Asymptotic Efficiency in High-Dimensional Covariance Estimation

Vladimir Koltchinskii

School of Mathematics
Georgia Institute of Technology

ICM 2018
Rio de Janeiro

Estimation of Smooth Functionals

- $X^{(n)} \sim P_\theta^{(n)}, \theta \in \Theta$;
- for instance, $X^{(n)} = (X_1, \dots, X_n)$, $X_1, \dots, X_n$ i.i.d. $\sim P_\theta, \theta \in \Theta$;
- Θ a subset of a linear normed space;
- $f : \Theta \mapsto \mathbb{R}$ a (smooth) functional (more generally, a sequence of smooth functionals $f_n : \Theta \mapsto \mathbb{R}$);
- $f(\theta)$ to be estimated based on $X^{(n)}$;
- $\mathcal{L}$ 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 $\mathbb{R}_+$;
  - for some $c > 0$, $\ell(t) = O(e^{c|t|})$ as $t \to \infty$.

Asymptotic Efficiency: Definition

An estimator $T_n = T_n(X^{(n)})$ is *asymptotically efficient* for $\Theta_n \subset \Theta$, $n \geq 1$ with convergence rate $\sqrt{n}$ and (limit) variance $\sigma_f^2(\theta) > 0$ iff

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

$$\forall \ell \in \mathcal{L} \quad \sup_{\theta \in \Theta_n} \left| \mathbb{E}_\theta \ell \left(\frac{n^{1/2}(T_n(X^{(n)}) - f(\theta))}{\sigma_f(\theta)} \right) - \mathbb{E}\ell(Z) \right| \to 0 \text{ as } n \to \infty$$

and

$$\liminf_{n \to \infty} \inf_{\tilde{T}_n} \sup_{\theta \in \Theta_n} \frac{n \mathbb{E}_\theta (\tilde{T}_n(X^{(n)}) - f(\theta))^2}{\sigma_f^2(\theta)} \geq 1.$$

R.A. Fisher

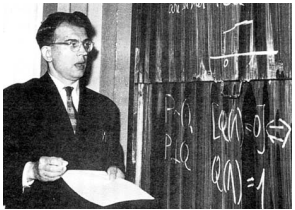
- The idea of asymptotic efficiency goes back to Ronald Aylmer Fisher (1922, 1925).

R.A. Fisher

- Fisher conjectured ("Fisher's program") that, for $\Theta \subset \mathbb{R}$, the maximum likelihood estimator (MLE) $\hat{\theta}_n$ is asymptotically efficient with convergence rate $\sqrt{n}$ and limit variance $I(\theta)^{-1}$, where $I(\theta)$ is the Fisher information.
- This happened to be not true for Fisher's initial definition of asymptotic efficiency even in the case of normal model $X_1, \ldots, X_n$ i.i.d. $\sim N(\theta; 1), \theta \in \mathbb{R}$ (Hodges' example of a superefficient estimator).

L. Le Cam and J. Hájek

- Contemporary definition of asymptotic efficiency is due, primarily, to Lucien Le Cam and Jaroslav Hájek (in the 50s-70s).



L. Le Cam and J. Hájek

- It follows from Hájek- Le Cam theory that, if Θ is an open subset of $\mathbb{R}^d$ and proper regularity assumptions hold (quadratic mean differentiability with nonsingular Fisher information matrix $I(\theta)$, Lipschitz condition on $\theta \mapsto p_\theta$), then $f(\hat{\theta}_n)$ ($\hat{\theta}_n$ being MLE) is an asymptotically efficient estimator of $f(\theta)$ for a smooth functional $f : \Theta \mapsto \mathbb{R}$ with convergence rate $\sqrt{n}$ and limit variance

$$\sigma_f^2(\theta) := \langle I(\theta)^{-1} f'(\theta), f'(\theta) \rangle.$$

Estimation of functionals in nonparametric and high-dimensional models

- For special models, such as nonparametric Gaussian shift model or nonparametric density estimation; often, for special functionals, including linear and quadratic.
  - Levit (1975, 1978), Ibragimov and Khasminskii (1981), Ibragimov, Nemirovski and Khasminskii (1987), Donoho and Liu (1987, 1991), Bickel and Ritov (1988), Donoho and Nussbaum (1990), Nemirovski (1990, 2000), Birgé and Massart (1995), Laurent (1996), Cai and Low (2005), Klemela (2006)
- Semiparametric efficiency.
  - Book by Bickel, Klaassen, Ritov and Wellner (1993)

