Smooth functionals of covariance operators: minimax estimation error bounds and effective rank

Vladimir Koltchinskii

School of Mathematics
Georgia Institute of Technology

Bedlewo, 2023

Covariance and sample covariance operators

- $X \sim N(0, \Sigma)$ a centered Gaussian r.v. in a separable Banach space E with the dual space E^*
- $\Sigma : E^* \mapsto E$ is the covariance operator of X :

$$\Sigma u := \mathbb{E}\langle X, u\rangle X, u \in E^*$$

- $X_1, \dots, X_n$ i.i.d. copies of X
- $\hat{\Sigma} = \hat{\Sigma}_n : E^* \mapsto E$ is the sample covariance operator based on $X_1, \dots, X_n$

$$\hat{\Sigma} u := n^{-1} \sum_{j=1}^{n} \langle X_j, u\rangle X_j, u \in E^*$$

Covariance operators and effective rank

- The effective rank of Σ : for $X \sim N(0, \Sigma)$,

$$\mathbf{r}(\Sigma) := \frac{\mathbb{E}\|X\|^2}{\|\Sigma\|} = \frac{\mathbb{E}\sup_{\|u\|,\|v\|\leq 1}\langle X, u\rangle\langle X, v\rangle}{\sup_{\|u\|,\|v\|\leq 1}\mathbb{E}\langle X, u\rangle\langle X, v\rangle} \geq 1$$

- $\mathbf{r}(\lambda\Sigma) = \mathbf{r}(\Sigma), \lambda > 0$
- $\mathbf{r}(\Sigma) \leq \mathrm{rank}(\Sigma) \leq \dim(E)$
- If $E = \mathbb{H}$ is a Hilbert space, then $\mathbf{r}(\Sigma) = \frac{\mathrm{tr}(\Sigma)}{\|\Sigma\|}$
- If $E = \mathbb{R}^d$ with the Euclidean norm and $\sigma(\Sigma) \subset [a^{-1}, a]$ for some $a \geq 1$ ("almost isotropic covariance"), then $\mathbf{r}(\Sigma) \asymp d$.

Gaussian Version of Dvoretzky's Theorem and Effective Rank

Theorem (Pisier (1986, 1989))

For all $\varepsilon \in (0,1)$, there exists $\eta(\varepsilon) > 0$ with the following property. If, in a Banach space E, there exists $X \sim N(0,\Sigma)$ with $\mathbf{r}(\Sigma) = r$, then E contains a subspace F of dimension $\sim \eta(\varepsilon) r$ which is $(1 \pm \varepsilon)$-isomorphic to ℓ_2^n.

Pisier (1989) called $\mathbf{r}(\Sigma)$ “the dimension of X", or “the concentration dimension" of X.

Bounding the sample covariance via effective rank

Theorem (K& Lounici (2014))

Let $X \sim N(0, \Sigma)$ in E and let $X_1, \dots, X_n$ be i.i.d. copies of X. Then

$$\mathbb{E}\|\hat{\Sigma} - \Sigma\| \asymp \|\Sigma\| \Bigl(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \vee \frac{\mathbf{r}(\Sigma)}{n}\Bigr),$$

and, for all $t \geq 1$ with probability $\geq 1 - e^{-t}$,

$$\Bigl| \|\hat{\Sigma} - \Sigma\| - \mathbb{E}\|\hat{\Sigma} - \Sigma\| \Bigr| \lesssim \|\Sigma\| \Bigl(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \vee 1\Bigr) \sqrt{\frac{t}{n}} \vee \|\Sigma\| \frac{t}{n}.$$

- Earlier results: in the case $E = \mathbb{R}^d$ with the Euclidean norm, bounds with $\log d$-factors based on non-commutative Khintchine inequalities by Lust-Piquard and Pisier (Vershynin, around 2011-2012).
- Alternative proof of concentration bound: Adamczak (2014)

Functionals of covariance

- $L(E^*, E)$ the space of bounded symmetric linear operators from E^* into E equipped with the operator norm
- $f : L(E^*, E) \mapsto \mathbb{R}$ a functional
- $f(\Sigma)$ to be estimated based on i.i.d. $X_1, \dots, X_n \sim N(0, \Sigma)$
- For $a > 0$ and $r \geq 1$,

$$\mathcal{S}(a, r) := \{\Sigma : \|\Sigma\| \leq a, \mathbf{r}(\Sigma) \leq r\}$$

- Find the size of

$$\sup_{\|f\|_{C^s} \leq 1} \inf_{T_n} \sup_{\Sigma \in \mathcal{S}(a,r)} \Big\| T_n(X_1, \dots, X_n) - f(\Sigma) \Big\|_{L_2(\mathbb{P}_\Sigma)}.$$

- Is it possible to construct an estimator with $\sqrt{n}$-rate? Asymptotically normal estimator? Asymptotically efficient estimator?

