\Sigma \in \mathbb{R}^{q\times q}_{\mathrm{sym},+} : \det(\Sigma) = 1\big\}$ if and only if

$$Q(\mathbb{W}) \ < \ \frac{\dim(\mathbb{W})}{q}$$

for any linear subspace $\mathbb{W}$ of $\mathbb{R}^q$ with $1 \le \dim(\mathbb{W}) < q$. The set of all distributions Q which satisfy the latter constraints and $Q(\{0\}) = 0$ is denoted by $\mathcal{Q}_0$.

The functional $\mathbf{\Sigma}_0$ satisfies a restricted equivariance property: For any $Q \in \mathcal{Q}_0$ and any nonsingular matrix $B \in \mathbb{R}^{q\times q}$ with $|\det(B)| = 1$, equation (2) holds true with $\mathbf{\Sigma}_0$ in place of $\mathbf{\Sigma}$. This implies that $\mathbf{\Sigma}_0(Q) = I_q$ if Q is spherically symmetric. Moreover, if $Q^B = Q$ with $B = \mathrm{diag}(\xi_1, \ldots, \xi_q)$ and $\xi \in \{-1, 1\}^q$, then (3) is satisfied with $\mathbf{\Sigma}_0$ in place of $\mathbf{\Sigma}$.

The functional $\mathbf{\Sigma}_0$ is also differentiable in the following sense: For any distribution $Q \in \mathcal{Q}_0$ there exists a bounded, continuous and even function $J : \mathbb{R}^q \setminus \{0\} \to \mathbb{R}^{q\times q}_{\mathrm{sym}}$ such that $\int_{\mathbb{R}^q} J\, dQ = 0$, $\mathrm{trace}\big(\mathbf{\Sigma}_0(Q)^{-1} J\big) \equiv 0$, and for any distribution $\check{Q} \in \mathcal{Q}$,

$$\mathbf{\Sigma}_0(\check{Q}) \ = \ \mathbf{\Sigma}_0(Q) + \int_{\mathbb{R}^q} J\, d\check{Q} + o\Big(\Big\|\int_{\mathbb{R}^q} J\, d\check{Q}\Big\|\Big)$$

as $\check{Q} \to Q$ weakly. Again, the function $J = J_Q$ is rather complicated in general. But in case of a spherically symmetric distribution $Q \in \mathcal{Q}_0$,

$$J(y) \ = \ (q+2)\big(\|y\|^{-2} yy^\top - q^{-1} I_q\big), \quad y \in \mathbb{R}^q \setminus \{0\}.$$

2.3 Symmetrized M-functionals of scatter

Now we consider a general distribution P on $\mathbb{R}^q$ and want to define its scatter matrix without having to specify a center of P. To this end we consider the symmetrized distribution $Q = Q(P)$ as defined in (1). Then the symmetrized version of the functional of scatter $\boldsymbol{\Sigma}$ is given by

$$\boldsymbol{\Sigma}^{\mathrm{s}}(P) \ := \ \boldsymbol{\Sigma}(Q(P)).$$

Here we assume that P belongs to the family $\mathcal{P}$ of all probability distributions on $\mathbb{R}^q$ such that $Q(P) \in \mathcal{Q}$. In case of an M-functional $\boldsymbol{\Sigma}$ with underlying function ρ, a sufficient condition for $P \in \mathcal{P}$ is that

$$P(H) \ = \ 0 \quad \text{for any hyperplane } H \subset \mathbb{R}^q. \tag{5}$$

Analogously, one may define the symmetrized version of Tyler's functional $\boldsymbol{\Sigma}_0$ via $\boldsymbol{\Sigma}_0^{\mathrm{s}}(P) := \boldsymbol{\Sigma}_0(Q(P))$, where we assume that P belongs to the family $\mathcal{P}_0$ of all probability distributions on $\mathbb{R}^p$ such that $Q(P) \in \mathcal{Q}_0$. Again, condition (5) is sufficient for that.

As to the benefits of symmetrization, suppose that P is elliptically symmetric with unknown center $\mu_* \in \mathbb{R}^q$ and unknown scatter matrix $\Sigma_* \in \mathbb{R}^{q\times q}_{\mathrm{sym},+}$. That means, the distribution of $\Sigma_*^{-1/2}(X_1 - \mu_*)$ is spherically symmetric. Then $Q(P)$ is elliptically symmetric with center 0 and the same scatter matrix Σ_*. Note that Σ_* is defined only up to positive multiples. This is no problem as long as one is mainly interested in the shape matrix $\mathrm{shape}(\Sigma_*)$, where

$$\mathrm{shape}(\Sigma) \ := \ \det(\Sigma)^{-1/q}\,\Sigma$$

for $\Sigma \in \mathbb{R}^{q\times q}_{\mathrm{sym},+}$, that is, $\mathrm{shape}(\Sigma)$ is a positive multiple of Σ with determinant one. For instance, in connection with principal components, regression coefficients and correlation measures, multiplying Σ_* with a positive scalar has no impact. Our specific choice of $\mathrm{shape}(\Sigma)$ is justified by Paindaveine (2008).

Symmetrization has a second, even more important advantage: Consider an arbitrary distribution P, not necessarily symmetric in any sense. Suppose that a random vector $X \sim P$ may be written as $X = [X_{\mathrm{a}}^\top, X_{\mathrm{b}}^\top]^\top$ with independent subvectors $X_{\mathrm{a}} \in \mathbb{R}^{q(\mathrm{a})}$ and $X_{\mathrm{b}} \in \mathbb{R}^{q(\mathrm{b})}$. Then $\boldsymbol{\Sigma}^{\mathrm{s}}(P)$ is block-diagonal in the sense that

$$\boldsymbol{\Sigma}^{\mathrm{s}}(P) \ = \ \begin{bmatrix} \boldsymbol{\Sigma}_{\mathrm{a}}(P) & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\mathrm{b}}(P) \end{bmatrix}$$

with certain matrices $\boldsymbol{\Sigma}_{\mathrm{a}}(P) \in \mathbb{R}^{q(\mathrm{a})\times q(\mathrm{a})}_{\mathrm{sym},+}$ and $\boldsymbol{\Sigma}_{\mathrm{b}}(P) \in \mathbb{R}^{q(\mathrm{b})\times q(\mathrm{b})}_{\mathrm{sym},+}$.

At this point, it is not clear whether $\hat{Q}_n$ or $\hat{Q}_{n,d}$ belongs to $\mathcal{Q}$. As explained in Section 5, $\hat{Q}_n$ and $\hat{Q}_{n,d}$ converge weakly in probability to Q, uniformly in $1 \le d \le (n-1)/2$. Thus, $\mathbb{P}(\hat{Q}_n \in \mathcal{Q})$ and $\min_{1\le d\le (n-1)/2} \mathbb{P}(\hat{Q}_{n,d} \in \mathcal{Q})$ converge to 1. The same conclusion is true for $(\boldsymbol{\Sigma}_0, \mathcal{Q}_0)$ in place of $(\boldsymbol{\Sigma}, \mathcal{Q})$, if we assume that P has no atoms, that is, if $P(\{x\}) = 0$ for any $x \in \mathbb{R}^q$. Here is also a non-asymptotic result for M-estimators of scatter in case of smooth distributions P:

Proposition 1. *Suppose that P satisfies (5). With probability one, $\hat{Q}_n(\{0\}) = \hat{Q}_{n,d}(\{0\}) = 0$, and in the case of $n > q$,*

$$\hat{Q}_n(\mathbb{W}), \hat{Q}_{n,d}(\mathbb{W}) \ < \ \frac{\dim(\mathbb{W})}{q}$$

for arbitrary linear subspaces $\mathbb{W}$ of $\mathbb{R}^q$ with $1 \le \dim(\mathbb{W}) < q$ and $1 \le d \le (n-1)/2$.

This theorem implies that for the M-functional $\boldsymbol{\Sigma}(\cdot)$, the symmetrized M-estimators $\boldsymbol{\Sigma}(\hat{Q}_n)$ and $\boldsymbol{\Sigma}(\hat{Q}_{n,d})$ are well-defined almost surely for $1 \le d \le (n-1)/2$, provided that $n > q$ and P satisfies (5). The same conclusion is true for Tyler's M-functional $\boldsymbol{\Sigma}_0(\cdot)$ in place of $\boldsymbol{\Sigma}(\cdot)$.

3 Asymptotic Expansions and Distributions

In what follows, $\boldsymbol{\Sigma}(\cdot)$ denotes either a linear equivariant and differentiable scatter functional or Tyler's functional $\boldsymbol{\Sigma}_0(\cdot)$. In addition to $\hat{Q}_n$ and $\hat{Q}_{n,d}$, we consider the usual empirical distribution of the observations X_i,

$$\hat{P}_n \ := \ n^{-1} \sum_{i=1}^n \delta_{X_i}.$$

Theorem 2. *Suppose that $\boldsymbol{\Sigma}(Q)$ is well-defined for $Q = Q(P)$. With $J = J_Q$, define*

$$H_1(x) \ := \ \mathbb{E}\, J(x - X_1) \quad \text{and} \quad H_2(x,y) \ := \ J(x-y) - H_1(x) - H_1(y)$$

for $x, y \in \mathbb{R}^q$, where $J(0) := 0$ in connection with Tyler's functional. Let $\boldsymbol{G}_1$ and $\boldsymbol{G}_2$ be two stochastically independent Gaussian random matrices in $\mathbb{R}^{q\times q}_{\mathrm{sym}}$ such that $\mathbb{E}\,\boldsymbol{G}_1 = \mathbb{E}\,\boldsymbol{G}_2 = 0$, and

$$\begin{aligned} \mathbb{E}\big(\langle A, \boldsymbol{G}_1\rangle^2\big) \ &= \ \mathbb{E}\big(\langle A, H_1(X_1)\rangle^2\big), \\ \mathbb{E}\big(\langle A, \boldsymbol{G}_2\rangle^2\big) \ &= \ \mathbb{E}\big(\langle A, H_2(X_1, X_2)\rangle^2\big) \end{aligned}$$

for all matrices $A \in \mathbb{R}^{q\times q}_{\mathrm{sym}}$. If $(n-1)/2 \ge d(n) \to \infty$, then

$$\left.\begin{matrix} \boldsymbol{\Sigma}(\hat{Q}_n) \\ \boldsymbol{\Sigma}(\hat{Q}_{n,d(n)}) \end{matrix}\right\} \ = \ \boldsymbol{\Sigma}(Q) + 2\int_{\mathbb{R}^q} H_1 \, d\hat{P}_n + o_p(n^{-1/2}).$$