Estimation of functionals in nonparametric and high-dimensional models

- High-dimensional models.
  - Semi-parametric efficiency of regularization based estimators: van de Geer, Bühlmann, Ritov and Dezeure (2014), Javanmard and Montanari (2014), C.-H. Zhang and S.S. Zhang (2014), Jankova and van de Geer (2016)
  - minimax optimal rates of estimation of special functionals in high-dimensional models (under sparsity and other complexity assumptions): Cai and Low (2005), Collier, Comminges and Tsybakov (2017)

Gaussian shift model: Ibragimov, Nemirovski and Khasminskii (1987), Nemirovski (1990, 2000)

Gaussian shift model: Ibragimov, Nemirovski and Khasminskii (1987), Nemirovski (1990, 2000)

- Gaussian shift model:

$$dX^{(n)}(t) = \theta(t)dt + \frac{1}{\sqrt{n}}dw(t), t \in [0,1],\ \theta \in \Theta \subset L_2([0,1]).$$

- Kolmogorov diameters:

$$d_m(\Theta) := \inf_{L \subset L_2([0,1]), \dim(L) \leq m} \sup_{\theta \in \Theta} \|\theta - P_L\theta\|.$$

Gaussian shift model: Ibragimov, Nemirovski and Khasminskii (1987), Nemirovski (1990, 2000)

- Assumptions on Θ :
  - $\Theta \subset U := \{\theta : \|\theta\| \leq 1\}$
  - For some $\beta > 0$,

$$d_m(\Theta) \lesssim m^{-\beta}, m \geq 1.$$

- Problem: Let $f : \Theta \mapsto \mathbb{R}$ be a functional of "smoothness" $s > 0$. Is there a threshold $s(\beta) > 0$ such that efficient estimation of $f(\theta)$ with $\sqrt{n}$-rate is possible for $s > s(\beta)$?

Smoothness

- For a symmetric k-linear form $M(h_1, \dots, h_k), h_1, \dots, h_k \in L_2([0,1])$, let $\|M\|$ be its operator norm, $\|M\|_{HS}$ be its Hilbert–Schmidt norm and, for $0 \le j \le k$, let $\|M\|_{(j)}$ be its "mixed" norm

$$\|M\|_{(j)} := \sup_{\|h_1\| \le 1, \dots, \|h_j\| \le 1} \|M(h_1, \dots, h_j, \cdot, \dots, \cdot)\|_{HS}.$$

- Let $s := k + \gamma, \gamma \in (0,1], k \ge 0$. For k times Frèchet differentiable f, denote

$$\|f\|_{\tilde{C}^s} := \max_{0 \le j \le k-1} \sup_{\theta \in U} \|f^{(j)}(\theta)\|_{HS} \bigvee \sup_{\theta \in U} \|f^{(k)}(\theta)\|_{(1)}$$
$$\bigvee \sup_{\theta, \theta' \in U} \frac{\|f^{(k)}(\theta) - f^{(k)}(\theta')\|}{\|\theta - \theta'\|^{\gamma}}.$$

- If $k \le 2$, $\|f\|_{\tilde{C}^s} = \|f\|_{C^s}$.

Efficient estimation of smooth functionals

Theorem (Ibragimov-Nemirovski-Khasminskii)

Let $s := k + \gamma, \gamma \in (0,1], k \geq 0$ *and* $\|f\|_{\tilde{C}^s} < \infty$. *If either*

$$k \leq 2 \text{ and } s > \frac{1}{2\beta} + 1, \text{ or } k \geq 3 \text{ and } s > \frac{1}{2\beta},$$

then there exists an asymptotically efficient estimator of $f(\theta)$ *with convergence rate* $\sqrt{n}$ *and limit variance* $\sigma_f^2(\theta) := \|f'(\theta)\|^2$.

Efficient estimation of smooth functionals

- The estimation method was based on development of unbiased estimators of Hilbert–Schmidt polynomials and on Taylor expansion of $f(\theta)$ around an estimator $\hat{\theta}$ with optimal non-parametric error rate.
- Nemirovski (1990, 2000) proved that the smoothness thresholds for efficient estimation in the above theorem are sharp.

Asymptotic normality for functionals of covariance

- Girko (1987–): asymptotically normal estimators of a number of special functionals (such as $\log\det(\Sigma) = \mathrm{tr}(\log\Sigma)$, Stieltjes transform of spectral function of Σ : $\mathrm{tr}((I + t\Sigma)^{-1})$), ... Based on martingale CLT; also continued by Serdobolski in the 90s.
- Asymptotic normality of log-determinant $\log\det(\hat{\Sigma})$ has been studied by many authors (see, e.g., Cai, Liang and Zhou (2015) for a recent result)
- Asymptotic normality of $\mathrm{tr}(f(\hat{\Sigma}))$ for a smooth function $f : \mathbb{R} \mapsto \mathbb{R}$: (linear spectral statistic). Common topic in random matrix theory (both for Wigner and for Wishart matrices): Bai and Silverstein (2004), Lytova and Pastur (2009), Sosoe and Wong (2015)
- Efficient estimation of linear functionals of principal components: Koltchinskii, Löffler and Nickl (2017)

Smooth functionals of covariance (Koltchinskii (2017))

Let $X, X_1, \dots, X_n$ be i.i.d. Gaussian vectors with values in $\mathbb{R}^d$, with $\mathbb{E}X = 0$ and with covariance operator $\Sigma = \mathbb{E}(X \otimes X) \in \mathcal{C}_+^d$, $\mathcal{C}_+^d$ being the cone of positively semi-definite operators (covariance operators) in $\mathbb{R}^d$.