Hölder smoothness

- F is a Banach space (say, $F = L(E^*, E)$), $\Theta \subset F$ an open subset
- $g : \Theta \mapsto F_1$, F_1 is a Banach space

$$\|g\|_{L_\infty(\Theta)} := \sup_{x \in \Theta} \|g(x)\|, \ \|g\|_{\mathrm{Lip}(\Theta)} := \sup_{x, x' \in \Theta, x \neq x'} \frac{\|g(x) - g(x')\|}{\|x - x'\|}$$

$$\|g\|_{\mathrm{Lip}_\rho(\Theta)} := \sup_{x, x' \in \Theta, x \neq x'} \frac{\|g(x) - g(x')\|}{\|x - x'\|^\rho}, \rho \in (0, 1].$$

- $g : \Theta \mapsto \mathbb{R}$ k times Fréchet differentiable for some $k \geq 0$
- For $s = k + \rho$ with $\rho \in (0, 1]$, define

$$\|g\|_{C^s(\Theta)} := \max\Big(\|g\|_{L_\infty}, \max_{0 \leq j \leq k-1} \|g^{(j)}\|_{\mathrm{Lip}}, \|g^{(k)}\|_{\mathrm{Lip}_\rho}\Big).$$

- Note: norms of the derivatives are defined as the operator norms (of multilinear forms).
- $C^s(\Theta) := \{g : \Theta \mapsto \mathbb{R} : \|g\|_{C^s(\Theta)} < \infty\}$

Example of Hölder smoothness: functions of operators

- $E = E^* = \mathbb{H}$, $\mathbb{H}$ is a Hilbert space
- $\mathcal{L}_{sa}(\mathbb{H})$ is the space of self-adjoint operators in $\mathbb{H}$ equipped with the operator norm
- $g : \mathbb{R} \mapsto \mathbb{R}$ induces $\mathcal{L}_{sa}(\mathbb{H}) \ni A \mapsto g(A) \in \mathcal{L}_{sa}(\mathbb{H})$
- for a compact operator A,

$$A = \sum_{\lambda \in \sigma(A)} \lambda P_\lambda \Longrightarrow g(A) = \sum_{\lambda \in \sigma(A)} g(\lambda) P_\lambda$$

- K (2017):

$$\|g(\cdot)\|_{C^s(\mathcal{L}_{sa}(\mathbb{H}))} \lesssim_s \|g\|_{B^s_{\infty,1}(\mathbb{R})}$$

- Then, for $f(A) := \mathrm{tr}(g(A)B)$, $A \in \mathcal{L}_{sa}(\mathbb{H})$ and $\|B\|_1 < \infty$

$$\|f\|_{C^s(\mathcal{L}_{sa}(\mathbb{H}))} \lesssim_s \|g\|_{B^s_{\infty,1}(\mathbb{R})} \|B\|_1$$

Example of Hölder smoothness: functions of operators

- $\Sigma_0 : \mathbb{H} \mapsto \mathbb{H}$ a covariance operator with eigenvalues $\|\Sigma_0\| = \lambda_1 = \dots = \lambda_l > \lambda_{l+1} \geq \dots$.
- $g_l := \lambda_l - \lambda_{l+1}$ the spectral gap
- $U = B(\Sigma_0; \delta) := \{A : \|A - \Sigma_0\| < \delta\}, \delta < g_l/8.$
- $P(A)$ the orthogonal projection onto the linear span of eigenvectors corresponding to the top l eigenvalues of A
- The function $U \ni A \mapsto P(A)$ is C^∞ and

$$\|P^{(k)}\|_{L_\infty(U)} \lesssim_k g_l^{-k}, k \geq 0$$

- For $f(A) := \langle P(A), B \rangle$,

$$\|f^{(k)}\|_{L_\infty(U)} \lesssim_k \|B\|_1 g_l^{-k}, k \geq 0$$

A classical problem

- $X_1, \ldots, X_n$ i.i.d. $\sim P_\theta, \theta \in \Theta$, $\Theta \subset \mathbb{R}^d$ an open set, $d \geq 1$, $n \to \infty$
- $\{P_\theta : \theta \in \Theta\}$ a regular statistical model with non-singular Fisher information $I(\theta)$
- $f : \Theta \mapsto \mathbb{R}$ a continuously differentiable function
- $\hat{\theta}_n$ maximum likelihood estimator (MLE) based on $X_1, \ldots, X_n$
- Then, uniformly in Θ,

$$n\mathbb{E}_\theta(f(\hat{\theta}_n) - f(\theta))^2 \to \sigma_f^2(\theta) := \langle I(\theta)^{-1} f'(\theta), f'(\theta)\rangle$$

  and

$$\sqrt{n}(f(\hat{\theta}_n) - f(\theta)) \xrightarrow{d} N(0; \sigma_f^2(\theta)) \text{ as } n \to \infty$$

- Moreover, $f(\hat{\theta}_n)$ is a locally asymptotically minimax estimator of $f(\theta)$ in the sense of Hájek and Le Cam:

$$\lim_{c\to\infty} \liminf_{n\to\infty} \inf_{T_n} \sup_{\|\theta-\theta_0\|\leq cn^{-1/2}} n\mathbb{E}_\theta(T_n(X_1, \ldots, X_n) - f(\theta))^2 \geq \sigma_f^2(\theta_0).$$