For fixed integers $d \ge 1$,

$$\boldsymbol{\Sigma}(\hat{Q}_{n,d}) \ = \ \boldsymbol{\Sigma}(\hat{Q}_n) + \boldsymbol{M}_{n,d} + o_p(n^{-1/2}),$$

where

$$\boldsymbol{M}_{n,d} \ := \ (nd)^{-1} \sum_{i=1}^n \sum_{j=i+1}^{i+d} H_2(X_i, X_j).$$

Moreover,

$$\Big(\sqrt{n}\int_{\mathbb{R}^q} H_1 \, d\hat{P}_n, \ \sqrt{nd}\,\boldsymbol{M}_{n,d}\Big) \ \to_{\mathcal{L}} \ (\boldsymbol{G}_1, \boldsymbol{G}_2).$$

In particular, as $d(n) \to \infty$,

$$\left.\begin{matrix} \sqrt{n}\big(\boldsymbol{\Sigma}(\hat{Q}_n) - \boldsymbol{\Sigma}(Q)\big) \\ \sqrt{n}\big(\boldsymbol{\Sigma}(\hat{Q}_{n,d(n)}) - \boldsymbol{\Sigma}(Q)\big) \end{matrix}\right\} \ \to_{\mathcal{L}} \ 2\boldsymbol{G}_1,$$

whereas for fixed integers $d \geq 1$,

$$\sqrt{n}\big(\boldsymbol{\Sigma}(\hat{Q}_{n,d}) - \boldsymbol{\Sigma}(Q)\big) \ \to_{\mathcal{L}} \ 2\boldsymbol{G}_1 + d^{-1/2}\boldsymbol{G}_2.$$

It remains to explain the asymptotic properties of the alternative estimator $d^{-1}\sum_{\ell=1}^d \boldsymbol{\Sigma}(\hat{Q}_{n,1}^{(\ell)})$, where $\hat{Q}_{n,1}^{(\ell)}$ is defined as $\hat{Q}_{n,1}$ with $(X_{\Pi^{(\ell)}(i)})_{i=1}^n$ in place of $(X_i)_{i=1}^n$. Here $\Pi^{(1)}, \ldots, \Pi^{(d)}$ are independent random permutations of $\{1, 2, \ldots, n\}$, and independent from the data $(X_i)_{i=1}^n$.

Theorem 3. *For fixed* $d \geq 1$ *and* $1 \leq \ell \leq d$,

$$\boldsymbol{\Sigma}(\hat{Q}_{n,1}^{(\ell)}) \ = \ \boldsymbol{\Sigma}(Q) + 2\int_{\mathbb{R}^q} H_1 \, d\hat{P}_n + M_{n,1}^{(\ell)} + o_p(n^{-1/2}),$$

where

$$M_{n,1}^{(\ell)} \ := \ n^{-1}\sum_{i=1}^n H_2(X_{\Pi^{(\ell)}(i)}, X_{\Pi^{(\ell)}(i+1)})$$

with $\Pi^{(\ell)}(n+1) := \Pi^{(\ell)}(1)$. *Moreover,*

$$\sqrt{n}\Big(\int_{\mathbb{R}^q} H_1 \, d\hat{P}_n, \, M_{n,1}^{(1)}, \ldots, M_{n,1}^{(d)}\Big) \ \to_{\mathcal{L}} \ \big(\boldsymbol{G}_1, \, \boldsymbol{G}_2^{(1)}, \ldots, \boldsymbol{G}_2^{(d)}\big)$$

with independent random matrices $\boldsymbol{G}_1$ *and* $\boldsymbol{G}_2^{(1)}, \ldots, \boldsymbol{G}_2^{(d)}$, *where* $\boldsymbol{G}_1$ *and* $\boldsymbol{G}_2^{(\ell)}$ *have the same distribution as* $\boldsymbol{G}_1$ *and* $\boldsymbol{G}_2$, *respectively, in Theorem 2. In particular,*

$$\sqrt{n}\Big(d^{-1}\sum_{\ell=1}^d \boldsymbol{\Sigma}(\hat{Q}_{n,1}^{(\ell)}) - \boldsymbol{\Sigma}(Q)\Big) \ \to_{\mathcal{L}} \ 2\boldsymbol{G}_1 + d^{-1/2}\boldsymbol{G}_2.$$

This theorem shows that averaging $\boldsymbol{\Sigma}(\hat{Q}_{n,1}^{(\ell)})$ over $\ell = 1, \ldots, d$ is asymptotically equivalent to computing $\boldsymbol{\Sigma}(\hat{Q}_{n,d})$. One could guess that averaging over $d(n)$ random permutations with $d(n) \to \infty$ leads to an estimator with the same asymptotic distributions as $\boldsymbol{\Sigma}(\hat{Q}_n)$. But this is not obvious, because the average of $d(n)$ random variables which are uniformly of order $o_p(1)$ need not be of order $o_p(1)$ too.

4 Numerical Illustration

The computations are based on Partial Newton algorithms proposed by Dümbgen et al. (2016). They are implemented in the R package *fastM* by Dümbgen et al. (2014) which is publicly available on CRAN.

As explained in Section 2.3, in numerous applications one is mainly interested in the scatter matrix up to positive scalars. Thus we illustrate the previous results with the shape matrix $\boldsymbol{H} := \mathrm{shape}(\boldsymbol{\Sigma}(Q))$ and its estimators

$$\begin{aligned}
\hat{\boldsymbol{H}}_n \ &:= \ \mathrm{shape}(\boldsymbol{\Sigma}(\hat{Q}_n)), \\
\hat{\boldsymbol{H}}_{n,d} \ &:= \ \mathrm{shape}(\boldsymbol{\Sigma}(\hat{Q}_{n,d})), \\
\hat{\boldsymbol{H}}_{n,d}^{\mathrm{rand}} \ &:= \ \mathrm{shape}\Big(d^{-1}\sum_{\ell=1}^d \boldsymbol{\Sigma}(\hat{Q}_{n,1}^{(\ell)})\Big).
\end{aligned}$$

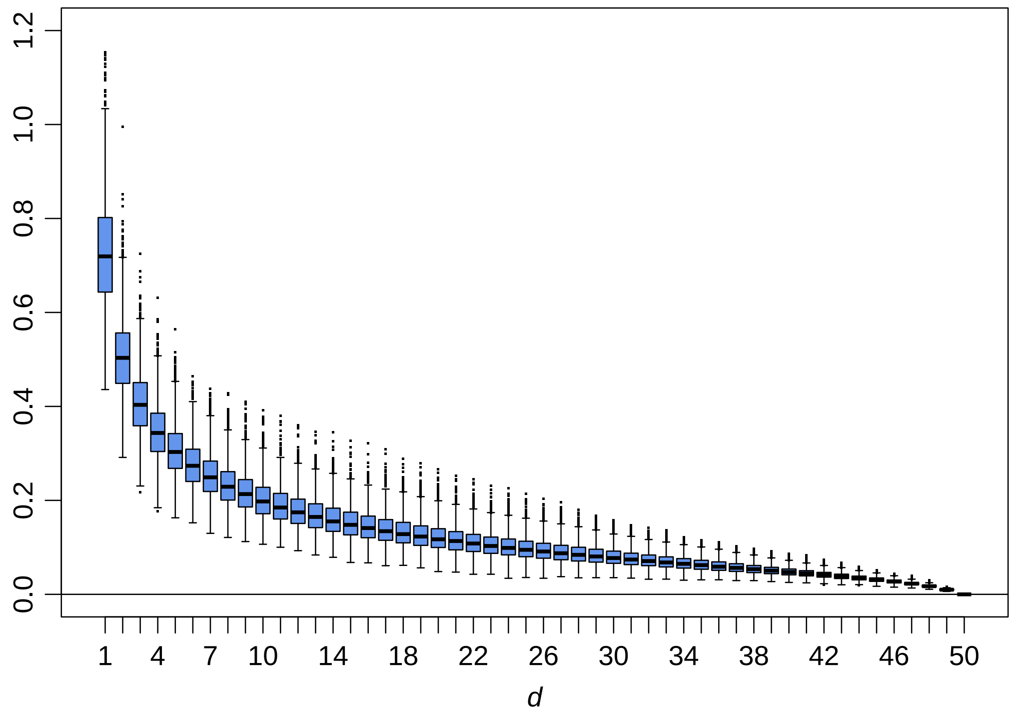

Figure 1: $(q, n) = (10, 100)$: Relative approximation errors $D(\hat{\boldsymbol{H}}_{n,d}, \hat{\boldsymbol{H}}_n)/D(\hat{\boldsymbol{H}}_n, \boldsymbol{H})$.

On the one hand, we look at the approximation errors, that is, the distances between $\hat{\boldsymbol{H}}_{n,d}, \hat{\boldsymbol{H}}_{n,d}^{\text{rand}}$ and the full estimator $\hat{\boldsymbol{H}}_n$. The distance between two matrices $\Sigma_1, \Sigma_2 \in \mathbb{R}^{q\times q}_{\text{sym},+}$ is measured by the so-called geodesic distance

$$D(\Sigma_1, \Sigma_2) \;:=\; \Big(\sum_{j=1}^{q} \log[\lambda_j(\Sigma_1^{-1}\Sigma_2)]^2\Big),$$

where $\lambda_1(\cdot) \geq \cdots \geq \lambda_q(\cdot)$ are the ordered real eigenvalues of a matrix $A \in \mathbb{R}^{q\times q}$. On the other hand, we look at the estimation errors, that is, the distances between the estimators $\hat{\boldsymbol{H}}_n$, $\hat{\boldsymbol{H}}_{n,d}$, $\hat{\boldsymbol{H}}_{n,d}^{\text{rand}}$ and the true shape matrix $\boldsymbol{H}$.

