Proc. Int. Cong. of Math. – 2018
Rio de Janeiro, Vol. 3 (2891–2912)

ASYMPTOTIC EFFICIENCY IN HIGH-DIMENSIONAL COVARIANCE ESTIMATION

Vladimir Koltchinskii

Abstract

We discuss recent results on asymptotically efficient estimation of smooth functionals of covariance operator Σ of a mean zero Gaussian random vector X in a separable Hilbert space based on n i.i.d. observations of this vector. We are interested in functionals that are of importance in high-dimensional statistics such as linear forms of eigenvectors of Σ (principal components) as well as in more general functionals of the form $\langle f(\Sigma), B\rangle$, where $f : \mathbb{R} \mapsto \mathbb{R}$ is a sufficiently smooth function and B is an operator with nuclear norm bounded by a constant. In the case when X takes values in a finite-dimensional space of dimension $d \le n^{\alpha}$ for some $\alpha \in (0,1)$ and f belongs to Besov space $B^{s}_{\infty,1}(\mathbb{R})$ for $s > \frac{1}{1-\alpha}$, we develop asymptotically normal estimators of $\langle f(\Sigma), B\rangle$ with $\sqrt{n}$ convergence rate and prove asymptotic minimax lower bounds showing their asymptotic efficiency.

1 Introduction

Let $X_1, \dots, X_n$ be i.i.d. random variables sampled from unknown distribution $P_\theta, \theta \in \Theta$. Assume that the parameter space Θ is a subset of a linear normed space and the goal is to estimate $f(\theta)$ for a smooth functional $f : \Theta \mapsto \mathbb{R}$ based on observations $X_1, \dots, X_n$. Let $\mathfrak{L}$ be the set of loss functions $\ell : \mathbb{R} \mapsto \mathbb{R}_+$ such that $\ell(0) = 0$, $\ell(-t) = \ell(t), t \in \mathbb{R}$, ℓ is convex and increasing on $\mathbb{R}_+$ and for some $c > 0$, $\ell(t) = O(e^{c|t|})$ as $t \to \infty$. Let Z be a standard normal random variable.

Definition 1. *An estimator $T_n = T_n(X_1, \dots, X_n)$ will be called asymptotically efficient with respect to $\Theta_n \subset \Theta, n \ge 1$ with convergence rate $\sqrt{n}$ and (limit) variance $\sigma_f^2(\theta) > 0$*

The author was supported in part by NSF grants DMS-1509739 and CCF-1523768.
MSC2010: primary 62H12; secondary 62G20, 62H25, 60B20.
Keywords: asymptotic efficiency, sample covariance, bootstrap, effective rank, concentration inequalities, normal approximation.

iff the following properties hold:

$$(1)\qquad \sup_{\theta\in\Theta_n}\sup_{x\in\mathbb{R}}\left|\mathbb{P}_\theta\left\{\frac{n^{1/2}(T_n(X_1,\dots,X_n)-f(\theta))}{\sigma_f(\theta)}\le x\right\}-\mathbb{P}\{Z\le x\}\right|\to 0,$$

for all $\ell\in\mathcal{L}$,

$$(2)\qquad \sup_{\theta\in\Theta_n}\left|\mathbb{E}_\theta\ell\left(\frac{n^{1/2}(T_n(X_1,\dots,X_n)-f(\theta))}{\sigma_f(\theta)}\right)-\mathbb{E}\ell(Z)\right|\to 0 \text{ as } n\to\infty$$

and

$$(3)\qquad \liminf_{n\to\infty}\inf_{\tilde T_n}\sup_{\theta\in\Theta_n}\frac{n\mathbb{E}_\theta(\tilde T_n(X_1,\dots,X_n)-f(\theta))^2}{\sigma_f^2(\theta)}\ge 1$$

with the infimum in (3) *being over all estimators* $\tilde T_n$.

A similar definition can be also used for more general models in which the data $X^{(n)}$ is sampled from a distribution $P_\theta^{(n)}$, $\theta\in\Theta$ as well as in the case of a sequence of smooth functions $f_n:\Theta_n\mapsto\mathbb{R}$.

The idea of asymptotically efficient estimation (initially understood as asymptotically normal estimation with the smallest possible limit variance) goes back to Fisher [1922, 1925]. Fisher conjectured ("Fisher's program") that, under suitable regularity of statistical model, the maximal likelihood method would yield asymptotically efficient estimators with the optimal limit variance being the reciprocal of the Fisher information. The difficulties with implementing Fisher's program became apparent in the early 50s when Hodges developed a well known counterexample of a superefficient estimator in a regular statistical model. The development of contemporary view of asymptotic efficiency is due to several authors, in particular, to Le Cam and Hàjek (LeCam [1953] and Hájek [1972]). For regular finite-dimensional models, asymptotically efficient estimators of smooth functions $f(\theta)$ could be obtained from the maximum likelihood estimator $\hat\theta$ using the Delta Method: for a continuously differentiable function f, $f(\hat\theta)-f(\theta)=\langle f'(\theta),\hat\theta-\theta\rangle+o_{\mathbb{P}}(n^{-1/2})$, implying that $n^{1/2}(f(\hat\theta)-f(\theta))$ is asymptotically normal $N(0;\sigma_f^2(\theta))$ with the limit variance $\sigma_f^2(\theta)=\langle I(\theta)^{-1}f'(\theta),f'(\theta)\rangle$, $I(\theta)$ being the Fisher information matrix. The optimality of the limit variance is usually proved using convolution and local asymptotic minimax theorems (Hàjek, Le Cam). It could be also proved using van Trees inequality (see Gill and Levit [1995]) leading to bounds similar to (3).

Due to slow convergence rates of estimation of infinite-dimensional parameters in nonparametric statistics, it becomes important to identify low-dimensional features of these parameters that admit asymptotically efficient estimation with parametric $\sqrt{n}$-rate.

Such features are often represented by smooth functionals of infinite-dimensional parameters. Early references on asymptotically efficient estimation of smooth functionals include Levit [1975, 1978] and Ibragimov and Khasminskii [1981] with a number of further publications for the last decades on estimation of linear, quadratic and more general smooth functionals and with connections to extensive literature on efficiency in semiparametric estimation (see Bickel, Klaassen, Ritov, and Wellner [1993], Giné and Nickl [2016] and references therein). Ibragimov, Nemirovski and Khasminskii in Ibragimov, Nemirovski, and Khasminskii [1986] and Nemirovski in Nemirovski [1990, 2000] systematically studied the problem of estimation of general smooth functionals of unknown parameter of Gaussian shift model. In this model (also known as Gaussian sequence model), the parameter of interest is a "signal" $\theta \in \Theta$, where Θ is a bounded subset of a separable Hilbert space. Given an orthonormal basis $\{e_k : k \geq 1\}$ of $\mathbb{H}$, the data consists of observations $X_k = \langle \theta, e_k \rangle + \sigma Z_k, k \geq 1$, where $\{Z_k\}$ are i.i.d. $N(0,1)$ r.v. and σ is a small parameter characterizing the level of the noise (we will set $\sigma := n^{-1/2}$). In Ibragimov, Nemirovski, and Khasminskii [1986] and Nemirovski [1990, 2000], two different notions of smoothness of a functional f were used, with control of the derivatives either in the operator norm, or in the Hilbert–Schmidt norm (of multilinear forms). The complexity of estimation problem was characterized by the rate of decay of Kolmogorov diameters of set Θ defined as $d_m(\Theta) := \inf_{L \subset \mathbb{H}, \dim(L) \leq m} \sup_{\theta \in \Theta} \|\theta - P_L \theta\|, m \geq 1$, P_L being the orthogonal projection on subspace L. Assuming that $d_m(\Theta) \lesssim m^{-\beta}, m \geq 1$ for some $\beta > 0$, it was proved that efficient estimation (with a somewhat different definition of efficiency than Definition 1) of a smooth functional f on $\mathbb{H}$ is possible for smoothness parameter $s > s(\beta)$, where $s(\beta)$ is a threshold depending on the rate of decay β of Kolmogorov diameters. The estimation method was based on Taylor expansions of $f(\theta)$ around an estimator $\hat{\theta}$ with an optimal nonparametric rate, which allowed to reduce the problem to estimation of polynomial functions on $\mathbb{H}$. Nemirovski [1990, 2000] also proved that efficient estimation is impossible for some functionals f of smoothness $s < s(\beta)$.

More recently, estimation problems for functionals of unknown parameters have been studied in various models of high-dimensional statistics, including semi-parametric efficiency of regularization-based estimators (such as LASSO) van de Geer, Bühlmann, Ritov, and Dezeure [2014], Javanmard and Montanari [2014], C.-H. Zhang and S. S. Zhang [2014], Janková and van de Geer [2016] as well as minimax optimal rates of estimation of special functionals (in particular, linear and quadratic) Cai and Low [2005b], Cai and Low [2005a], Collier, Comminges, and Tsybakov [2017].

In this paper, we are primarily interested in the problem of estimation of smooth functionals of unknown covariance operator Σ based on a sample of size n of i.i.d. mean zero Gaussian random variables with covariance Σ. In this problem, the maximum likelihood

estimator is the sample covariance $\hat{\Sigma}$. By standard Hàjek-LeCam theory, plug-in estimator $h(\hat{\Sigma})$ is an asymptotically efficient estimator of a smooth functional $h(\Sigma)$ in the finite-dimensional case. The problem of asymptotically efficient estimation of general smooth functionals of covariance operators in high-dimensional setting (when the dimension d of the space is allowed to grow with the sample size n) has not been systematically studied. However, there are many results on asymptotic normality in this type of problems. In the 80s–90s, Girko developed asymptotically normal (but hardly asymptotically efficient) estimators of many special functionals of covariance matrices in high-dimensional setting (see Girko [1987], Girko [1995] and references therein). Central limit theorems for so called linear spectral statistics $\mathrm{tr}(f(\hat{\Sigma}))$ have been studied in random matrix theory with a number of deep results both in the case of high-dimensional sample covariance (or Wishart matrices) and in other random matrix models such as Wigner matrices, see, e.g., Bai and Silverstein [2004], Lytova and Pastur [2009]. However, these results do not have straightforward statistical implications since $\mathrm{tr}(f(\hat{\Sigma}))$ does not "concentrate" around the corresponding population parameter (with the exception of some special functionals of this form such as *log-determinant* $\log\det(\Sigma) = \mathrm{tr}(\log\Sigma)$, for which $\log\det(\hat{\Sigma})$ (with a simple bias correction) provides an asymptotically normal estimator (see Girko [1987] and Cai, Liang, and Zhou [2015])). More recent references include Fan, Rigollet, and Wang [2015] where optimal error rates in estimation of several special functionals of covariance under sparsity assumptions were studied and Gao and Zhou [2016] where Bernstein-von Mises type theorems for functionals of covariance were proved.