- Given a smooth function $f : \mathbb{R} \mapsto \mathbb{R}$ and a linear operator $B : \mathbb{R}^d \mapsto \mathbb{R}^d$ with $\|B\|_1 \le 1$, estimate $\langle f(\Sigma), B\rangle$ based on $X_1, \dots, X_n$.
- More precisely, we are interested in finding asymptotically efficient estimators of $\langle f(\Sigma), B\rangle$ with $\sqrt{n}$-convergence rate in the case when $d = d_n \to \infty$.
- Suppose $d_n \le n^\alpha$ for some $\alpha > 0$. Is there $s(\alpha)$ such that for all $s > s(\alpha)$ and for all functions f of smoothness s, asymptotically efficient estimation is possible?

Sample Covariance Operator

- Let
$$\hat{\Sigma} := n^{-1} \sum_{j=1}^{n} X_j \otimes X_j$$
be the sample covariance based on $(X_1, \dots, X_n)$.
- Let
$$\mathcal{S}_{a,d} := \left\{ \Sigma \in \mathcal{C}_+^d : a^{-1} I_d \preceq \Sigma \preceq a I_d \right\}, a > 1.$$

- If $\Sigma \in \mathcal{S}_{a,d}$, then
$$\mathbb{E}\|\hat{\Sigma} - \Sigma\| \asymp_a \|\Sigma\| \left(\sqrt{\frac{d}{n}} \bigvee \frac{d}{n} \right)$$
and, for all $t \geq 1$ with probability at least $1 - e^{-t}$,
$$\|\hat{\Sigma} - \Sigma\| \lesssim_a \|\Sigma\| \left(\sqrt{\frac{d}{n}} \bigvee \frac{d}{n} \bigvee \sqrt{\frac{t}{n}} \bigvee \frac{t}{n} \right).$$

Operator Differentiability

- Let $f \in C^1(\mathbb{R})$ and let $f^{[1]}(\lambda, \mu)$ be the Loewner kernel:

$$f^{[1]}(\lambda, \mu) := \frac{f(\lambda) - f(\mu)}{\lambda - \mu}, \ \lambda \neq \mu; \quad f^{[1]}(\lambda, \lambda) := f'(\lambda).$$

- $A \mapsto f(A)$ is Fréchet differentiable at $A = \sum_{\lambda \in \sigma(A)} \lambda P_\lambda$ with derivative

$$Df(A; H) = \sum_{\lambda, \mu \in \sigma(A)} f^{[1]}(\lambda, \mu) P_\lambda H P_\mu.$$

Limit Variance and Loss Functions

- Let
$$\sigma_f^2(\Sigma; B) := 2\|\Sigma^{1/2} Df(\Sigma; B)\Sigma^{1/2}\|_2^2.$$

- Loss functions. Recall that $\mathcal{L}$ is the class of functions $\ell : \mathbb{R} \mapsto \mathbb{R}_+$ such that
  - $\ell(0) = 0$
  - $\ell(-t) = \ell(t), t \in \mathbb{R}$
  - ℓ is convex and nondecreasing on $\mathbb{R}_+$
  - For some $c > 0$, $\ell(t) = O(e^{ct})$ as $t \to \infty$

Efficient Estimation of $\langle f(\Sigma), B\rangle$

Theorem

Suppose, for some $\alpha \in (0,1)$, $d_n \le n^\alpha$ and for some $s > \frac{1}{1-\alpha}$, $f \in B^s_{\infty,1}(\mathbb{R})$. Then, there exists an estimator $h(\hat{\Sigma})$ such that for all $\sigma_0 > 0$

$$\sup_{\Sigma,B}\sup_{x\in\mathbb{R}}\left|\mathbb{P}_\Sigma\left\{\frac{n^{1/2}\Big(\langle h(\hat{\Sigma}),B\rangle - \langle f(\Sigma),B\rangle\Big)}{\sigma_f(\Sigma,B)} \le x\right\} - \Phi(x)\right| \to 0$$

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

$$\sup_{\Sigma,B}\left|\mathbb{E}_\Sigma \ell\left(\frac{n^{1/2}\Big(\langle h(\hat{\Sigma}),B\rangle - \langle f(\Sigma),B\rangle\Big)}{\sigma_f(\Sigma,B)}\right) - \mathbb{E}\ell(Z)\right| \to 0,$$

where suprema are taken over all Σ, B such that $\Sigma \in \mathcal{S}_{a,d_n}$, $\|B\|_1 \le 1$, $\sigma_f(\Sigma;B) \ge \sigma_0$.