We simulated 2000 times a data set of size $n = 100$ in dimension $q = 10$, where each observation X_i had independent components with standard exponential distribution. The scatter functional was the M-functional with $\rho(s) = \rho_1(s) = (q + 1)\log(s + 1)$ for $s \geq 0$. In this particular example, $\boldsymbol{\Sigma}(P)$ is *not* a multiple of I_q, but the symmetrized distributions $Q = Q(P)$ yields $\boldsymbol{H} = I_q$. Figures 1 and 2 show box-and-whiskers plots of the resulting relative approximation errors

$$D(\hat{\boldsymbol{H}}_{n,d}, \hat{\boldsymbol{H}}_n)/D(\hat{\boldsymbol{H}}_n, \boldsymbol{H}) \quad \text{and} \quad D(\hat{\boldsymbol{H}}_{n,d}^{\text{rand}}, \hat{\boldsymbol{H}}_n)/D(\hat{\boldsymbol{H}}_n, \boldsymbol{H}),$$

respectively, for $1 \leq d \leq 49$. Figure 3 shows these ratios in one plot for $1 \leq d \leq 15$.

Figures 4, 5 and 6 are analogous, this time with the relative estimation errors

$$D(\hat{\boldsymbol{H}}_{n,d}, \boldsymbol{H})/D(\hat{\boldsymbol{H}}_n, \boldsymbol{H}) \quad \text{and} \quad D(\hat{\boldsymbol{H}}_{n,d}^{\text{rand}}, \boldsymbol{H})/D(\hat{\boldsymbol{H}}_n, \boldsymbol{H}).$$

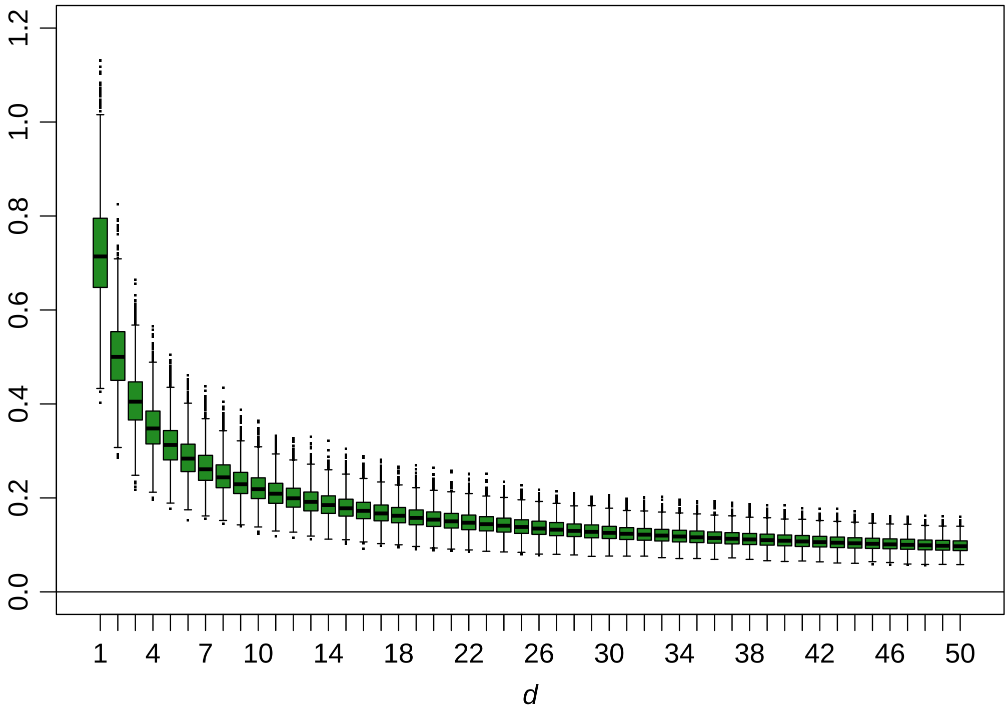

Figure 2: $(q, n) = (10, 100)$: Relative approximation errors $D(\hat{\boldsymbol{H}}_{n,d}^{\text{rand}}, \hat{\boldsymbol{H}}_n)/D(\hat{\boldsymbol{H}}_n, \boldsymbol{H})$.

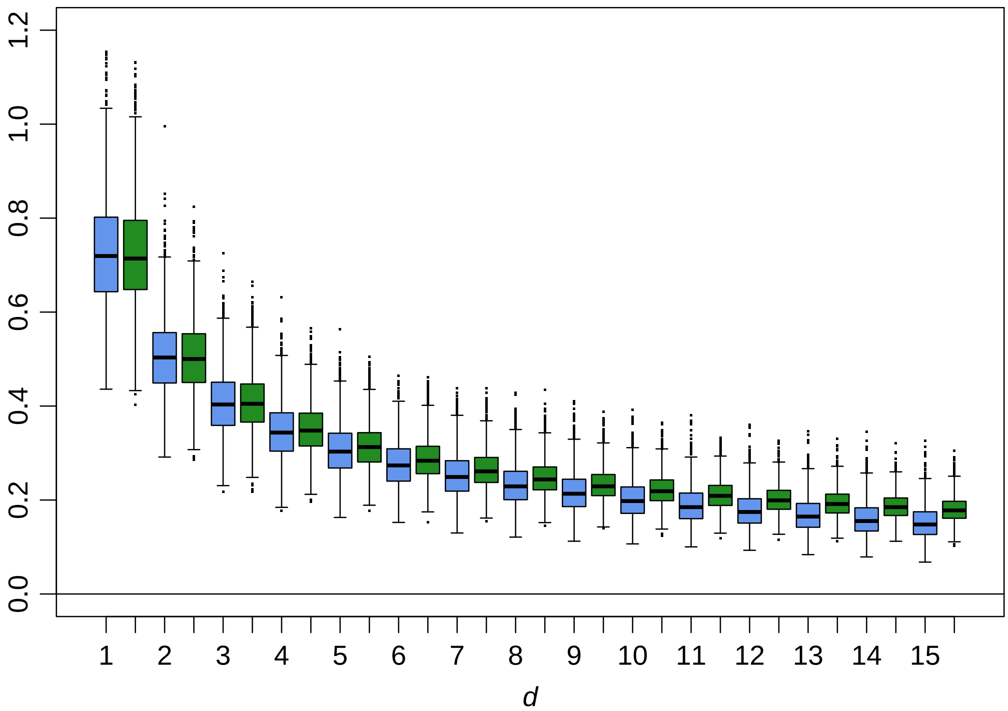

Figure 3: $(q, n) = (10, 100)$: Relative approximation errors $D(\hat{\boldsymbol{H}}_{n,d}, \hat{\boldsymbol{H}}_n)/D(\hat{\boldsymbol{H}}_n, \boldsymbol{H})$ (blue) and $D(\hat{\boldsymbol{H}}_{n,d}^{\text{rand}}, \hat{\boldsymbol{H}}_n)/D(\hat{\boldsymbol{H}}_n, \boldsymbol{H})$ (green).

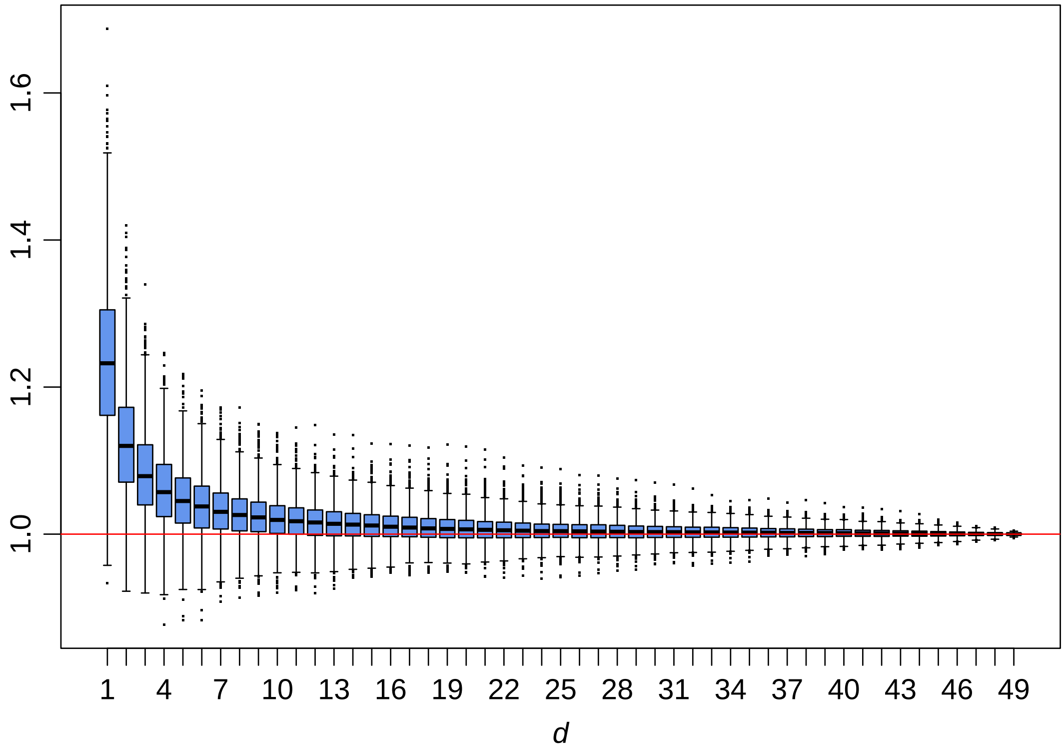

Figure 4: $(q, n) = (10, 100)$: Relative estimation errors $D(\hat{\boldsymbol{H}}_{n,d}, \boldsymbol{H})/D(\hat{\boldsymbol{H}}_n, \boldsymbol{H})$.

The median of the estimation error $D(\hat{\boldsymbol{H}}_n, \boldsymbol{H})$ in the simulations was equal to 1.1643. Interestingly, the relative estimation errors approach 1 more quickly than the relative approximation arrors approach 0. With respect to relative estimation error, a value $d = 10$, say, seems to be sufficient, although the approximation errors for this value are still substantial.