In what follows, $\mathfrak{B}(\mathbb{H})$ is the space of bounded linear operators in a separable Hilbert space $\mathbb{H}$. $\mathfrak{B}(\mathbb{H})$ is usually equipped with the operator norm denoted by $\|\cdot\|$. Let $\mathfrak{B}_{sa}(\mathbb{H})$ be the subspace of bounded self-adjoint operators. Denote by $\mathfrak{C}_+(\mathbb{H})$ the cone of self-adjoint positively semi-definite nuclear operators in $\mathbb{H}$ (the covariance operators). We use notation A^* for the adjoint operator of A, $\mathrm{rank}(A)$ for the rank of A, $\mathrm{tr}(A)$ for the trace of a trace class operator A, and $\|A\|_p$ for the Schatten p-norm of A : $\|A\|_p^p := \mathrm{tr}(|A|^p), |A| = (A^*A)^{1/2}, p \in [1,\infty]$. In particular, $\|A\|_1$ is the nuclear norm, $\|A\|_2$ is the Hilbert–Schmidt norm and $\|A\|_\infty = \|A\|$ is the operator norm of A. The inner product notation $\langle\cdot,\cdot\rangle$ is used for the inner product in the underlying Hilbert space $\mathbb{H}$, for the Hilbert–Schmidt inner product between the operators and also for linear functionals on the spaces of operators (for instance, $\langle A, B\rangle$, where A is a bounded operator and B is a nuclear operator, is a value of such a linear functional on the space of bounded operators). Given $u, v \in \mathbb{H}$, $u \otimes v$ denotes the tensor product of vectors u and v : $(u \otimes v)x := u\langle v, x\rangle, x \in \mathbb{H}$. Notation $A \preceq B$ means that operator $B - A$ is positively semi-definite.

We also use the following notations: given $a, b \geq 0$, $a \lesssim b$ means that $a \leq cb$ for a numerical constant $c > 0$; $a \gtrsim b$ is equivalent to $b \lesssim a$; $a \asymp b$ is equivalent to $a \lesssim b$ and $b \lesssim a$. Sometimes, constants in the above relationships depend on some parameter(s).

In such cases, the signs $\lesssim$, $\gtrsim$ and $\asymp$ are provided with subscripts: $a \lesssim_\gamma b$ means that $a \le c_\gamma b$ for a constant $c_\gamma > 0$.

2 Effective rank and estimation of linear functionals of principal components

Let X be a centered Gaussian random variable in a separable Hilbert space $\mathbb{H}$ with covariance operator $\Sigma = \mathbb{E}(X \otimes X)$ and let $X_1, \dots, X_n$ be a sample of n independent observations of X. The sample covariance operator is defined as $\hat{\Sigma} := n^{-1} \sum_{j=1}^n X_j \otimes X_j$. In the finite-dimensional case, it is well known that the operator norm error $\|\hat{\Sigma} - \Sigma\|$ could be controlled in terms of the dimension $d = \dim(\mathbb{H})$ of the space $\mathbb{H}$. In particular (see, e.g., Vershynin [2012]), for all $t \ge 1$ with probability at least $1 - e^{-t}$

$$\|\hat{\Sigma} - \Sigma\| \lesssim \|\Sigma\| \left(\sqrt{\frac{d}{n}} \bigvee \frac{d}{n} \bigvee \sqrt{\frac{t}{n}} \bigvee \frac{t}{n} \right), \tag{4}$$

which implies $\mathbb{E}\|\hat{\Sigma} - \Sigma\| \asymp \|\Sigma\| \left(\sqrt{\frac{d}{n}} \vee \frac{d}{n} \right)$. These bounds are sharp if the covariance operator is *isotropic* ($\Sigma = cI$ for a constant $c > 0$), or, more generally, it is of *isotropic type,* meaning that $c_1 I \preceq \Sigma \preceq c_2 I$ for some constants $0 < c_1 \le c_2 < \infty$ (it is assumed that c, c_1, c_2 are dimension free). The last condition holds, for instance, for well known *spiked covariance model* introduced by Johnstone [2001] (see also Johnstone and Lu [2009] and Paul [2007]). If the space $\mathbb{H}$ is infinite-dimensional (or it is finite-dimensional, but the covariance operator Σ is not of isotropic type), bound (4) is no longer sharp and other complexity parameters become relevant in covariance estimation problem. In particular, Vershynin [2012] suggested to use in such cases so called *effective rank* $\mathbf{r}(\Sigma) := \frac{\mathrm{tr}(\Sigma)}{\|\Sigma\|}$ instead of the dimension. Clearly, $\mathbf{r}(\Sigma) \le \mathrm{rank}(\Sigma) \le \dim(\mathbb{H})$. The next result was proved by Koltchinskii and Lounici [2017a] and it shows that $\mathbf{r}(\Sigma)$ is a natural complexity parameter in covariance estimation (at least, in the Gaussian case).

Theorem 1. *The following expectation bound holds:*

$$\mathbb{E}\|\hat{\Sigma} - \Sigma\| \asymp \|\Sigma\| \left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \bigvee \frac{\mathbf{r}(\Sigma)}{n} \right). \tag{5}$$

Moreover, for all $t \ge 1$, the following concentration inequality holds with probability at least $1 - e^{-t}$:

$$\left| \|\hat{\Sigma} - \Sigma\| - \mathbb{E}\|\hat{\Sigma} - \Sigma\| \right| \lesssim \|\Sigma\| \left(\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \bigvee 1 \right) \sqrt{\frac{t}{n}} \bigvee \frac{t}{n} \right). \tag{6}$$

Note that bounds (5) and (6) were proved in Koltchinskii and Lounici [2017a] in a more general setting of estimation of covariance operator of a Gaussian random variable in a separable Banach space with effective rank defined as $\mathbf{r}(\Sigma) := \frac{(\mathbb{E}\|X\|)^2}{\|\Sigma\|}$. These bounds show that the "relative operator norm error" $\frac{\|\hat{\Sigma}-\Sigma\|}{\|\Sigma\|}$ is controlled by the ratio $\frac{\mathbf{r}(\Sigma)}{n}$ and that condition $\mathbf{r}(\Sigma) = o(n)$ is necessary and sufficient for the operator norm consistency of the sample covariance. In view of these results, it became natural to study concentration and normal approximation properties of various statistics represented by functionals of sample covariance in a dimension free framework in which the effective rank $\mathbf{r}(\Sigma)$ is allowed to be large (although satisfying the condition $\mathbf{r}(\Sigma) = o(n)$, which ensures that $\hat{\Sigma}$ is a small perturbation of Σ). This was done in Koltchinskii and Lounici [2016] in the case of bilinear forms of spectral projection operators of $\hat{\Sigma}$ (empirical spectral projections) and in Koltchinskii and Lounici [2017c,b] in the case of their squared Hilbert–Schmidt error. It turned out that naive plug-in estimators (such as bilinear forms of empirical spectral projections) are not $\sqrt{n}$-consistent (unless $\mathbf{r}(\Sigma) = o(n)$) due to their substantial bias and bias reduction becomes crucial for asymptotically efficient estimation. We briefly discuss below the approach to this problem developed by Koltchinskii and Lounici [2016], Koltchinskii, Löffler, and Nickl [2017].

Let $\sigma(\Sigma)$ be the spectrum of Σ and let $\lambda(\Sigma) = \sup(\sigma(\Sigma)) = \|\Sigma\|$ be its largest eigenvalue. Let $g(\Sigma) := \operatorname{dist}(\lambda(\Sigma); \sigma(\Sigma) \setminus \{\lambda(\Sigma)\})$ be the gap between $\lambda(\Sigma)$ and the rest of the spectrum. Suppose $\lambda(\Sigma)$ has multiplicity 1 and let $P(\Sigma) = \theta(\Sigma) \otimes \theta(\Sigma)$ be the corresponding one-dimensional spectral projection. Here $\theta(\Sigma)$ is the unit eigenvector corresponding to $\lambda(\Sigma)$ (defined up to its sign). Given $u \in \mathbb{H}$, our goal is to estimate the linear functional $\langle \theta(\Sigma), u\rangle$ based on i.i.d. observations $X_1, \dots, X_n$ sampled from $N(0;\Sigma)$ (note that the value of this functional is also defined only up to its sign, so, essentially, we can estimate only its absolute value). If $\theta(\hat{\Sigma})$ denotes a unit eigenvector of sample covariance $\hat{\Sigma}$ that corresponds to its top eigenvalue $\lambda(\hat{\Sigma}) = \|\hat{\Sigma}\|$, then $\langle \theta(\hat{\Sigma}), u\rangle$ is the plug-in estimator of $\langle \theta(\Sigma), u\rangle$. Without loss of generality, we assume in what follows that $\theta(\hat{\Sigma})$ and $\theta(\Sigma)$ are properly aligned in the sense that $\langle \theta(\hat{\Sigma}), \theta(\Sigma)\rangle \geq 0$ (which allows us indeed to view $\langle \theta(\hat{\Sigma}), u\rangle$ as an estimator of $\langle \theta(\Sigma), u\rangle$). It was shown in Koltchinskii and Lounici [2016], that the quantity

$$b(\Sigma) = b_n(\Sigma) := \mathbb{E}_\Sigma \langle \theta(\hat{\Sigma}), \theta(\Sigma)\rangle^2 - 1 \in [-1, 0]$$

characterizes the size of the bias of estimator $\langle \theta(\hat{\Sigma}), u\rangle$. In particular, the results of Koltchinskii and Lounici [2016] and Koltchinskii, Löffler, and Nickl [2017] imply that $\langle \theta(\hat{\Sigma}), u\rangle$ "concentrates" around the value $\sqrt{1 + b(\Sigma)}\langle \theta(\Sigma), u\rangle$ rather than around the value of the functional $\langle \theta(\Sigma), u\rangle$ itself. To state this result more precisely, consider the spectral representation $\Sigma = \sum_{\lambda \in \sigma(\Sigma)} \lambda P_\lambda$ with eigenvalues λ and corresponding orthogonal spectral

projections P_λ. Define

$$C(\Sigma) := \sum_{\lambda \neq \lambda(\Sigma)} (\lambda(\Sigma) - \lambda)^{-1} P_\lambda \text{ and } \sigma^2(\Sigma; u) := \lambda(\Sigma)\langle \Sigma C(\Sigma) u, C(\Sigma) u\rangle.$$

For $u \in \mathbb{H}$, $r > 1$, $a > 1$ and $\sigma_0 > 0$, define the following class of covariance operators in $\mathbb{H}$: $\mathcal{S}(r, a, \sigma_0, u) := \left\{\Sigma : \mathbf{r}(\Sigma) \leq r, \frac{\|\Sigma\|}{g(\Sigma)} \leq a, \sigma^2(\Sigma; u) \geq \sigma_0^2\right\}$. Note that additional conditions on r, a, σ_0, u might be needed for the class $\mathcal{S}(r, a, \sigma_0, u)$ to be nonempty.