Efficient Estimation of $\langle f(\Sigma), B\rangle$: A Lower Bound

Theorem

Let $f \in B^1_{\infty,1}(\mathbb{R})$ and let $\{B_n\}$ be a sequence of operators with $\|B_n\|_1 \le 1$. Suppose $d_n \ge 1, a > 1, \sigma_0 > 0$ are such that, for some $1 < a' < a$ and $\sigma_0' > \sigma_0$ and for all large enough n,

$$\left\{\Sigma \in \mathcal{S}_{a',d_n}, \sigma_f(\Sigma; B_n) \ge \sigma_0'\right\} \neq \emptyset.$$

Then, the following bound holds:

$$\liminf_n \inf_{T_n} \sup_{\Sigma \in \mathcal{S}_{a,d_n}, \sigma_f(\Sigma;B_n) \ge \sigma_0} \frac{n\mathbb{E}_\Sigma\Big(T_n(X_1,\dots,X_n) - \langle f(\Sigma), B_n\rangle\Big)^2}{\sigma_f^2(\Sigma; B_n)} \ge 1,$$

where the infimum is taken over all estimators $T_n = T_n(X_1, \dots, X_n)$.

Operator Theory Tools: Bounds on the Remainder of Taylor Expansion for Operator Functions

Lemma

Let

$$S_f(A;H) = f(A+H) - f(A) - (Df)(A;H)$$

be the remainder of differentiation. If, for some $s \in [1,2]$, $f \in B^s_{\infty,1}(\mathbb{R})$, *then the following bounds hold:*

$$\|S_f(A;H)\| \le 2^{3-s}\|f\|_{B^s_{\infty,1}}\|H\|^s$$

and

$$\|S_f(A;H) - S_f(A;H')\| \le 2^{1+s}\|f\|_{B^s_{\infty,1}}(\|H\| \vee \|H'\|)^{s-1}\|H' - H\|.$$

The proof is based on Littlewood-Paley decomposition of f and on operator versions of Bernstein inequalities for entire functions of exponential type (as in the work by Peller (1985, 2006), Aleksandrov and Peller (2016) on operator Lipschitz and operator differentiable functions).

Perturbation Theory: Application to Functions of Sample Covariance

- $$\langle f(\hat{\Sigma}) - f(\Sigma), B\rangle = \langle Df(\Sigma; \hat{\Sigma} - \Sigma), B\rangle + \langle S_f(\Sigma; \hat{\Sigma} - \Sigma), B\rangle$$
- The linear term $\langle Df(\Sigma; \hat{\Sigma} - \Sigma), B\rangle$ is of the order $O(n^{-1/2})$ and $n^{1/2}\langle Df(\Sigma; \hat{\Sigma} - \Sigma), B\rangle$ is close in distribution to $N(0; \sigma_f^2(\Sigma; B))$.
- For $s \in (1, 2]$, $\|S_f(\Sigma; \hat{\Sigma} - \Sigma)\| \lesssim \|f\|_{B^s_{\infty,1}} \|\hat{\Sigma} - \Sigma\|^s$, implying that

$$|\langle S_f(\Sigma; \hat{\Sigma} - \Sigma), B\rangle| \le \|B\|_1 \|S_f(\Sigma; \hat{\Sigma} - \Sigma)\|$$

$$= O\Big(\Big(\frac{d}{n}\Big)^{s/2}\Big) = O(n^{(1-\alpha)s/2}) = o(n^{-1/2})$$

and, similarly,

$$|\langle \mathbb{E} f(\hat{\Sigma}) - f(\Sigma), B\rangle| = |\langle \mathbb{E} S_f(\Sigma; \hat{\Sigma} - \Sigma), B\rangle| = o(n^{-1/2})$$

provided that $s > \frac{1}{1-\alpha}, \alpha \in (0, 1/2)$. In this case, $h(\hat{\Sigma}) = f(\hat{\Sigma})$.