We did the same simulations and calculations for sample size $n = 400$ instead of $n = 100$. Figures 7 and 8 show the resulting relative approximation errors and relative estimation errors. This time, the median of $D(\hat{\boldsymbol{H}}_n, \boldsymbol{H})$ was only 0.5662. But note that the relative errors are similarly distributed for both sample sizes. The main difference seems to be that with increasing sample size the differences between $\hat{\boldsymbol{H}}_{n,d}$ and $\hat{\boldsymbol{H}}_{n,d}^{\mathrm{rand}}$ become smaller.

The simulation results are coherent with the asymptotic theory and confirm our claim that moderately large values of d yield already estimators with similar precision as the full symmetrized M-estimators. Therefore for larger sample sizes, computational costs are no longer a hindrance to apply symmetrized scatter matrices in practice.

5 Proofs

Proof of Proposition 1. Some of our arguments are similar to parts of Section 8.2 of Dümbgen et al. (2015), but for the reader's convenience, we present a complete and self-contained proof here.

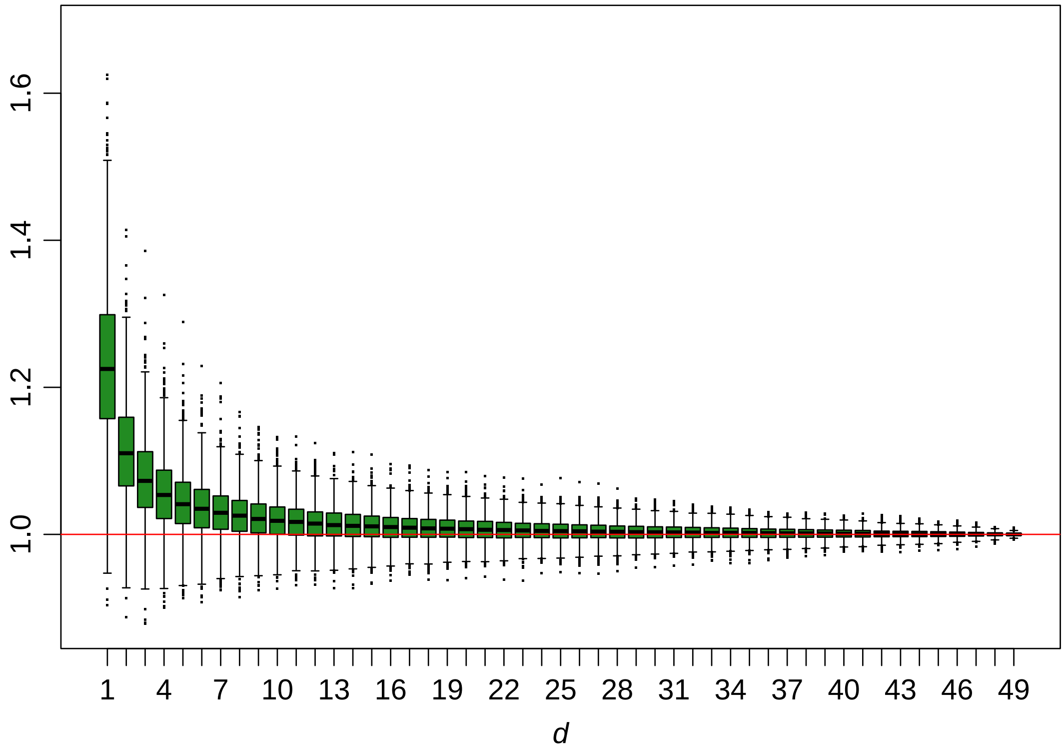

Figure 5: $(q, n) = (10, 100)$: Relative estimation errors $D(\hat{\boldsymbol{H}}^{\text{rand}}_{n,d}, \boldsymbol{H})/D(\hat{\boldsymbol{H}}_n, \boldsymbol{H})$.

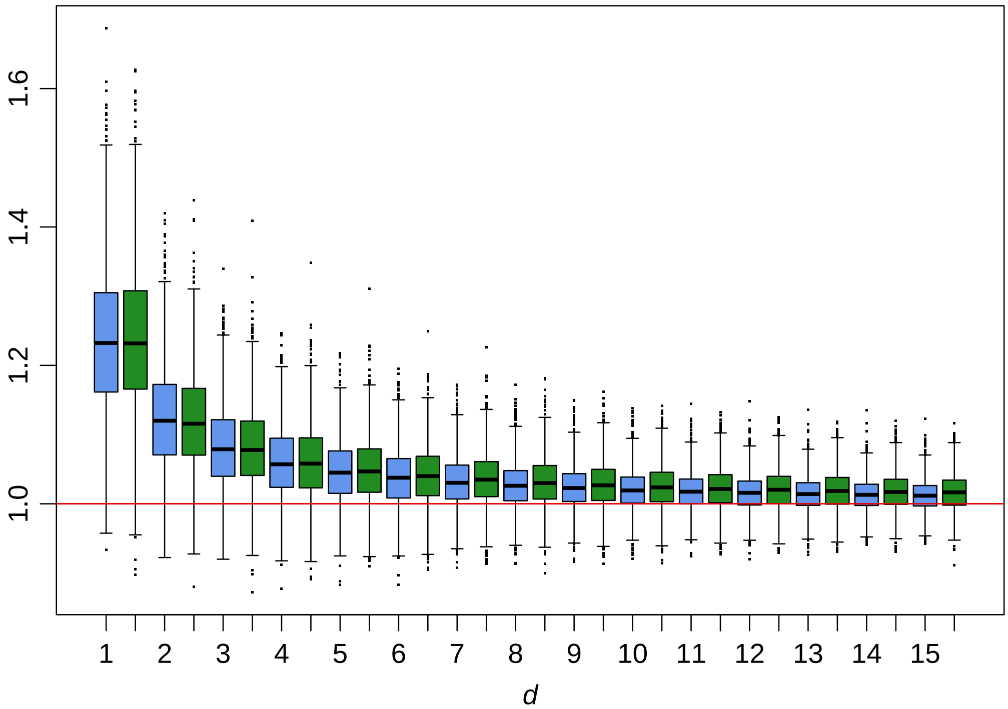

Figure 6: $(q, n) = (10, 100)$: Relative estimation errors $D(\hat{\boldsymbol{H}}_{n,d}, \boldsymbol{H})$ (blue) and $D(\hat{\boldsymbol{H}}^{\text{rand}}_{n,d}, \boldsymbol{H})$ (green).

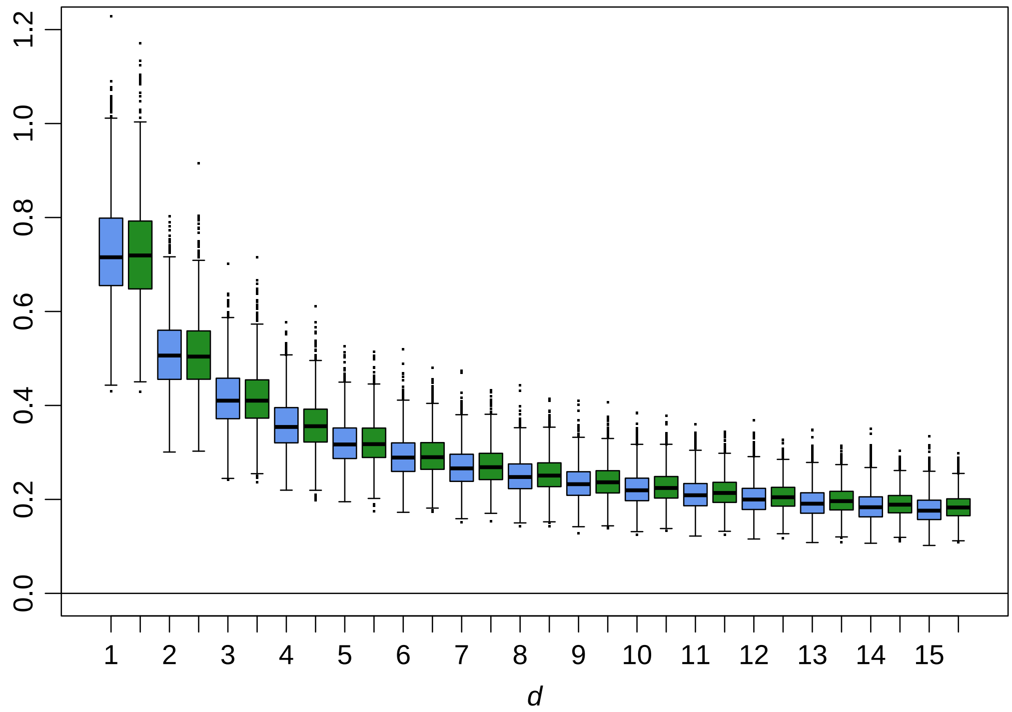

Figure 7: $(q, n) = (10, 400)$: Relative approximation errors $D(\hat{\boldsymbol{H}}_{n,d}, \hat{\boldsymbol{H}}_n)/D(\hat{\boldsymbol{H}}_n, \boldsymbol{H})$ (blue) and $D(\hat{\boldsymbol{H}}^{\text{rand}}_{n,d}, \hat{\boldsymbol{H}}_n)/D(\hat{\boldsymbol{H}}_n, \boldsymbol{H})$ (green).

