Estimation of Functionals of High-Dimensional and Infinite-Dimensional Parameters: Bias Reduction and Concentration

Vladimir Koltchinskii

School of Mathematics
Georgia Institute of Technology

INFORMS 2025

Table of contents

Functional estimation problem

Estimation of functionals

- $X_1, \ldots, X_n$ i.i.d. $\sim P_\theta, \theta \in \Theta$ in a space S
- $\Theta \subset E$, E a linear normed space
- $\mathcal{F}$ a class of functionals $f : \Theta \mapsto \mathbb{R}$
- For given $f \in \mathcal{F}$, the goal is to estimate $f(\theta)$ based on the observations $X_1, \ldots, X_n$

Estimation of functionals: more specific problems

- Given a class $\mathcal{F}$ of functionals $f : \Theta \mapsto \mathbb{R}$, what is the size of the "optimal risk" of functional estimation in the class $\mathcal{F}$

$$\delta_n(\Theta; \mathcal{F}) := \sup_{f\in\mathcal{F}} \inf_{T_n} \sup_{\theta\in\Theta} \mathbb{E}_\theta (T_n(f; X_1, \dots, X_n) - f(\theta))^2?$$

- Is there a "universal" estimation method $T_n(f; X_1, \dots, X_n), f \in \mathcal{F}$ for which the "optimal risk" $\delta_n(\Theta; \mathcal{F})$ is attained?
- Let $\mathcal{F} = \mathcal{F}_s$ be a class of functionals of "smoothness" $s > 0$. Is there a smoothness threshold $s(\Theta, n)$ such that, for all $s \geq s(\Theta; n)$,

$$\delta_n(\Theta; \mathcal{F}_s) = O(n^{-1})$$

and, in this case, is it possible to construct asymptotically efficient estimators of $f(\theta)$ with $\sqrt{n}$ convergence rate?

Estimation of functionals: more specific problems

- Given a class $\mathcal{F}$ of functionals $f : \Theta \mapsto \mathbb{R}$, what is the size of the "optimal risk" of functional estimation in the class $\mathcal{F}$

$$\delta_n(\Theta; \mathcal{F}) := \sup_{f\in\mathcal{F}} \inf_{T_n} \sup_{\theta\in\Theta} \mathbb{E}_\theta (T_n(f; X_1, \dots, X_n) - f(\theta))^2?$$

- Is there a "universal" estimation method $T_n(f; X_1, \dots, X_n), f \in \mathcal{F}$ for which the "optimal risk" $\delta_n(\Theta; \mathcal{F})$ is attained?
- Let $\mathcal{F} = \mathcal{F}_s$ be a class of functionals of "smoothness" $s > 0$. Is there a smoothness threshold $s(\Theta, n)$ such that, for all $s \geq s(\Theta; n)$,

$$\delta_n(\Theta; \mathcal{F}_s) = O(n^{-1})$$

and, in this case, is it possible to construct asymptotically efficient estimators of $f(\theta)$ with $\sqrt{n}$ convergence rate?

Estimation of functionals: more specific problems

- Given a class $\mathcal{F}$ of functionals $f : \Theta \mapsto \mathbb{R}$, what is the size of the "optimal risk" of functional estimation in the class $\mathcal{F}$

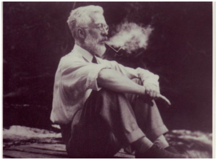
$$\delta_n(\Theta; \mathcal{F}) := \sup_{f\in\mathcal{F}} \inf_{T_n} \sup_{\theta\in\Theta} \mathbb{E}_\theta(T_n(f; X_1, \dots, X_n) - f(\theta))^2?$$

- Is there a "universal" estimation method $T_n(f; X_1, \dots, X_n), f \in \mathcal{F}$ for which the "optimal risk" $\delta_n(\Theta; \mathcal{F})$ is attained?
- Let $\mathcal{F} = \mathcal{F}_s$ be a class of functionals of "smoothness" $s > 0$. Is there a smoothness threshold $s(\Theta, n)$ such that, for all $s \geq s(\Theta; n)$,

$$\delta_n(\Theta; \mathcal{F}_s) = O(n^{-1})$$

and, in this case, is it possible to construct asymptotically efficient estimators of $f(\theta)$ with $\sqrt{n}$ convergence rate?

A classical problem (going back to Ronald Fisher)



- $X_1, \dots, X_n$ 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 density p_θ and non-singular Fisher information matrix $I(\theta)$,

$$I(\theta) := \mathbb{E}_\theta \frac{\partial}{\partial \theta} \log p_\theta(X) \otimes \frac{\partial}{\partial \theta} \log p_\theta(X)$$

- $f : \Theta \mapsto \mathbb{R}$ a continuously differentiable function, $f(\theta)$ to be estimated based on i.i.d. $X_1, \dots, X_n \sim P_\theta$

A classical problem (going back to Ronald Fisher)

- Maximum likelihood estimator (MLE) based on $X_1, \dots, X_n$:

$$\hat{\theta}_n := \mathrm{argmax}_{\theta \in \Theta} \prod_{j=1}^{n} p_\theta(X_j)$$

- $f(\hat{\theta}_n)$ the plug-in estimator of $f(\theta)$
- Uniformly in Θ,

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

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

implying that $\delta_n(\Theta, \{f\}) = O(n^{-1})$.

A classical problem (going back to Ronald Fisher)

- Maximum likelihood estimator (MLE) based on $X_1, \dots, X_n$:

$$\hat{\theta}_n := \mathrm{argmax}_{\theta \in \Theta} \prod_{j=1}^{n} p_\theta(X_j)$$

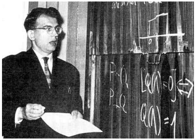
- $f(\hat{\theta}_n)$ the plug-in estimator of $f(\theta)$
- Uniformly in Θ,

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

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

implying that $\delta_n(\Theta, \{f\}) = O(n^{-1})$.

Local asymptotic minimaxity



$f(\hat{\theta}_n)$ is a locally asymptotically minimax estimator of $f(\theta)$ in the sense of the following bound due to Hájek and Le Cam:

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

Estimation of functionals in a Gaussian white noise model

- 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:

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

- Kolmogorov widths: for some $\beta > 0$,

$$d_m(\Theta) := \inf_{\dim(L)\leq m} \sup_{\theta\in\Theta} \|\theta - P_L\theta\|_{L_2} \lesssim m^{-\beta}$$

- the existence of a "smoothness threshold" $s(\beta)$ such that the efficient estimation with $\sqrt{n}$-rate is possible when $s > s(\beta)$ and is impossible (for some functionals of smoothness s) otherwise.

Estimation of functionals in a Gaussian white noise model

- 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:

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

- Kolmogorov widths: for some $\beta > 0$,

$$d_m(\Theta) := \inf_{\dim(L)\leq m} \sup_{\theta\in\Theta} \|\theta - P_L\theta\|_{L_2} \lesssim m^{-\beta}$$

- the existence of a "smoothness threshold" $s(\beta)$ such that the efficient estimation with $\sqrt{n}$-rate is possible when $s > s(\beta)$ and is impossible (for some functionals of smoothness s) otherwise.

Estimation of functionals in a Gaussian white noise model

- 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:

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

- Kolmogorov widths: for some $\beta > 0$,

$$d_m(\Theta) := \inf_{\dim(L) \le m} \sup_{\theta \in \Theta} \|\theta - P_L\theta\|_{L_2} \lesssim m^{-\beta}$$

- the existence of a "smoothness threshold" $s(\beta)$ such that the efficient estimation with $\sqrt{n}$-rate is possible when $s > s(\beta)$ and is impossible (for some functionals of smoothness s) otherwise.

Hölder smoothness

- $\Theta \subset E$, E is a Banach space
- $g : \Theta \mapsto F$, F 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

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

Hölder smoothness

- $\Theta \subset E$, E is a Banach space
- $g : \Theta \mapsto F$, F 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

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

Hölder smoothness

- Note: Fréchet derivative $g^{(j)}(x)$ is a j-linear form on E : $g^{(j)}(x)[h_1, \dots, h_j]$ is linear w.r.t. its j variables $h_1, \dots, h_j \in E$
- the norms of the derivatives are defined as the operator norms (of multilinear forms):

$$\|g^{(j)}(x)\| = \sup_{\|h_1\|\leq 1, \dots, \|h_j\|\leq 1} |g^{(j)}(x)[h_1, \dots, h_j]|.$$

- $C^s(\Theta) := \{g : \Theta \mapsto \mathbb{R} : \|g\|_{C^s(\Theta)} < \infty\}$
- Taylor expansion: for $g \in C^s$, $s = k + \rho, k \geq 1, \rho \in (0, 1]$

$$g(x+h) = \sum_{j=0}^{k} \frac{g^{(j)}(x)[h, \dots, h]}{j!} + S_g^{(k)}(x, h),$$

where $|S_g^{(k)}(x, h)| \lesssim \|g^{(k)}\|_{\mathrm{Lip}_\rho} \|h\|^s$.

Hölder smoothness

- Note: Fréchet derivative $g^{(j)}(x)$ is a j-linear form on E : $g^{(j)}(x)[h_1, \dots, h_j]$ is linear w.r.t. its j variables $h_1, \dots, h_j \in E$
- the norms of the derivatives are defined as the operator norms (of multilinear forms):

$$\|g^{(j)}(x)\| = \sup_{\|h_1\|\leq 1, \dots, \|h_j\|\leq 1} |g^{(j)}(x)[h_1, \dots, h_j]|.$$

- $C^s(\Theta) := \{g : \Theta \mapsto \mathbb{R} : \|g\|_{C^s(\Theta)} < \infty\}$
- Taylor expansion: for $g \in C^s$, $s = k + \rho, k \geq 1, \rho \in (0, 1]$

$$g(x+h) = \sum_{j=0}^{k} \frac{g^{(j)}(x)[h, \dots, h]}{j!} + S_g^{(k)}(x, h),$$

where $|S_g^{(k)}(x, h)| \lesssim \|g^{(k)}\|_{\mathrm{Lip}_\rho} \|h\|^s$.

A local minimax lower bound (K& Li, 2024)

Theorem

Let $\{P_\theta : \theta \in \Theta\}, \Theta \subset \mathbb{R}^d$ *be a statistical model. Suppose that, for some* $r > 0$ *and* $\theta_0 \in \Theta$,

$$B_{\ell_\infty}(\theta_0, \frac{r}{\sqrt{n}}) := \left\{\theta : \|\theta - \theta_0\|_{\ell_\infty} \leq \frac{r}{\sqrt{n}}\right\} \subset \Theta,$$

and, for some constant $C > 0$ *and for all* $\theta \in B_{\ell_\infty}(\theta_0, \frac{r}{\sqrt{n}})$,

$$K(P_\theta \| P_{\theta_0}) := \mathbb{E}_\theta \log \frac{p_\theta(X)}{p_{\theta_0}(X)} \leq C^2 \|\theta - \theta_0\|^2.$$

Finally, suppose that $r \leq \frac{\gamma}{C}$ *for a sufficiently small* $\gamma > 0$. *Then, for all* $s > 0$,

$$\sup_{\|f\|_{C^s} \leq 1} \inf_{T_n} \sup_{\theta \in B_{\ell_\infty}(\theta_0, \frac{r}{\sqrt{n}})} \mathbb{E}_\theta (T_n(X_1, \dots, X_n) - f(\theta))^2 \gtrsim \left(\frac{r^2}{n} + \left(r^2 \frac{d}{n}\right)^s\right) \wedge 1.$$

Phase transition in convergence rates

Assume $r \asymp 1$ and let $\alpha \in (0,1)$. Then

$$\sup_{\|f\|_{C^s}\leq 1} \inf_{T_n} \sup_{\theta\in B_{\ell_\infty}(\theta_0, \frac{r}{\sqrt{n}})} \mathbb{E}_\theta (T_n(X_1,\dots,X_n) - f(\theta))^2$$

$$\gtrsim \left(\frac{1}{n} + \left(\frac{d}{n}\right)^s\right) \wedge 1 \asymp \begin{cases} n^{-1} & d \leq n^\alpha, s \geq \frac{1}{1-\alpha} \\ n^{-s(1-\alpha)} & d \geq n^\alpha, s < \frac{1}{1-\alpha}. \end{cases}$$

Another local minimax lower bound (K& Li, 2024)

- $\dim(E) < \infty$, $\Theta \subset E$ an open subset
- $\{P_\theta : \theta \in \Theta\}$ a regular statistical model, p_θ density of P_θ
- $\frac{\partial}{\partial\theta} \log p_\theta \in E^*$ the score function
- $I(\theta) : E \mapsto E^*$ the Fisher information:

$$I(\theta) := \mathbb{E}_\theta \frac{\partial}{\partial\theta} \log p_\theta(X) \otimes \frac{\partial}{\partial\theta} \log p_\theta(X)$$

- $I(\theta)$ is invertible
- For differentiable $f : \Theta \mapsto \mathbb{R}$, define $\sigma_f^2(\theta) := \langle I(\theta)^{-1} f'(\theta), f'(\theta)\rangle$
- For $\theta_0 \in \Theta$ and $\delta > 0$, let $\omega_I(\theta_0, \delta) := \sup_{\theta\in\Theta, \|\theta-\theta_0\|\leq\delta} \|I(\theta) - I(\theta_0)\|$
- for $f \in C^1(\Theta)$, $\omega_{f'}(\theta_0, \delta) := \sup_{\theta\in\Theta, \|\theta-\theta_0\|\leq\delta} \|f'(\theta) - f'(\theta_0)\|$.