Theorem 2. *Let $u \in \mathbb{H}$, $a > 1$ and $\sigma_0 > 0$. Suppose that $r_n > 1$ and $r_n = o(n)$ as $n \to \infty$. Then*

$$\sup_{\Sigma \in \mathcal{S}(r_n, a, \sigma_0, u)} \sup_{x \in \mathbb{R}} \left| \mathbb{P}_\Sigma \left\{ \frac{\sqrt{n}(\langle \theta(\hat{\Sigma}), u\rangle - \sqrt{1 + b(\Sigma)}\langle \theta(\Sigma), u\rangle)}{\sigma(\Sigma; u)} \leq x \right\} - \mathbb{P}\{Z \leq x\} \right| \to 0$$

and, for all $\ell \in \mathcal{L}$,

$$\sup_{\Sigma \in \mathcal{S}(r_n, a, \sigma_0, u)} \left| \mathbb{E}_\Sigma \ell \left(\frac{\sqrt{n}(\langle \theta(\hat{\Sigma}), u\rangle - \sqrt{1 + b(\Sigma)}\langle \theta(\Sigma), u\rangle)}{\sigma(\Sigma; u)} \right) - \mathbb{E}\ell(Z) \right| \to 0 \text{ as } n \to \infty.$$

It was also proved in Koltchinskii, Löffler, and Nickl [2017] that $b(\Sigma) \asymp \frac{\mathbf{r}(\Sigma)}{n}$. This implies that the "bias" $(\sqrt{1 + b(\Sigma)} - 1)\langle \theta(\Sigma), u\rangle$ of estimator $\langle \theta(\hat{\Sigma}), u\rangle$ is asymptotically negligible (of the order $o(n^{-1/2})$) if $\mathbf{r}(\Sigma) = o(\sqrt{n})$, which yields the following result:

Corollary 1. *Let $u \in \mathbb{H}$, $a > 1$ and $\sigma_0 > 0$. Suppose that $r_n > 1$ and $r_n = o(\sqrt{n})$ as $n \to \infty$, and that $\mathcal{S}(r, a', \sigma_0', u) \neq \varnothing$ for some $r > 1, a' < a, \sigma_0' > \sigma_0$. Then $\langle \theta(\hat{\Sigma}), u\rangle$ is an asymptotically efficient estimator of $\langle \theta(\Sigma), u\rangle$ with respect to $\mathcal{S}(r_n, a, \sigma_0, u)$ with convergence rate $\sqrt{n}$ and variance $\sigma^2(\Sigma; u)$.*

On the other hand, it was shown in Koltchinskii, Löffler, and Nickl [ibid.] that, under the assumptions $r_n = o(n)$ and $\frac{r_n}{n^{1/2}} \to \infty$ as $n \to \infty$,

$$\lim_{n \to \infty} \sup_{\Sigma \in \mathcal{S}(r_n, a, \sigma_0, u)} \mathbb{P}_\Sigma \left\{ |\langle \theta(\hat{\Sigma}), u\rangle - \langle \theta(\Sigma), u\rangle| \geq c\|u\| \frac{r_n}{n} \right\} = 1$$

for some constant $c = c(a; \sigma_0) > 0$, implying that $\langle \theta(\hat{\Sigma}), u\rangle$ is not even $\sqrt{n}$-consistent estimator of $\langle \theta(\Sigma), u\rangle$ when effective rank is larger than $\sqrt{n}$. Clearly, slower convergence rate is due to a large bias of the estimator $\langle \theta(\hat{\Sigma}), u\rangle$ when the complexity of the problem becomes large (the effective rank exceeds $\sqrt{n}$), and bias reduction is crucial to construct a $\sqrt{n}$-consistent estimator in this case. In Koltchinskii and Lounici [2016], a method of bias reduction based on estimation of bias parameter $b(\Sigma)$ was developed. For simplicity,

assume that the sample size is even $n = 2n'$ and split the sample into two equal parts, each of size n'. Let $\hat{\Sigma}^{(1)}, \hat{\Sigma}^{(2)}$ be the sample covariances based on these two subsamples and let $\theta(\hat{\Sigma}^{(1)}), \theta(\hat{\Sigma}^{(2)})$ be their top principal components. Since for all $u \in \mathbb{H}$ and for $i = 1, 2$, $\langle \theta(\hat{\Sigma}^{(i)}), u\rangle$ "concentrates" around $\sqrt{1 + b_{n'}(\Sigma)}\langle \theta(\Sigma), u\rangle$ and $\theta(\hat{\Sigma}^{(i)}), i = 1, 2$ are independent, it is not hard to check that $\langle \theta(\hat{\Sigma}^{(1)}), \theta(\hat{\Sigma}^{(2)})\rangle$ "concentrates" around $1 + b_{n'}(\Sigma)$. Thus, $\hat{b} := \langle \theta(\hat{\Sigma}^{(1)}), \theta(\hat{\Sigma}^{(2)})\rangle - 1$ could be used as an estimator of $b_{n'}(\Sigma)$. It was proved in Koltchinskii and Lounici [2016] that $\hat{b} - b_{n'}(\Sigma) = o_{\mathbb{P}}(n^{-1/2})$ provided that $\mathbf{r}(\Sigma) = o(n)$, and this led to a bias corrected estimator $(1 + \hat{b})^{-1/2}\langle \theta(\hat{\Sigma}^{(1)}), u\rangle$ of linear functional $\langle \theta(\Sigma), u\rangle$, which was proved to be asymptotically normal with convergence rate $\sqrt{n}$. This approach was further developed in Koltchinskii, Löffler, and Nickl [2017], where a more subtle version of sample split yielded an asymptotically efficient estimator of the functional $\langle \theta(\Sigma), u\rangle$. Let $m = m_n = o(n)$ as $n \to \infty$, $m < n/3$. Split the sample $X_1, \dots, X_n$ into three disjoint subsamples, one of size $n' = n'_n := n - 2m > n/3$ and two others of size m. Let $\hat{\Sigma}^{(1)}, \hat{\Sigma}^{(2)}, \hat{\Sigma}^{(3)}$ be the sample covariances based on these three subsamples and let $\theta(\hat{\Sigma}^{(j)}), j = 1, 2, 3$ be the corresponding top principal components. Denote

$$\hat{d} := \frac{|\langle \theta(\hat{\Sigma}^{(1)}), \theta(\hat{\Sigma}^{(2)})\rangle|}{|\langle \theta(\hat{\Sigma}^{(2)}), \theta(\hat{\Sigma}^{(3)})\rangle|^{1/2}} \text{ and } \hat{\theta} := \frac{\theta(\hat{\Sigma}^{(1)})}{\hat{d} \vee (1/2)}.$$

Theorem 3. *Let $u \in \mathbb{H}$, $a > 1$ and $\sigma_0 > 0$. Suppose that $r_n > 1$ and $r_n = o(n)$ as $n \to \infty$. Suppose also that $\mathcal{S}(r, a', \sigma'_0, u) \neq \varnothing$ for some $r > 1, a' < a, \sigma'_0 > \sigma_0$. Take $m = m_n$ such that $m_n = o(n)$ and $nr_n = o(m_n^2)$ as $n \to \infty$. Then $\langle \hat{\theta}, u\rangle$ is an asymptotically efficient estimator of $\langle \theta(\Sigma), u\rangle$ with respect to $\mathcal{S}(r_n, a, \sigma_0, u)$ with convergence rate $\sqrt{n}$ and variance $\sigma^2(\Sigma, u)$.*

The approach to bias reduction and efficient estimation described above is based on rather special structural properties of the bias of empirical spectral projections. In the following sections, we discuss a much more general approach applicable to broader classes of problems.

3 Normal approximation bounds for plug-in estimators of smooth functionals

Let $f : \mathbb{R} \mapsto \mathbb{R}$ be a smooth function and let B be a nuclear operator in $\mathbb{H}$. The goal is to estimate functionals of the form $\langle f(\Sigma), B\rangle, \Sigma \in \mathcal{C}_+(\mathbb{H})$ based on i.i.d. observations $X_1, \dots, X_n$ sampled from the Gaussian distribution with mean zero and covariance operator Σ. We will first consider a simple plug-in estimator $\langle f(\hat{\Sigma}), B\rangle$. To study its properties, we will rely on several results on Fréchet differentiability of operator functions $\mathcal{B}_{sa}(\mathbb{H}) \ni A \mapsto f(A) \in \mathcal{B}_{sa}(\mathbb{H})$ with respect to the operator norm as well as on bounds

on the remainders of their Taylor expansions. These results could be found in operator theory literature (see, in particular, Aleksandrov and Peller [2016]).

We will need the definition of Besov spaces (see Triebel [1983] for more details). Consider a C^∞ function $w \geq 0$ in $\mathbb{R}$ with $\mathrm{supp}(w) \subset [-2,2]$, satisfying the assumptions $w(t) = 1, |t| \leq 1$ and $w(-t) = w(t), t \in \mathbb{R}$. Let $w_0(t) := w(t/2) - w(t), t \in \mathbb{R}$ (implying that $\mathrm{supp}(w_0) \subset \{t : 1 \leq |t| \leq 4\}$). For $w_j(t) := w_0(2^{-j}t), t \in \mathbb{R}$, we have $\mathrm{supp}(w_j) \subset \{t : 2^j \leq |t| \leq 2^{j+2}\}, j = 0, 1, \dots$ and also $w(t) + \sum_{j\geq 0} w_j(t) = 1, t \in \mathbb{R}$. Let $W, W_j, j \geq 1$ be functions in Schwartz space $\mathcal{S}(\mathbb{R})$ defined by their Fourier transforms: $w(t) = (\mathcal{F}W)(t), w_j(t) = (\mathcal{F}W_j)(t), t \in \mathbb{R}, j \geq 0$. For a tempered distribution $f \in \mathcal{S}'(\mathbb{R})$, define its Littlewood-Paley decomposition as the set of functions $f_0 := f * W, f_n := f * W_{n-1}, n \geq 1$ with compactly supported Fourier transforms. By Paley-Wiener Theorem, f_n can be extended to an entire function of exponential type 2^{n+1} (for all $n \geq 0$). It is also well known that $\sum_{n\geq 0} f_n = f$ with convergence of the series in the space $\mathcal{S}'(\mathbb{R})$. Define $B^s_{\infty,1}$-Besov norm as

$$\|f\|_{B^s_{\infty,1}} := \sum_{n\geq 0} 2^{ns} \|f_n\|_{L_\infty(\mathbb{R})}, s \in \mathbb{R}$$

and let $B^s_{\infty,1}(\mathbb{R}) := \{f \in \mathcal{S}'(\mathbb{R}) : \|f\|_{B^s_{\infty,1}} < +\infty\}$ be the corresponding (inhomogeneous) Besov space. It is easy to check that, for $s \geq 0$, the series $\sum_{n\geq 0} f_n$ converges uniformly in $\mathbb{R}$ and the space $B^s_{\infty,1}(\mathbb{R})$ is continuously embedded in the space $C_u(\mathbb{R})$ of all bounded uniformly continuous functions equipped with the uniform norm $\|\cdot\|_{L_\infty(\mathbb{R})}$.