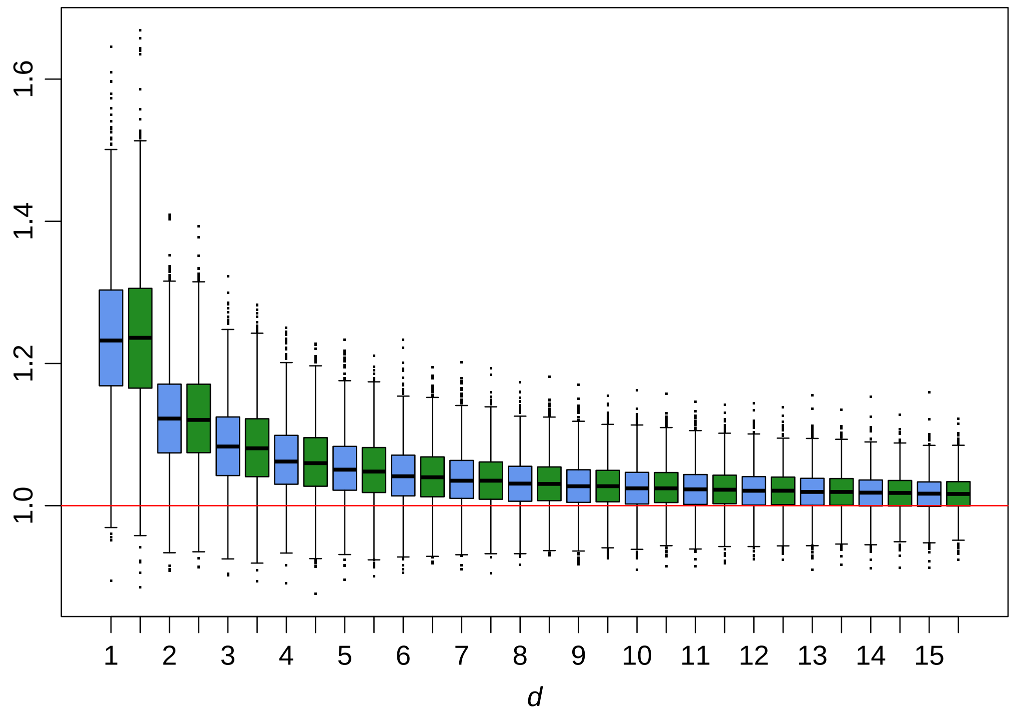

Figure 8: $(q, n) = (10, 400)$: Relative estimation errors $D(\hat{\boldsymbol{H}}_{n,d}, \boldsymbol{H})/D(\hat{\boldsymbol{H}}_n, \boldsymbol{H})$ (blue) and $D(\hat{\boldsymbol{H}}^{\text{rand}}_{n,d}, \boldsymbol{H})/D(\hat{\boldsymbol{H}}_n, \boldsymbol{H})$ (green).

Step 0. For arbitrary different indices $i, j \in \{1, \ldots, n\}$, the vector $X_j - X_i \neq 0$ almost surely, because $\mathbb{P}(X_j - X_i = 0) = \mathbb{E}\,\mathbb{P}(X_j \in \{X_i\} \,|\, X_i) = 0$. Hence, $\hat{Q}_n(\{0\}) = \hat{Q}_{n,d}(\{0\}) = 0$ for $1 \le d \le (n-1)/2$.

Step 1. Let $\mathcal{S}$ be the set of all index sets $\{i, j\}$, $1 \le i < j \le n$. Let $\mathcal{E}_o \subset \mathcal{S}$, and set $V_o := \bigcup_{E \in \mathcal{E}_o} E$. Suppose that the graph $(V_o, \mathcal{E}_o)$ is connected. That means, for arbitrary $\{i, j\} \in \mathcal{E}$, there exist $T \in \mathbb{N}$ and indices $i_0, i_1, \ldots, i_T$ in V_o such that $i_0 = i$, $i_T = j$, and $\{i_{t-1}, i_t\} \in \mathcal{E}_o$ for $1 \le t \le T$. Then for any index $i_o \in V_o$, the following three linear spaces are identical:

$$
\begin{aligned}
\mathbb{W}_1 &:= \operatorname{span}(X_i - X_{i_o} : i \in V_o), \\
\mathbb{W}_2 &:= \operatorname{span}(X_j - X_i : \{i, j\} \in \mathcal{E}_o), \\
\mathbb{W}_3 &:= \operatorname{span}(X_j - X_i : i, j \in V_o).
\end{aligned}
$$

The inclusions $\mathbb{W}_1, \mathbb{W}_2 \subset \mathbb{W}_3$ are obvious. On the other hand, for $i, j \in V_o$, the vector $X_j - X_i = (X_j - X_{i_o}) - (X_i - X_{i_o}) \in \mathbb{W}_1$, whence $\mathbb{W}_3 \subset \mathbb{W}_1$. Finally, by connectednes of $(V_o, \mathcal{E}_o)$, for arbitrary different indices $i, j \in V_o$, there exist $T \in \mathbb{N}$ and indices $i_0, i_1, \ldots, i_T \in V_o$ such that $i_0 = i$, $i_T = j$, and $\{i_{t-1}, i_t\} \in \mathcal{E}_o$ for $1 \le t \le T$. Hence, $X_j - X_i = \sum_{t=1}^T (X_{i_t} - X_{i_{t-1}}) \in \mathbb{W}_2$, and this shows that $\mathbb{W}_3 \subset \mathbb{W}_2$.

Step 2. Let $\mathcal{E}$ be an arbitrary subset of $\mathcal{S}$, and let $V := \bigcup_{E \in \mathcal{E}} E$. Let $V_1, \ldots, V_M$ be the $M \ge 1$ maximal connected components of the graph $(V, \mathcal{E})$. That means, $\mathcal{E} = \bigcup_{m=1}^M \mathcal{E}_m$ with sets $\mathcal{E}_m \subset \mathcal{S}$ such that the sets $V_m := \bigcup_{E \in \mathcal{E}_m} E$ are disjoint, and each subgraph $(V_m, \mathcal{E}_m)$ is connected. Then, the linear space

$$
\mathbb{W} \;:=\; \operatorname{span}(X_i - X_j : \{i, j\} \in \mathcal{E})
$$

has almost surely dimension

$$
\dim(\mathbb{W}) \;=\; \min(S, q) \quad \text{with} \quad S \;:=\; \sum_{m=1}^M (\#V_m - 1).
$$

To verify this, fix an arbitrary point $i_m \in V_m$ for $1 \le m \le M$. Then Step 1 shows that

$$
\mathbb{W} \;=\; \sum_{m=1}^M \operatorname{span}(X_j - X_{i_m} : j \in V_m \setminus \{i_m\}),
$$

and it suffices to show that in case of $S \le q$, the vectors $X_j - X_{i_m}$, $j \in V_m \setminus \{i_m\}$, $1 \le m \le M$, are almost surely linearly independent. But this can be shown by induction: Let $\bigl\{\{i_m, j\} : 1 \le m \le M, j \in V_m \setminus \{i_m\}\bigr\} = \{(k_1, \ell_1), \ldots, (k_S, \ell_S)\}$ with $k_1, \ldots, k_S \in \{i_1, \ldots, i_M\}$ and $\ell_s \in \bigcup_{m=1}^M V_m \setminus \{i_m\}$. Then, by Step 0, $X_{\ell_1} - X_{k_1} \neq 0$ almost surely, and for $1 \le s < S$ and $\mathbb{W}_s := \operatorname{span}(X_{\ell_r} - X_{k_r} : 1 \le r \le s)$,

$$
\begin{aligned}
&\mathbb{P}(X_{\ell_{s+1}} - X_{k_{s+1}} \notin \mathbb{W}_s) \\
&\quad = \; \mathbb{E}\,\mathbb{P}\bigl(X_{\ell_{s+1}} \notin X_{k_{s+1}} + \mathbb{W}_s \,\big|\, X_i : i \in \{i_1, \ldots, i_M\} \cup \{\ell_1, \ldots, \ell_s\}\bigr) \;=\; 0.
\end{aligned}
$$

Step 3. With $(V, \mathcal{E})$ and its subgraphs $(V_m, \mathcal{E}_m)$, $1 \le m \le M$, as in Step 2,

$$\#\mathcal{E} \ \le \ \sum_{m=1}^{M} \binom{\#V_m}{2} \ \le \ \binom{S+1}{2}.$$

The first inequality is a consequence of $\#\mathcal{E}_m \le \binom{\#V_m}{2}$ for $1 \le m \le M$. The second inequality follows from the fact that the mapping

$$\mathcal{E} \ni \{i,j\} \ \mapsto \ \begin{cases} \{i,j\} & \text{for } i,j \in V_m \setminus \{i_m\}, 1 \le m \le M, \\ \{0,j\} & \text{for } i = i_m, j \in V_m \setminus \{i_m\}, 1 \le m \le M, \end{cases}$$

is injective, and the images are subsets of $\{0\} \cup \bigcup_{m=1}^{M} V_m \setminus \{i_m\}$ with two elements.

For a fixed integer $d \ge 1$ with $d \le (n-1)/2$, let $\mathcal{S}_d$ be the subset of all $\{i,j\} \in \mathcal{S}$ such that $0 < j - i \le d$ or $j - i \ge n - d$. That means, for any $i \in \{1, \ldots, n\}$ there are exactly $2d$ indices $j \in \{1, \ldots, n\}$ such that $\{i,j\} \in \mathcal{S}_d$. Then

$$\#(\mathcal{E} \cap \mathcal{S}_d) \ \le \ Sd \quad \text{unless } M = 1 \text{ and } V = \{1, 2, \ldots, n\}.$$

To see this, note that in case of $M > 1$ or $V \ne \{1, 2, \ldots, n\}$, all sets V_m are different from $\{1, \ldots, n\}$. For a given $m \in \{1, \ldots, M\}$, let $k \in \{1, \ldots, n\} \setminus V_m$. To get an upper bound for $\#(\mathcal{E}_m \cap \mathcal{S}_d)$, we may assume without loss of generality that $k = n$. Otherwise, we could transform $\{1, 2, \ldots, n\}$ with the permutation $i \mapsto T(i) := 1_{[i \le k]}(i+n-k) + 1_{i>k}(i-k)$, because $\{i,j\} \in \mathcal{S}_d$ if and only if $\{T(i), T(j)\} \in \mathcal{S}_d$. Now, if $i_0 < i_1 < \cdots < i_{q_m} < n$ are the elements of V_m, then

$$\begin{aligned}
\#(\mathcal{E}_m \cap \mathcal{S}_d) &= \#\big\{\{i_a, i_b\} : 0 \le a < b \le q_m, i_b - i_a \le d \text{ or } i_b - i_a \ge n - d\big\} \\
&\le \#\big\{\{a,b\} : 0 \le a < b \le q_m, b - a \le d\big\} \\
&\quad + \ \#\big\{\{i,j\} : 1 \le i < j < n, j - i \ge n - d\big\} \\
&= \#\big\{\{a, a+c\} : 1 \le c \le d, 0 \le a \le q_m - c\big\} \\
&\quad + \ \#\big\{\{i,j\} : 1 \le i < d, n - d + i \le j < n\big\} \\
&\le \sum_{c=1}^{d} (q_m + 1 - c) + \sum_{i=1}^{d-1} (d - i) \\
&= q_m d - \sum_{c'=0}^{d-1} c' + \sum_{i'=1}^{d-1} i' \ = \ q_m d \ = \ (\#V_m - 1)d.
\end{aligned}$$