Another local minimax lower bound (K& Li, 2024)

- $\dim(E) < \infty$, $\Theta \subset E$ an open subset
- $\{P_\theta : \theta \in \Theta\}$ a regular statistical model, p_θ density of P_θ
- $\frac{\partial}{\partial\theta} \log p_\theta \in E^*$ the score function
- $I(\theta) : E \mapsto E^*$ the Fisher information:

$$I(\theta) := \mathbb{E}_\theta \frac{\partial}{\partial\theta} \log p_\theta(X) \otimes \frac{\partial}{\partial\theta} \log p_\theta(X)$$

- $I(\theta)$ is invertible
- For differentiable $f : \Theta \mapsto \mathbb{R}$, define $\sigma_f^2(\theta) := \langle I(\theta)^{-1} f'(\theta), f'(\theta) \rangle$
- For $\theta_0 \in \Theta$ and $\delta > 0$, let $\omega_I(\theta_0, \delta) := \sup_{\theta\in\Theta, \|\theta-\theta_0\|\leq\delta} \|I(\theta) - I(\theta_0)\|$
- for $f \in C^1(\Theta)$, $\omega_{f'}(\theta_0, \delta) := \sup_{\theta\in\Theta, \|\theta-\theta_0\|\leq\delta} \|f'(\theta) - f'(\theta_0)\|$.

Another local minimax lower bound

Theorem

Let $f \in C^1(\Theta)$. Suppose that, for $\theta_0 \in \Theta$ and $\delta > 0$, $B(\theta_0, \delta) \subset \Theta$ and, for all $\theta \in B(\theta_0, \delta)$, there exists $I(\theta)^{-1}$. Then, for some $D \geq 2$,

$$\inf_{T_n} \sup_{\|\theta-\theta_0\|\leq\delta} \frac{n\mathbb{E}_\theta(T_n(X_1,\dots,X_n) - f(\theta))^2}{\sigma_f^2(\theta)}$$

$$\geq 1 - D\|I(\theta_0)\|\|I(\theta_0)^{-1}\|\Bigl(\frac{\omega_{f'}(\theta_0,\delta)}{\|f'(\theta_0)\|} + \|I(\theta_0)^{-1}\|\,\omega_I(\theta_0,\delta) + \frac{\|I(\theta_0)^{-1}\|}{\delta^2 n}\Bigr).$$

Local asymptotic minimaxity

If $\|I(\theta_0)\| \lesssim 1$, $\|I(\theta_0)^{-1}\| \lesssim 1$, $\|f'(\theta_0)\| \gtrsim 1$, $\omega_{f'}(\theta_0, \delta) \to 0$ and $\omega_I(\theta_0, \delta) \to 0$ as $\delta \to 0$, then

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

Higher order bias reduction methods

Higher order bias reduction in functional estimation

Example: the square of the norm, standard normal model.

- $X_1, \ldots, X_n$ i.i.d. $\sim N(\theta, I_d), \theta \in \mathbb{R}^d$, $f(\theta) := \|\theta\|^2$
- MLE: $\bar{X}_n := \frac{X_1 + \cdots + X_n}{n}$
- $\|\bar{X}_n\|^2 = \|\theta\|^2 + 2\langle \bar{X}_n - \theta, \theta \rangle + \|\bar{X}_n - \theta\|^2$
- Bias: $\mathbb{E}_\theta \|\bar{X}_n\|^2 - \|\theta\|^2 = \frac{d}{n}$
- Bias reduction: $\hat{T}_n := \hat{T}_n(X_1, \ldots, X_n) := \|\bar{X}_n\|^2 - \frac{d}{n}$
- Mean squared error:

$$\mathbb{E}_\theta(\hat{T}_n - \|\theta\|^2)^2 = \frac{\|\theta\|^2}{n} + \frac{3d}{n^2}$$

- If $d \lesssim n$, $\mathbb{E}_\theta(\hat{T}_n - \|\theta\|)^2 \asymp n^{-1}$. Moreover,
  - if $d = o(n)$, then $n^{1/2}(\hat{T}_n - \|\theta\|^2) \xrightarrow{d} N(0, \|\theta\|^2)$
  - if $\frac{d}{n} \to \infty$, then $\frac{n}{\sqrt{d}}(\hat{T}_n - \|\theta\|^2) \to N(0, 3)$.

Higher order bias reduction in functional estimation

Example: the square of the norm, standard normal model.

- $X_1, \dots, X_n$ i.i.d. $\sim N(\theta, I_d), \theta \in \mathbb{R}^d$, $f(\theta) := \|\theta\|^2$
- MLE: $\bar{X}_n := \frac{X_1 + \cdots + X_n}{n}$
- $\|\bar{X}_n\|^2 = \|\theta\|^2 + 2\langle \bar{X}_n - \theta, \theta \rangle + \|\bar{X}_n - \theta\|^2$
- Bias: $\mathbb{E}_\theta \|\bar{X}_n\|^2 - \|\theta\|^2 = \frac{d}{n}$
- Bias reduction: $\hat{T}_n := \hat{T}_n(X_1, \dots, X_n) := \|\bar{X}_n\|^2 - \frac{d}{n}$
- Mean squared error:

$$\mathbb{E}_\theta (\hat{T}_n - \|\theta\|^2)^2 = \frac{\|\theta\|^2}{n} + \frac{3d}{n^2}$$

- If $d \lesssim n$, $\mathbb{E}_\theta (\hat{T}_n - \|\theta\|)^2 \asymp n^{-1}$. Moreover,
  - if $d = o(n)$, then $n^{1/2}(\hat{T}_n - \|\theta\|^2) \xrightarrow{d} N(0, \|\theta\|^2)$
  - if $\frac{d}{n} \to \infty$, then $\frac{n}{\sqrt{d}}(\hat{T}_n - \|\theta\|^2) \to N(0, 3)$.

Higher order bias reduction in functional estimation

A review of the bias reduction methods: Jiao and Han, *IEEE Trans. Information Theory*, 2020.

- Taylor expansions based methods: given a “base estimator" $\hat{\theta}$ and functional $f \in C^s$ for some $s = k + \rho$, $k \geq 1$, $\rho \in (0,1]$,

$$f(\theta) = \sum_{j=0}^{k} \frac{f^{(j)}(\hat{\theta})[\theta - \hat{\theta}, \ldots, \theta - \hat{\theta}]}{j!} + R =: p(\theta) + R,$$

where $p(\theta)$ is a polynomial of degree k and

$$|R| \lesssim \|f\|_{C^s} \|\hat{\theta} - \theta\|^s$$

Need to construct an estimator of $p(\theta)$ with a small bias.

Higher order bias reduction in functional estimation

Approximate solution of " bias equation": given a "base estimator" $\hat{\theta}$ and a smooth functional f, find a functional $g : \Theta \mapsto \mathbb{R}$ such that

$$\mathbb{E}_\theta g(\hat{\theta}) \approx f(\theta), \theta \in \Theta.$$

More precisely, we want to find g such that the bias

$$\mathbb{E}_\theta g(\hat{\theta}) - f(\theta)$$

of estimator $g(\hat{\theta})$ is small and g is "smooth enough" to have a decent concentration of $g(\hat{\theta}) - \mathbb{E}_\theta g(\hat{\theta})$.

Higher order bias reduction in functional estimation

Aggregation of plug-in estimators for different sample sizes: given a "base estimator" $\hat{\theta}_n$ and a smooth functional f, find $1 \leq n_1 < n_2 \cdots < n_k \leq n$ and $C_1, \ldots, C_k$ such that the bias of estimator

$$T_f(X_1, \ldots, X_n) := \sum_{j=1}^{k} C_j f(\hat{\theta}_{n_j})$$

is small. Note that $f(\hat{\theta}_{n_j})$ are often replaced by the corresponding U-statistics yielding a "jackknife" estimator.

Iterative bias reduction and bootstrap chains

Higher order bias reduction via iterative solution of "bias equation"

- $X_1, \dots, X_n$ i.i.d. $\sim P_\theta, \theta \in \Theta, \Theta \subset E$
- $\hat\theta \in \Theta$ an estimator of θ based on $X_1, \dots, X_n$
- $\sup_{\theta\in\Theta} \mathbb{E}_\theta \|\hat\theta - \theta\| \lesssim \sqrt{\frac{d}{n}}$
- d is "dimension" (or "complexity") of Θ
- Problem: given a smooth functional $f : \Theta \mapsto \mathbb{R}$, find a functional $g : \Theta \mapsto \mathbb{R}$ such that

$$\mathbb{E}_\theta g(\hat\theta) - f(\theta) = O(n^{-1/2}).$$

- We also want

$$g(\hat\theta) - \mathbb{E}_\theta g(\hat\theta) = O_{\mathbb{P}}(n^{-1/2})$$

that would lead to $\sqrt{n}$ convergence rate.

Higher order bias reduction via iterative solution of "bias equation"

- $X_1, \dots, X_n$ i.i.d. $\sim P_\theta, \theta \in \Theta, \Theta \subset E$
- $\hat{\theta} \in \Theta$ an estimator of θ based on $X_1, \dots, X_n$
- $\sup_{\theta \in \Theta} \mathbb{E}_\theta \|\hat{\theta} - \theta\| \lesssim \sqrt{\frac{d}{n}}$
- d is "dimension" (or "complexity") of Θ
- Problem: given a smooth functional $f : \Theta \mapsto \mathbb{R}$, find a functional $g : \Theta \mapsto \mathbb{R}$ such that

$$\mathbb{E}_\theta g(\hat{\theta}) - f(\theta) = O(n^{-1/2}).$$

- We also want

$$g(\hat{\theta}) - \mathbb{E}_\theta g(\hat{\theta}) = O_{\mathbb{P}}(n^{-1/2})$$

that would lead to $\sqrt{n}$ convergence rate.

An iterative method (iterated bootstrap Hall and Martin (1988))

- The bias of plug-in estimator $f(\hat{\theta})$ is
$$\mathbb{E}_\theta f(\hat{\theta}) - f(\theta) =: (\mathcal{B}f)(\theta).$$
- The first order bias correction yields an estimator
$$f(\hat{\theta}) - (\mathcal{B}f)(\hat{\theta}).$$
- The bias of estimator $(\mathcal{B}f)(\hat{\theta})$ of $(\mathcal{B}f)(\theta)$ is
$$\mathbb{E}_\theta(\mathcal{B}f)(\hat{\theta}) - (\mathcal{B}f)(\theta) = (\mathcal{B}^2 f)(\theta).$$
- The second order bias correction yields an estimator
$$f(\hat{\theta}) - (\mathcal{B}f)(\hat{\theta}) + (\mathcal{B}^2 f)(\hat{\theta}),$$
and so on.

An iterative method (iterated bootstrap Hall and Martin (1988))

- The bias of plug-in estimator $f(\hat{\theta})$ is
$$\mathbb{E}_\theta f(\hat{\theta}) - f(\theta) =: (\mathcal{B}f)(\theta).$$
- The first order bias correction yields an estimator
$$f(\hat{\theta}) - (\mathcal{B}f)(\hat{\theta}).$$
- The bias of estimator $(\mathcal{B}f)(\hat{\theta})$ of $(\mathcal{B}f)(\theta)$ is
$$\mathbb{E}_\theta(\mathcal{B}f)(\hat{\theta}) - (\mathcal{B}f)(\theta) = (\mathcal{B}^2 f)(\theta).$$
- The second order bias correction yields an estimator
$$f(\hat{\theta}) - (\mathcal{B}f)(\hat{\theta}) + (\mathcal{B}^2 f)(\hat{\theta}),$$
and so on.

An iterative method (iterated bootstrap Hall and Martin (1988))

- The bias of plug-in estimator $f(\hat{\theta})$ is
$$\mathbb{E}_\theta f(\hat{\theta}) - f(\theta) =: (\mathcal{B}f)(\theta).$$
- The first order bias correction yields an estimator
$$f(\hat{\theta}) - (\mathcal{B}f)(\hat{\theta}).$$
- The bias of estimator $(\mathcal{B}f)(\hat{\theta})$ of $(\mathcal{B}f)(\theta)$ is
$$\mathbb{E}_\theta (\mathcal{B}f)(\hat{\theta}) - (\mathcal{B}f)(\theta) = (\mathcal{B}^2 f)(\theta).$$
- The second order bias correction yields an estimator
$$f(\hat{\theta}) - (\mathcal{B}f)(\hat{\theta}) + (\mathcal{B}^2 f)(\hat{\theta}),$$
and so on.

An iterative method (iterated bootstrap Hall and Martin (1988))

- The bias of plug-in estimator $f(\hat{\theta})$ is

$$\mathbb{E}_{\theta} f(\hat{\theta}) - f(\theta) =: (\mathcal{B}f)(\theta).$$

- The first order bias correction yields an estimator

$$f(\hat{\theta}) - (\mathcal{B}f)(\hat{\theta}).$$

- The bias of estimator $(\mathcal{B}f)(\hat{\theta})$ of $(\mathcal{B}f)(\theta)$ is

$$\mathbb{E}_{\theta}(\mathcal{B}f)(\hat{\theta}) - (\mathcal{B}f)(\theta) = (\mathcal{B}^2 f)(\theta).$$

- The second order bias correction yields an estimator

$$f(\hat{\theta}) - (\mathcal{B}f)(\hat{\theta}) + (\mathcal{B}^2 f)(\hat{\theta}),$$

  and so on.