Previous results: high-dimensional and non-parametric models

- Ibragimov, Nemirovski and Khasminskii (1987), Nemirovski (1990, 2000)
- estimation of $f(\theta)$ for functionals f of smoothness s, θ being the parameter of infinite-dimensional Gaussian white noise model:

$$dX(t) = \theta(t)dt + n^{-1/2}dw(t), t \in [0,1], \theta \in \Theta \subset L_2([0,1])$$

- proved the existence of a “smoothness threshold" such that the efficient estimation with $\sqrt{n}$-rate is possible when s is above the threshold and is impossible (for some functionals) otherwise;
- the smoothness threshold depends on the rate of decay of Kolmogorov widths of the parameter set.

Previous results: high-dimensional and non-parametric models

- $E = \mathbb{R}^d$ equipped with the Euclidean norm

$$\tilde{\mathcal{S}}(a; d) := \{\Sigma : \sigma(\Sigma) \subset [1/a, a]\}, a \geq 1$$

- $\mathbf{r}(\Sigma) \asymp d, \ \Sigma \in \tilde{\mathcal{S}}(a; d)$
- K (2017, 2018) studied estimation of functionals $\langle g(\Sigma), B \rangle$ and determined the degree of smoothness of function $g : \mathbb{R} \mapsto \mathbb{R}$ needed for asymptotically efficient estimation
- K& Zhilova (2019) developed estimators $\hat{T}_f(X_1, \dots, X_n)$ for general Hölder smooth functionals f such that

$$\sup_{\|f\|_{C^s} \leq 1} \sup_{\Sigma \in \tilde{\mathcal{S}}(a,d)} \left\| \hat{T}_f - f(\Sigma) \right\|_{L_2(\mathbb{P}_\Sigma)} \lesssim_a \left(\frac{1}{\sqrt{n}} \vee \left(\sqrt{\frac{d}{n}} \right)^s \right) \wedge 1$$

- If $d \leq n^\alpha$ for some $\alpha \in (0, 1)$ and $s \geq \frac{1}{1-\alpha}$, then the L_2-error rate is $O(n^{-1/2})$ and, for $s > \frac{1}{1-\alpha}$, $\hat{T}_f$ is asymptotically efficient

Previous results: high-dimensional and non-parametric models

The error rate

$$\left(\frac{1}{\sqrt{n}} \vee \left(\sqrt{\frac{d}{n}}\right)^s\right) \wedge 1$$

is typical in the problems of this type:

- finite-dimensional and infinite dimensional random shift models with Gaussian and Poincaré noise: K& Zhilova (2018), K& Zhilova (2020)
- general high-dimensional parametric models with an estimator admitting normal approximation: K (2020)
- high-dimensional log-concave location models: K and Wahl (2021)

The estimation method in these papers was based on higher order iterative bias reduction (“iterated bootstrap", Hall and Martin (1988))

Loss functions and Orlicz norm

- $\psi : \mathbb{R}_+ \mapsto \mathbb{R}_+$ be a convex nondecreasing loss function such that $\psi(0) = 0$
- $\|\eta\|_{\psi} := \|\eta\|_{L_{\psi}(\mathbb{P})} := \inf\Big\{c > 0 : \mathbb{E}\psi\Big(\frac{|\eta|}{c}\Big) \leq 1\Big\}$
- For $\psi(u) := u^p, p \geq 1$, $\|\eta\|_{\psi} = \|\eta\|_{L_p}$
- Other choices: $\psi(u) = \psi_1(u) := e^u - 1$ (subexponential loss) and $\psi(u) = \psi_2(u) = e^{u^2} - 1$ (subgaussian loss). More generally, $\psi_\alpha(u) := e^{u^\alpha} - 1$, $\alpha \geq 1$.
- $\|\eta\|_{\psi_\alpha} \asymp \sup_{p\geq 1} p^{-1/\alpha}\|\eta\|_{L_p}, \alpha > 0$
- Given two loss functions ψ and φ, we write $\psi \preceq \varphi$ (ψ is dominated by φ) iff

$$\psi(u) \leq c_1\varphi(c_2 u), u \geq 0$$

for some constants $c_1, c_2 > 0$.

Upper bound on the L_ψ-risk (K (2022))

For $a > 0$, $U \subset L(E^*, E)$ and $s = k + \rho$, $k \geq 0, \rho \in (0, 1]$, define the weighted Hölder norm of f as

$$\|f\|_{C^{s,a}(U)} := \max_{0 \leq j \leq k} a^j \|f^{(j)}\|_{L_\infty(U)} \vee a^s \|f^{(k)}\|_{\mathrm{Lip}_\rho(U)}.$$