Step 4. Since there are only finitely many nonempty subsets $\mathcal{E}$ of $\mathcal{S}$, we may conclude from Step 2 that for *any* nonempty set $\mathcal{E} \subset \mathcal{S}$, the dimension of $\mathrm{span}(X_j - X_i : \{i,j\} \in \mathcal{E})$ is given by $S = S(\mathcal{E})$ as defined in Step 2. Now we consider an arbitrary linear subspace $\mathbb{W}$ of $\mathbb{R}^q$ with dimension $q' < q$ such that $\mathcal{E} = \mathcal{E}(\mathbb{W}) := \big\{\{i,j\} \in \mathcal{S} : X_j - X_i \in \mathbb{W}\big\}$ is nonempty. Then Step 3 implies that

$$\hat{Q}_n(\mathbb{W}) \ \le \ \binom{n}{2}^{-1} \binom{q'+1}{2} \quad \text{and} \quad \hat{Q}_{n,d}(\mathbb{W}) \ \le \ \frac{q'}{n}.$$

But

$$\binom{n}{2}^{-1}\binom{q'+1}{2} = \frac{q'}{q}\frac{q(q'+1)}{n(n-1)} \le \frac{q'}{q}\frac{q^2}{n(n-1)} \quad \text{and} \quad \frac{q'}{n} = \frac{q'}{q}\frac{q}{n}.$$

Both factors $q^2/(n(n-1))$ and q/n are strictly smaller than 1 if and only if $n > q$. This proves our claim about $\hat{Q}_n$ and $\hat{Q}_{n,d}$. $\square$

Some facts about complete and balanced incomplete U-statistics. Let us first recollect some well-known facts about U-statistics of order two (Serfling, 1980; Lee, 1990), with obvious adaptations to vector-valued kernels and the particular distributions $\hat{Q}_n$ and $\hat{Q}_{n,d}$. For some integer $r \ge 1$, let $f : \mathbb{R}^q \to \mathbb{R}^r$ be measurable such that $\mathbb{E}(\|f(X_1 - X_2)\|^2) < \infty$. With the symmetrized function $f^{\mathrm{s}}(x) := 2^{-1}\big(f(x) + f(-x)\big)$, define $f_0 := \mathbb{E}\, f(X_1 - X_2) = \mathbb{E}\, f^{\mathrm{s}}(X_1 - X_2)$ and

$$f_1(x) := \mathbb{E}\, f^{\mathrm{s}}(x - X_1) - f_0, \quad f_2(x, y) := f^{\mathrm{s}}(x - y) - f_0 - f_1(x) - f_1(y)$$

for $x, y \in \mathbb{R}^q$. Then the covariance matrices $\Gamma := \mathrm{Var}(f(X_1 - X_2))$, $\Gamma^{\mathrm{s}} := \mathrm{Var}(f^{\mathrm{s}}(X_1 - X_2))$, $\Gamma_1 := \mathrm{Var}(f_1(X_1))$ and $\Gamma_2 := \mathrm{Var}(f_2(X_1, X_2))$ satisfy the (in)equalities

$$\Gamma \ge \Gamma^{\mathrm{s}} = 2\Gamma_1 + \Gamma_2.$$

Here and subsequently, inequalities between symmetric matrices refer to the Loewner partial order on $\mathbb{R}^{q\times q}_{\mathrm{sym}}$. The random vectors $f_1(X_i)$, $1 \le i \le n$, and $f_2(X_i, X_j)$, $1 \le i < j \le n$, are centered and uncorrelated, and

$$\begin{aligned} U_n &:= \int_{\mathbb{R}^q} f\, d\hat{Q}_n = f_0 + 2\int_{\mathbb{R}^q} f_1\, d\hat{P}_n + M_n, \\ U_{n,d} &:= \int_{\mathbb{R}^q} f\, d\hat{Q}_{n,d} = f_0 + 2\int_{\mathbb{R}^q} f_1\, d\hat{P}_n + M_{n,d}, \end{aligned}$$

where

$$M_n := \binom{n}{2}^{-1} \sum_{1\le i<j\le n} f_2(X_i, X_j), \quad M_{n,d} := (nd)^{-1} \sum_{i=1}^{n} \sum_{j=i+1}^{i+d} f_2(X_i, X_j).$$

Moreover, $\mathbb{E}(U_n) = \mathbb{E}(U_{n,d}) = f_0$, and

$$\left.\begin{aligned} n\,\mathrm{Var}(U_n) &= 4\Gamma_1 + n\,\mathrm{Var}(M_n) &&= 4\Gamma_1 + 2(n-1)^{-1}\Gamma_2 \\ n\,\mathrm{Var}(U_{n,d}) &= 4\Gamma_1 + n\,\mathrm{Var}(M_{n,d}) &&= 4\Gamma_1 + d^{-1}\Gamma_2 \end{aligned}\right\} \le 2\Gamma. \tag{6}$$

The final ingredient for the proof of Theorem 2 is a result about the asymptotic joint distribution of $\int_{\mathbb{R}^q} f_1\, d\hat{P}_n$ and $M_{n,d}$.

Proposition 4. *For any fixed $d \ge 1$, the random pair $\big(\sqrt{n}\int_{\mathbb{R}^q} f_1\, d\hat{P}_n, \sqrt{nd}\, M_{n,d}\big)$ converges in distribution to $\mathcal{N}_r(0, \Gamma_1) \otimes \mathcal{N}_r(0, \Gamma_2)$.*

Proof of Proposition 4. The proof of this result uses a standard trick for sequences of m-dependent random variables, in our case, $m = d + 1$. For a fixed number $k \geq d$, let

$$
\begin{aligned}
S_n &:= n^{-1/2} \sum_{i=1}^{n} f_1(X_i) \ = \ \sqrt{n} \int_{\mathbb{R}^q} f_1 \, d\hat{P}_n, \\
S_n^k &:= n^{-1/2} \sum_{\ell=1}^{\lfloor n/k \rfloor} Y_\ell^k \quad \text{with} \quad Y_\ell^k := \sum_{i=\ell k - k + 1}^{\ell k} f_1(X_i), \\
T_n &:= (nd)^{-1/2} \sum_{i=1}^{n} \sum_{j=i+1}^{i+d} f_2(X_i, X_j) \ = \ \sqrt{nd} M_{n,d}, \\
T_n^k &:= (nd)^{-1/2} \sum_{\ell=1}^{\lfloor n/k \rfloor} Z_\ell^k \quad \text{with} \quad Z_\ell^k := \sum_{i=\ell k - k + 1}^{\ell k} \ \sum_{j=i+1}^{\min(i+d, \ell k)} f_2(X_i, X_j).
\end{aligned}
$$

The random pairs (Y_ℓ^k, Z_ℓ^k), $\ell \geq 1$, are independent and identically distributed with $\mathbb{E}(Y_\ell^k) = \mathbb{E}(Z_\ell^k) = 0$ and

$$
\mathrm{Var}(Y_\ell^k) \ = \ k\,\Gamma_1, \quad \mathrm{Var}(Z_\ell^k) \ = \ \Big(k - \frac{d-1}{2}\Big) d\,\Gamma_2, \quad \mathrm{Cov}(Y_\ell^k, Z_\ell^k) \ = \ 0.
$$

Consequently, it follows from the multivariate central limit theorem and Slutzky's lemma that

$$
(S_n^k, T_n^k) \ \to_{\mathcal{L}} \ \mathcal{N}_r(0, \Gamma_1) \otimes \mathcal{N}_r\Big(0, \Big(1 - \frac{d-1}{2k}\Big)\Gamma_2\Big),
$$

and the distribution on the right hand side converges weakly to $\mathcal{N}_r(0, \Gamma_1) \otimes \mathcal{N}_r(0, \Gamma_2)$ as $k \to \infty$. Moreover,

$$
\begin{aligned}
\mathbb{E}(\|S_n - S_n^k\|^2) &\leq \frac{k-1}{n} \, \mathrm{trace}(\Gamma_1) \ \to \ 0, \\
\mathbb{E}(\|T_n - T_n^k\|^2) &\leq \Big(\frac{d-1}{2k} + \frac{k-1}{nd}\Big) \mathrm{trace}(\Gamma_2) \ \to \ \frac{d-1}{2k} \, \mathrm{trace}(\Gamma_2),
\end{aligned}
$$

and the right hand side converges to 0 as $k \to \infty$. This implies that (S_n, T_n) converges in distribution to $\mathcal{N}_r(0, \Gamma_1) \otimes \mathcal{N}_r(0, \Gamma_2)$. □

Proof of Theorem 2. Let $\check{Q}_n$ stand for $\hat{Q}_n$, $\hat{Q}_{n,d(n)}$ with $(n-1)/2 \geq d(n) \to \infty$, or $\hat{Q}_{n,d}$ with fixed $d \geq 1$. For any bounded, continuous function $f : \mathbb{R}^q \to \mathbb{R}$,

$$
\mathbb{E} \Big| \int_{\mathbb{R}^q} f \, d\check{Q}_n - \int_{\mathbb{R}^q} f \, dQ \Big| \ \leq \ (2/n)^{1/2} \|f\|_\infty.
$$

This follows from inequality (6) applied to real-valued functions. This implies that $d_{\mathcal{L}}(\check{Q}_n, Q) \to_p 0$. In particular,

$$
\mathbf{\Sigma}(\check{Q}_n) \ = \ \mathbf{\Sigma}(Q) + \int_{\mathbb{R}^q} J \, d\check{Q}_n + o\Big(\Big\| \int_{\mathbb{R}^q} J \, d\check{Q}_n \Big\|\Big).
$$

We may identify $\mathbb{R}^{q \times q}_{\mathrm{sym}}$ with $\mathbb{R}^r$, where $r = q(q+1)/2$. Then $\int_{\mathbb{R}^q} J \, d\check{Q}_n$ is a (complete or balanced incomplete) U-statistic with vector-valued kernel function, and it follows from boundedness of $J(\cdot)$ with $\int_{\mathbb{R}^q} J\,dQ = 0$ and the general considerations about U-statistics that $\int_{\mathbb{R}^q} J\,d\check{Q}_n = O_p(n^{-1/2})$. Consequently,

$$\mathbf{\Sigma}(\check{Q}_n) \;=\; \mathbf{\Sigma}(Q) + \int_{\mathbb{R}^q} J\,d\check{Q}_n + o_p(n^{-1/2}),$$

so we may replace $\mathbf{\Sigma}(\check{Q}_n) - \mathbf{\Sigma}(Q)$ with the matrix-valued U-statistic $\int_{\mathbb{R}^q} J\,d\check{Q}_n$. But then the assertions of Theorem 2 are direct consequences of the general considerations about U-statistics and Proposition 4. $\square$

For the proof of Theorem 3 we need a variation of Proposition 4 for the random vectors

$$\tilde{M}_{n,1} \;:=\; n^{-1} \sum_{i=1}^{n} f_2(X_{\Pi(i)}, X_{\Pi(i+1)}),$$

where Π is uniformly distributed on the set of all permutations of $\{1, 2, \ldots, n\}$, independent from $(X_i)_{i=1}^n$, and $\Pi(n+1) := \Pi(1)$.