It was proved by Peller [1985] that, for all $f \in B^1_{\infty,1}(\mathbb{R})$, the mapping $\mathfrak{B}_{sa}(\mathbb{H}) \ni A \mapsto f(A) \in \mathfrak{B}_{sa}(\mathbb{H})$ is Fréchet differentiable with respect to the operator norm (in fact, Peller used homogeneous Besov spaces). Let $Df(A;H) = Df(A)(H)$ denote its derivative at A in direction H. If $A \in \mathfrak{B}_{sa}(\mathbb{H})$ is a compact operator with spectral representation $A = \sum_{\lambda\in\sigma(A)} \lambda P_\lambda$ with eigenvalues λ and spectral projections P_λ, then $Df(A;H) = \sum_{\lambda,\mu\in\sigma(A)} f^{[1]}(\lambda,\mu) P_\lambda H P_\mu$, where $f^{[1]}(\lambda,\mu) := \frac{f(\lambda)-f(\mu)}{\lambda-\mu}, \lambda \neq \mu$, $f^{[1]}(\lambda,\mu) := f'(\lambda), \lambda = \mu$ is Loewner kernel (there are also extensions of this formula for more general operators with continuous spectrum with double operator integrals instead of the sums Peller [2006] and Aleksandrov and Peller [2016]). If $f \in B^s_{\infty,1}(\mathbb{R})$ for some $s \in (1,2]$, then the first order Taylor expansion $f(A+H) = f(A) + Df(A;H) + S_f(A;H)$ holds with the following bound on the remainder: $\|S_f(A;H)\| \lesssim_s \|f\|_{B^s_{\infty,1}} \|H\|^s, H \in \mathfrak{B}_{sa}(\mathbb{H})$ (see Koltchinskii [2017] for the proof fully based on methods of Aleksandrov and Peller [2016]). Applying the Taylor expansion to $f(\hat{\Sigma})$ and using the bound on the remainder along with Theorem 1, we get that

$$f(\hat{\Sigma}) - f(\Sigma) = Df(\Sigma; \hat{\Sigma} - \Sigma) + S_f(\Sigma; \hat{\Sigma} - \Sigma)$$

with $\|S_f(\Sigma;\hat{\Sigma}-\Sigma)\| = o_{\mathbb{P}}(n^{-1/2})$ provided that $\mathbf{r}(\Sigma) = o(n^{1-1/s})$. It is also easy to check that

$$\sqrt{n}\langle Df(\Sigma;\hat{\Sigma}-\Sigma), B\rangle = n^{-1/2}\sum_{j=1}^{n}\langle Df(\Sigma; X_j\otimes X_j - \Sigma), B\rangle$$

is asymptotically normal $N(0;\sigma_f^2(\Sigma,B))$, $\sigma_f^2(\Sigma;B) := 2\|\Sigma^{1/2}Df(\Sigma;B)\Sigma^{1/2}\|_2^2$, which, along with asymptotic negligibility of the remainder, implies the asymptotic normality of $\sqrt{n}(\langle f(\hat{\Sigma}), B\rangle - \langle f(\Sigma), B\rangle)$ with the same limit mean and variance. Similar rather standard perturbation analysis (most often, based on holomorphic functional calculus rather than on more sophisticated tools of Aleksandrov and Peller [2016]) has been commonly used, especially, in applications to PCA, in the case of finite-dimensional problems of bounded dimension, see, e.g., Anderson [2003]. It, however, fails as soon as the effective rank is sufficiently large (above $n^{1-1/s}$ for functions of smoothness $s\in(1,2]$) since the remainder $S_f(\Sigma;\hat{\Sigma}-\Sigma)$ of Taylor expansion is not asymptotically negligible. It turns out, that in this case $\langle f(\hat{\Sigma}), B\rangle$ is still a $\sqrt{n}$-consistent and asymptotically normal estimator of its own expectation $\langle \mathbb{E}_\Sigma f(\hat{\Sigma}), B\rangle$, but the bias $\langle \mathbb{E}_\Sigma f(\hat{\Sigma}) - f(\Sigma), B\rangle$ is no longer asymptotically negligible. In fact, the bias is equal to $\langle \mathbb{E}_\Sigma S_f(\Sigma;\hat{\Sigma}-\Sigma), B\rangle$, which is upper bounded by $\lesssim \|f\|_{B^s_{\infty,1}}\|B\|_1(\frac{\mathbf{r}(\Sigma)}{n})^{s/2}$. This bound is sharp for typical smooth functions. For instance, if $f(x)=x^2$ and $B = u\otimes u$, it is easy to check that

$$\sup_{\|u\|\le 1}|\langle \mathbb{E}_\Sigma f(\hat{\Sigma}) - f(\Sigma), u\otimes u\rangle| \asymp \|\Sigma\|^2\frac{\mathbf{r}(\Sigma)}{n},$$

and the bias is not asymptotically negligible if $\mathbf{r}(\Sigma)\ge n^{1/2}$. Moreover, if $\frac{\mathbf{r}(\Sigma)}{\sqrt{n}}\to\infty$, the plug-in estimator $\langle f(\hat{\Sigma}), u\otimes u\rangle$ of $\langle f(\Sigma), u\otimes u\rangle$ is not $\sqrt{n}$-consistent (for some u with $\|u\|\le 1$).

The next result (see also Koltchinskii [2017]) shows asymptotic normality (with $\sqrt{n}$-rate) of $\langle f(\hat{\Sigma}), B\rangle$ as an estimator of its own expectation. Define

$$\mathcal{G}_{f,B}(r;a;\sigma_0) := \left\{\Sigma : \mathbf{r}(\Sigma)\le r, \|\Sigma\|\le a, \sigma_f^2(\Sigma;B)\ge\sigma_0^2\right\},\quad r>1, a>0, \sigma_0^2>0.$$

Theorem 4. *Let $f\in B^s_{\infty,1}(\mathbb{R})$ for some $s\in(1,2]$ and let B be a nuclear operator. For any $a>0,\sigma_0^2>0$ and $r_n>1$ such that $r_n = o(n)$ as $n\to\infty$,*

(7)
$$\sup_{\Sigma\in\mathcal{G}_{f,B}(r_n;a;\sigma_0)}\sup_{x\in\mathbb{R}}\left|\mathbb{P}_\Sigma\left\{\frac{n^{1/2}\langle f(\hat{\Sigma})-\mathbb{E}_\Sigma f(\hat{\Sigma}), B\rangle}{\sigma_f(\Sigma;B)}\le x\right\} - \mathbb{P}\{Z\le x\}\right| \to 0 \text{ as } n\to\infty.$$

The proof of this result is based on the following simple representation

$$\langle f(\hat{\Sigma})-\mathbb{E}_\Sigma f(\hat{\Sigma}), B\rangle = \frac{1}{n}\sum_{j=1}^{n}\langle Df(\Sigma;X_j\otimes X_j-\Sigma), B\rangle + \langle S_f(\Sigma;\hat{\Sigma}-\Sigma)-\mathbb{E}_\Sigma S_f(\Sigma;\hat{\Sigma}-\Sigma), B\rangle.$$

For the first term in the right hand side, it is easy to prove a normal approximation bound based on Berry-Esseen inequality. The main part of the proof deals with the second term, the centered remainder of Taylor expansion $\langle S_f(\Sigma;\hat\Sigma-\Sigma),B\rangle - \mathbb{E}\langle S_f(\Sigma;\hat\Sigma-\Sigma),B\rangle$. For this term, the following bound was proved using Gaussian concentration inequality: for all $t \geq 1$ with probability at least $1-e^{-t}$,

$$(8)\quad |\langle S_f(\Sigma;\hat\Sigma-\Sigma) - \mathbb{E}S_f(\Sigma;\hat\Sigma-\Sigma),B\rangle| \lesssim_s \|f\|_{B^s_{\infty,1}}\|B\|_1\|\Sigma\|^s\Big(\Big(\frac{\mathbf{r}(\Sigma)}{n}\Big)^{(s-1)/2}\bigvee\Big(\frac{\mathbf{r}(\Sigma)}{n}\Big)^{s-1/2}\bigvee\Big(\frac{t}{n}\Big)^{(s-1)/2}\bigvee\Big(\frac{t}{n}\Big)^{s-1/2}\Big)\sqrt{\frac{t}{n}}.$$

It implies that the centered remainder is of the order $\Big(\frac{\mathbf{r}(\Sigma)}{n}\Big)^{(s-1)/2}\sqrt{\frac{1}{n}}$, which is $o(n^{-1/2})$ as soon as $\mathbf{r}(\Sigma)=o(n)$.

If $\Sigma \in \mathcal{G}_{f,B}(r_n;a;\sigma_0)$ with $r_n = o(n^{1-1/s})$, the bias $\langle \mathbb{E}_\Sigma f(\hat\Sigma) - f(\Sigma),B\rangle$ is of the order $o(n^{-1/2})$ and plug-in estimator $\langle f(\hat\Sigma),B\rangle$ of $\langle f(\Sigma),B\rangle$ is asymptotically normal with $\sqrt{n}$-rate. The following corollary of Theorem 4 holds.

Corollary 2. *Let $f \in B^s_{\infty,1}(\mathbb{R})$ for some $s \in (1,2]$. and let B be a nuclear operator. Let $a>0,\sigma_0^2>0$ and let $r_n>1$ be such that $r_n\to\infty$ and $r_n = o(n^{1-\frac{1}{s}})$ as $n\to\infty$. Suppose that $\mathcal{G}_{f,B}(r;a';\sigma_0')\neq\varnothing$ for some $r>1, a'<a, \sigma_0'>\sigma_0$. Then $\langle f(\hat\Sigma),B\rangle$ is an asymptotically efficient estimator of $\langle f(\Sigma),B\rangle$ with respect to $\mathcal{G}_{f,B}(r_n;a;\sigma_0)$ with convergence rate $\sqrt{n}$ and variance $\sigma_f^2(\Sigma;B)$.*

Thus, as soon as $r_n = o(n^{1/2})$ and f is sufficiently smooth, *the plug-in estimator is asymptotically efficient*. However, as we have already pointed out above, this conclusion is not true if $r_n \geq n^{1/2}$ regardless of the degree of smoothness of f (even in the case of function $f(x)=x^2$). Moreover, not only asymptotic efficiency, but even $\sqrt{n}$-consistency of the plug-in estimator does not hold in this case, and the problem of asymptotically efficient estimation of functionals $\langle f(\Sigma),B\rangle$ becomes much more complicated. In the following sections, we outline a solution of this problem with the dimension of the space rather than the effective rank playing the role of complexity parameter. The idea of our approach is to try to find a function g on the space $\mathcal{B}_{sa}(\mathbb{H})$ of self-adjoint operators that solves approximately the equation $\mathbb{E}_\Sigma g(\hat\Sigma) = f(\Sigma)$ with an error of the order $o(n^{-1/2})$. If such solution g is sufficiently smooth, it could be possible to prove an analog of normal approximation of Theorem 4 for estimator $\langle g(\hat\Sigma),B\rangle$. Since the bias of this estimator is asymptotically negligible, it would be possible to show asymptotic normality of $\langle g(\hat\Sigma),B\rangle$ as an estimator of $\langle f(\Sigma),B\rangle$.