Operators $\mathcal{T}$ and $\mathcal{B}$

- $$(\mathcal{T}g)(\theta) := \mathbb{E}_\theta g(\hat\theta) = \int_\Theta g(t)P(\theta; dt), \theta \in \Theta,$$

  where

  $$P(\theta; A) := \mathbb{P}_\theta\{\hat\theta \in A\}, A \subset \Theta$$

  is a Markov kernel.
- Want to find an approximate solution g of the integral equation $\mathcal{T}g = f$ such that

  $$(\mathcal{T}g)(\theta) = f(\theta) + O(n^{-1/2}), \theta \in \Theta.$$

Neumann series

- Let $\mathcal{B} := \mathcal{T} - \mathcal{I}$. Informally, $\mathcal{T}g = f$ implies ("Neumann series")

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

- Define

$$f_k(\theta) := \sum_{j=0}^{k} (-1)^j (\mathcal{B}^j f)(\theta) = f(\theta) + \sum_{j=1}^{k} (-1)^j (\mathcal{B}^j f)(\theta), \theta \in \Theta.$$

- Then, the bias of estimator $f_k(\hat{\theta})$ is

$$\mathbb{E}_\theta f_k(\hat{\theta}) - f(\theta) = (-1)^k (\mathcal{B}^{k+1} f)(\theta), \theta \in \Theta.$$

- If $(\mathcal{B}^{k+1} f)(\theta) = 0, \theta \in \Theta$, then $f_k(\hat{\theta})$ is an unbiased estimator of $f(\theta)$.

Neumann series

- Let $\mathcal{B} := \mathcal{T} - \mathcal{I}$. Informally, $\mathcal{T}g = f$ implies ("Neumann series")

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

- Define

$$f_k(\theta) := \sum_{j=0}^{k} (-1)^j (\mathcal{B}^j f)(\theta) = f(\theta) + \sum_{j=1}^{k} (-1)^j (\mathcal{B}^j f)(\theta), \theta \in \Theta.$$

- Then, the bias of estimator $f_k(\hat{\theta})$ is

$$\mathbb{E}_\theta f_k(\hat{\theta}) - f(\theta) = (-1)^k (\mathcal{B}^{k+1} f)(\theta), \theta \in \Theta.$$

- If $(\mathcal{B}^{k+1} f)(\theta) = 0, \theta \in \Theta$, then $f_k(\hat{\theta})$ is an unbiased estimator of $f(\theta)$.

Equivariant estimators of location parameter

- P a probability distribution in a Banach space E
- $P_\theta(A) = P(A - \theta), A \subset E, \theta \in E$ the location parameter
- $X_1, \dots X_n$ i.i.d. $\sim P_\theta, \theta \in E$
- $X_j = \theta + \eta_j, j = 1, \dots, n, \quad \eta_1, \dots, \eta_n$ i.i.d. $\sim P$
- $\hat{\theta}_n(X_1, \dots, X_n)$ an equivariant estimator of location parameter θ : $\hat{\theta}_n(X_1 + a, \dots, X_n + a) = \hat{\theta}_n(X_1, \dots, X_n) + a, a \in E$
- Random shift model

$$\hat{\theta}_n(X_1, \dots, X_n) = \theta + \underbrace{\hat{\theta}_n(\eta_1, \dots, \eta_n)}_{=:\xi} = \theta + \xi$$

- WLOG: $\mathbb{E}\xi = 0 \iff \hat{\theta}_n$ is an unbiased estimator

Equivariant estimators of location parameter

- P a probability distribution in a Banach space E
- $P_\theta(A) = P(A-\theta), A \subset E, \theta \in E$ the location parameter
- $X_1, \dots X_n$ i.i.d. $\sim P_\theta, \theta \in E$
- $X_j = \theta + \eta_j, j = 1, \dots, n, \quad \eta_1, \dots, \eta_n$ i.i.d. $\sim P$
- $\hat\theta_n(X_1, \dots, X_n)$ an equivariant estimator of location parameter θ : $\hat\theta_n(X_1 + a, \dots, X_n + a) = \hat\theta_n(X_1, \dots, X_n) + a, a \in E$
- Random shift model

$$\hat\theta_n(X_1, \dots, X_n) = \theta + \underbrace{\hat\theta_n(\eta_1, \dots, \eta_n)}_{=:\xi} = \theta + \xi$$

- WLOG: $\mathbb{E}\xi = 0 \iff \hat\theta_n$ is an unbiased estimator

Equivariant estimators of location parameter

- P a probability distribution in a Banach space E
- $P_\theta(A) = P(A - \theta), A \subset E, \theta \in E$ the location parameter
- $X_1, \dots X_n$ i.i.d. $\sim P_\theta, \theta \in E$
- $X_j = \theta + \eta_j, j = 1, \dots, n, \quad \eta_1, \dots, \eta_n$ i.i.d. $\sim P$
- $\hat\theta_n(X_1, \dots, X_n)$ an equivariant estimator of location parameter θ : $\hat\theta_n(X_1 + a, \dots, X_n + a) = \hat\theta_n(X_1, \dots, X_n) + a, a \in E$
- Random shift model

$$\hat\theta_n(X_1, \dots, X_n) = \theta + \underbrace{\hat\theta_n(\eta_1, \dots, \eta_n)}_{=:\xi} = \theta + \xi$$

- WLOG: $\mathbb{E}\xi = 0 \iff \hat\theta_n$ is an unbiased estimator

A random shift model, polynomials

- $X = \theta + \xi$, $\theta \in E$, ξ a mean zero r.v. in E, $\hat{\theta}(X) = X$
- $(\mathcal{T}f)(\theta) = \mathbb{E}f(\theta + \xi), \theta \in E$
- If f is a polynomial, then $\mathcal{T}f, \mathcal{B}f$ are also polynomials and

$$\deg(\mathcal{T}f) = \deg(f),\ \deg(\mathcal{B}f) = \deg(f) - 2.$$

  - Indeed, if $f(\theta) = M[\theta, \dots, \theta]$, where $M[x_1, \dots, x_m]$ is an m-linear form, then

$$f(\theta + \xi) = f(\theta) + M[\xi, \theta, \dots, \theta] + \dots + M[\theta, \dots, \theta, \xi] + R(\theta),$$

    where $\deg(R) = m - 2$.

A random shift model, polynomials

- $X = \theta + \xi$, $\theta \in E$, ξ a mean zero r.v. in E, $\hat{\theta}(X) = X$
- $(\mathcal{T}f)(\theta) = \mathbb{E}f(\theta + \xi), \theta \in E$
- If f is a polynomial, then $\mathcal{T}f, \mathcal{B}f$ are also polynomials and

$$\deg(\mathcal{T}f) = \deg(f),\ \deg(\mathcal{B}f) = \deg(f) - 2.$$

- Indeed, if $f(\theta) = M[\theta, \dots, \theta]$, where $M[x_1, \dots, x_m]$ is an m-linear form, then

$$f(\theta + \xi) = f(\theta) + M[\xi, \theta, \dots, \theta] + \dots + M[\theta, \dots, \theta, \xi] + R(\theta),$$

where $\deg(R) = m - 2$.

A random shift model, polynomials

- $X = \theta + \xi$, $\theta \in E$, ξ a mean zero r.v. in E, $\hat{\theta}(X) = X$
- $(\mathcal{T}f)(\theta) = \mathbb{E}f(\theta + \xi), \theta \in E$
- If f is a polynomial, then $\mathcal{T}f, \mathcal{B}f$ are also polynomials and

$$\deg(\mathcal{T}f) = \deg(f),\ \deg(\mathcal{B}f) = \deg(f) - 2.$$

  - Indeed, if $f(\theta) = M[\theta, \dots, \theta]$, where $M[x_1, \dots, x_m]$ is an m-linear form, then

$$f(\theta + \xi) = f(\theta) + M[\xi, \theta, \dots, \theta] + \cdots + M[\theta, \dots, \theta, \xi] + R(\theta),$$

    where $\deg(R) = m - 2$.

Example: a random shift model, polynomials

- Let $\mathcal{P}_m$ be the space of all bounded polynomials on E of degree $\leq m$
- Then $\mathcal{T} : \mathcal{P}_m \mapsto \mathcal{P}_m$, $\mathcal{B} : \mathcal{P}_m \mapsto \mathcal{P}_{m-2}$
- If $k > m/2$, then $\mathcal{B}^k f = 0, f \in \mathcal{P}_m$
- Operator $\mathcal{T} : \mathcal{P}_m \mapsto \mathcal{P}_m$ is invertible with

$$\mathcal{T}^{-1} = \sum_{0 \leq k \leq m/2} (-1)^k \mathcal{B}^k$$

- If $k + 1 > m/2$, then $\mathrm{f}_k(X)$ is an unbiased estimator of $f(\theta)$ for all $f \in \mathcal{P}_m$

Example: a random shift model, polynomials

- Let $\mathcal{P}_m$ be the space of all bounded polynomials on E of degree $\leq m$
- Then $\mathcal{T} : \mathcal{P}_m \mapsto \mathcal{P}_m$, $\mathcal{B} : \mathcal{P}_m \mapsto \mathcal{P}_{m-2}$
- If $k > m/2$, then $\mathcal{B}^k f = 0, f \in \mathcal{P}_m$
- Operator $\mathcal{T} : \mathcal{P}_m \mapsto \mathcal{P}_m$ is invertible with

$$\mathcal{T}^{-1} = \sum_{0 \leq k \leq m/2} (-1)^k \mathcal{B}^k$$

- If $k + 1 > m/2$, then $\mathrm{f}_k(X)$ is an unbiased estimator of $f(\theta)$ for all $f \in \mathcal{P}_m$

Example: a random shift model, polynomials

- Let $\mathcal{P}_m$ be the space of all bounded polynomials on E of degree $\leq m$
- Then $\mathcal{T} : \mathcal{P}_m \mapsto \mathcal{P}_m$, $\mathcal{B} : \mathcal{P}_m \mapsto \mathcal{P}_{m-2}$
- If $k > m/2$, then $\mathcal{B}^k f = 0, f \in \mathcal{P}_m$
- Operator $\mathcal{T} : \mathcal{P}_m \mapsto \mathcal{P}_m$ is invertible with

$$\mathcal{T}^{-1} = \sum_{0 \leq k \leq m/2} (-1)^k \mathcal{B}^k$$

- If $k + 1 > m/2$, then $\mathrm{f}_k(X)$ is an unbiased estimator of $f(\theta)$ for all $f \in \mathcal{P}_m$

A random shift model: tensoriazation

Let $E = \mathbb{R}^d$, $\xi = (\xi^{(1)}, \dots, \xi^{(d)})$ with independent components.

- Tensorization: if

$$(g_1 \otimes \cdots \otimes g_d)(x_1, \dots, x_d) = g_1(x_1) \dots g_d(x_d), (x_1, \dots, x_d) \in \mathbb{R}^d,$$

then

$$\mathcal{T}(g_1 \otimes \cdots \otimes g_d) = \mathcal{T} g_1 \otimes \cdots \otimes \mathcal{T} g_d.$$

Example: Gaussian shift model, polynomials

- $E = \mathbb{R}^d$, $\xi = (\xi^{(1)}, \dots, \xi^{(d)})$ with independent components $\xi^{(j)} \sim N(0, \sigma_j^2)$
- Assume that $d = 1$ and, for simplicity, that $\xi = Z \sim N(0, 1)$. If f is a polynomial of degree m, then

$$(\mathcal{T} f)(\theta) = \mathbb{E} f(\theta + Z) = \sum_{j=0}^{m} \frac{\mathbb{E} f^{(j)}(Z)}{j!} \theta^j$$

- Let $\partial := \frac{d}{dx}$, $\partial^* = x - \partial$. Then, for $\gamma(dx) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$,

$$\langle \partial \varphi, \psi \rangle_{L_2(\gamma)} = \langle \varphi, \partial^* \psi \rangle_{L_2(\gamma)}$$

- Therefore,

$$\mathbb{E} f^{(j)}(Z) = \langle \partial^j f, 1 \rangle_{L_2(\gamma)} = \langle f, (\partial^*)^j 1 \rangle_{L_2(\gamma)} = \langle f, H_j \rangle_{L_2(\gamma)},$$

where $H_j := (\partial^*)^j 1$ are Hermite polynomials.