Proposition 5. *Let $d_{\mathcal{L}}(\cdot,\cdot)$ be a metric on the space of probability distributions on $\mathbb{R}^r$ which metrizes weak convergence. Then,*

$$d_{\mathcal{L}}\Big(\mathcal{L}\big(\sqrt{n}\,\tilde{M}_{n,1} \,\big|\, (X_i)_{i=1}^n\big), \mathcal{N}_r(0, \Gamma_2)\Big) \;\to_p\; 0.$$

Proof of Proposition 5. By means of the Cramér–Wold device, it suffices to consider the case $r = 1$. Then the random variable $\sqrt{n}\,\tilde{M}_{n,1}$ can be written as $\sum_{i=1}^n A_{\Pi(i),\Pi(i+1)}$ with the random matrix

$$A = A(X_1, \ldots, X_n) \;:=\; \big(n^{-1/2} 1_{[i \neq j]} f_2(X_i, X_j)\big)_{i,j=1}^n \in \mathbb{R}^{n \times n}_{\mathrm{sym}}.$$

As explained in the appendix, there exist permutations $B = B(\cdot \,|\, \Pi)$ and $B^* = B^*(\cdot \,|\, \Pi)$ such that

$$\sum_{i=1}^{n} A_{\Pi(i),\Pi(i+1)} \;=\; \sum_{i=1}^{n} A_{i,B^*(i)},$$

while B is uniformly distributed on the set of all permutations of $\{1, \ldots, n\}$, and

$$\mathbb{E}\big(\#\big\{i \in \{1, \ldots, n\} : B(i) \neq B^*(i)\big\}\big) \;\leq\; 1 + \log(n).$$

Consequently,

$$\sqrt{n}\tilde{M}_{n,1} \;=\; \sum_{i=1}^{n} A_{i,B(i)} + R_n,$$

where $R_n := \sum_{i=1}^n (A_{i,B^*(i)} - A_{i,B(i)})$ satisfies

$$\mathbb{E}\,|R_n| \leq\; 2(1 + \log(n))n^{-1/2}\,\mathbb{E}\,|f_2(X_1, X_2)| \;\to\; 0.$$

Hence, it suffices to show that the conditional distribution of $\sum_{i=1}^n A_{i,B(i)}$, given $(X_i)_{i=1}^n$, converges weakly in probability to $\mathcal{N}(0, \Gamma_2)$. Distributions of this type have been investigated by Hoeffding (1951). It follows from Hoeffding's results and elementary inequalities presented in

Section A.2 that it suffices to verify the following three properties of the random symmetric matrices $A = A(X_1, \ldots, X_n)$:

$$\mathbb{E}\Big|n^{-1} \sum_{i,j=1}^{n} A_{i,j}^2 - \Gamma_2\Big| \;\to\; 0, \tag{7}$$

$$\mathbb{E}\Big(\sum_{i=1}^{n} \bar{A}_i^2\Big) \;\to\; 0, \tag{8}$$

$$\mathbb{E}\Big(n^{-1} \sum_{i,j=1}^{n} A_{i,j}^2 \min(|A_{i,j}|, 1)\Big) \;\to\; 0, \tag{9}$$

where $\bar{A}_i := n^{-1} \sum_{j=1}^{n} A_{i,j}$.

For an arbitrary threshold $c > 0$, we split $f_2(x,y)^2$ into the bounded function $g_c(x,y) := f_2(x,y)^2 1_{[f_2(x,y)^2 \le c]}$ and the remainder $h_c(x,y) := f_2(x,y)^2 1_{[f_2(x,y)^2 > c]}$. Then the left-hand side of (7) equals

$$\begin{aligned}
&\mathbb{E}\Big|n^{-2} \sum_{i,j=1}^{n} 1_{[i \ne j]} f_2(X_i, X_j)^2 - \Gamma_2\Big| \\
&\le\; n^{-1}\Gamma_2 + 2\,\mathbb{E}\,h_c(X_1, X_2) + n^{-2}\,\mathbb{E}\Big|\sum_{i,j=1}^{n} \big(g_c(X_i, X_j) - \mathbb{E}\,g_c(X_1, X_2)\big)\Big| \\
&\le\; n^{-1}\Gamma_2 + 2\,\mathbb{E}\,h_c(X_1, X_2) + n^{-2}\,\mathrm{Var}\Big(\sum_{i,j=1}^{n} g_c(X_i, X_j)\Big)^{1/2} \\
&\le\; n^{-1}\Gamma_2 + 2\,\mathbb{E}\,h_c(X_1, X_2) + cn^{-1/2} \\
&\to\; 2\,\mathbb{E}\,h_c(X_1, X_2).
\end{aligned}$$

The last inequality follows from the facts that

$$\mathrm{Cov}\big(g_c(X_i, X_j), g_c(X_k, X_\ell)\big) \begin{cases} = \; 0 & \text{if } \{i,j\} \cap \{k,\ell\} = \emptyset, \\ \le \; c^2/4 & \text{else,} \end{cases}$$

and that the number of quadruples (i,j,k,ℓ) with $\{i,j\} \cap \{k,\ell\} \ne \emptyset$ is smaller than $4n^3$. Since by dominated convergence, $\mathbb{E}\,h_c(X_1, X_2) \to 0$ as $c \to \infty$, Condition (7) is satisfied.

The left-hand side of (8) equals $n\,\mathbb{E}(\bar{A}_1^2)$, and $\bar{A}_1$ is the sum of the uncorrelated, centered random variables $f_2(X_1, X_j)$, $2 \le j \le n$, times $n^{-3/2}$. Consequently,

$$n\,\mathbb{E}(\bar{A}_1^2) \;\le\; n^{-1}\,\mathbb{E}\big(f_2(X_1, X_2)^2\big) \;\to\; 0.$$

Finally, the left-hand side of (9) is not larger than

$$\mathbb{E}\big(f_2(X_1, X_2)^2 \min\{n^{-1/2}|f_2(X_1, X_2)|, 1\}\big) \;\to\; 0$$

by dominated convergence. $\square$

Proof of Theorem 3. Since $(\hat{P}_n, \hat{Q}_{n,1}^{(\ell)}, M_{n,1}^{(\ell)})$ has the same distribution as $(\hat{P}_n, \hat{Q}_{n,1}, M_{n,1})$, the first assertion is a direct consequence of Theorem 2 with $d = 1$. The second part is a consequence of the central limit theorem, applied to $\int_{\mathbb{R}^q} H_1 \, d\hat{P}_n$, and Proposition 5, applied to $\sqrt{n}\, M_{n,1}^{(\ell)}$. The final statement is a consequence of the first and second part and the continuous mapping theorem. $\square$

A Auxiliary results

A.1 A particular coupling of random permutations

Preparations. For an integer $n \geq 1$, let $\mathcal{S}_n$ be the set of all permutations of $\langle n \rangle := \{1, 2, \ldots, n\}$. A cycle in $\mathcal{S}_n$ is a permutation $\sigma \in \mathcal{S}_n$ such that for $m \geq 1$ pairwise different points $a_1, \ldots, a_m \in \langle n \rangle$,

$$a_1 \ \mapsto \ a_2 \mapsto \ \cdots \ \mapsto \ a_m \ \mapsto \ a_1,$$

while $\sigma(i) = i$ for $i \in \langle n \rangle \setminus \{a_1, \ldots, a_m\}$. (In case of $m = 1$, $\sigma(i) = i$ for all $i \in \langle n \rangle$.) We write

$$\sigma \ = \ (a_1, \ldots, a_m)_{\mathrm{c}}$$

for this mapping and note that it has m equivalent representations

$$\sigma \ = \ (a_1, \ldots, a_m)_{\mathrm{c}} \ = \ (a_2, \ldots, a_m, a_1)_{\mathrm{c}} \ = \ \cdots \ = \ (a_m, a_1, \ldots, a_{m-1})_{\mathrm{c}}.$$

Any permutation $\sigma \in \mathcal{S}_n$ can be written as

$$\sigma \ = \ (a_{11}, \ldots, a_{1m(1)})_{\mathrm{c}} \circ \cdots \circ (a_{k1}, \ldots, a_{km(k)})_{\mathrm{c}},$$

where the sets $\{a_{j1}, \ldots, a_{jm(j)}\}$, $1 \leq j \leq k$, form a partition of $\langle n \rangle$. Note that the cycles $(a_{j1}, \ldots, a_{jm(j)})_{\mathrm{c}}$, $1 \leq j \leq m$, commute. This representation of σ as a combination of cycles is unique if we require, for instance, that

$$a_{jm(j)} \ = \ \min\{a_{j1}, \ldots, a_{jm(j)}\} \quad \text{for } 1 \leq j \leq k$$

and

$$a_{1m(1)} \ < \ \cdots \ < \ a_{km(k)}.$$

In what follows, let $\mathcal{S}_n^*$ be the set of all permutations $\sigma \in \mathcal{S}_n$ consisting of just one cycle, i.e.

$$\sigma \ = \ (a_1, a_2, \ldots, a_n)_{\mathrm{c}}$$

with pairwise different numbers $a_1, a_2, \ldots, a_n \in \langle n \rangle$.