4 Bootstrap Chain bias reduction and asymptotically efficient estimation

Assume that $d := \dim(\mathbb{H})$ is finite (in what follows, $d = d_n$ could grow with n). It also will be assumed that the covariance operator Σ is of isotropic type. The following integral operator on the cone $\mathcal{C}_+(\mathbb{H})$ will be crucial in our approach:

$$\mathcal{T}g(\Sigma) := \mathbb{E}_\Sigma g(\hat{\Sigma}) = \int_{\mathcal{C}_+(\mathbb{H})} g(S)P(\Sigma; dS), \Sigma \in \mathcal{C}_+(\mathbb{H}).$$

Here $P(\Sigma;\cdot)$ is the distribution of the sample covariance $\hat{\Sigma}$ based on n i.i.d. observations sampled from $N(0;\Sigma)$. Clearly, $P(\Sigma;\cdot)$ is a rescaled Wishart distribution and $P(\cdot;\cdot)$ is a Markov kernel on the cone $\mathcal{C}_+(\mathbb{H})$. We will call $\mathcal{T}$ *the Wishart operator* and view it as an operator acting on bounded measurable functions on the cone $\mathcal{C}_+(\mathbb{H})$ with values either in $\mathbb{R}$ or in $\mathcal{B}_{sa}(\mathbb{H})$. Such operators are well known in the theory of Wishart matrices (see, e.g., James [1961], Letac and Massam [2004]). To obtain an unbiased estimator $g(\hat{\Sigma})$ of $f(\Sigma)$, one needs to solve the integral equation $\mathcal{T}g(\Sigma) = f(\Sigma), \Sigma \in \mathcal{C}_+(\mathbb{H})$ (*the Wishart equation*). Denoting $\mathcal{B} := \mathcal{T} - \mathcal{I}$, $\mathcal{I}$ being the identity operator, one can write the solution of the Wishart equation as a formal Neumann series $g(\Sigma) = (\mathcal{I} + \mathcal{B})^{-1} f(\Sigma) = \sum_{j=0}^{\infty}(-1)^j \mathcal{B}^j f(\Sigma)$. We will use its partial sums to define approximate solutions of the Wishart equation:

$$f_k(\Sigma) := \sum_{j=0}^{\infty}(-1)^j \mathcal{B}^j f(\Sigma), \Sigma \in \mathcal{C}_+(\mathbb{H}), k \geq 0,$$

with $f_k(\hat{\Sigma})$ for a properly chosen k being an estimator of $f(\Sigma)$. Note that its bias is

$$\mathbb{E}_\Sigma f_k(\hat{\Sigma}) - f(\Sigma) = (-1)^k \mathcal{B}^{k+1} f(\Sigma), \Sigma \in \mathcal{C}_+(\mathbb{H})$$

and, to justify this approach to bias reduction, one has to show that, for smooth enough functions f and large enough value of k, $\langle \mathcal{B}^{k+1} f(\Sigma), B\rangle$ is of the order $o(n^{-1/2})$. Note that a similar approach was recently discussed by Jiao, Han, and Weissman [2017] in the case of a problem of estimation of smooth function of parameter of binomial model $B(n;\theta), \theta \in [0,1]$. If $\hat{\theta}$ denotes the frequency, then $Tg(\theta) = \mathbb{E}_\theta g(\hat{\theta})$ is a Bernstein polynomial approximation of function g and bounds on $\mathcal{B}^{k+1} f(\theta)$ were deduced in Jiao, Han, and Weissman [ibid.] from some results of classical approximation theory (see, e.g., Totik [1994]).

We describe below our approach in Koltchinskii [2017] based on a Markov chain interpretation of the problem. To this end, consider a Markov chain $\hat{\Sigma}^{(0)} = \Sigma \to \hat{\Sigma}^{(1)} = \hat{\Sigma} \to \hat{\Sigma}^{(2)} \to \dots$ in the cone $\mathcal{C}_+(\mathbb{H})$ with transition probability kernel $P(\cdot;\cdot)$. Note that

for any $t \geq 1$, $\hat{\Sigma}^{(t)}$ can be viewed as the sample covariance based on n i.i.d. observations sampled from normal distribution $N(0; \hat{\Sigma}^{(t-1)})$, conditionally on $\hat{\Sigma}^{(t-1)}$. In other words, the Markov chain $\hat{\Sigma}^{(t)}, t = 0, 1, 2, \ldots$ is an outcome of iterative parametric bootstrap procedure and it will be called in what follows *the Bootstrap Chain*. By bound (4), conditionally on $\hat{\Sigma}^{(t-1)}$, with a high probability $\|\hat{\Sigma}^{(t)} - \hat{\Sigma}^{(t-1)}\| \lesssim \|\hat{\Sigma}^{(t-1)}\|\sqrt{\frac{d}{n}}$, implying that the Bootstrap Chain moves with "small steps", provided that $d = o(n)$. Now observe that $\mathcal{T}^k f(\Sigma) = \mathbb{E}_\Sigma f(\hat{\Sigma}^{(k)})$ and, by Newton's binomial formula,

(9)
$$\mathcal{B}^k f(\Sigma) = (\mathcal{T} - \mathcal{I})^k f(\Sigma) = \sum_{j=0}^k (-1)^{k-j} \binom{k}{j} \mathcal{T}^j f(\Sigma) = \mathbb{E}_\Sigma \sum_{j=0}^k (-1)^{k-j} \binom{k}{j} f(\hat{\Sigma}^{(j)}).$$

Note that to compute functions $\mathcal{B}^k f(\hat{\Sigma}), k \geq 1$ (which is needed to compute the estimator $f_k(\hat{\Sigma})$) one can use bootstrap: $\mathcal{B}^k f(\hat{\Sigma}) = \mathbb{E}_{\hat{\Sigma}} \sum_{j=0}^k (-1)^{k-j} \binom{k}{j} f(\hat{\Sigma}^{(j+1)})$ since the Bootstrap Chain now starts with $\hat{\Sigma}^{(0)} = \hat{\Sigma}$, and it can be approximated by the average of Monte Carlo simulations of $\sum_{j=0}^k (-1)^{k-j} \binom{k}{j} f(\hat{\Sigma}^{(j+1)})$.

Denote $F(j) := f(\hat{\Sigma}^{(j)}), j \geq 0$ and $\Delta F(j) := F(j+1) - F(j), j \geq 0$. Then $\sum_{j=0}^k (-1)^{k-j} \binom{k}{j} f(\hat{\Sigma}^{(j)}) = \Delta^k F(0)$ is the k-th order difference of sequence $F(j), j \geq 0$ at $j = 0$ (in other words, the k-th order difference of function f on the Markov chain $\{\hat{\Sigma}^{(t)}\}$). It is well known that, for a k times continuously differentiable function f in $\mathbb{R}$ the k-th order difference $\Delta_h^k f(x)$, where $\Delta_h f(x) := f(x+h) - f(x)$, is of the order $O(h^k)$ as $h \to 0$. Since the chain $\{\hat{\Sigma}^{(t)}\}$ moves with steps $\asymp \sqrt{\frac{d}{n}}$, it becomes plausible that, on average, $\sum_{j=0}^k (-1)^{k-j} \binom{k}{j} f(\hat{\Sigma}^{(j)})$ would be of the order $O((\frac{d}{n})^{k/2})$ for functions of smoothness k. The justification of this heuristic will be discussed in some detail in the next section and it is based on the development of certain integral representations of functions $\mathcal{B}^k f(\Sigma), k \geq 1$ that rely on properties of orthogonally invariant functions on the cone $\mathcal{C}_+(\mathbb{H})$. These representations are then used to obtain bounds on operators $\mathcal{B}^k f(\Sigma)$ and on the bias $\mathbb{E}_\Sigma f_k(\hat{\Sigma}) - f(\Sigma)$ of estimator $f(\Sigma)$, to study smoothness properties of functions $\mathcal{B}^k f(\Sigma)$ and $f_k(\Sigma)$ that allow us to prove concentration bounds on the remainder $\langle S_{f_k}(\Sigma; \hat{\Sigma} - \Sigma), B\rangle$ of the first order Taylor expansion of $\langle f_k(\hat{\Sigma}), B\rangle$ and, finally, to establish normal approximation bounds for $\langle f_k(\hat{\Sigma}) - f(\Sigma), B\rangle$. This leads to the following result.

For given $d > 1, a > 0$ and $\sigma_0^2 > 0$, let $\mathcal{S}_{f,B}(d; a; \sigma_0)$ be the set of all covariance operators in d-dimensional space $\mathbb{H}$ such that $\|\Sigma\| \leq a$, $\|\Sigma^{-1}\| \leq a$ and $\sigma_f^2(\Sigma; B) \geq \sigma_0^2$.

Theorem 5. *Suppose that, for some $\alpha \in (0,1)$, $1 \leq d_n \leq n^\alpha, n \geq 1$. Let $B = B_n$ be a self-adjoint operator with $\|B\|_1 \leq 1$. Let $f \in B^s_{\infty,1}(\mathbb{R})$ for some $s > \frac{1}{1-\alpha}$, and let k be an integer number such that, for some $\beta \in (0,1]$, $\frac{1}{1-\alpha} < k + 1 + \beta \leq s$. Finally,*

suppose that $\mathcal{S}_{f,B}(d_n; a'; \sigma_0') \neq \varnothing$ *for some* $a' < a, \sigma_0' > \sigma_0$ *and for all large enough* n. *Then* $\langle f_k(\hat{\Sigma}), B\rangle$ *is an asymptotically efficient estimator of* $\langle f(\Sigma), B\rangle$ *with respect to* $\mathcal{S}_{f,B}(d_n; a; \sigma_0)$ *with convergence rate* $\sqrt{n}$ *and variance* $\sigma_f^2(\Sigma; B)$.