Perturbation Theory: Application to Functions of Sample Covariance

- The bounds are sharp, for instance, for $f(x) = x^2, B = u \otimes u, s = 2, d = n^\alpha$

$$\sup_{\|u\| \le 1} |\langle \mathbb{E} f(\hat{\Sigma}) - f(\Sigma), u \otimes u \rangle| = \sup_{\|u\| \le 1} |\langle \mathbb{E} S_f(\Sigma; \hat{\Sigma} - \Sigma), u \otimes u \rangle| =$$

$$= \frac{\|\mathrm{tr}(\Sigma)\Sigma + \Sigma^2\|}{n} \asymp \|\Sigma\|^2 \frac{d}{n} \asymp \|\Sigma\|^2 n^{\alpha - 1}$$

$$\sup_{\|u\| \le 1} |\langle \mathbb{E} f(\hat{\Sigma}) - f(\Sigma), u \otimes u \rangle| = o(n^{-1/2})$$

iff $\alpha < 1/2$.

- What if $d_n \ge n^{1/2}, d_n = o(n)$?

Normal Approximation for Smooth Functions of Sample Covariance

Theorem

Let $f \in B^s_{\infty,1}(\mathbb{R})$ for some $s \in (1,2]$ and let B be a linear operator with $\|B\|_1 \le 1$. Suppose $a > 0, \sigma_0 > 0$ and

$$d_n = o(n) \quad \text{as} \quad n \to \infty.$$

Then

$$\sup_{\Sigma \in \mathcal{S}_{a,d_n}, \sigma_f(\Sigma;B) \ge \sigma_0} \sup_{x \in \mathbb{R}} \left| \mathbb{P}_\Sigma \left\{ \frac{n^{1/2} \left\langle f(\hat{\Sigma}) - \mathbb{E}_\Sigma f(\hat{\Sigma}), B \right\rangle}{\sigma_f(\Sigma, B)} \le x \right\} - \Phi(x) \right| \to 0.$$

Perturbation Theory for Functions of Sample Covariance

- $$\langle f(\hat{\Sigma}) - f(\Sigma), B\rangle = \langle Df(\Sigma; \hat{\Sigma} - \Sigma), B\rangle + \langle S_f(\Sigma; \hat{\Sigma} - \Sigma), B\rangle$$

implies that

$$\langle f(\hat{\Sigma}) - \mathbb{E}f(\hat{\Sigma}), B\rangle =$$

$$= \langle Df(\Sigma)(\hat{\Sigma} - \Sigma), B\rangle + \langle S_f(\Sigma; \hat{\Sigma} - \Sigma) - \mathbb{E}S_f(\Sigma; \hat{\Sigma} - \Sigma), B\rangle$$

$$= \langle Df(\Sigma)(B), \hat{\Sigma} - \Sigma\rangle + \langle S_f(\Sigma; \hat{\Sigma} - \Sigma) - \mathbb{E}S_f(\Sigma; \hat{\Sigma} - \Sigma), B\rangle$$

Perturbation Theory for Functions of Sample Covariance

- The linear term

$$\langle Df(\Sigma)B, \hat{\Sigma} - \Sigma \rangle$$

$$= n^{-1} \sum_{j=1}^{n} \langle Df(\Sigma; B)X_j, X_j \rangle - \mathbb{E}\langle Df(\Sigma, B)X, X \rangle$$

  is of the order $O(n^{-1/2})$ and it is approximated by a normal distribution using Berry-Esseen bound.

- The centered remainder

$$\left\langle S_f(\Sigma; \hat{\Sigma} - \Sigma) - \mathbb{E}S_f(\Sigma; \hat{\Sigma} - \Sigma), B \right\rangle$$

  is of the order $o(n^{-1/2})$ when $d_n = o(n)$ and it is controlled using Gaussian concentration inequalities.

Concentration of the Remainder

Theorem

Suppose that $f \in B^s_{\infty,1}(\mathbb{R})$ and also that $d \lesssim n$. Then, there exists a constant $C = C_s > 0$ such that, for all $t \geq 1$, with probability at least $1 - e^{-t}$

$$\left|\left\langle S_f(\Sigma; \hat{\Sigma} - \Sigma) - \mathbb{E}S_f(\Sigma; \hat{\Sigma} - \Sigma), B \right\rangle\right|$$

$$\leq C\|f\|_{B^s_{\infty,1}}\|B\|_1\|\Sigma\|^s\left(\left(\frac{d}{n}\right)^{(s-1)/2} \bigvee \left(\frac{t}{n}\right)^{(s-1)/2} \bigvee \left(\frac{t}{n}\right)^{s-1/2}\right)\sqrt{\frac{t}{n}}.$$