Theorem

Let $a > 0, r \geq 1$ and $\Sigma_0 \in \mathcal{S}(a, r)$. Suppose, for a sufficiently large $C > 0$, $Ca\sqrt{\frac{r}{n}} < \delta$. Let $U := B(\Sigma_0, 2\delta) := \{\Sigma : \|\Sigma - \Sigma_0\| < 2\delta\}$ and let $s = k + \rho$, $k \geq 0, \rho \in (0, 1]$. Then, there exists an estimator $T_f(X_1, \dots, X_n)$ of $f(\Sigma)$ such that, for all $\psi \preceq \psi_1$

$$\sup_{\|f\|_{C^{s,a}(U)} \leq 1} \sup_{\Sigma \in \mathcal{S}(a,r), \|\Sigma - \Sigma_0\| < \delta} \Big\| T_f(X_1, \dots, X_n) - f(\Sigma) \Big\|_{L_\psi(\mathbb{P}_\Sigma)}$$

$$\lesssim_s \frac{1}{\sqrt{n}} \vee \left(\sqrt{\frac{r}{n}}\right)^s.$$

Minimax lower bound (K (2022))

- $E = \mathbb{H}$ is a separable Hilbert space.
- Σ_0 is a spiked covariance of rank $d \iff \sigma(\Sigma_0) = \{\lambda, \mu, 0\}$ with $\lambda > \mu > 0$, λ of multiplicity 1, μ of multiplicity $d-1$.

Theorem

Let $\Sigma_0 \in \mathcal{S}(a, r)$ be a spiked covariance of rank $[r]$ with $\lambda := \gamma_1 a, \mu := \gamma_2 a, 0 < \gamma_2 < \gamma_1$. Denote $\kappa := \gamma_2 \wedge (\gamma_1 - \gamma_2) \wedge (1 - \gamma_1)$ and suppose that $c_1 \gamma_1 a \sqrt{\frac{r}{n}} < \delta \leq c_2 \kappa a \wedge 1$ for a sufficiently large c_1 and sufficiently small c_2. Let $U := B(\Sigma_0, 2\delta)$ and let $s > 0$. Then

$$\sup_{\|f\|_{C^{s,a}(U)} \leq 1} \inf_T \sup_{\Sigma \in S(a,r), \|\Sigma - \Sigma_0\| < \delta} \Big\| T(X_1, \dots, X_n) - f(\Sigma) \Big\|_{L_2(\mathbb{P}_\Sigma)}$$

$$\asymp_{s, \gamma_1, \gamma_2} \frac{1}{\sqrt{n}} \vee \left(\sqrt{\frac{r}{n}}\right)^s.$$

The lower bound: a sketch of the proof

- $\Sigma_0 = \lambda(u \otimes u) + \mu(P_L - u \otimes u)$
  $L \subset \mathbb{H}, \dim(L) = d := [r], u \in L, \|u\| = 1$
  (L could be identified with $\mathbb{R}^d$)
- $\theta(\Sigma), \Sigma \in U$ the eigenvector corresponding to the top eigenvalue of Σ, $\langle \theta(\Sigma), u \rangle \geq 0$, $U \ni \Sigma \mapsto \theta(\Sigma)$ is C^∞
- We will construct certain least favorable functionals

$$f_k(\Sigma) := h_k(\theta(\Sigma)), \Sigma \in U, k = 1, \dots, d,$$

  where $h_k : \mathbb{H} \mapsto \mathbb{R}$, $\|h_k\|_{C^s} \lesssim 1$
- For these functionals, $\|f_k\|_{C^{s,a}(U)} \leq 1$ and we will show that

$$\max_{1 \leq k \leq d} \inf_{T_k} \sup_{\Sigma \in S(a,r), \|\Sigma - \Sigma_0\| < \delta} \Big\| T_k(X_1, \dots, X_n) - f_k(\Sigma) \Big\|_{L_2(\mathbb{P}_\Sigma)}$$

$$\gtrsim_{s,\gamma_1,\gamma_2} \left(\sqrt{\frac{r}{n}}\right)^s.$$

Well separated subsets

- $\exists B \subset \{-1,1\}^d : \mathrm{card}(B) \geq \frac{e^{d/8}}{2}$, $|\langle u, \omega\rangle| < 2$ and
$$h(\omega,\omega') := \sum_{j=1}^d I(\omega_j \neq \omega'_j) \geq \frac{d}{4}, \omega,\omega' \in B, \omega \neq \omega'.$$
- Let $\varepsilon \asymp \sqrt{\frac{d}{n}}$
- $\Theta_\varepsilon = \{\theta_\omega : \omega \in B\}$, $\theta_\omega := \frac{t_\omega}{\|t_\omega\|}$, $t_\omega := \varepsilon \frac{\omega}{\sqrt{d}} + \sqrt{1-\varepsilon^2} u, \omega \in B$
$$\frac{\varepsilon}{2\sqrt{d}}\sqrt{h(\omega,\omega')} \leq \|\theta_\omega - \theta_{\omega'}\| \leq \frac{8\varepsilon}{\sqrt{d}}\sqrt{h(\omega,\omega')}, \omega,\omega' \in B,$$
implying that
$$\frac{\varepsilon}{4} \leq \|\theta_\omega - \theta_{\omega'}\| \leq 8\varepsilon, \omega,\omega' \in B, \omega \neq \omega'$$