Note that for $\alpha \in (0, 1/2)$ and $s > \frac{1}{1-\alpha}$, we can set $k = 0$. In this case, $f_k(\hat{\Sigma}) = f(\hat{\Sigma})$ is a standard plug-in estimator (see also Corollary 2). For $\alpha = \frac{1}{2}$, the assumption $s > 2$ is needed and, to satisfy the condition $k + 1 + \beta > \frac{1}{1-\alpha} = 2$, we should choose $k = 1$. The bias correction becomes nontrivial in this case. For larger values of α, more smoothness of f and more iterations k in the bias reduction method are needed.

5 Wishart operators and orthogonally invariant functions

In this section, we outline our approach to the proof of Theorem 5 (see Koltchinskii [2017] for further details). The idea is to represent function f in the form $f(x) = x\psi'(x), x \in \mathbb{R}$, where ψ is a smooth function in the real line. Consider now the functional $g(\Sigma) := \mathrm{tr}(\psi(\Sigma))$. Then, g is Fréchet differentiable with derivative $Dg(\Sigma) = \psi'(\Sigma)$ and

$$f(\Sigma) = \Sigma^{1/2} Dg(\Sigma) \Sigma^{1/2} =: \mathfrak{D}g(\Sigma). \tag{10}$$

The functional $g(\Sigma)$ is orthogonally invariant which allowed us to develop integral representations of functions $\mathfrak{B}^k \mathfrak{D}g(\Sigma)$ and use them to study analytic properties of functions $\mathfrak{B}^k f(\Sigma)$ and $f_k(\Sigma)$.

As in the previous section, we assume that $\mathbb{H}$ is a finite-dimensional inner product space of dimension $\dim(\mathbb{H}) = d$. Recall that $\mathcal{T} g(\Sigma) = \mathbb{E}_\Sigma g(\hat{\Sigma}), \Sigma \in \mathcal{C}_+(\mathbb{H})$. We will view $\mathcal{T}$ as an operator from the space $L_\infty(\mathcal{C}_+(\mathbb{H}))$ into itself, $L_\infty(\mathcal{C}_+(\mathbb{H}))$ being the space of uniformly bounded Borel measurable real valued functions on the cone $\mathcal{C}_+(\mathbb{H})$. Alternatively, $\mathcal{T}$ can be viewed as an operator from $L_\infty(\mathcal{C}_+(\mathbb{H}); \mathcal{B}_{sa}(\mathbb{H}))$ into $L_\infty(\mathcal{C}_+(\mathbb{H}); \mathcal{B}_{sa}(\mathbb{H}))$ (the space of uniformly bounded Borel measurable functions from $\mathcal{C}_+(\mathbb{H})$ into $\mathcal{B}_{sa}(\mathbb{H})$).

A function $g \in L_\infty(\mathcal{C}_+(\mathbb{H}))$ is called *orthogonally invariant* iff, for all orthogonal transformations U of $\mathbb{H}$, $g(U\Sigma U^{-1}) = g(\Sigma), \Sigma \in \mathcal{C}_+(\mathbb{H})$. Any such function g could be represented as a symmetric function φ of eigenvalues $\lambda_1(\Sigma), \dots, \lambda_d(\Sigma)$ of $\Sigma : g(\Sigma) = \varphi(\lambda_1(\Sigma), \dots, \lambda_d(\Sigma))$. A typical example is orthogonally invariant function $g(\Sigma) = \mathrm{tr}(\psi(\Sigma)) = \sum_{j=1}^d \psi(\lambda_j(\Sigma))$ for a function of real variable ψ. Denote by $L_\infty^O(\mathcal{C}_+(\mathbb{H}))$ the subspace of all orthogonally invariant functions from $L_\infty(\mathcal{C}_+(\mathbb{H}))$. Clearly, $L_\infty^O(\mathcal{C}_+(\mathbb{H}))$ is an algebra. It is easy to see that operators $\mathcal{T}$ and $\mathcal{B} = \mathcal{T} - \mathcal{I}$ map the space $L_\infty^O(\mathcal{C}_+(\mathbb{H}))$ into itself.

A function $g \in L_\infty(\mathcal{C}_+(\mathbb{H}); \mathcal{B}_{sa}(\mathbb{H})$ is called *orthogonally equivariant* iff, for all orthogonal transformations U, $g(U\Sigma U^{-1}) = Ug(\Sigma)U^{-1}, \Sigma \in \mathcal{C}_+(\mathbb{H})$. A function $g : \mathcal{C}_+(\mathbb{H}) \mapsto \mathcal{B}_{sa}(\mathbb{H})$ will be called differentiable (continuously differentiable) in $\mathcal{C}_+(\mathbb{H})$

with respect to the operator norm iff there exists a uniformly bounded, Lipschitz and differentiable (continuously differentiable) extension of g to an open set G, $\mathcal{C}_+(\mathbb{H}) \subset G \subset \mathcal{B}_{sa}(\mathbb{H})$. If $g : \mathcal{C}_+(\mathbb{H}) \mapsto \mathbb{R}$ is orthogonally invariant and continuously differentiable in $\mathcal{C}_+(\mathbb{H})$ with derivative Dg, then it is easy to check that Dg is orthogonally equivariant.

We will use some simple properties of operators $\mathcal{T}$ and $\mathcal{B} = \mathcal{T} - \mathcal{I}$ acting in the space $L_\infty^O(\mathcal{C}_+(\mathbb{H}))$ of uniformly bounded orthogonally invariant functions (and its subspaces). These properties are well known in the literature on Wishart distribution (at least, in the case of orthogonally invariant polynomials, see, e.g., Letac and Massam [2004]). Define the following differential operator $\mathcal{D}g(\Sigma) := \Sigma^{1/2} Dg(\Sigma)\Sigma^{1/2}$ acting on continuously differentiable functions in $\mathcal{C}_+(\mathbb{H})$. It turns out that operators $\mathcal{T}$ and $\mathcal{D}$ commute (and, as a consequence, $\mathcal{B}$ and $\mathcal{D}$ also commute).

Proposition 1. *If $g \in L_\infty^O(\mathcal{C}_+(\mathbb{H}))$ is continuously differentiable in $\mathcal{C}_+(\mathbb{H})$ with a uniformly bounded derivative Dg, then, for all $\Sigma \in \mathcal{C}_+(\mathbb{H})$, $\mathcal{D}\mathcal{T}g(\Sigma) = \mathcal{T}\mathcal{D}g(\Sigma)$ and $\mathcal{D}\mathcal{B}g(\Sigma) = \mathcal{B}\mathcal{D}g(\Sigma)$.*

Let W be the sample covariance based on i.i.d. standard normal random variables $Z_1, \dots, Z_n$ in $\mathbb{H}$ (in other words, nW has standard Wishart distribution) and let $W_1, \dots, W_k, \dots$ be i.i.d. copies of W. The next proposition provides representations of operators $\mathcal{T}^k$ and $\mathcal{B}^k$ that will be used in what follows.

Proposition 2. *For all $g \in L_\infty^O(\mathcal{C}_+(\mathbb{H}))$ and for all $k \geq 1$,*

$$\mathcal{T}^k g(\Sigma) = \mathbb{E} g(W_k^{1/2} \dots W_1^{1/2} \Sigma W_1^{1/2} \dots W_k^{1/2}) \tag{11}$$

and

$$\mathcal{B}^k g(\Sigma) = \mathbb{E} \sum_{I \subset \{1,\dots,k\}} (-1)^{k-|I|} g(A_I^* \Sigma A_I), \tag{12}$$

where $A_I := \prod_{i \in I} W_i^{1/2}$. If, in addition, g is continuously differentiable in $\mathcal{C}_+(\mathbb{H})$ with a uniformly bounded derivative Dg, then

$$D\mathcal{B}^k g(\Sigma) = \mathbb{E} \sum_{I \subset \{1,\dots,k\}} (-1)^{k-|I|} A_I \, Dg(A_I^* \Sigma A_I) A_I^*, \tag{13}$$

and, for all $\Sigma \in \mathcal{C}_+(\mathbb{H})$,

$$\mathcal{D}\mathcal{T}^k g(\Sigma) = \mathcal{T}^k \mathcal{D}g(\Sigma) \text{ and } \mathcal{D}\mathcal{B}^k g(\Sigma) = \mathcal{B}^k \mathcal{D}g(\Sigma). \tag{14}$$

Finally,

$$\mathcal{B}^k \mathcal{D}g(\Sigma) = \mathcal{D}\mathcal{B}^k g(\Sigma) = \mathbb{E}\Big(\sum_{I \subset \{1,\dots,k\}} (-1)^{k-|I|} \Sigma^{1/2} A_I \, Dg(A_I^* \Sigma A_I) A_I^* \Sigma^{1/2} \Big). \tag{15}$$

Proof. Note that $\hat{\Sigma} \stackrel{d}{=} \Sigma^{1/2} W \Sigma^{1/2}$. It is easy to check that

$$W^{1/2}\Sigma W^{1/2} = U^{-1}\Sigma^{1/2} W \Sigma^{1/2} U$$

where U is an orthogonal operator. Since g is orthogonally invariant, we have

$$\mathfrak{T} g(\Sigma) = \mathbb{E}_{\Sigma} g(\hat{\Sigma}) = \mathbb{E} g(W^{1/2}\Sigma W^{1/2}). \tag{16}$$

Recall that orthogonal invariance of g implies orthogonal invariance of $\mathfrak{T} g$ and, by induction, of $\mathfrak{T}^k g$ for all $k \geq 1$. Then, also by induction, (16) implies that

$$\mathfrak{T}^k g(\Sigma) = \mathbb{E} g(W_k^{1/2} \dots W_1^{1/2} \Sigma W_1^{1/2} \dots W_k^{1/2}).$$

For $I \subset \{1, \dots, k\}$ with $|I| = \mathrm{card}(I) = j$ and $A_I = \prod_{i \in I} W_i^{1/2}$, it follows that $\mathfrak{T}^j g(\Sigma) = \mathbb{E} g(A_I^* \Sigma A_I)$. In view of (9), we easily get (12). If g is continuously differentiable in $\mathcal{C}_+(\mathbb{H})$ with a uniformly bounded derivative Dg, then (12) implies (13). Finally, it follows from (13) that the derivatives $D\mathfrak{B}^k g, k \geq 1$ are continuous and uniformly bounded in $\mathcal{C}_+(\mathbb{H})$. Similar property holds for the derivatives $D\mathfrak{T}^k g, k \geq 1$ (as a consequence of (11) and the properties of g). Therefore, (14) follows from Proposition 1 by induction. Formula (15) follows from (14) and (13).