Note: the centered remainder is $o_{\mathbb{P}}(n^{-1/2})$ provided that

$$d = d_n = o(n).$$

Wishart Operators and Bias Reduction

- Our next goal is to find an estimator $g(\hat{\Sigma})$ of $f(\Sigma)$ with a small bias $\mathbb{E}_\Sigma g(\hat{\Sigma}) - f(\Sigma)$ (of the order $o(n^{-1/2})$) and such that

$$n^{1/2}(\langle g(\hat{\Sigma}), B\rangle - \langle \mathbb{E}_\Sigma g(\hat{\Sigma}), B\rangle)$$

  is asymptotically normal.
- To this end, one has to find a sufficiently smooth approximate solution of the equation

$$\mathbb{E}_\Sigma g(\hat{\Sigma}) = f(\Sigma), \Sigma \in \mathcal{C}_+^d.$$

Wishart Operators and Bias Reduction

- $$\mathcal{T}g(\Sigma) := \mathbb{E}_\Sigma g(\hat{\Sigma}) = \int_{\mathcal{C}_+^d} g(V)P(\Sigma; dV), \Sigma \in \mathcal{C}_+^d$$

- $P(\Sigma; dV)$ is a Markov kernel on $\mathcal{C}_+^d$,

$$P(\Sigma; A) := \mathbb{P}_\Sigma\{\hat{\Sigma} \in A\}, A \subset \mathcal{C}_+^d.$$

- Such operators have been often studied in the theory of Wishart matrices (James (1961), Gross and Richards (1987), Letac and Massam (2004)) and, more generally, in the literature on analysis on symmetric cones (Faraut and Koranyi (1994))

Wishart Operators and Bias Reduction

- To find an estimator of $f(\Sigma)$ with a small bias, one needs to solve (approximately) the following integral equation ("the Wishart equation")

$$\mathcal{T} g(\Sigma) = f(\Sigma), \Sigma \in \mathcal{C}_+^d.$$

- Bias operator: $\mathcal{B} := \mathcal{T} - \mathcal{I}$.
- Formally, the solution of the Wishart equation is given by Neumann series:

$$g(\Sigma) = (\mathcal{I} + \mathcal{B})^{-1} f(\Sigma) = (\mathcal{I} - \mathcal{B} + \mathcal{B}^2 - \dots) f(\Sigma)$$

Wishart Operators and Bias Reduction

- Given a smooth function $f : \mathbb{R} \mapsto \mathbb{R}$, define

$$f_k(\Sigma) := \sum_{j=0}^{k} (-1)^j \mathcal{B}^j f(\Sigma) := f(\Sigma) + \sum_{j=1}^{k} (-1)^j \mathcal{B}^j f(\Sigma)$$

- Then

$$\mathbb{E}_\Sigma f_k(\hat{\Sigma}) - f(\Sigma) = (\mathcal{I} + \mathcal{B}) f_k(\Sigma) - f(\Sigma) = (-1)^k \mathcal{B}^{k+1} f(\Sigma).$$

- Asymptotically efficient estimator is $h(\hat{\Sigma}) = f_k(\hat{\Sigma})$, where k is an integer such that, for some $\beta \in (0, 1]$, $\frac{1}{1-\alpha} < k + 1 + \beta \leq s$.

Bootstrap Chain

- $$\mathcal{T}g(\Sigma) := \mathbb{E}_\Sigma g(\hat{\Sigma}) = \int_{\mathcal{C}_+^d} g(V)P(\Sigma; dV), \Sigma \in \mathcal{C}_+^d$$

- $$\mathcal{T}^k g(\Sigma) = \mathbb{E}_\Sigma g(\hat{\Sigma}^{(k)}), \Sigma \in \mathcal{C}_+^d,$$

  where

  $$\hat{\Sigma}^{(0)} = \Sigma \to \hat{\Sigma}^{(1)} = \hat{\Sigma} \to \hat{\Sigma}^{(2)} \to \dots$$

  is a Markov chain in $\mathcal{C}_+^d$ with transition probability kernel P.
- Note that $\hat{\Sigma}^{(j+1)}$ is the sample covariance based on n i.i.d. observations $\sim N(0; \hat{\Sigma}^{(j)})$ (conditionally on $\hat{\Sigma}^{(j)}$)
- Conditionally on $\hat{\Sigma}^{(j)}$, with a “high probability",

$$\|\hat{\Sigma}^{(j+1)} - \hat{\Sigma}^{(j)}\| \lesssim \|\hat{\Sigma}^{(j)}\|\sqrt{\frac{d}{n}}$$

k-th order difference

- k-th order difference along the Markov chain:

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

- $\sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} f(\hat{\Sigma}^{(j)})$ is the k-th order difference of f along the trajectory of the Bootstrap Chain. For f of smoothness k, is it of the order $\left(\sqrt{\frac{d}{n}}\right)^k$?
- This is similar to the k-th order differences in $\mathbb{R}$: if f is C^k, then

$$\Delta_h^k f(x) = O(h^k) \text{ as } h \to 0,$$

where $\Delta_h f(x) := f(x + h) - f(x)$.