The coupling. The standardized cycle representation of $\sigma \in \mathcal{S}_n$ gives rise to a particular mapping $\mathcal{S}_n \ni \pi \mapsto (\sigma, \sigma^*) \in \mathcal{S}_n \times \mathcal{S}_n^*$ such that $\pi \mapsto \sigma$ is bijective. For fixed $\pi \in \mathcal{S}_n$ and any index $i \in \langle n \rangle$ let

$$M_i \ := \ \langle n \rangle \setminus \{\pi(s) : 1 \le s < i\},$$

i.e. $\langle n \rangle = M_1 \supset M_2 \supset \cdots \supset M_n = \{\pi(n)\}$, and $\#M(i) = n + 1 - i$. Let $1 \le t_1 < t_2 < \cdots < t_k = n$ be those indices i such that $\pi(i) = \min(M_i)$. Then

$$\sigma \ := \ \bigl(\pi(1), \ldots, \pi(t_1)\bigr)_{\rm c} \circ \bigl(\pi(t_1 + 1), \ldots, \pi(t_2)\bigr)_{\rm c} \circ \cdots \circ \bigl(\pi(t_{k-1} + 1), \ldots, \pi(t_k)\bigr)_{\rm c}$$

defines a permutation of $\langle n \rangle$ with standardized cycle representation. This is essentially the construction used by Feller (1945) to investigate the number of cycles of a random permuation. Moreover,

$$\sigma^* \ := \ \bigl(\pi(1), \pi(2), \ldots, \pi(n)\bigr)_{\rm c}$$

defines a permutation in $\mathcal{S}_n^*$ such that

$$\bigl\{i \in \langle n \rangle : \sigma(i) \ne \sigma^*(i)\bigr\} \ = \ \begin{cases} \emptyset & \text{if } k = 1, \\ \{t_1, \ldots, t_k\} & \text{if } k \ge 2. \end{cases}$$

Suppose that π is a random permutation with uniform distribution on $\mathcal{S}_n$. Then σ is a random permutation with uniform distribution on $\mathcal{S}_n$ too, because $\pi \mapsto \sigma$ is a bijection. Since the conditional distribution of $\pi(i)$, given $(\pi(s))_{1 \le s < i}$, is the uniform distribution on M_i, the random variables

$$Y_i \ := \ 1_{[\pi(i) = \min(M_i)]}, \quad i \in \langle n \rangle,$$

are stochastically independent Bernoulli random variables with $\mathbb{P}(Y_i = 1) = (n + 1 - i)^{-1} = 1 - \mathbb{P}(Y_i = 0)$. Consequently,

$$\mathbb{E}\bigl(\#\bigl\{i \in \langle n \rangle : \sigma(i) \ne \sigma^*(i)\bigr\}\bigr) \ \le \ \sum_{i=1}^n (n + 1 - i)^{-1} \ = \ 1 + \sum_{j=2}^n j^{-1} \ \le \ 1 + \log(n),$$

because $j^{-1} \le \int_{j-1}^j x^{-1}\, dx = \log(j) - \log(j - 1)$ for $2 \le j \le n$.

A.2 Some inequalities related to Lindeberg type conditions

In connection with Gaussian approximations and Stein's method, see Stein (1986) or Barbour and Chen (2005), the quantity

$$L(X) \ := \ \mathbb{E}\bigl(X^2 \min(|X|, 1)\bigr)$$

for a square-integrable random variable X plays an important role. Elementary considerations show that

$$h(x) \ \le \ x^2 \min(|x|, 1) \ \le \ \sqrt{2}\, h(x) \quad \text{with} \quad h(x) \ := \ \frac{|x|^3}{\sqrt{1 + x^2}}$$

for arbitrary $x \in \mathbb{R}$. Moreover, $h : \mathbb{R} \to [0, \infty)$ is an even, convex function such that $h(2x) \le 8h(x)$. Consequently, for arbitrary $x, y \in \mathbb{R}$, Jensen's inequality implies that

$$\begin{aligned} (x+y)^2 \min(|x+y|, 1) \ &\le \ \sqrt{2}\ \mathbb{E}\, h(x+y) \\ &\le \ 2^{-1/2}\bigl(h(2x) + h(2y)\bigr) \\ &\le \ \sqrt{32}\ \mathbb{E}\, h(x) + \sqrt{32}\ \mathbb{E}\, h(y) \\ &\le \ 6x^2 \min(|x|, 1) + 6y^2 \min(|y|, 1) \ \le \ 6x^2 \min(|x|, 1) + 6y^2. \end{aligned}$$

For a symmetric matrix $A \in \mathbb{R}^{n \times n}$, we define its row means $\bar{A}_i := n^{-1} \sum_{j=1}^n A_{ij}$ and its overall mean $\bar{A} := n^{-2} \sum_{i,j=1}^n A_{ij}$. Let $\tilde{A} := (A_{ij} - \bar{A}_i - \bar{A}_j + \bar{A})_{i,j=1}^n$. Then elementary calculations and the previous inequalities reveal that

$$0 \ \le \ n^{-1} \sum_{i,j=1}^n A_{ij}^2 - n^{-1} \sum_{i,j=1}^n \tilde{A}_{ij}^2 \ \le \ 2 \sum_{i=1}^n \bar{A}_i^2$$

and

$$n^{-1} \sum_{i,j=1}^n \tilde{A}_{ij}^2 \min(|\tilde{A}_{ij}|, 1) \ \le \ 6n^{-1} \sum_{i,j=1}^n A_{ij}^2 \min(|A_{ij}|, 1) + 12 \sum_{i=1}^n \bar{A}_i^2.$$

Acknowledgement. We thank Sara Taskinen for stimulating discussions. Constructive comments of three referees are gratefully acknowledged.

References

BARBOUR, A. D. and CHEN, L. H. Y. (eds.) (2005). *An introduction to Stein's method*, vol. 4 of *Lecture Notes Series. Institute for Mathematical Sciences. National University of Singapore*. Singapore University Press, Singapore; World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ. Lectures from the Meeting on Stein's Method and Applications: a Program in Honor of Charles Stein held at the National University of Singapore, Singapore, July 28–August 31, 2003.

BLOM, G. (1976). Some properties of incomplete U-statistics. *Biometrika* **63** 573–580.

BROWN, B. M. and KILDEA, D. G. (1978). Reduced U-statistics and the Hodges-Lehmann estimator. *Ann. Statist.* **6** 828–835.

DUDLEY, R. M., SIDENKO, S. and WANG, Z. (2009). Differentiability of t-functionals of location and scatter. *Ann. Statist.* **37** 939–960.

DÜMBGEN, L. (1998). On Tyler's M-functional of scatter in high dimension. *Ann. Inst. Statist. Math.* **50** 471–491.

DÜMBGEN, L., NORDHAUSEN, K. and SCHUHMACHER, H. (2014). *fastM: Fast Computation of Multivariate M-estimators*. R package.

DÜMBGEN, L., NORDHAUSEN, K. and SCHUHMACHER, H. (2016). New algorithms for M-estimation of multivariate scatter and location. *J. Multivar. Analysis* **144** 200–217.

DÜMBGEN, L., PAULY, M. and SCHWEIZER, T. (2015). M-functionals of multivariate scatter. *Stat. Surv.* **9** 32–105.

FELLER, W. (1945). The fundamental limit theorems in probability. *Bull. Amer. Math. Soc.* **51** 800–832.

HOEFFDING, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.* **19** 293–325.

HOEFFDING, W. (1951). A combinatorial Central Limit Theorem. *Ann. Math. Statist.* **22** 558–566.

KENT, J. T. and TYLER, D. E. (1991). Redescending M-estimates of multivariate location and scatter. *Ann. Statist.* **19** 2102–2119.

LEE, A. J. (1990). *U-Statistics – Theory and Practice*, vol. 110 of *Statistics: Textbooks and Monographs*. Marcel Dekker, Inc., New York.

MIETTINEN, J., NORDHAUSEN, K., TASKINEN, S. and TYLER, D. E. (2016). On the computation of symmetrized M-estimators of scatter. In *Recent Advances in Robust Statistics: Theory and Applications* (C. Agostinelli, A. Basu, P. Filzmoser and D. Mukherjee, eds.). Springer, India.

MUANDET, K., FUKUMIZU, K., SRIPERUMBUDUR, B. and SCHÖLKOPF, B. (2017). Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning* **10** 1–141.

NORDHAUSEN, K., OJA, H. and OLLILA, E. (2008). Robust independent component analysis based on two scatter matrices. *Austrian J. Statist.* **37** 91–100.

NORDHAUSEN, K. and TYLER, D. E. (2015). A cautionary note on robust covariance plug-in methods. *Biometrika* **102** 573–588.

PAINDAVEINE, D. (2008). A canonical definition of shape. *Statist. Probab. Lett.* **78** 2240–2247.

SERFLING, R. J. (1980). *Approximation theorems of mathematical statistics*. John Wiley & Sons, Inc., New York. Wiley Series in Probability and Mathematical Statistics.

SIRKIÄ, S., TASKINEN, S. and OJA, H. (2007). Symmetrised M-estimators of multivariate scatter. *J. Multivar. Anal.* **98** 1611–1629.

STEIN, C. (1986). *Approximate computation of expectations*, vol. 7 of *Institute of Mathematical Statistics Lecture Notes—Monograph Series*. Institute of Mathematical Statistics, Hayward, CA.

TYLER, D. E. (1987). A distribution-free M-estimator of multivariate scatter. *Ann. Statist.* **15** 234–251.

TYLER, D. E., CRITCHLEY, F., DÜMBGEN, L. and OJA, H. (2009). Invariant coordinate selection (with discussion). *J. Royal Statist. Soc. B* **71** 549–592.