Example: Gaussian shift model, Hermite polynomials

- Thus $(\mathcal{T} f)(\theta) = \sum_{j=0}^{m} \frac{\langle f, H_j\rangle_{L_2(\gamma)}}{j!} \theta^j$.
- Recall also that $\langle H_k, H_j\rangle_{L_2(\gamma)} = k!\delta_{kj}$ and $f = \sum_{j=0}^{m} \frac{\langle f, H_j\rangle_{L_2(\gamma)}}{j!} H_j$.
- Thus

$$\sum_j c_j H_j \xrightarrow{\mathcal{T}} \sum_j c_j \theta^j$$

and, for a polynomial $f = \sum_j c_j \theta^j$, the solution of the equation $\mathcal{T} g = f$ is $g = \mathcal{T}^{-1} f = \sum_j c_j H_j$.
- If $\deg(f) = m$, then the unbiased estimator of $f(\theta)$ is

$$f_k(X) = \sum_j c_j H_j(X), \ k+1 > m/2.$$

Example: Gaussian shift model, Hermite polynomials

- Thus $(\mathcal{T}f)(\theta) = \sum_{j=0}^{m} \frac{\langle f, H_j\rangle_{L_2(\gamma)}}{j!}\theta^j$.
- Recall also that $\langle H_k, H_j\rangle_{L_2(\gamma)} = k!\delta_{kj}$ and $f = \sum_{j=0}^{m} \frac{\langle f, H_j\rangle_{L_2(\gamma)}}{j!} H_j$.
- Thus
$$\sum_j c_j H_j \xrightarrow{\mathcal{T}} \sum_j c_j \theta^j$$
and, for a polynomial $f = \sum_j c_j \theta^j$, the solution of the equation $\mathcal{T}g = f$ is $g = \mathcal{T}^{-1} f = \sum_j c_j H_j$.
- If $\deg(f) = m$, then the unbiased estimator of $f(\theta)$ is
$$f_k(X) = \sum_j c_j H_j(X),\ k + 1 > m/2.$$

Example: Gaussian shift model, Hermite polynomials

- Thus $(\mathcal{T} f)(\theta) = \sum_{j=0}^{m} \frac{\langle f, H_j\rangle_{L_2(\gamma)}}{j!}\theta^j$.
- Recall also that $\langle H_k, H_j\rangle_{L_2(\gamma)} = k!\delta_{kj}$ and $f = \sum_{j=0}^{m} \frac{\langle f, H_j\rangle_{L_2(\gamma)}}{j!} H_j$.
- Thus

$$\sum_j c_j H_j \xrightarrow{\mathcal{T}} \sum_j c_j \theta^j$$

  and, for a polynomial $f = \sum_j c_j \theta^j$, the solution of the equation $\mathcal{T} g = f$ is $g = \mathcal{T}^{-1} f = \sum_j c_j H_j$.
- If $\deg(f) = m$, then the unbiased estimator of $f(\theta)$ is

$$f_k(X) = \sum_j c_j H_j(X),\ k + 1 > m/2.$$

Example: Gaussian shift model, Hermite polynomials

- Thus $(\mathcal{T}f)(\theta) = \sum_{j=0}^{m} \frac{\langle f, H_j\rangle_{L_2(\gamma)}}{j!}\theta^j$.
- Recall also that $\langle H_k, H_j\rangle_{L_2(\gamma)} = k!\delta_{kj}$ and $f = \sum_{j=0}^{m} \frac{\langle f, H_j\rangle_{L_2(\gamma)}}{j!} H_j$.
- Thus

$$\sum_j c_j H_j \xrightarrow{\mathcal{T}} \sum_j c_j \theta^j$$

  and, for a polynomial $f = \sum_j c_j\theta^j$, the solution of the equation $\mathcal{T}g = f$ is $g = \mathcal{T}^{-1}f = \sum_j c_j H_j$.
- If $\deg(f) = m$, then the unbiased estimator of $f(\theta)$ is

$$f_k(X) = \sum_j c_j H_j(X),\ k+1 > m/2.$$

Example: Gaussian shift model, Hermite polynomials

More generally, by tensorization, for $d \geq 1$ and for a polynomial

$$f(\theta) = \sum_{k_1,\dots,k_d} c_{k_1,\dots,k_d} \theta_1^{k_1} \dots \theta_d^{k_d}, \theta = (\theta_1, \dots, \theta_d)$$

of degree m, the unbiased estimator of $f(\theta)$ is

$$f_k(X) = \sum_{k_1,\dots,k_d} c_{k_1,\dots,k_d} \sigma_1^{k_1} \dots \sigma_d^{k_d} H_{k_1}\Big(\frac{X^{(1)}}{\sigma_1}\Big) \dots H_{k_d}\Big(\frac{X^{(d)}}{\sigma_d}\Big),$$

$$k + 1 > m/2.$$

Smooth functionals in random shift model: representation of operators $\mathcal{T}^k$ and $\mathcal{B}^k$

Theorem

- *Let* $\xi_1, \xi_2, \dots$ *be i.i.d. copies of* ξ. *For all uniformly bounded functionals* f *and* $k \geq 1$,

$$(\mathcal{T}^k f)(\theta) = \mathbb{E} f\Big(\theta + \sum_{j=1}^{k} \xi_j\Big), \theta \in E.$$

- *Suppose* $f \in C^k(E)$ *for some* $k \geq 1$. *Then*

$$(\mathcal{B}^k f)(\theta) = \mathbb{E} f^{(k)}\Big(\theta + \sum_{j=1}^{k} U_j \xi_j\Big)[\xi_1, \dots, \xi_k], \theta \in E,$$

$U_1, \dots, U_k \sim U[0,1]$ *being i.i.d. independent of* $\xi_1, \dots, \xi_k$.

Proof

- Define $\varphi(t_1, \dots, t_k) := f\Big(\theta + \sum_{i=1}^k t_i \xi_i\Big), (t_1, \dots, t_k) \in [0, 1]^k$.
- For all $j \leq k$ and for all $(t_1, \dots, t_k) \in \{0, 1\}^k$ with $\sum_{i=1}^k t_i = j$,

$$(\mathcal{T}^j f)(\theta) = \mathbb{E}\varphi(t_1, \dots, t_k).$$

-

$$(\mathcal{B}^k f)(\theta) = ((\mathcal{T} - \mathcal{I})^k f)(\theta) = \sum_{j=0}^k (-1)^{k-j} \binom{k}{j} (\mathcal{T}^j f)(\theta)$$

$$= \sum_{j=0}^k (-1)^{k-j}\, \mathbb{E} \sum_{(t_1,\dots,t_k)\in\{0,1\}^k, \sum_{i=1}^j t_i = j} \varphi(t_1, \dots, t_k).$$

$$= \mathbb{E} \sum_{(t_1,\dots,t_k)\in\{0,1\}^k} (-1)^{k-\sum_{i=1}^k t_i} \varphi(t_1, \dots, t_k).$$

Proof

- Define $\varphi(t_1, \dots, t_k) := f\Big(\theta + \sum_{i=1}^k t_i \xi_i\Big), (t_1, \dots, t_k) \in [0, 1]^k$.
- For all $j \leq k$ and for all $(t_1, \dots, t_k) \in \{0, 1\}^k$ with $\sum_{i=1}^k t_i = j$,

$$(\mathcal{T}^j f)(\theta) = \mathbb{E}\varphi(t_1, \dots, t_k).$$

-

$$(\mathcal{B}^k f)(\theta) = ((\mathcal{T} - \mathcal{I})^k f)(\theta) = \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} (\mathcal{T}^j f)(\theta)$$

$$= \sum_{j=0}^{k} (-1)^{k-j}\, \mathbb{E} \sum_{(t_1,\dots,t_k)\in\{0,1\}^k, \sum_{i=1}^{j} t_i = j} \varphi(t_1, \dots, t_k).$$

$$= \mathbb{E} \sum_{(t_1,\dots,t_k)\in\{0,1\}^k} (-1)^{k-\sum_{i=1}^k t_i} \varphi(t_1, \dots, t_k).$$

Proof

- Define $\varphi(t_1,\dots,t_k) := f\Big(\theta + \sum_{i=1}^k t_i\xi_i\Big), (t_1,\dots,t_k) \in [0,1]^k$.
- For all $j \le k$ and for all $(t_1,\dots,t_k) \in \{0,1\}^k$ with $\sum_{i=1}^k t_i = j$,

$$(\mathcal{T}^j f)(\theta) = \mathbb{E}\varphi(t_1,\dots,t_k).$$

-

$$(\mathcal{B}^k f)(\theta) = ((\mathcal{T}-\mathcal{I})^k f)(\theta) = \sum_{j=0}^k (-1)^{k-j}\binom{k}{j}(\mathcal{T}^j f)(\theta)$$

$$= \sum_{j=0}^k (-1)^{k-j}\,\mathbb{E}\sum_{(t_1,\dots,t_k)\in\{0,1\}^k, \sum_{i=1}^j t_i = j} \varphi(t_1,\dots,t_k).$$

$$= \mathbb{E}\sum_{(t_1,\dots,t_k)\in\{0,1\}^k} (-1)^{k-\sum_{i=1}^k t_i}\varphi(t_1,\dots,t_k).$$

Proof

- Define $\varphi(t_1, \dots, t_k) := f\Big(\theta + \sum_{i=1}^k t_i \xi_i\Big), (t_1, \dots, t_k) \in [0,1]^k$.
- For all $j \le k$ and for all $(t_1, \dots, t_k) \in \{0,1\}^k$ with $\sum_{i=1}^k t_i = j$,

$$(\mathcal{T}^j f)(\theta) = \mathbb{E}\varphi(t_1, \dots, t_k).$$

-

$$(\mathcal{B}^k f)(\theta) = ((\mathcal{T} - \mathcal{I})^k f)(\theta) = \sum_{j=0}^k (-1)^{k-j} \binom{k}{j} (\mathcal{T}^j f)(\theta)$$

$$= \sum_{j=0}^k (-1)^{k-j}\, \mathbb{E} \sum_{(t_1,\dots,t_k)\in\{0,1\}^k, \sum_{i=1}^j t_i = j} \varphi(t_1, \dots, t_k).$$

$$= \mathbb{E} \sum_{(t_1,\dots,t_k)\in\{0,1\}^k} (-1)^{k - \sum_{i=1}^k t_i} \varphi(t_1, \dots, t_k).$$

Proof

- Define $\varphi(t_1, \ldots, t_k) := f\left(\theta + \sum_{i=1}^k t_i \xi_i\right), (t_1, \ldots, t_k) \in [0,1]^k$.
- For all $j \le k$ and for all $(t_1, \ldots, t_k) \in \{0,1\}^k$ with $\sum_{i=1}^k t_i = j$,

$$(\mathcal{T}^j f)(\theta) = \mathbb{E}\varphi(t_1, \ldots, t_k).$$

-

$$(\mathcal{B}^k f)(\theta) = ((\mathcal{T} - \mathcal{I})^k f)(\theta) = \sum_{j=0}^k (-1)^{k-j} \binom{k}{j} (\mathcal{T}^j f)(\theta)$$

$$= \sum_{j=0}^k (-1)^{k-j}\, \mathbb{E} \sum_{(t_1,\ldots,t_k)\in\{0,1\}^k, \sum_{i=1}^j t_i = j} \varphi(t_1, \ldots, t_k).$$

$$= \mathbb{E} \sum_{(t_1,\ldots,t_k)\in\{0,1\}^k} (-1)^{k - \sum_{i=1}^k t_i} \varphi(t_1, \ldots, t_k).$$

Proof

- Denote

$$\Delta^{(i)}\varphi(t_1,\dots,t_k) := \varphi(t_1,\dots,t_k)\big|_{t_i=1} - \varphi(t_1,\dots,t_k)\big|_{t_i=0}.$$

- Then

$$\sum_{(t_1,\dots,t_k)\in\{0,1\}^k} (-1)^{k-\sum_{i=1}^k t_i}\varphi(t_1,\dots,t_k)$$
$$= \Delta^{(1)}\dots\Delta^{(k)}\varphi(t_1,\dots,t_k)$$
$$= \int_0^1\cdots\int_0^1 \frac{\partial^k \varphi(t_1,\dots,t_k)}{\partial t_1\dots\partial t_k}dt_1\dots dt_k.$$

- For $f \in C^k(E)$,

$$\frac{\partial^k \varphi(t_1,\dots,t_k)}{\partial t_1\dots\partial t_k} = f^{(k)}\Big(\theta + \sum_{j=1}^k t_j\xi_j\Big)[\xi_1,\dots,\xi_k].$$

Proof

- Denote

$$\Delta^{(i)}\varphi(t_1,\dots,t_k) := \varphi(t_1,\dots,t_k)\Big|_{t_i=1} - \varphi(t_1,\dots,t_k)\Big|_{t_i=0}.$$

- Then

$$\sum_{(t_1,\dots,t_k)\in\{0,1\}^k} (-1)^{k-\sum_{i=1}^k t_i}\varphi(t_1,\dots,t_k)$$
$$= \Delta^{(1)}\dots\Delta^{(k)}\varphi(t_1,\dots,t_k)$$
$$= \int_0^1\cdots\int_0^1 \frac{\partial^k\varphi(t_1,\dots,t_k)}{\partial t_1\dots\partial t_k}dt_1\dots dt_k.$$

- For $f\in C^k(E)$,

$$\frac{\partial^k\varphi(t_1,\dots,t_k)}{\partial t_1\dots\partial t_k} = f^{(k)}\Bigl(\theta+\sum_{j=1}^k t_j\xi_j\Bigr)[\xi_1,\dots,\xi_k].$$

Proof

- Denote

$$\Delta^{(i)}\varphi(t_1,\dots,t_k) := \varphi(t_1,\dots,t_k)\big|_{t_i=1} - \varphi(t_1,\dots,t_k)\big|_{t_i=0}.$$

- Then

$$\sum_{(t_1,\dots,t_k)\in\{0,1\}^k} (-1)^{k-\sum_{i=1}^k t_i}\varphi(t_1,\dots,t_k)$$
$$= \Delta^{(1)}\dots\Delta^{(k)}\varphi(t_1,\dots,t_k)$$
$$= \int_0^1\cdots\int_0^1 \frac{\partial^k\varphi(t_1,\dots,t_k)}{\partial t_1\dots\partial t_k}dt_1\dots dt_k.$$

- For $f\in C^k(E)$,

$$\frac{\partial^k\varphi(t_1,\dots,t_k)}{\partial t_1\dots\partial t_k} = f^{(k)}\Bigl(\theta+\sum_{j=1}^k t_j\xi_j\Bigr)[\xi_1,\dots,\xi_k].$$

- Therefore,

$$(\mathcal{B}^k f)(\theta) = \mathbb{E}\Delta^{(1)} \dots \Delta^{(k)} \varphi$$
$$= \mathbb{E} \int_0^1 \cdots \int_0^1 f^{(k)}\Big(\theta + \sum_{j=1}^k t_j \xi_j\Big)[\xi_1, \dots, \xi_k] dt_1 \dots dt_k.$$