A bound on $\mathcal{B}^k f(\Sigma)$

Theorem

Suppose that $f \in B^k_{\infty,1}(\mathbb{R})$ and that $k \le d \le n$. Then, for some $C > 1$,

$$\|\mathcal{B}^k f(\Sigma)\| \le C^{k^2} \|f\|_{B^k_{\infty,1}} (\|\Sigma\|^{k+1} \vee \|\Sigma\|) \left(\frac{d}{n}\right)^{k/2}.$$

Bounds on the bias of $f_k(\hat{\Sigma})$

Corollary

Suppose $f \in B^{k+1}_{\infty,1}(\mathbb{R})$ and $k+1 \le d \le n$. Then, for some $C > 1$,

$$\|\mathbb{E}_\Sigma f_k(\hat{\Sigma}) - f(\Sigma)\| \le C^{(k+1)^2} \|f\|_{B^{k+1}_{\infty,1}} (\|\Sigma\|^{k+2} \vee \|\Sigma\|) \left(\frac{d}{n}\right)^{(k+1)/2}.$$

If, for some $\alpha \in (1/2, 1)$, $d \le n^\alpha$ and $k+1 > \frac{1}{1-\alpha}$, then

$$\|\mathbb{E}_\Sigma f_k(\hat{\Sigma}) - f(\Sigma)\| = o(n^{-1/2}).$$

Sketch of the proof: reduction to orthogonally invariant functions

$g : \mathcal{C}_+^d \mapsto \mathbb{R}$ is orthogonally invariant function iff for all orthogonal transformations $U : \mathbb{R}^d \mapsto \mathbb{R}^d$:

$$g(U\Sigma U^{-1}) = g(\Sigma)$$

Proposition

If g is orthogonally invariant, then $\mathcal{T}g$ is orthogonally invariant.

Sketch of the proof: reduction to orthogonally invariant functions ("lifting")

Lifting operator: $\mathcal{D}g(\Sigma) := \Sigma^{1/2} Dg(\Sigma)\Sigma^{1/2}$

Proposition (Commutativity)

For all smooth orthogonally invariant g,

$$\mathcal{D}\mathcal{T}^k g = \mathcal{T}^k \mathcal{D}g \text{ and } \mathcal{D}\mathcal{B}^k g = \mathcal{B}^k \mathcal{D}g, k \geq 0.$$

Sketch of the proof: reduction to orthogonally invariant functions ("lifting")

- $f(x) = x\psi'(x)$
- $g(\Sigma) := \mathrm{tr}(\psi(\Sigma))$, g is orthogonally invariant
- $Dg(\Sigma) = \psi'(\Sigma)$
- $\mathcal{D}g(\Sigma) = f(\Sigma)$
- Hence

$$\mathcal{B}^k f(\Sigma) = \mathcal{B}^k \mathcal{D} g(\Sigma) = \mathcal{D}\mathcal{B}^k g(\Sigma)$$

Sketch of the proof: a representation for $\mathcal{T}^k g(\Sigma)$

Proposition

Let $W := n^{-1} \sum_{j=1}^n Z_j \otimes Z_j$, $Z_1, \dots, Z_n$ i.i.d. standard normal in $\mathbb{R}^d$, $W_1, \dots, W_k$ i.i.d. copies of W. Suppose g is orthogonally invariant. Then

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

Proof.

g is orthogonally invariant and $\Sigma^{1/2} W \Sigma^{1/2} = U(W^{1/2} \Sigma W^{1/2}) U^{-1}$ imply $\mathcal{T} g(\Sigma) = \mathbb{E}_\Sigma g(\hat{\Sigma}) = \mathbb{E} g(\Sigma^{1/2} W \Sigma^{1/2}) = \mathbb{E} g(W^{1/2} \Sigma W^{1/2})$.

$$\mathcal{T}^2 g(\Sigma) = \mathcal{T}(\mathcal{T} g(\Sigma)) = \mathbb{E}_{W_1} (\mathcal{T} g)(W_1^{1/2} \Sigma W_1^{1/2})$$
$$= \mathbb{E}_{W_1} \mathbb{E}_{W_2} g(W_2^{1/2} W_1^{1/2} \Sigma W_1^{1/2} W_2^{1/2}) = \mathbb{E} g(W_2^{1/2} W_1^{1/2} \Sigma W_1^{1/2} W_2^{1/2}), \dots$$

Sketch of the proof: a representation for $\mathcal{B}^k g(\Sigma)$

- $$\mathcal{T}^j g(\Sigma) = \mathbb{E} g(A_I^* \Sigma A_I),$$

  where $A_I := \prod_{i \in I} W_i^{1/2}, |I| = j$

- $$\begin{aligned}\mathcal{B}^k g(\Sigma) = (\mathcal{T} - \mathcal{I})^k g(\Sigma) &= \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} \mathcal{T}^j g(\Sigma) \\ &= \sum_{j=0}^{k} (-1)^{k-j} \sum_{I \subset \{1,\dots,k\}, |I|=j} \mathbb{E} g(A_I^* \Sigma A_I) \\ &= \mathbb{E} \sum_{I \subset \{1,\dots,k\}} (-1)^{k-|I|} g(A_I^* \Sigma A_I).\end{aligned}$$