Well separated subsets

- $\Sigma_\theta := \lambda(\theta \otimes \theta) + \mu(P_L - \theta \otimes \theta)$, $\theta \in L$, $\|\theta\| = 1$
- Then

$$\|\Sigma_{\theta_\omega} - \Sigma_0\| < \delta, \omega \in B$$

- Moreover, for all $\omega, \omega' \in B$,

$$K(N(0, \Sigma_{\theta_\omega})^{\otimes n} \| N(0, \Sigma_{\theta_{\omega'}})^{\otimes n}) \lesssim n\varepsilon^2 \lesssim \log \mathrm{card}(B),$$

implying that for $X_1, \dots, X_n$ i.i.d. $\sim N(0, \Sigma_{\theta_\omega}), \omega \in B$

$$\inf_{\hat\theta} \max_{\omega \in B} \mathbb{E}_{\Sigma_{\theta_\omega}} \|\hat\theta - \theta_\omega\|^2 \gtrsim \varepsilon^2.$$

Nemirovski's bump functionals

- $\varphi : \mathbb{R} \mapsto [0,1]$, φ is C^∞, $\mathrm{supp}(\varphi) \in [-1,1]$, $\varphi(0) > 0$
- $\phi(u) := \varphi(\|u\|^2), u \in \mathbb{H}$
- $$h_k(\theta) := \sum_{\omega \in B} \omega_k \varepsilon^s \phi\Big(\frac{\theta - \theta_\omega}{c\varepsilon}\Big), k = 1, \dots, d,$$

  $c > 0$ small enough
- Note that "bump functions" $\phi\Big(\frac{\theta-\theta_\omega}{c\varepsilon}\Big), \omega \in B$ have disjoint supports and $\|h_k\|_{C^s} \lesssim 1$.
- $h_k(\theta_\omega) = \omega_k \varepsilon^s \varphi(0), k = 1, \dots, d, \omega \in B$
- The values $h_k(\theta_\omega), k = 1, \dots, d$ provide a "coding" for θ_ω

Back to least favorable functionals

- $f_k(\Sigma) := h_k(\theta(\Sigma)), \Sigma \in U, k = 1, \dots, d$
- $f_k(\Sigma_{\theta_\omega}) = \omega_k \varepsilon^s \varphi(0), k = 1, \dots, d, \omega \in B$
- Define

$$\tau^2(\omega, \omega') := \frac{1}{d} \sum_{k=1}^{d} (f_k(\Sigma_{\theta_\omega}) - f_k(\Sigma_{\theta_{\omega'}}))^2, \omega, \omega' \in B.$$

- Then

$$\tau^2(\omega, \omega') \asymp \varepsilon^{2s} \frac{h(\omega, \omega')}{d}, \omega, \omega' \in B,$$

implying that

$$\tau^2(\omega, \omega') \asymp \varepsilon^{2(s-1)} \|\theta_\omega - \theta_{\omega'}\|^2, \omega, \omega' \in B.$$

Back to lower bounds

- If there exist estimators $\hat{T}_k$ of $f_k(\Sigma_{\theta_\omega}), k = 1, \dots, d$ based on i.i.d. $X_1, \dots, X_n \sim N(0, \Sigma_{\theta_\omega}), \omega \in B$ with

$$\delta^2 := \max_{1 \le k \le d} \max_{\omega \in B} \Big\| \hat{T}_k(X_1, \dots, X_n) - f_k(\Sigma_{\theta_\omega}) \Big\|_{L_2(\mathbb{P}_{\Sigma_{\theta_\omega}})},$$

then

$$\max_{\omega \in B} \mathbb{E}_{\Sigma_{\theta_\omega}} \frac{1}{d} \sum_{k=1}^{d} (\hat{T}_k - f_k(\Sigma_{\theta_\omega}))^2 \le \delta^2.$$

- Based on $\hat{T}_k$, it is not hard to construct estimator $\hat{\theta}$ such that

$$\varepsilon^2 \lesssim \max_{\omega \in B} \mathbb{E}_{\Sigma_{\theta_\omega}} \|\hat{\theta} - \theta_\omega\|^2 \lesssim \varepsilon^{2(s-1)} \delta^2,$$

implying that $\delta \gtrsim \varepsilon^s \asymp \left(\sqrt{\frac{d}{n}}\right)^s \asymp \left(\sqrt{\frac{r}{n}}\right)^s.$

Estimation method: linear aggregation of plug-in estimators with different sample sizes

- $f : L(E^*, E) \mapsto \mathbb{R}$, $f \in C^s$, $s = k + \rho$, $k \geq 2$, $\rho \in (0, 1]$
- $1 \leq n_1, \ldots, n_k \leq n$ sample sizes
- $$T_f(X_1, \ldots, X_n) := \sum_{j=1}^{k} C_j f(\hat{\Sigma}_{n_j})$$
- The goal is to choose n_j so that the biases of plug-in estimators $f(\hat{\Sigma}_{n_j}), j = 1, \ldots, k$ “almost cancel out."