□

Random shift model: smoothness and bias reduction

Theorem

If $s = k + 1 + \rho$, $k \geq 0$, $\rho \in (0, 1]$ *and suppose* $f \in C^s(E)$. *Then*

$$\|\mathcal{B}^k f\|_{C^{1+\rho}} \lesssim \|f\|_{C^s} \Big(\mathbb{E}\|\xi\|\Big)^k.$$

If, in addition $\mathbb{E}\|\xi\| \lesssim 1$, *then*

$$\|f_k\|_{C^{1+\rho}} \lesssim_k \|f\|_{C^s}.$$

Theorem

Let $s = k + 1 + \rho$, $\rho \in (0, 1]$ *and suppose* $f \in C^s(E)$. *Then*

$$|\mathbb{E}_\theta f_k(X) - f(\theta)| \lesssim_s \|f\|_{C^s} \Big(\mathbb{E}\|\xi\|\Big)^k \mathbb{E}\|\xi\|^{1+\rho}$$
$$\lesssim_s \|f\|_{C^s} \Big(\mathbb{E}^{1/2}\|\xi\|^2\Big)^s.$$

Back to equivariant estimator of location parameter

- $\hat{\theta}_n$ equivariant estimator of location parameter θ
- $\xi = \hat{\theta}_n - \theta$
- If $\mathbb{E}\|\hat{\theta}_n - \theta\|^2 \lesssim \frac{d}{n}$, then

$$|\mathbb{E}_\theta f_k(\hat{\theta}_n) - f(\theta)| \lesssim_s \|f\|_{C^s}\left(\sqrt{\frac{d}{n}}\right)^s.$$

General model: bootstrap chain, K (2017, 2018)

- Recall that $\hat{\theta}$ is an estimator of parameter θ based on $X_1, \dots, X_n$ i.i.d. $\sim P_\theta, \theta \in \Theta \subset E$
-
$$\mathbb{E}_\theta \|\hat{\theta} - \theta\| \lesssim \sqrt{\frac{d}{n}},$$
where d is the dimension (or other “complexity" parameter)
-
$$(\mathcal{T}g)(\theta) := \mathbb{E}_\theta g(\hat{\theta}) = \int_\Theta g(t) P(\theta; dt), \theta \in \Theta,$$
where $P(\theta; A) := \mathbb{P}_\theta\{\hat{\theta} \in A\}, A \subset \Theta.$
- $\hat{\theta}^{(0)} = \theta \to \hat{\theta}^{(1)} = \hat{\theta} \to \hat{\theta}^{(2)} \to \dots$ the Markov chain starting at $\theta \in \Theta$ with transition probability kernel $P(\theta; A), \theta \in \Theta, A \subset \Theta.$

General model: bootstrap chain, K (2017, 2018)

- Recall that $\hat{\theta}$ is an estimator of parameter θ based on $X_1, \dots, X_n$ i.i.d. $\sim P_\theta, \theta \in \Theta \subset E$
- $$\mathbb{E}_\theta \|\hat{\theta} - \theta\| \lesssim \sqrt{\frac{d}{n}},$$
  where d is the dimension (or other “complexity" parameter)
- $$(\mathcal{T} g)(\theta) := \mathbb{E}_\theta g(\hat{\theta}) = \int_\Theta g(t) P(\theta; dt), \theta \in \Theta,$$
  where $P(\theta; A) := \mathbb{P}_\theta\{\hat{\theta} \in A\}, A \subset \Theta.$
- $\hat{\theta}^{(0)} = \theta \to \hat{\theta}^{(1)} = \hat{\theta} \to \hat{\theta}^{(2)} \to \dots$ the Markov chain starting at $\theta \in \Theta$ with transition probability kernel $P(\theta; A), \theta \in \Theta, A \subset \Theta.$

General model: bootstrap chain, K (2017, 2018)

- Recall that $\hat{\theta}$ is an estimator of parameter θ based on $X_1, \dots, X_n$ i.i.d. $\sim P_\theta, \theta \in \Theta \subset E$
- $$\mathbb{E}_\theta \|\hat{\theta} - \theta\| \lesssim \sqrt{\frac{d}{n}},$$
  where d is the dimension (or other “complexity" parameter)
- $$(\mathcal{T} g)(\theta) := \mathbb{E}_\theta g(\hat{\theta}) = \int_\Theta g(t) P(\theta; dt), \theta \in \Theta,$$
  where $P(\theta; A) := \mathbb{P}_\theta\{\hat{\theta} \in A\}, A \subset \Theta.$
- $\hat{\theta}^{(0)} = \theta \to \hat{\theta}^{(1)} = \hat{\theta} \to \hat{\theta}^{(2)} \to \dots$ the Markov chain starting at $\theta \in \Theta$ with transition probability kernel $P(\theta; A), \theta \in \Theta, A \subset \Theta.$

Bootstrap chain

- $(\mathcal{T}^k f)(\theta) = \mathbb{E}_\theta f(\hat{\theta}^{(k)}), k \geq 0.$
-

$$(\mathcal{B}^k f)(\theta) = (\mathcal{T} - \mathcal{I})^k f(\theta) = \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} (\mathcal{T}^j f)(\theta)$$

$$= \mathbb{E}_\theta \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} f(\hat{\theta}^{(j)}).$$

- Thus, $(\mathcal{B}^k f)(\theta)$ is the expectation of the k-th order difference of f along the bootstrap chain
- Note also that

$$f_k(\theta) = \mathbb{E}_\theta \sum_{j=0}^{k} (-1)^j \binom{k+1}{j+1} f(\hat{\theta}^{(j)}).$$

Bootstrap chain

- $(\mathcal{T}^k f)(\theta) = \mathbb{E}_\theta f(\hat{\theta}^{(k)}), k \geq 0.$
- $$(\mathcal{B}^k f)(\theta) = (\mathcal{T} - \mathcal{I})^k f(\theta) = \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} (\mathcal{T}^j f)(\theta)$$
$$= \mathbb{E}_\theta \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} f(\hat{\theta}^{(j)}).$$
- Thus, $(\mathcal{B}^k f)(\theta)$ is the expectation of the k-th order difference of f along the bootstrap chain
- Note also that
$$f_k(\theta) = \mathbb{E}_\theta \sum_{j=0}^{k} (-1)^j \binom{k+1}{j+1} f(\hat{\theta}^{(j)}).$$

Bootstrap chain

- $(\mathcal{T}^k f)(\theta) = \mathbb{E}_\theta f(\hat{\theta}^{(k)}), k \geq 0.$
-

$$(\mathcal{B}^k f)(\theta) = (\mathcal{T} - \mathcal{I})^k f(\theta) = \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} (\mathcal{T}^j f)(\theta)$$

$$= \mathbb{E}_\theta \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} f(\hat{\theta}^{(j)}).$$

- Thus, $(\mathcal{B}^k f)(\theta)$ is the expectation of the k-th order difference of f along the bootstrap chain
- Note also that

$$f_k(\theta) = \mathbb{E}_\theta \sum_{j=0}^{k} (-1)^j \binom{k+1}{j+1} f(\hat{\theta}^{(j)}).$$

Bootstrap chain

- $(\mathcal{T}^k f)(\theta) = \mathbb{E}_\theta f(\hat{\theta}^{(k)}), k \geq 0.$
- $$(\mathcal{B}^k f)(\theta) = (\mathcal{T} - \mathcal{I})^k f(\theta) = \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} (\mathcal{T}^j f)(\theta)$$
$$= \mathbb{E}_\theta \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} f(\hat{\theta}^{(j)}).$$
- Thus, $(\mathcal{B}^k f)(\theta)$ is the expectation of the k-th order difference of f along the bootstrap chain
- Note also that
$$f_k(\theta) = \mathbb{E}_\theta \sum_{j=0}^{k} (-1)^j \binom{k+1}{j+1} f(\hat{\theta}^{(j)}).$$

Bootstrap chain

- Recall that for $f \in C^k(\mathbb{R})$, $(\Delta_h f)(x) := f(x+h) - f(x)$ and

$$(\Delta_h^k f)(x) = \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} f(x+jh) = f^{(k)}(x) h^k + o(h^k) \text{ as } h \to 0.$$

- Note that $\mathbb{E}_{\hat{\theta}^{(k)}} \|\hat{\theta}^{(k+1)} - \hat{\theta}^{(k)}\| \lesssim \sqrt{\frac{d}{n}}$.
- Question. Suppose $d \lesssim n$. Is it true that, for $f \in C^k$,

$$|(\mathcal{B}^k f)(\theta)| \lesssim \left(\sqrt{\frac{d}{n}}\right)^k ?$$

Bootstrap chain

- Recall that for $f \in C^k(\mathbb{R})$, $(\Delta_h f)(x) := f(x+h) - f(x)$ and

$$(\Delta_h^k f)(x) = \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} f(x+jh) = f^{(k)}(x) h^k + o(h^k) \text{ as } h \to 0.$$

- Note that $\mathbb{E}_{\hat{\theta}^{(k)}} \|\hat{\theta}^{(k+1)} - \hat{\theta}^{(k)}\| \lesssim \sqrt{\frac{d}{n}}$.
- Question. Suppose $d \lesssim n$. Is it true that, for $f \in C^k$,

$$|(\mathcal{B}^k f)(\theta)| \lesssim \left(\sqrt{\frac{d}{n}}\right)^k ?$$

Bootstrap chain

- Recall that for $f \in C^k(\mathbb{R})$, $(\Delta_h f)(x) := f(x+h) - f(x)$ and

$$(\Delta_h^k f)(x) = \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} f(x+jh) = f^{(k)}(x) h^k + o(h^k) \text{ as } h \to 0.$$

- Note that $\mathbb{E}_{\hat{\theta}^{(k)}} \|\hat{\theta}^{(k+1)} - \hat{\theta}^{(k)}\| \lesssim \sqrt{\frac{d}{n}}$.
- Question. Suppose $d \lesssim n$. Is it true that, for $f \in C^k$,

$$|(\mathcal{B}^k f)(\theta)| \lesssim \left(\sqrt{\frac{d}{n}}\right)^k ?$$

Random homotopies (K& Zhilova, 2021)

Definition

A stochastic process $H(\theta; t), \theta \in \Theta, t \in [0, 1]$ with values in Θ will be called a random homotopy between θ and $\hat{\theta}$ iff

(i) $H(\theta; 0) = \theta, \theta \in \Theta$

(ii) $H(\theta; 1) \overset{d}{=} \hat{\theta}(X_1, \dots, X_n), X_1, \dots, X_n \text{ i.i.d. } \sim P_\theta, \theta \in \Theta.$

In addition, some smoothness assumptions on $H(\theta; t), \theta \in \Theta, t \in [0, 1]$ are needed.

Examples of random homotopies

Example 1 (location model)

- $X = \theta + \eta$, $\theta \in E$, $\eta \sim P$ a r.v. in E, E a linear normed space
- $X_j = \theta + \eta_j, j = 1, \dots, n$ i.i.d. copies of X
- $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ an equivariant estimator of θ
- $\hat{\theta} = \theta + \xi$, where $\xi := \hat{\theta}(\eta_1, \dots, \eta_n)$
- $H(\theta; t) = \theta + t\xi$, $\theta \in E, t \in [0, 1]$
- $\frac{d}{dt} H(\theta; t) = \hat{\theta} - \theta = \xi$

Examples of random homotopies

Example 2 (normal model)

- $X_1, \dots, X_n$ i.i.d. $\sim N(\mu, \Sigma)$, $\theta = (\mu, \Sigma) \in \mathbb{R}^d \times \mathcal{C}_+^d$.
- $\hat\theta = (\hat\mu, \hat\Sigma), \hat\mu := \bar X, \hat\Sigma := \frac{1}{n-1}\sum_{j=1}^n (X_j - \bar X) \otimes (X_j - \bar X)$
- Let $Z_1, \dots, Z_n$ be i.i.d. $\sim N(0; I_d)$

$$H(\theta; t) := (1-t)(\mu, \Sigma) + t(\mu + \Sigma^{1/2}\hat\mu_Z, \Sigma^{1/2}\hat\Sigma_Z\Sigma^{1/2})$$

- $\frac{d}{dt}H(\theta; t) \overset{d}{=} \hat\theta - \theta$

Examples of random homotopies

Example 3 (Moser coupling)

- S compact Riemannian manifold, P normalized Riemannian volume
- $X_1, \dots, X_n$ i.i.d. $\sim P_\theta, \theta \in \Theta$, $P_\theta(dx) = p_\theta(x)P(dx)$, $\Theta \subset E$ a convex set,
- $(x, \theta) \mapsto p_\theta(x)$ smooth, p_θ bounded away from 0
- Moser coupling: Let u_θ be a solution of Poisson equation: $\Delta u_\theta = 1 - p_\theta$ and let

$$v_\theta(t; x) := \frac{\nabla u_\theta(x)}{1 - t + tp_\theta(x)}, x \in S, t \in [0, 1]$$

- The vector field v_θ generates a flow $T_0^t(x)$ on the manifold. Let $g_\theta(x) = T_0^1(x), x \in S$
- Then (Moser): $P_\theta = P \circ g_\theta^{-1}$

Examples of random homotopies (Moser coupling)

Example 3 (cont.)

- Let $U_1, \ldots, U_n$ i.i.d. $\sim P$
- $(X_1, \ldots, X_n) \overset{d}{=} (g_\theta(U_1), \ldots, g_\theta(U_n))$
- $\hat{\theta}(X_1, \ldots, X_n) \overset{d}{=} \hat{\theta}(g_\theta(U_1), \ldots, g_\theta(U_n))$
- Let

$$H(\theta; t) := (1-t)\theta + t\hat{\theta}(g_\theta(U_1), \ldots, g_\theta(U_n)), \theta \in \Theta, t \in [0,1]$$

- $\frac{d}{dt} H(\theta; t) \overset{d}{=} \hat{\theta} - \theta$
- A similar construction could be used when S is not compact and is equipped with probability measure $P(dx) = e^{-V(x)}dx$.