□ □

The following functions provide the linear interpolation between the identity operator I and operators $W_1^{1/2}, \dots, W_k^{1/2}$:

$$V_j(t_j) := I + t_j(W_j^{1/2} - I), t_j \in [0,1], 1 \leq j \leq k.$$

Note that for all $j = 1, \dots, k, t_j \in [0,1]$, $V_j(t_j) \in \mathcal{C}_+(\mathbb{H})$. Let

$$R = R(t_1, \dots, t_k) = V_1(t_1) \dots V_k(t_k), \quad L = L(t_1, \dots, t_k) = V_k(t_k) \dots V_1(t_1) = R^*$$

and define

$$S = S(t_1, \dots, t_k) = L(t_1, \dots, t_k)\Sigma R(t_1, \dots, t_k), (t_1, \dots, t_k) \in [0,1]^k,$$

$$\varphi(t_1, \dots, t_k) := \Sigma^{1/2} R(t_1, \dots, t_k) Dg(S(t_1, \dots, t_k)) L(t_1, \dots, t_k)\Sigma^{1/2}, (t_1, \dots, t_k) \in [0,1]^k.$$

The following representation is basic in the analysis of functions $\mathfrak{B}^k \mathfrak{D} g(\Sigma)$.

Proposition 3. *Suppose $g \in L_\infty^O(\mathcal{C}_+(\mathbb{H}))$ is $k+1$ times continuously differentiable function with uniformly bounded derivatives $D^j g$, $j = 1, \dots, k+1$. Then the function φ is k times continuously differentiable in $[0,1]^k$ and*

$$\mathfrak{B}^k \mathfrak{D} g(\Sigma) = \mathbb{E} \int_0^1 \cdots \int_0^1 \frac{\partial^k \varphi(t_1, \dots, t_k)}{\partial t_1 \dots \partial t_k} dt_1 \dots dt_k. \tag{17}$$

Proof. For $\phi : [0,1]^k \mapsto \mathbb{R}$, define finite difference operators

$$\Delta_i \phi(t_1, \dots, t_k) := \phi(t_1, \dots, t_{i-1}, 1, t_{i+1}, \dots, t_k) - \phi(t_1, \dots, t_{i-1}, 0, t_{i+1}, \dots, t_k).$$

Then $\Delta_1 \dots \Delta_k \phi$ is given by the following formula

$$\Delta_1 \dots \Delta_k \phi = \sum_{(t_1,\dots,t_k)\in\{0,1\}^k} (-1)^{k-(t_1+\dots+t_k)} \phi(t_1, \dots, t_k). \tag{18}$$

It is well known that, if ϕ is k times continuously differentiable in $[0,1]^k$, then

$$\Delta_1 \dots \Delta_k \phi = \int_0^1 \cdots \int_0^1 \frac{\partial^k \phi(t_1, \dots, t_k)}{\partial t_1 \dots \partial t_k} dt_1 \dots dt_k. \tag{19}$$

Formula (19) also holds for vector- and operator-valued functions ϕ. Identities (15) and (18) imply that

$$\mathfrak{B}^k \mathfrak{D} g(\Sigma) = \mathbb{E} \Delta_1 \dots \Delta_k \varphi. \tag{20}$$

Since Dg is k times continuously differentiable and functions $S(t_1, \dots, t_k)$, $R(t_1, \dots, t_k)$ are polynomials with respect to $t_1, \dots, t_k$, the function φ is k times continuously differentiable in $[0,1]^k$. Representation (17) follows from (20) and (19). □ □

Representation (17) implies a bound on $\|\mathfrak{B}^k \mathfrak{D} g(\Sigma)\|$ of the order $O\big(\big(\frac{d}{n}\big)^{k/2}\big)$.

Theorem 6. *Suppose $k \le d \le n$ and let $g \in L_\infty^O(\mathcal{C}_+(\mathbb{H}))$ be a $k+1$ times continuously differentiable function with uniformly bounded derivatives $D^j g$, $j = 1, \dots, k+1$. Then the following bound holds for some constant $C > 0$:*

$$\|\mathfrak{B}^k \mathfrak{D} g(\Sigma)\| \le C^{k^2} \max_{1\le j\le k+1} \|D^j g\|_{L_\infty} (\|\Sigma\|^{k+1} \vee \|\Sigma\|) \left(\frac{d}{n}\right)^{k/2}, \ \Sigma \in \mathcal{C}_+(\mathbb{H}).^1 \tag{21}$$

The proof is based on deriving the following bound on the partial derivative $\frac{\partial^k \varphi(t_1,\dots,t_k)}{\partial t_1 \dots \partial t_k}$ in (17):

$$\left\| \frac{\partial^k \varphi(t_1, \dots, t_k)}{\partial t_1 \dots \partial t_k} \right\| \le 3^k 2^{k(2k+1)} \max_{1\le j\le k+1} \|D^j g\|_{L_\infty} (\|\Sigma\|^{k+1} \vee \|\Sigma\|) \prod_{i=1}^k (1+\delta_i)^{2k+1} \delta_i, \tag{22}$$

[1]Note that j-th derivative $D^j g(\Sigma)$ can be viewed as symmetric j-linear form $D^j g(\Sigma)(H_1, \dots, H_j)$, $H_1, \dots, H_j \in \mathfrak{B}_{sa}(\mathbb{H})$. The space of such j-linear forms $M(H_1, \dots, H_j)$ is equipped with operator norm: $\|M\| := \sup_{\|H_1\|,\dots,\|H_j\|\le 1} |M(H_1, \dots, H_j)|$. The L_∞-norm $\|D^j g\|_{L_\infty}$ is then defined as $\|D^j g\|_{L_\infty} := \sup_{\Sigma\in\mathcal{C}_+(\mathbb{H})} \|D^j g(\Sigma)\|$.

where $\delta_i := \|W_i - I\|$. Substituting (22) in (17), using independence of r.v. δ_i and bound (4), one can complete the proof.

Representation (17) can be also used to study differentiability of function $\mathfrak{B}^k \mathfrak{D} g(\Sigma)$ and to obtain bounds on the remainder of its Taylor expansion. In view of representation (10) and properties of operators $\mathfrak{T}, \mathfrak{B}, \mathfrak{D}$ (see Proposition 2), this could be further used to prove concentration bounds for the remainder of first order Taylor expansion $\langle S_{f_k}(\Sigma; \hat{\Sigma} - \Sigma), B\rangle$, to prove normal approximation bounds for $\langle f_k(\hat{\Sigma}) - \mathbb{E}_\Sigma f_k(\hat{\Sigma}), B\rangle$ and bounds on the bias $\langle \mathbb{E}_\Sigma f_k(\hat{\Sigma}) - f(\Sigma), B\rangle$, leading to the proof of Theorem 5 (see Koltchinskii [2017] for more details).

6 Open Problems

We discuss below several open problems related to estimation of smooth functionals of covariance.

1. It would be of interest to study asymptotically efficient estimation of functionals $\langle f(\Sigma), B\rangle$ in a dimension-free framework with effective rank playing the role of complexity parameter and in the classes of covariance operators not necessarily of isotropic type. The question is whether a version of Theorem 5 holds for the class $\mathfrak{S}_{f,B}(r_n; a; \sigma_0)$ (instead of $\mathfrak{S}_{f,B}(d_n; a; \sigma_0)$) with $r_n \le n^\alpha, \alpha \in (0,1)$. The main difficulty is to understand how to control k-th order difference $\sum_{j=0}^k (-1)^{k-j} \binom{k}{j} f(\hat{\Sigma}^{(j)})$ of smooth function f along the trajectory of Bootstrap Chain in this case (compare with the approach outlined in Section 5).

2. Another problem is to understand whether the smoothness threshold $s > \frac{1}{1-\alpha}$ for asymptotically efficient estimation of functionals $\langle f(\Sigma), B\rangle$ is sharp (a similar problem was solved in Ibragimov, Nemirovski, and Khasminskii [1986] and Nemirovski [2000] in the case of Gaussian shift model).

3. It would be also of interest to study minimax optimal convergence rates of estimation of functionals $\langle f(\Sigma), B\rangle$ in the case when the nuclear norm of operator B is not bounded by a constant. This includes, for instance, functionals $\mathrm{tr}(f(\Sigma))$ (the case of $B = I$). In such problems, the $\sqrt{n}$-convergence rate no longer holds (see, for instance, the example of estimation of log-determinant Cai, Liang, and Zhou [2015] for which the rate becomes of the order $\asymp \sqrt{\frac{n}{d}}$). A more general problem is to study minimax optimal convergence rate of estimation of smooth orthogonally invariant functionals of Σ. The Bootstrap Chain bias reduction could be still relevant in such problems.

4. One more problem is to study estimation of smooth functionals under further "complexity" constraints (such as smoothness or sparsity) on the set of possible covariance operators. For instance, if $\mathfrak{S} \subset \mathfrak{C}_+(\mathbb{H})$ is a set of covariance operators and $\mathfrak{M}$ is a family of

finite-dimensional subspaces of $\mathbb{H}$, the complexity of $\mathcal{S}$ could characterized by quantities

$$d_m(\mathcal{S};\mathfrak{M}) := \inf_{L\in\mathfrak{M},\dim(L)\le m}\sup_{\Sigma\in\mathcal{S}}\|\Sigma - P_L\Sigma P_L\|, m\ge 1.$$

Assuming that the dimension $d = \dim(\mathbb{H})$ satisfies the condition $d \le n^{\alpha}$ for some $\alpha > 0$ and $d_m(\mathcal{S};\mathfrak{M}) \lesssim m^{-\beta}$ for some $\beta > 0$, the question is to determine threshold $s(\alpha,\beta)$ such that asymptotically efficient estimation is possible for functionals $\langle f(\Sigma), B\rangle$ of smoothness $s > s(\alpha,\beta)$ (and impossible for some functionals of smoothness $s < s(\alpha,\beta)$).

5. Asymptotically efficient estimator $\langle f_k(\hat{\Sigma}), B\rangle$ in Theorem 5 is based on an approximate solution of Wishart integral equation $\mathcal{T}g(\Sigma) = f(\Sigma)$. The Wishart operator $\mathcal{T}$ is well studied in the literature on Wishart distribution (see, e.g., James [1961], Letac and Massam [2004]). In particular, it is known that zonal polynomials James [1961] are its eigenfuctions. It would be of interest to study other approaches to regularized approximate solution of Wishart equation (and corresponding estimators of such functionals as $\langle f(\Sigma), B\rangle$) that would use more directly the spectral properties of operator $\mathcal{T}$ (and could require the tools from analysis on symmetric cones Faraut and Korányi [1994] and Gross and Richards [1987]).

References

A. B. Aleksandrov and V. V. Peller (2016). "Operator Lipschitz functions". *Uspekhi Mat. Nauk* 71.4(430), pp. 3–106. arXiv: 1611.01593. MR: 3588921 (cit. on pp. 2899, 2900).

T. W. Anderson (2003). *An introduction to multivariate statistical analysis*. Third. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, pp. xx+721. MR: 1990662 (cit. on p. 2900).