Bias reduction via linear aggregation of plug-in estimators (Jiao and Han (2021), for binomial model)

- $Y \sim P$ a r.v. in a Banach space F with unknown mean $\mathbb{E}Y$
- In our case, $F := L(E^*, E)$, $Y := X \otimes X$, $\mathbb{E}Y = \Sigma$
- $f : F \mapsto \mathbb{R}$
- Goal: estimate $f(\mathbb{E}Y)$ based on i.i.d. observations $Y_1, \dots, Y_n$ of Y
- Let $k \geq 2$ and let $n/c \leq n_1 < \cdots < n_k \leq n$ for some $c > 1$.
- Let

$$\hat{T}_f(Y_1, \dots, Y_n) := \sum_{j=1}^{k} C_j f(\bar{Y}_{n_j}),$$

where

$$C_j := \prod_{i \neq j} \frac{n_j}{n_j - n_i}, j = 1, \dots, k.$$

Bias reduction via linear aggregation of plug-in estimators (Jiao and Han (2021))

Proposition

The following properties hold:

- $\sum_{j=1}^{k} C_j = 1$
- $\sum_{j=1}^{k} \frac{C_j}{n_j^l} = 0, l = 1, \dots, k.$

Assumption

Suppose that $\sum_{j=1}^{k} |C_j| \lesssim_k 1.$

Clearly, for this assumption to hold, one needs $n_{j+1} - n_j \asymp n$.

Bounding the bias

Proposition

Let $k \geq 2$ and let f be k times Fréchet differentiable with $f^{(k)} \in \mathrm{Lip}_\rho(F)$ for some $\rho \in (0,1]$. Then

$$\left|\mathbb{E}\hat{T}_f(Y_1,\dots,Y_n) - f(\mathbb{E}Y)\right| \lesssim_{k,\rho} \|f^{(k)}\|_{\mathrm{Lip}_\rho} \max_{1\leq j\leq k} \mathbb{E}\|\bar{Y}_{n_j} - \mathbb{E}Y\|^{k+\rho}.$$

Moreover, if f is a polynomial of degree k, then $\hat{T}_f(Y_1,\dots,Y_n)$ is an unbiased estimator of $f(\mathbb{E}Y)$.

Bias decomposition

Proposition

Let $k \geq 2, \rho \in (0,1]$ and suppose that f is k times Fréchet differentiable with $f^{(k)} \in \mathrm{Lip}_\rho(F)$. Suppose also that $\mathbb{E}\|Y\|^{k+\rho} < \infty$. Then

$$\mathbb{E}f(\bar{Y}_n) - f(\mathbb{E}Y) = \sum_{l=1}^{k} \frac{\beta_{l,k}(P)}{n^l} + R,$$

where $\beta_{l,k}(P), l = 1, \dots, k$ do not depend on n and

$$|R| \lesssim \|f^{(k)}\|_{\mathrm{Lip}_\rho} \mathbb{E}\|\bar{Y}_n - \mathbb{E}Y\|^{k+\rho}.$$

If f is a polynomial of degree k, then $R = 0$.

Jackknife type estimators

- $\mathcal{F}_{\rm sym}$ the σ-algebra generated by symmetric functions of $Y_1, \dots, Y_n$
- $$\check{T}_f(Y_1, \dots, Y_n) := \mathbb{E}(\hat{T}_f(Y_1, \dots, Y_n)|\mathcal{F}_{\rm sym}) = \sum_{j=1}^{k} C_j U_n f(\bar{Y}_{n_j}),$$

  where, for $h(Y_1, \dots, Y_m), m \le n$,

  $$(U_n h)(Y_1, \dots, Y_n) := \mathbb{E}(h(Y_1, \dots, Y_m)|\mathcal{F}_{\rm sym}) = \frac{1}{\binom{n}{m}} \sum_{1 \le j_1 < \dots < j_m \le n} h(Y_{j_1}, \dots, Y_{j_k}).$$

Jackknife type estimators

Note that

- $\mathbb{E}\check{T}_f(Y_1, \dots, Y_n) = \mathbb{E}\hat{T}_f(Y_1, \dots, Y_n)$
- for all $p \geq 1$,

$$
\begin{aligned}
&\|\check{T}_f(Y_1, \dots, Y_n) - \mathbb{E}\check{T}_f(Y_1, \dots, Y_n)\|_{L_p} \\
&\leq \|\hat{T}_f(Y_1, \dots, Y_n) - \mathbb{E}\hat{T}_f(Y_1, \dots, Y_n)\|_{L_p} \\
&\lesssim_k \max_{1 \leq j \leq k} \|f(\bar{Y}_{n_j}) - \mathbb{E}f(\bar{Y}_{n_j})\|_{L_p}.
\end{aligned}
$$

Jackknife type estimators

Moreover, let

$$S_f(x,h) := f(x+h) - f(x) - \langle h, f'(x)\rangle, x, h \in F$$

be the remainder of the first order Taylor expansion.