Random homotopies and representation of bootstrap chain

- $F_1 : \Theta \times [0,1]^k \mapsto \Theta, F_2 : \Theta \times [0,1]^l \mapsto \Theta$
- $F_2 \bullet F_1 : \Theta \times [0,1]^{k+l} \mapsto \Theta,$

$$(F_2 \bullet F_1)(\theta, (t, s)) := F_2(F_1(\theta; t); s), t \in [0,1]^k, s \in [0,1]^l.$$

- Let H be a random homotopy between θ and $\hat{\theta}$ and $H_1, \dots, H_k$ be its i.i.d. copies. Define

$$G_k := H_k \bullet \cdots \bullet H_1 : \Theta \times [0,1]^k \mapsto \Theta.$$

$$G_3(\theta; t_1, t_2, t_3) = H_3(H_2(H_1(\theta; t_1); t_2); t_3).$$

Random homotopies and representation of bootstrap chain

- $F_1 : \Theta \times [0,1]^k \mapsto \Theta, F_2 : \Theta \times [0,1]^l \mapsto \Theta$
- $F_2 \bullet F_1 : \Theta \times [0,1]^{k+l} \mapsto \Theta,$

$$(F_2 \bullet F_1)(\theta, (t, s)) := F_2(F_1(\theta; t); s), t \in [0,1]^k, s \in [0,1]^l.$$

- Let H be a random homotopy between θ and $\hat{\theta}$ and $H_1, \dots, H_k$ be its i.i.d. copies. Define

$$G_k := H_k \bullet \cdots \bullet H_1 : \Theta \times [0,1]^k \mapsto \Theta.$$

$$G_3(\theta; t_1, t_2, t_3) = H_3(H_2(H_1(\theta; t_1); t_2); t_3).$$

Random homotopies and representation of bootstrap chain

Proposition

(i) Let $\tilde{\theta}_k := G_k(\theta; 1, \dots, 1), k \geq 1$ and $\tilde{\theta}_0 = \theta$. Then

$$\{\tilde{\theta}_k : k \geq 0\} \overset{d}{=} \{\hat{\theta}_k : k \geq 0\}.$$

(ii) Moreover, for $j \leq k$,

$$\hat{\theta}_j \overset{d}{=} G_k(\theta; t_1, \dots, t_k),\ (t_1, \dots, t_k) \in \{0, 1\}^k, \sum_{i=1}^k t_i = j.$$

Random homotopies and representation of operators $\mathcal{T}^k, \mathcal{B}^k$

Then

- $$(\mathcal{T}^k f)(\theta) = \mathbb{E}_\theta f(\hat{\theta}^{(k)}) = \mathbb{E} f(G_k(\theta; 1, \dots, 1))$$

  and for all $1 \leq j \leq k$, $(t_1, \dots, t_k) \in \{0, 1\}^k$ with $\sum_{i=1}^k t_i = j$,

  $$(\mathcal{T}^j f)(\theta) = \mathbb{E}_\theta f(\hat{\theta}^{(j)}) = \mathbb{E} f(G_k(\theta; t_1, \dots, t_k))$$

- $$(\mathcal{B}^k f)(\theta) = \mathbb{E} \frac{\partial^k f(G_k(\theta; U_1, \dots, U_k))}{\partial t_1 \dots \partial t_k}, \ U_1, \dots, U_k \text{ i.i.d. } U[0, 1].$$

Proof of representation of $\mathcal{B}^k f$

$$(\mathcal{B}^k f)(\theta) = (\mathcal{T} - \mathcal{I})^k f(\theta)$$

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

$$= \sum_{j=0}^{k} (-1)^{k-j} \sum_{(t_1,\dots,t_k)\in\{0,1\}^k, \sum_i t_i = j} \mathbb{E} f(G_k(\theta; t_1, \dots, t_k))$$

$$= \mathbb{E} \sum_{(t_1,\dots,t_k)\in\{0,1\}^k} (-1)^{k-\sum_{i=1}^k t_i} f(G_k(\theta; t_1, \dots, t_k))$$

$$= \mathbb{E} \Delta^{(1)} \dots \Delta^{(k)} f(G_k(\theta; t_1, \dots, t_k))$$

$$= \mathbb{E} \int_0^1 \cdots \int_0^1 \frac{\partial^k f(G_k(\theta; t_1, \dots, t_k))}{\partial t_1 \dots \partial t_k} dt_1 \dots dt_k$$

Francesco Faà di Bruno

- Faà di Bruno's Formula:

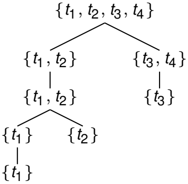
$$\frac{d^n}{dx^n} f(g(x)) = \sum \frac{n!}{k_1! \dots k_n!} f^{(k_1+\cdots+k_n)}(g(x)) \prod_{j=1}^{n} \left(\frac{g^{(j)}(x)}{j!} \right)^{k_j},$$

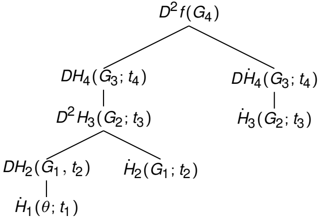
where the sum is over all $(k_1, \dots, k_n)$, $k_j \geq 0$, $\sum_{j=1}^n jk_j = n$.

- Cavaliere Francesco Faà di Bruno, *Sullo sviluppo delle Funzioni, Annali di Scienze Matematiche e Fisiche 6 (1855) 479-480*

A representation formula for $\mathcal{B}^k f$

$$(\mathcal{B}^k f)(\theta) = \sum_{\tau \in \mathcal{T}_k} \mathbb{E}\partial_\tau f(G_k(\theta; U_1, \dots, U_k)), \theta \in \Theta.$$





For this tree,

$$\partial_\tau f(G_4) = D^2 f(G_4)[DH_4(G_3; t_4)[D^2 H_3(G_2; t_3)[DH_2(G_1; t_2)[\dot{H}_1(\theta; t_1)], \dot{H}_2(G_1; t_2)]], D\dot{H}_4(G_3; t_4)[\dot{H}_3(G_2; t_3)]].$$

Bounds on operators $\mathcal{B}^k$

If functional f and random homotopy H are sufficiently smooth, such representations of $(\mathcal{B}^k f)(\theta)$ could be used to prove upper bounds

$$|(\mathcal{B}^k f)(\theta)| \lesssim$$
$$\|f\|_{C^k}\biggl(\mathbb{E}\Bigl(\sup_{t\in[0,1]}\|H(\cdot;t)\|_{C^{k-1}}\vee 1\Bigr)^k \sup_{t\in[0,1]}\Bigl\|\frac{d}{dt}H(\cdot;t)\Bigr\|_{C^{k-1}}\biggr)^k$$

of the order $O\Bigl(\Bigl(\sqrt{\frac{d}{n}}\Bigr)^k\Bigr)$ provided that (with a high probability)

$$\sup_{t\in[0,1]}\Bigl\|\frac{d}{dt}H(\cdot;t)\Bigr\|_{C^{k-1}} \lesssim \sqrt{\frac{d}{n}}.$$

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

Moreover, the following bound on the bias of estimator $f_k(\hat{\theta})$ holds

$$|\mathbb{E}_\theta f_k(\hat{\theta}) - f(\theta)| \lesssim_s \|f\|_{C^s}\left(\sqrt{\frac{d}{n}}\right)^s$$

and one can also study smoothness properties of functional f_k that are crucial in deriving concentration and normal approximation bounds for $f_k(\hat{\theta})$.

Risk bounds in functional estimation via iterative bias reduction method

Loss functions and Orlicz norm

- $\psi : \mathbb{R}_+ \mapsto \mathbb{R}_+$ be a convex nondecreasing loss function such that $\psi(0) = 0$
- $\|\eta\|_\psi := \|\eta\|_{L_\psi(\mathbb{P})} := \inf\left\{c > 0 : \mathbb{E}\psi\left(\frac{|\eta|}{c}\right) \leq 1\right\}$
  - 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, $\psi_\alpha(u) := e^{u^\alpha} - 1, \alpha \geq 1$
  - Note that

$$\|\eta\|_{\psi_\alpha} \asymp \sup_{p\geq 1} p^{-1/\alpha}\|\eta\|_{L_p}, \alpha \geq 1$$

  - The right-hand side could be used to define ψ_α-norm for $\alpha < 1$
- Given two loss functions ψ and φ, we write $\psi \prec \varphi$ (ψ is dominated by φ) iff

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

for some constants $c_1, c_2 > 0$.

Loss functions and Orlicz norm

- $\psi : \mathbb{R}_+ \mapsto \mathbb{R}_+$ be a convex nondecreasing loss function such that $\psi(0) = 0$
- $\|\eta\|_\psi := \|\eta\|_{L_\psi(\mathbb{P})} := \inf\left\{c > 0 : \mathbb{E}\psi\left(\frac{|\eta|}{c}\right) \leq 1\right\}$
  - 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, $\psi_\alpha(u) := e^{u^\alpha} - 1, \alpha \geq 1$
  - Note that
  $$\|\eta\|_{\psi_\alpha} \asymp \sup_{p \geq 1} p^{-1/\alpha} \|\eta\|_{L_p}, \alpha \geq 1$$
  - The right-hand side could be used to define ψ_α-norm for $\alpha < 1$
- Given two loss functions ψ and φ, we write $\psi \prec \varphi$ (ψ is dominated by φ) iff
$$\psi(u) \leq c_1 \varphi(c_2 u), u \geq 0$$
for some constants $c_1, c_2 > 0$.

Loss functions and Orlicz norm

- $\psi : \mathbb{R}_+ \mapsto \mathbb{R}_+$ be a convex nondecreasing loss function such that $\psi(0) = 0$
- $\|\eta\|_\psi := \|\eta\|_{L_\psi(\mathbb{P})} := \inf\left\{c > 0 : \mathbb{E}\psi\left(\frac{|\eta|}{c}\right) \leq 1\right\}$
  - 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, $\psi_\alpha(u) := e^{u^\alpha} - 1, \alpha \geq 1$
  - Note that
$$\|\eta\|_{\psi_\alpha} \asymp \sup_{p\geq 1} p^{-1/\alpha}\|\eta\|_{L_p}, \alpha \geq 1$$
  - The right-hand side could be used to define ψ_α-norm for $\alpha < 1$
- Given two loss functions ψ and φ, we write $\psi \prec \varphi$ (ψ is dominated by φ) iff
$$\psi(u) \leq c_1\varphi(c_2 u), u \geq 0$$
for some constants $c_1, c_2 > 0$.

Loss functions and Orlicz norm

- $\psi : \mathbb{R}_+ \mapsto \mathbb{R}_+$ be a convex nondecreasing loss function such that $\psi(0) = 0$
- $\|\eta\|_\psi := \|\eta\|_{L_\psi(\mathbb{P})} := \inf\left\{c > 0 : \mathbb{E}\psi\left(\frac{|\eta|}{c}\right) \leq 1\right\}$
  - 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, $\psi_\alpha(u) := e^{u^\alpha} - 1$, $\alpha \geq 1$
  - Note that

$$\|\eta\|_{\psi_\alpha} \asymp \sup_{p\geq 1} p^{-1/\alpha}\|\eta\|_{L_p}, \alpha \geq 1$$

  - The right-hand side could be used to define ψ_α-norm for $\alpha < 1$
- Given two loss functions ψ and φ, we write $\psi \prec \varphi$ (ψ is dominated by φ) iff

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

for some constants $c_1, c_2 > 0$.

Loss functions and Orlicz norm

- $\psi : \mathbb{R}_+ \mapsto \mathbb{R}_+$ be a convex nondecreasing loss function such that $\psi(0) = 0$
- $\|\eta\|_\psi := \|\eta\|_{L_\psi(\mathbb{P})} := \inf\left\{c > 0 : \mathbb{E}\psi\left(\frac{|\eta|}{c}\right) \leq 1\right\}$
  - 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, $\psi_\alpha(u) := e^{u^\alpha} - 1, \alpha \geq 1$
  - Note that

$$\|\eta\|_{\psi_\alpha} \asymp \sup_{p\geq 1} p^{-1/\alpha}\|\eta\|_{L_p}, \alpha \geq 1$$

  - The right-hand side could be used to define ψ_α-norm for $\alpha < 1$
- Given two loss functions ψ and φ, we write $\psi \prec \varphi$ (ψ is dominated by φ) iff

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

for some constants $c_1, c_2 > 0$.

Loss functions and Orlicz norm

- $\psi : \mathbb{R}_+ \mapsto \mathbb{R}_+$ be a convex nondecreasing loss function such that $\psi(0) = 0$
- $\|\eta\|_\psi := \|\eta\|_{L_\psi(\mathbb{P})} := \inf\left\{c > 0 : \mathbb{E}\psi\left(\frac{|\eta|}{c}\right) \leq 1\right\}$
  - 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, $\psi_\alpha(u) := e^{u^\alpha} - 1$, $\alpha \geq 1$
  - Note that

$$\|\eta\|_{\psi_\alpha} \asymp \sup_{p \geq 1} p^{-1/\alpha} \|\eta\|_{L_p}, \alpha \geq 1$$

  - The right-hand side could be used to define ψ_α-norm for $\alpha < 1$
- Given two loss functions ψ and φ, we write $\psi \prec \varphi$ (ψ is dominated by φ) iff

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

for some constants $c_1, c_2 > 0$.