Sketch of the proof: a representation for $\mathcal{B}^k g(\Sigma)$

- Denote $V_j(t_j) := I + t_j(W_j^{1/2} - I), t_j \in [0,1]$
- $S(t_1, \dots, t_k) := V_1(t_1) \dots V_k(t_k)$
- $\phi(t_1, \dots, t_k) := g(S(t_1, \dots, t_k)^* \Sigma S(t_1, \dots, t_k)), (t_1, \dots, t_k) \in [0,1]^k$
- $g(A_I^* \Sigma A_I) = \phi(t_1, \dots, t_k)$, for $(t_1, \dots, t_k) \in \{0,1\}^k, I = \{i : t_i = 1\}$.
-

$$\sum_{I \subset \{1,\dots,k\}} (-1)^{k-|I|} g(A_I^* \Sigma A_I) = \sum_{(t_1,\dots,t_k) \in \{0,1\}^k} (-1)^{k - \sum_{i=1}^k t_i} \phi(t_1, \dots, t_k)$$
$$= \Delta_1 \dots \Delta_k \phi(t_1, \dots, t_k),$$

where $\Delta_i \phi(t_1, \dots, t_k) := \phi(t_1, \dots, 1, \dots, t_k) - \phi(t_1, \dots, 0, \dots, t_k)$.

Sketch of the proof: a representation of $\mathcal{B}^k g(\Sigma)$

- $$\begin{aligned}\mathcal{B}^k g(\Sigma) &= \mathbb{E}\Delta_1 \dots \Delta_k \phi \\ &= \mathbb{E}\int_0^1 \cdots \int_0^1 \frac{\partial^k \phi(t_1, \dots, t_k)}{\partial t_1 \dots \partial t_k} dt_1 \dots dt_k\end{aligned}$$

- Integral representation:

$$\mathcal{B}^k f(\Sigma) = \mathcal{D}\mathcal{B}^k g(\Sigma) = \mathbb{E}\int_0^1 \cdots \int_0^1 \mathcal{D}\frac{\partial^k \phi(t_1, \dots, t_k)}{\partial t_1 \dots \partial t_k} dt_1 \dots dt_k$$

Sketch of the proof: bounding partial derivatives

Lemma

$$\left\| \mathcal{D} \frac{\partial^k \phi(t_1, \dots, t_k)}{\partial t_1 \dots \partial t_k} \right\| \leq$$

$$\leq 3^k 2^{k(2k+1)} \max_{1 \leq j \leq k+1} \|D^j g\|_{L_\infty} (\|\Sigma\|^{k+1} \vee \|\Sigma\|) \prod_{i=1}^{k} \delta_i (1+\delta_i)^{2k+1},$$

where $\delta_i := \|W_i - I\|$.

It implies that

$$\|\mathcal{B}^k f(\Sigma)\| = \|\mathcal{D}\mathcal{B}^k g(\Sigma)\| \leq$$

$$\leq 3^k 2^{k(2k+1)} \max_{1 \leq j \leq k+1} \|D^j g\|_{L_\infty} (\|\Sigma\|^{k+1} \vee \|\Sigma\|) \left(\mathbb{E}\|W - I\| (1 + \|W - I\|)^{2k+1} \right)^k$$

$$\leq C^{k^2} \max_{1 \leq j \leq k+1} \|D^j g\|_{L_\infty} (\|\Sigma\|^{k+1} \vee \|\Sigma\|) \left(\frac{d}{n} \right)^{k/2}.$$

Open Problems

- Optimality of smoothness threshold $s > \frac{1}{1-\alpha}$ for efficient estimation.
- Asymptotically efficient estimation of $\langle f(\Sigma), B\rangle$ in dimension-free framework (with complexity of estimation problem characterized by the "effective rank" $\mathbf{r}(\Sigma) := \frac{\mathrm{tr}(\Sigma)}{\|\Sigma\|}$). See Koltchinskii, Loeffler and Nickl (2017) for the problem of estimation of linear functionals of principal components in this framework.
- Estimation of functionals $\langle f(\Sigma), B\rangle$ without the constraint $\|B\|_1 \leq 1$, in particular, $\mathrm{tr}(f(\Sigma))$.
- Estimation of more general smooth functionals of covariance.
- Estimation of smooth functionals under further structural assumptions on Σ (e.g., sparse models, etc).