Proposition

For all $p \geq 1$,

$$\Big\| \check{T}_f(Y_1,\dots,Y_n) - \mathbb{E}\check{T}_f(Y_1,\dots,Y_n) - \langle \bar{Y}_n - \mathbb{E}Y, f'(\mathbb{E}Y)\rangle \Big\|_{L_p}$$

$$\lesssim_k \max_{1\leq j\leq k} \Big\| S_f(\mathbb{E}Y, \bar{Y}_{n_j} - \mathbb{E}Y) - \mathbb{E}S_f(\mathbb{E}Y, \bar{Y}_{n_j} - \mathbb{E}Y) \Big\|_{L_p}.$$

Reduction to concentration bounds

Thus, the problem reduces to:

- Bounds on

$$\mathbb{E}\|\bar{Y}_n - \mathbb{E}Y\|^{k+\rho};$$

- Bounds on

$$\|f(\bar{Y}_n) - \mathbb{E}f(\bar{Y}_n)\|_{L_p};$$

- Bounds on

$$\Big\|S_f(\mathbb{E}Y, \bar{Y}_n - \mathbb{E}Y) - \mathbb{E}S_f(\mathbb{E}Y, \bar{Y}_n - \mathbb{E}Y)\Big\|_{L_p}.$$

Concentration bounds for sample covariance

- For $s \geq 1$,

$$\mathbb{E}\|\hat{\Sigma}_n - \Sigma\|^s \lesssim_s \|\Sigma\|^s \Big(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \vee \frac{\mathbf{r}(\Sigma)}{n}\Big)^s.$$

- For a Lipschitz functional $f : L(E^*, E) \mapsto \mathbb{R}$ and for all $p \geq 1$,

$$\|f(\hat{\Sigma}_n) - \mathbb{E}f(\hat{\Sigma}_n)\|_{L_p} \lesssim \|f\|_{\mathrm{Lip}}\|\Sigma\| \Bigg(\Big(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \vee 1\Big)\sqrt{\frac{p}{n}} + \frac{p}{n}\Bigg).$$

- Let $f' \in \mathrm{Lip}_\rho(L(E^*, E))$ for some $\rho \in (0, 1]$. Suppose $\mathbf{r}(\Sigma) \lesssim n$. Then, for all $p \geq 1$,

$$\Big\|S_f(\Sigma, \hat{\Sigma}_n - \Sigma) - \mathbb{E}S_f(\Sigma, \hat{\Sigma}_n - \Sigma)\Big\|_{L_p}$$
$$\lesssim \|f'\|_{\mathrm{Lip}_\rho}\|\Sigma\|^{1+\rho}\Bigg(\sqrt{\frac{p}{n}}\Big(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}}\Big)^\rho + \Big(\frac{p}{n}\Big)^{(1+\rho)/2} + \Big(\frac{p}{n}\Big)^{1+\rho}\Bigg).$$

Estimators

- Linear aggregation

$$T_f^{(1)}(X_1, \dots, X_n)$$

$$:= \hat{T}_f(X_1 \otimes X_1, \dots, X_n \otimes X_n) = \sum_{j=1}^{k} C_j f(\hat{\Sigma}_{n_j})$$

- Jackknife

$$T_f^{(2)}(X_1, \dots, X_n)$$

$$:= \check{T}_f(X_1 \otimes X_1, \dots, X_n \otimes X_n) = \sum_{j=1}^{k} C_j U_n f(\hat{\Sigma}_{n_j})$$

A local risk bound (K (2022))

Theorem

Suppose $\mathbf{r}(\Sigma) \lesssim n$. *Let* $f : L(E^*, E) \mapsto \mathbb{R}$ *be Lipschitz and* k *times Fréchet differentiable in an open ball* $U = B(\Sigma, \delta)$ *of radius* $\delta > 0$ *with* $\|f^{(k)}\|_{\mathrm{Lip}_\rho(U)} < \infty$ *for some* $k \geq 2$ *and* $\rho \in (0, 1]$. *Suppose also that, for a sufficiently large* $C > 0$, $C\|\Sigma\|\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} < \delta$. *Then, for* $i = 1, 2$ *and for all* $\psi \preceq \psi_1$

$$\Big\| T_f^{(i)}(X_1, \dots, X_n) - f(\Sigma) \Big\|_{\psi}$$

$$\lesssim_{k,\rho} \|f\|_{\mathrm{Lip}} \frac{\|\Sigma\|}{\sqrt{n}} + \|f^{(k)}\|_{\mathrm{Lip}_\rho(U)} \|\Sigma\|^{k+\rho} \left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}}\right)^{k+\rho}$$

$$+ \max_{2 \leq j \leq k} \|f^{(j)}(\Sigma)\| \left(\frac{\|\Sigma\|}{\sqrt{n}}\right)^{j} \exp\left\{-cn\left(\frac{\delta^2}{\|\Sigma\|^2} \wedge \frac{\delta}{\|\Sigma\|}\right)\right\}$$

with some constant $c > 0$.