Gaussian location model

- $X_j = \theta + \xi_j, j = 1, \ldots, n, \theta \in E$, E is a separable Banach space
- $\xi, \xi_1, \ldots, \xi_n$ i.i.d. mean zero Gaussian r.v. in E with covariance operator Σ
- The level of noise could be characterized by its "weak variance"

$$\sup_{\|u\|\leq 1} \mathbb{E}\langle \xi, u\rangle^2 = \sup_{\|u\|,\|v\|\leq 1} \langle \Sigma u, v\rangle = \|\Sigma\|$$

and its "strong variance

$$\mathbb{E}\|\xi\|^2 = \mathbb{E} \sup_{\|u\|,\|v\|\leq 1} \langle \xi, u\rangle\langle \xi, v\rangle$$

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

the effective rank of Σ (could be "large")

Gaussian location model

- $X_j = \theta + \xi_j, j = 1, \dots, n, \theta \in E$, E is a separable Banach space
- $\xi, \xi_1, \dots, \xi_n$ i.i.d. mean zero Gaussian r.v. in E with covariance operator Σ
- The level of noise could be characterized by its "weak variance"

$$\sup_{\|u\|\leq 1} \mathbb{E}\langle \xi, u\rangle^2 = \sup_{\|u\|,\|v\|\leq 1} \langle \Sigma u, v\rangle = \|\Sigma\|$$

and its "strong variance

$$\mathbb{E}\|\xi\|^2 = \mathbb{E} \sup_{\|u\|,\|v\|\leq 1} \langle \xi, u\rangle\langle \xi, v\rangle$$

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

the effective rank of Σ (could be "large")

Gaussian location model: risk bounds (K& Zhilova (2019))

We use iterative bias reduction (with k iterations) based on $\bar{X}_n$ to estimate $f(\theta)$ by $f_k(\bar{X}_n)$, $f_k := \sum_{j=0}^k (-1)^j \mathcal{B}^j f$.

Theorem

Let $s = k + 1 + \rho$ for some $\rho \in (0, 1]$ and suppose $f \in C^s(E)$. Then, for all $\psi \prec \psi_2$,

$$\sup_{\theta \in E} \|f_k(\bar{X}_n) - f(\theta)\|_{L_\psi(\mathbb{P}_\theta)} \lesssim \|f\|_{C^s} \Big(\Big(\frac{\|\Sigma\|^{1/2}}{\sqrt{n}} \vee \Big(\frac{\mathbb{E}^{1/2}\|\xi\|^2}{\sqrt{n}}\Big)^s\Big) \wedge 1\Big).$$

In particular,

$$\sup_{\theta \in E} \mathbb{E}_\theta (f_k(\bar{X}_n) - f(\theta))^2 \lesssim \|f\|^2_{C^s} \Big(\Big(\frac{\|\Sigma\|}{n} \vee \Big(\frac{\mathbb{E}\|\xi\|^2}{n}\Big)^s\Big) \wedge 1\Big).$$

Gaussian concentration

- $g : E \mapsto \mathbb{R}$ locally Lipschitz
- Local Lipschitz constant:

$$(Lg)(x) := \inf_{U \ni x} \sup_{x_1, x_2 \in U, x_1 \neq x_2} \frac{|g(x_1) - g(x_2)|}{\|x_1 - x_2\|}.$$

- $\xi \sim N(0, \Sigma)$ in E
- For all $p \geq 1$,

$$\|g(\xi) - \mathbb{E}g(\xi)\|_{L_p} \lesssim \|\Sigma\|^{1/2}\sqrt{p}\|(Lg)(\xi)\|_{L_p}.$$

- For a Lipschitz function g, this implies

$$\|g(\xi) - \mathbb{E}g(\xi)\|_{\psi_2} \lesssim \|g\|_{\mathrm{Lip}}\|\Sigma\|^{1/2}.$$

Gaussian concentration

- $g : E \mapsto \mathbb{R}$ locally Lipschitz
- Local Lipschitz constant:

$$(Lg)(x) := \inf_{U \ni x} \sup_{x_1, x_2 \in U, x_1 \neq x_2} \frac{|g(x_1) - g(x_2)|}{\|x_1 - x_2\|}.$$

- $\xi \sim N(0, \Sigma)$ in E
- For all $p \geq 1$,

$$\|g(\xi) - \mathbb{E}g(\xi)\|_{L_p} \lesssim \|\Sigma\|^{1/2}\sqrt{p}\|(Lg)(\xi)\|_{L_p}.$$

- For a Lipschitz function g, this implies

$$\|g(\xi) - \mathbb{E}g(\xi)\|_{\psi_2} \lesssim \|g\|_{\mathrm{Lip}}\|\Sigma\|^{1/2}.$$

Gaussian concentration

- $g : E \mapsto \mathbb{R}$ locally Lipschitz
- Local Lipschitz constant:

$$(Lg)(x) := \inf_{U \ni x} \sup_{x_1, x_2 \in U, x_1 \neq x_2} \frac{|g(x_1) - g(x_2)|}{\|x_1 - x_2\|}.$$

- $\xi \sim N(0, \Sigma)$ in E
- For all $p \geq 1$,

$$\|g(\xi) - \mathbb{E}g(\xi)\|_{L_p} \lesssim \|\Sigma\|^{1/2}\sqrt{p}\|(Lg)(\xi)\|_{L_p}.$$

- For a Lipschitz function g, this implies

$$\|g(\xi) - \mathbb{E}g(\xi)\|_{\psi_2} \lesssim \|g\|_{\mathrm{Lip}}\|\Sigma\|^{1/2}.$$

Gaussian location model: proof of risk bounds

- Recall that

$$\|f_k\|_{\mathrm{Lip}} \leq \|f_k\|_{C^{1+\rho}} \lesssim_k \|f\|_{C^s}$$

and

$$|\mathbb{E}_\theta f_k(\bar{X}_n) - f(\theta)| \lesssim_s \|f\|_{C^s} \biggl(\frac{\mathbb{E}^{1/2}\|\xi\|^2}{\sqrt{n}}\biggr)^s.$$

- Note that $\bar{\xi}_n = \frac{\xi_1+\dots+\xi_n}{n} \sim N(0, n^{-1}\Sigma)$. Thus, by Gaussian concentration,

$$\|f_k(\bar{X}_n) - \mathbb{E}f_k(\bar{X}_n)\|_{\psi_2} = \|f_k(\theta+\bar{\xi}_n) - \mathbb{E}f_k(\theta+\bar{\xi}_n)\|_{\psi_2} \lesssim_k \|f\|_{C^s} \frac{\|\Sigma\|^{1/2}}{\sqrt{n}}.$$

Gaussian location model: proof of risk bounds

- Recall that

$$\|f_k\|_{\mathrm{Lip}} \leq \|f_k\|_{C^{1+\rho}} \lesssim_k \|f\|_{C^s}$$

and

$$|\mathbb{E}_\theta f_k(\bar{X}_n) - f(\theta)| \lesssim_s \|f\|_{C^s}\Big(\frac{\mathbb{E}^{1/2}\|\xi\|^2}{\sqrt{n}}\Big)^s.$$

- Note that $\bar{\xi}_n = \frac{\xi_1+\dots+\xi_n}{n} \sim N(0, n^{-1}\Sigma)$. Thus, by Gaussian concentration,

$$\|f_k(\bar{X}_n) - \mathbb{E}f_k(\bar{X}_n)\|_{\psi_2} = \|f_k(\theta+\bar{\xi}_n) - \mathbb{E}f_k(\theta+\bar{\xi}_n)\|_{\psi_2} \lesssim_k \|f\|_{C^s}\frac{\|\Sigma\|^{1/2}}{\sqrt{n}}.$$

Gaussian location model: phase transition in convergence rates

- If $\|\Sigma\| \lesssim 1$, $\mathbf{r}(\Sigma) \lesssim n^{\alpha}$ for some $\alpha \in (0,1)$ and $s \geq \frac{1}{1-\alpha}$, then

$$\sup_{\|f\|_{C^s} \leq 1} \sup_{\theta \in E} \mathbb{E}_{\theta}(f_k(\bar{X}_n) - f(\theta))^2 = O(n^{-1}).$$

  If $\|\Sigma\| \gtrsim 1$, $\mathbf{r}(\Sigma) \gtrsim n^{\alpha}$, then for $s < \frac{1}{1-\alpha}$, the convergence rate is slower than n^{-1}.

- In the case $E = \mathbb{R}^d, \xi = \sigma Z, Z \sim N(0; I_d)$, $\|\Sigma\| = \sigma^2$, $\mathbb{E}\|\xi\|^2 = \sigma^2 d$,

$$\sup_{\|f\|_{C^s} \leq 1} \inf_{T_n} \sup_{\|\theta\| \leq 1} \mathbb{E}_{\theta}(T_n - f(\theta))^2 \gtrsim \left(\left(\frac{\|\Sigma\|}{n} \bigvee \left(\frac{\mathbb{E}\|\xi\|^2}{n}\right)^s\right) \wedge 1\right)$$

$$= \left(\frac{\sigma^2}{n} \bigvee \left(\frac{\sigma^2 d}{n}\right)^s\right) \wedge 1$$

Gaussian location model: phase transition in convergence rates

- If $\|\Sigma\| \lesssim 1$, $\mathbf{r}(\Sigma) \lesssim n^{\alpha}$ for some $\alpha \in (0,1)$ and $s \geq \frac{1}{1-\alpha}$, then

$$\sup_{\|f\|_{C^s} \leq 1} \sup_{\theta \in E} \mathbb{E}_{\theta}(f_k(\bar{X}_n) - f(\theta))^2 = O(n^{-1}).$$

  If $\|\Sigma\| \gtrsim 1$, $\mathbf{r}(\Sigma) \gtrsim n^{\alpha}$, then for $s < \frac{1}{1-\alpha}$, the convergence rate is slower than n^{-1}.

- In the case $E = \mathbb{R}^d, \xi = \sigma Z, Z \sim N(0; I_d)$, $\|\Sigma\| = \sigma^2$, $\mathbb{E}\|\xi\|^2 = \sigma^2 d$,

$$\sup_{\|f\|_{C^s} \leq 1} \inf_{T_n} \sup_{\|\theta\| \leq 1} \mathbb{E}_{\theta}(T_n - f(\theta))^2 \gtrsim \left(\left(\frac{\|\Sigma\|}{n} \bigvee \left(\frac{\mathbb{E}\|\xi\|^2}{n}\right)^s\right) \wedge 1\right)$$

$$= \left(\frac{\sigma^2}{n} \bigvee \left(\frac{\sigma^2 d}{n}\right)^s\right) \wedge 1$$

Gaussian location model: normal approximation

Theorem

Let $s = k + 1 + \rho$ for some $\rho \in (0, 1]$ and suppose $f \in C^s(E)$. Suppose $\mathbb{E}\|\xi\|^2 \lesssim n$. Then

$$\|f_k(\bar{X}_n) - f(\theta) - \langle \bar{X}_n - \theta, f'(\theta)\rangle\|_{L_{\psi_1}(\mathbb{P}_\theta)}$$
$$\lesssim_s \|f\|_{C^s}\Big(\frac{\|\Sigma\|^{1/2}}{\sqrt{n}}\frac{\mathbb{E}^{1/2}\|\xi\|^2}{\sqrt{n}} + \Big(\frac{\mathbb{E}^{1/2}\|\xi\|^2}{\sqrt{n}}\Big)^s\Big).$$

Note that $\sqrt{n}\langle \bar{X}_n - \theta, f'(\theta)\rangle = \sigma_f(\theta)Z, Z \sim N(0, 1)$, where

$$\sigma_f^2(\theta) := \langle \Sigma f'(\theta), f'(\theta)\rangle.$$

Gaussian location model: normal approximation and asymptotic efficiency

- If $\|\Sigma\| \lesssim 1$, $\mathbf{r}(\Sigma) \lesssim n^{\alpha}$ for some $\alpha \in (0,1)$ and $s > \frac{1}{1-\alpha}$, then

$$W_{2,\mathbb{P}_\theta}\Big(\sqrt{n}(f_k(\bar{X}_n) - f(\theta)), \sigma_f(\theta) Z\Big) \to 0$$

  and

$$\sqrt{n}\|f_k(\bar{X}_n) - f(\theta)\|_{L_2(\mathbb{P}_\theta)} - \sigma_f(\theta) \to 0$$

  as $n \to \infty$, where $W_{2,\mathbb{P}_\theta}$ is the Wasserstein distance.
- The asymptotic efficiency of estimator $f_k(\bar{X}_n)$ follows from a nonasymptotic version of Hájek-Le Cam local minimax lower bound.