Z. D. Bai and Jack W. Silverstein (2004). "CLT for linear spectral statistics of large-dimensional sample covariance matrices". *Ann. Probab.* 32.1A, pp. 553–605. MR: 2040792 (cit. on p. 2894).

Peter J. Bickel, Chris A. J. Klaassen, Ya'acov Ritov, and Jon A. Wellner (1993). *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins Series in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, pp. xxii+560. MR: 1245941 (cit. on p. 2893).

T. Tony Cai, Tengyuan Liang, and Harrison H. Zhou (2015). "Law of log determinant of sample covariance matrix and optimal estimation of differential entropy for high-dimensional Gaussian distributions". *J. Multivariate Anal.* 137, pp. 161–172. MR: 3332804 (cit. on pp. 2894, 2908).

T. Tony Cai and Mark G. Low (2005a). "Nonquadratic estimators of a quadratic functional". *Ann. Statist.* 33.6, pp. 2930–2956. MR: 2253108 (cit. on p. 2893).

T. Tony Cai and Mark G. Low (2005b). "On adaptive estimation of linear functionals". *Ann. Statist.* 33.5, pp. 2311–2343. MR: 2211088 (cit. on p. 2893).

Olivier Collier, Laëtitia Comminges, and Alexandre B. Tsybakov (2017). "Minimax estimation of linear and quadratic functionals on sparsity classes". *Ann. Statist.* 45.3, pp. 923–958. MR: 3662444 (cit. on p. 2893).

Jianqing Fan, Philippe Rigollet, and Weichen Wang (2015). "Estimation of functionals of sparse covariance matrices". *Ann. Statist.* 43.6, pp. 2706–2737. MR: 3405609 (cit. on p. 2894).

Jacques Faraut and Adam Korányi (1994). *Analysis on symmetric cones*. Oxford Mathematical Monographs. Oxford Science Publications. The Clarendon Press, Oxford University Press, New York, pp. xii+382. MR: 1446489 (cit. on p. 2909).

R.A. Fisher (1922). "On the mathematical foundation of theoretical statistics". *Philosophical Transactions of the Royal Society of London, Series A* 222, pp. 309–368 (cit. on p. 2892).

– (1925). "Theory of statistical estimation". *Proceedings of the Cambridge Philosophical Society* 22, pp. 700–725 (cit. on p. 2892).

Chao Gao and Harrison H. Zhou (2016). "Bernstein–von Mises theorems for functionals of the covariance matrix". *Electron. J. Stat.* 10.2, pp. 1751–1806. MR: 3522660 (cit. on p. 2894).

S. van de Geer, P. Bühlmann, Ya. Ritov, and R. Dezeure (2014). "On asymptotically optimal confidence regions and tests for high-dimensional models". *Ann. Statist.* 42.3, pp. 1166–1202. MR: 3224285 (cit. on p. 2893).

R. D. Gill and B. Y. Levit (1995). "Applications of the Van Trees inequality: a Bayesian Cramér-Rao bound". *Bernoulli* 1.1-2, pp. 59–79. MR: 1354456 (cit. on p. 2892).

Evarist Giné and Richard Nickl (2016). *Mathematical foundations of infinite-dimensional statistical models*. Cambridge Series in Statistical and Probabilistic Mathematics, [40]. Cambridge University Press, New York, pp. xiv+690. MR: 3588285 (cit. on p. 2893).

V. L. Girko (1987). "An introduction to general statistical analysis". *Teor. Veroyatnost. i Primenen.* 32.2, pp. 252–265. MR: 902754 (cit. on p. 2894).

– (1995). *Statistical analysis of observations of increasing dimension*. Vol. 28. Theory and Decision Library. Series B: Mathematical and Statistical Methods. Translated from the Russian. Kluwer Academic Publishers, Dordrecht, pp. xxii+287. MR: 1473719 (cit. on p. 2894).

Kenneth I. Gross and Donald St. P. Richards (1987). "Special functions of matrix argument. I. Algebraic induction, zonal polynomials, and hypergeometric functions". *Trans. Amer. Math. Soc.* 301.2, pp. 781–811. MR: 882715 (cit. on p. 2909).

Jaroslav Hájek (1972). "Local asymptotic minimax and admissibility in estimation". In: *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, (Univ. California, Berkeley, Calif., 1970/1971), Vol. I: Theory of statistics*. Ed. by

Lucien Le Cam, Jerzy Neyman, and Elizabeth L. Scott. Univ. California Press, Berkeley, Calif., pp. 175–194. MR: 0400513 (cit. on p. 2892).

I. A. Ibragimov and R. Z. Khasminskii (1981). *Statistical estimation*. Vol. 16. Applications of Mathematics. Asymptotic theory, Translated from the Russian by Samuel Kotz. Springer-Verlag, New York-Berlin, pp. vii+403. MR: 620321 (cit. on p. 2893).

I. A. Ibragimov, A. S. Nemirovski, and R. Z. Khasminskii (1986). "Some problems of nonparametric estimation in Gaussian white noise". *Teor. Veroyatnost. i Primenen.* 31.3, pp. 451–466. MR: 866866 (cit. on pp. 2893, 2908).

Alan T. James (1961). "Zonal polynomials of the real positive definite symmetric matrices". *Ann. of Math. (2)* 74, pp. 456–469. MR: 0140741 (cit. on pp. 2902, 2909).

J. Janková and S. van de Geer (2016). "Semi-parametric efficiency bounds for high-dimensional models". To appear in *Annals of Statistics* (cit. on p. 2893).

Adel Javanmard and Andrea Montanari (2014). "Hypothesis testing in high-dimensional regression under the Gaussian random design model: asymptotic theory". *IEEE Trans. Inform. Theory* 60.10, pp. 6522–6554. MR: 3265038 (cit. on p. 2893).

Jiantao Jiao, Yanjun Han, and Tsachy Weissman (Sept. 2017). "Bias Correction with Jackknife, Bootstrap, and Taylor Series". arXiv: 1709.06183 (cit. on p. 2902).

Iain M. Johnstone (2001). "On the distribution of the largest eigenvalue in principal components analysis". *Ann. Statist.* 29.2, pp. 295–327. MR: 1863961 (cit. on p. 2895).

Iain M. Johnstone and Arthur Yu Lu (2009). "On consistency and sparsity for principal components analysis in high dimensions". *J. Amer. Statist. Assoc.* 104.486, pp. 682–693. MR: 2751448 (cit. on p. 2895).

Vladimir Koltchinskii (Oct. 2017). "Asymptotically Efficient Estimation of Smooth Functionals of Covariance Operators". arXiv: 1710.09072 (cit. on pp. 2899, 2900, 2902, 2904, 2908).

Vladimir Koltchinskii, Matthias Löffler, and Richard Nickl (Aug. 2017). "Efficient Estimation of Linear Functionals of Principal Components". arXiv: 1708.07642 (cit. on pp. 2896–2898).

Vladimir Koltchinskii and Karim Lounici (2016). "Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance". *Ann. Inst. Henri Poincaré Probab. Stat.* 52.4, pp. 1976–2013. MR: 3573302 (cit. on pp. 2896–2898).

– (2017a). "Concentration inequalities and moment bounds for sample covariance operators". *Bernoulli* 23.1, pp. 110–133. MR: 3556768 (cit. on pp. 2895, 2896).

– (2017b). "New asymptotic results in principal component analysis". *Sankhya A* 79.2, pp. 254–297. MR: 3707422 (cit. on p. 2896).

– (2017c). "Normal approximation and concentration of spectral projectors of sample covariance". *Ann. Statist.* 45.1, pp. 121–157. MR: 3611488 (cit. on p. 2896).

Lucien LeCam (1953). "On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates". *Univ. California Publ. Statist.* 1, pp. 277–329. MR: 0054913 (cit. on p. 2892).

Gérard Letac and Hélène Massam (2004). "All invariant moments of the Wishart distribution". *Scand. J. Statist.* 31.2, pp. 295–318. MR: 2066255 (cit. on pp. 2902, 2905, 2909).

B. Y. Levit (1975). "The efficiency of a certain class of nonparametric estimators". *Teor. Verojatnost. i Primenen.* 20.4, pp. 738–754. MR: 0403052 (cit. on p. 2893).

– (1978). "Asymptotically efficient estimation of nonlinear functionals". *Probl. Peredachi Inf. (Problems of Information Transmission)* 14.3, pp. 65–72. MR: 533450 (cit. on p. 2893).

A. Lytova and L. Pastur (2009). "Central limit theorem for linear eigenvalue statistics of random matrices with independent entries". *Ann. Probab.* 37.5, pp. 1778–1840. MR: 2561434 (cit. on p. 2894).

A. S. Nemirovski (1990). "Necessary conditions for efficient estimation of functionals of a nonparametric signal observed in white noise". *Teor. Veroyatnost. i Primenen.* 35.1, pp. 83–91. MR: 1050056 (cit. on p. 2893).

– (2000). "Topics in non-parametric statistics". In: *Lectures on probability theory and statistics (Saint-Flour, 1998)*. Vol. 1738. Lecture Notes in Math. Springer, Berlin, pp. 85–277. MR: 1775640 (cit. on pp. 2893, 2908).

Debashis Paul (2007). "Asymptotics of sample eigenstructure for a large dimensional spiked covariance model". *Statist. Sinica* 17.4, pp. 1617–1642. MR: 2399865 (cit. on p. 2895).

V. V. Peller (1985). "Hankel operators in the theory of perturbations of unitary and self-adjoint operators". *Funktsional. Anal. i Prilozhen.* 19.2, pp. 37–51, 96. MR: 800919 (cit. on p. 2899).

– (2006). "Multiple operator integrals and higher operator derivatives". *J. Funct. Anal.* 233.2, pp. 515–544. MR: 2214586 (cit. on p. 2899).

Vilmos Totik (1994). "Approximation by Bernstein polynomials". *Amer. J. Math.* 116.4, pp. 995–1018. MR: 1287945 (cit. on p. 2902).

Hans Triebel (1983). *Theory of function spaces*. Vol. 78. Monographs in Mathematics. Birkhäuser Verlag, Basel, p. 284. MR: 781540 (cit. on p. 2899).

Roman Vershynin (2012). "Introduction to the non-asymptotic analysis of random matrices". In: *Compressed sensing*. Cambridge Univ. Press, Cambridge, pp. 210–268. MR: 2963170 (cit. on p. 2895).

Cun-Hui Zhang and Stephanie S. Zhang (2014). "Confidence intervals for low dimensional parameters in high dimensional linear models". *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 76.1, pp. 217–242. MR: 3153940 (cit. on p. 2893).

Received 2017-11-29.

Vladimir Koltchinskii
School of Mathematics
Georgia Institute of Technology
Atlanta, GA 30332-0160
USA
vlad@math.gatech.edu