A global risk bound (K (2022))

Theorem

Let $f: L(E^*, E) \mapsto \mathbb{R}$ be Lipschits and, for some $k \geq 2$, let it be k times Fréchet differentiable with $\|f^{(k)}\|_{\mathrm{Lip}_\rho} < \infty$ for some $\rho \in (0, 1]$. Then, for $i = 1, 2$ and for all $\psi \preceq \psi_1$,

$$\Big\| T_f^{(i)}(X_1, \dots, X_n) - f(\Sigma) \|_\psi$$

$$\lesssim_s \|f\|_{\mathrm{Lip}} \left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \vee 1 \right) \frac{\|\Sigma\|}{\sqrt{n}} + \|f^{(k)}\|_{\mathrm{Lip}_\rho} \|\Sigma\|^{k+\rho} \left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \vee \frac{\mathbf{r}(\Sigma)}{n} \right)^{k+\rho}.$$

Normal approximation and efficiency

Theorem

Suppose that $\mathbf{r}(\Sigma) \lesssim n$. *If* $f : L(E^*, E) \mapsto \mathbb{R}$ *is* k *times Fréchet differentiable for some* $k \geq 2$ *with* $\|f'\|_{C^1} < \infty$ *and with* $\|f^{(k)}\|_{\mathrm{Lip}_\rho} < \infty$ *for some* $\rho \in (0, 1]$, *then, for all* $\beta \in [1/2, 1)$,

$$\left\| T_f^{(2)}(X_1, \ldots, X_n) - f(\Sigma) - \langle \hat{\Sigma}_n - \Sigma, f'(\Sigma) \rangle \right\|_{\psi_\beta}$$

$$\lesssim_{k,\rho} \|f'\|_{C^1} \frac{\|\Sigma\|^{1/\beta}}{\sqrt{n}} \left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \right)^{1/\beta - 1} + \|f^{(k)}\|_{\mathrm{Lip}_\rho} \|\Sigma\|^{k+\rho} \left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \right)^{k+\rho}.$$

Normal approximation and efficiency

Let $\sigma_f^2(\Sigma) := \mathrm{Var}(\langle X \otimes X, f'(\Sigma)\rangle)$. It follows that

- $$\Big|\sqrt{n}\Big\|T_f^{(2)}(X_1,\dots,X_n) - f(\Sigma)\Big\|_{L_2} - \sigma_f(\Sigma)\Big|$$
  $$\lesssim_{k,\rho} \|f'\|_{C^1}\|\Sigma\|^2\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} + \|f^{(k)}\|_{\mathrm{Lip}_\rho}\|\Sigma\|^{k+\rho}\sqrt{n}\Big(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}}\Big)^{k+\rho}.$$

- $$W_2\Big(\frac{\sqrt{n}(T_f^{(2)}(X_1,\dots,X_n) - f(\Sigma))}{\sigma_f(\Sigma)}, N(0,1)\Big) \lesssim_{k,\rho} \frac{\|\Sigma\|^2\|f'(\Sigma)\|^2}{\sigma_f^2(\Sigma)}\frac{1}{\sqrt{n}}$$
  $$+ \frac{\|f'\|_{C^1}\|\Sigma\|^2}{\sigma_f(\Sigma)}\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} + \frac{\|f^{(k)}\|_{\mathrm{Lip}_\rho}\|\Sigma\|^{k+\rho}}{\sigma_f(\Sigma)}\sqrt{n}\Big(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}}\Big)^{k+\rho}.$$

Example: an application to spectral projections

- $\Sigma_0 : \mathbb{H} \mapsto \mathbb{H}$ a covariance operator with eigenvalues $\|\Sigma_0\| = \lambda_1 = \cdots = \lambda_l > \lambda_{l+1} \geq \dots$.
- $g_l := \lambda_l - \lambda_{l+1}$ the spectral gap
- $U = B(\Sigma_0; \delta) := \{A : \|A - \Sigma_0\| < \delta\}, \delta < g_l/8.$
- $P(A)$ the orthogonal projection onto the linear span of eigenvectors corresponding to the top l eigenvalues of A
- $f(A) := \langle P(A), B \rangle, A \in U, \|B\|_1 < \infty$
- Then $f \in C^\infty(U)$

Example: an application to spectral projections

Corollary

Let $\gamma := \frac{\|\Sigma_0\|}{g_l}$ *and suppose that* $C\gamma\sqrt{\frac{r}{n}} \leq 1$ *for some* $r \geq 1$ *and some constant* $C > 0$. *Then, for* $i = 1, 2$, *for all* $k \geq 1$ *and for all* $\psi \preceq \psi_1$,

$$\sup_{\|\Sigma - \Sigma_0\| < \delta, \mathbf{r}(\Sigma) \leq r} \left\| T_{f,k}^{(i)}(X_1, \dots, X_n) - f(\Sigma) \right\|_{\psi} \lesssim_k \|B\|_1 \left(\frac{\gamma}{\sqrt{n}} + \left(\gamma \sqrt{\frac{r}{n}} \right)^{k+1} \right).$$

K.& Lounici (2016), K, Löffler and Nickl (2019): efficient estimators of linear functionals of principal components when $\mathbf{r}(\Sigma) = o(n)$ and the top eigenvalue of Σ is *simple* (i.e., $P(\Sigma)$ is a one-dimensional spectral projection).