Proof of normal approximation

- $S_g(x,h) := g(x+h) - g(x) - \langle h, g'(x)\rangle$
- Bound on local Lipschitz function of $E \ni h \mapsto S_g(x,h)$: if g' is Lipschitz,

$$(LS_g(x,\cdot))(h) = \|g'(x+h) - g'(x)\| \leq \|g'\|_{\mathrm{Lip}}\|h\|$$

- Concentration bound for $S_g(x,\xi) = g(x+\xi) - g(x) - \langle \xi, g'(x)\rangle$:

$$\|S_g(x,\xi) - \mathbb{E}S_g(x,\xi)\|_{L_p} \lesssim \|g'\|_{\mathrm{Lip}}\|\Sigma\|^{1/2}\sqrt{p}\|\|\xi\|\|_{L_p}$$
$$\lesssim \|g'\|_{\mathrm{Lip}}\|\Sigma\|^{1/2}\sqrt{p}\mathbb{E}\|\xi\| + \|g'\|_{\mathrm{Lip}}\|\Sigma\|^{1/2}\sqrt{p}\|\|\xi\| - \mathbb{E}\|\xi\|\|_{L_p}$$
$$\lesssim \|g'\|_{\mathrm{Lip}}\|\Sigma\|^{1/2}\mathbb{E}\|\xi\|\sqrt{p} + \|g'\|_{\mathrm{Lip}}\|\Sigma\|p.$$

Proof of normal approximation

- $S_g(x,h) := g(x+h) - g(x) - \langle h, g'(x)\rangle$
- Bound on local Lipschitz function of $E \ni h \mapsto S_g(x,h)$: if g' is Lipschitz,

$$(LS_g(x,\cdot))(h) = \|g'(x+h) - g'(x)\| \leq \|g'\|_{\rm Lip}\|h\|$$

- Concentration bound for $S_g(x,\xi) = g(x+\xi) - g(x) - \langle \xi, g'(x)\rangle$:

$$\|S_g(x,\xi) - \mathbb{E}S_g(x,\xi)\|_{L_p} \lesssim \|g'\|_{\rm Lip}\|\Sigma\|^{1/2}\sqrt{p}\|\|\xi\|\|_{L_p}$$
$$\lesssim \|g'\|_{\rm Lip}\|\Sigma\|^{1/2}\sqrt{p}\mathbb{E}\|\xi\| + \|g'\|_{\rm Lip}\|\Sigma\|^{1/2}\sqrt{p}\|\|\xi\| - \mathbb{E}\|\xi\|\|_{L_p}$$
$$\lesssim \|g'\|_{\rm Lip}\|\Sigma\|^{1/2}\mathbb{E}\|\xi\|\sqrt{p} + \|g'\|_{\rm Lip}\|\Sigma\|p.$$

Proof of normal approximation

It remains to observe that

$$f_k(\bar{X}_n) - \mathbb{E}_\theta f_k(\bar{X}_n) - \langle \bar{X}_n - \theta, f'(\theta)\rangle$$
$$= S_{f_k}(\theta, \bar{\xi}_n) - \mathbb{E} S_{f_k}(\theta, \bar{\xi}_n) + \langle \bar{X}_n - \theta, f_k'(\theta) - f'(\theta)\rangle,$$

$$\|S_{f_k}(\theta, \bar{\xi}_n) - \mathbb{E} S_{f_k}(\theta, \bar{\xi}_n)\|_{L_{\psi_1}} \lesssim \|f_k'\|_{\rm Lip} \frac{\|\Sigma\|^{1/2}}{\sqrt{n}} \mathbb{E}\|\bar{\xi}_n\| + \|f_k'\|_{\rm Lip} \frac{\|\Sigma\|}{n}$$
$$\lesssim \|f\|_{C^s} \frac{\|\Sigma\|^{1/2}}{\sqrt{n}} \frac{\mathbb{E}^{1/2}\|\xi\|^2}{\sqrt{n}},$$

$$\|\langle \bar{X}_n - \theta, f_k'(\theta) - f'(\theta)\rangle\|_{\psi_2} \lesssim_s \|f\|_{C^s} \frac{\|\Sigma\|^{1/2}}{\sqrt{n}} \frac{\mathbb{E}^{1/2}\|\xi\|^2}{\sqrt{n}}.$$

Functional estimation in log-concave location families (K& Wahl 21)

- $P(dx) = e^{-V(x)}dx$, $V : \mathbb{R}^d \mapsto \mathbb{R}$ convex
- $P_\theta(dx) := p_\theta(x)dx$, $p_\theta(x) := e^{-V(x-\theta)}, x \in \mathbb{R}^d, \theta \in \mathbb{R}^d$
- $X \sim P_\theta \iff X = \theta + \xi, \xi \sim P$, w.l.o.g. $\mathbb{E}\xi = 0$
- Let $\Sigma_\xi := \mathbb{E}(\xi \otimes \xi)$. Assume that $\|\Sigma_\xi\| \lesssim 1$
- $X_1, \dots, X_n$ i.i.d. copies of X
- Regularity Conditions:
  1. V is strictly convex and twice continuously differentiable
  2. for some constants $M, L > 0$, $\|V''\|_{L_\infty} \le M$, $\|V''\|_{\mathrm{Lip}} \le L$
  3. for some constant $m > 0$, $\mathcal{I} \succeq mI_d$, where

$$\mathcal{I} := \mathbb{E}V'(\xi) \otimes V'(\xi) = \mathbb{E}V''(\xi)$$

is the Fisher information.

Functional estimation in log-concave location families (K& Wahl 21)

- $P(dx) = e^{-V(x)}dx,\ V : \mathbb{R}^d \mapsto \mathbb{R}$ convex
- $P_\theta(dx) := p_\theta(x)dx,\ p_\theta(x) := e^{-V(x-\theta)}, x \in \mathbb{R}^d, \theta \in \mathbb{R}^d$
- $X \sim P_\theta \iff X = \theta + \xi, \xi \sim P$, w.l.o.g. $\mathbb{E}\xi = 0$
- Let $\Sigma_\xi := \mathbb{E}(\xi \otimes \xi)$. Assume that $\|\Sigma_\xi\| \lesssim 1$
- $X_1, \dots, X_n$ i.i.d. copies of X
- Regularity Conditions:
  1. V is strictly convex and twice continuously differentiable
  2. for some constants $M, L > 0$, $\|V''\|_{L_\infty} \le M$, $\|V''\|_{\mathrm{Lip}} \le L$
  3. for some constant $m > 0$, $\mathcal{I} \succeq mI_d$, where

$$\mathcal{I} := \mathbb{E}V'(\xi) \otimes V'(\xi) = \mathbb{E}V''(\xi)$$

is the Fisher information.

Functional estimation in log-concave location families (K& Wahl 21)

- $P(dx) = e^{-V(x)}dx,\ V : \mathbb{R}^d \mapsto \mathbb{R}$ convex
- $P_\theta(dx) := p_\theta(x)dx,\ p_\theta(x) := e^{-V(x-\theta)}, x \in \mathbb{R}^d, \theta \in \mathbb{R}^d$
- $X \sim P_\theta \iff X = \theta + \xi, \xi \sim P$, w.l.o.g. $\mathbb{E}\xi = 0$
- Let $\Sigma_\xi := \mathbb{E}(\xi \otimes \xi)$. Assume that $\|\Sigma_\xi\| \lesssim 1$
- $X_1, \dots, X_n$ i.i.d. copies of X
- Regularity Conditions:
  1. V is strictly convex and twice continuously differentiable
  2. for some constants $M, L > 0$, $\|V''\|_{L_\infty} \le M$, $\|V''\|_{\mathrm{Lip}} \le L$
  3. for some constant $m > 0, \mathcal{I} \succeq mI_d$, where

$$\mathcal{I} := \mathbb{E}V'(\xi) \otimes V'(\xi) = \mathbb{E}V''(\xi)$$

  is the Fisher information.

Minimax Risk Bound

Theorem

Suppose Regularity Conditions 1-3 hold. Then, for all $s > 0$,

$$\sup_{\|f\|_{C^s}\leq 1} \inf_{\hat{T}_n} \sup_{\theta\in\mathbb{R}^d} \|\hat{T}_n - f(\theta)\|_{L_2(\mathbb{P}_\theta)} \asymp \Big(\frac{1}{\sqrt{n}} \vee \Big(\sqrt{\frac{d}{n}}\Big)^s\Big) \wedge 1.$$

Phase transition in convergence rate

- If $d \lesssim n^{\alpha}$ for some $\alpha \in (0,1)$ and $s \geq \frac{1}{1-\alpha}$, then

$$\sup_{\|f\|_{C^s}\leq 1} \inf_{\hat{T}_n} \sup_{\theta \in \mathbb{R}^d} \|\hat{T}_n - f(\theta)\|_{L_2(\mathbb{P}_\theta)} \asymp n^{-1/2}.$$

- If $d \asymp n^{\alpha}$ for some $\alpha \in (0,1)$ and $s < \frac{1}{1-\alpha}$, then

$$\sup_{\|f\|_{C^s}\leq 1} \inf_{\hat{T}_n} \sup_{\theta \in \mathbb{R}^d} \|\hat{T}_n - f(\theta)\|_{L_2(\mathbb{P}_\theta)} \asymp n^{-(1-\alpha)s/2} \gg n^{-1/2}.$$

Construction of functional estimator for a log-concave location family

- Let $\hat{\theta}_n$ be the maximum likelihood estimator:

$$\hat{\theta}_n := \mathrm{argmax}_{\theta\in\mathbb{R}^d} \prod_{j=1}^{n} p_\theta(X_j) = \mathrm{argmin}_{\theta\in\mathbb{R}^d} \frac{1}{n}\sum_{j=1}^{n} V(X_j-\theta).$$

- $\hat{\theta}_n$ is used as a base estimator to define functionals $f_k(\theta)$
- $f_k(\hat{\theta}_n)$ is used as an estimator of $f(\theta)$.

The main bound

Theorem

Suppose Regularity Conditions 1-3 hold and $d \le \gamma n$, where $\gamma := c\Big(\frac{m}{M} \wedge \frac{m^2}{L\sqrt{M}}\Big)^2$ with a small enough $c > 0$. Let $f \in C^s$ for some $s = k + 1 + \rho, k \ge 0, \rho \in (0, 1]$. Then

$$\sup_{\theta\in\mathbb{R}^d} \Big\| f_k(\hat{\theta}_n) - f(\theta) - n^{-1}\sum_{j=1}^n \langle V'(\xi_j), \mathcal{I}^{-1} f'(\theta)\rangle \Big\|_{L_{\psi_{2/3}}(\mathbb{P}_\theta)}$$

$$\lesssim_{L,M,m,s} \|f\|_{C^s}\Big(\sqrt{\frac{c(V)}{n}}\Big(\frac{d}{n}\Big)^{\rho/2} + \Big(\sqrt{\frac{d}{n}}\Big)^s\Big),$$

$c(V)$ is the Poincaré constant of $dP = e^{-V}dx$.

Poincaré constant and KLS-conjecture

- For a probability measure P in $\mathbb{R}^d$, the Poincaré constant $c(P)$ is defined as

$$c(P) := \inf\Big\{C > 0 : \mathrm{Var}_P(g(\xi)) \leq C\mathbb{E}_P\|\nabla g(\xi)\|^2 : g \text{ locally Lipschitz}\Big\}$$

- KLS-conjecture: for a log-concave distribution P,

$$c(P) \lesssim \|\Sigma\|$$

- It is known (Yuansi Chen, 2020) that, for a log-concave P and for all $\epsilon > 0$,

$$c(P) \lesssim_\epsilon d^\epsilon \|\Sigma\|$$

Poincaré constant and KLS-conjecture

- For a probability measure P in $\mathbb{R}^d$, the Poincaré constant $c(P)$ is defined as

$$c(P) := \inf\Big\{C > 0 : \mathrm{Var}_P(g(\xi)) \leq C\mathbb{E}_P\|\nabla g(\xi)\|^2 : g \text{ locally Lipschitz}\Big\}$$

- KLS-conjecture: for a log-concave distribution P,

$$c(P) \lesssim \|\Sigma\|$$

- It is known (Yuansi Chen, 2020) that, for a log-concave P and for all $\epsilon > 0$,

$$c(P) \lesssim_\epsilon d^\epsilon \|\Sigma\|$$

Poincaré constant and KLS-conjecture

- For a probability measure P in $\mathbb{R}^d$, the Poincaré constant $c(P)$ is defined as

$$c(P) := \inf\Big\{C > 0 : \mathrm{Var}_P(g(\xi)) \leq C\mathbb{E}_P\|\nabla g(\xi)\|^2 : g \text{ locally Lipschitz}\Big\}$$

- KLS-conjecture: for a log-concave distribution P,

$$c(P) \lesssim \|\Sigma\|$$

- It is known (Yuansi Chen, 2020) that, for a log-concave P and for all $\epsilon > 0$,

$$c(P) \lesssim_\epsilon d^\epsilon \|\Sigma\|$$

Efficiency

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

$$\sup_{\theta\in\mathbb{R}^d}\left|\|f_k(\hat\theta) - f(\theta)\|_{L_2(\mathbb{P}_\theta)} - \frac{\sigma_f(\theta)}{\sqrt{n}}\right|$$
$$\lesssim \|f\|_{C^s}\left(\sqrt{\frac{c(V)}{n}}\left(\frac{d}{n}\right)^{\rho/2} + \left(\sqrt{\frac{d}{n}}\right)^s\right) = o(n^{-1/2})$$

provided that $d \leq n^\alpha$ for some $\alpha \in (0,1)$ and $s > \frac{1}{1-\alpha}$. Moreover, under the same assumptions,

$$\sup_{\|f\|_{C^s}\leq 1}\ \sup_{\theta\in\mathbb{R}^d} W_{2,\mathbb{P}_\theta}\left(\sqrt{n}(f_k(\hat\theta) - f(\theta)), \sigma_f(\theta)Z\right) \to 0$$

as $n \to \infty$, where $Z \sim N(0,1)$ and $W_{2,\mathbb{P}_\theta}$ is the Wasserstein distance.

Remarks

- One can also construct estimators $f_k(\hat{\theta}_n)$ based on $\hat{\theta}_n := \bar{X}_n$ rather than the MLE. In this case, the upper bound

$$\sup_{\|f\|_{C^s}\leq 1} \inf_{\hat{T}_n} \sup_{\theta\in\mathbb{R}^d} \|f_k(\bar{X}_n) - f(\theta)\|_{L_2(\mathbb{P}_\theta)} \lesssim \left(\frac{1}{\sqrt{n}} \vee \left(\sqrt{\frac{d}{n}}\right)^s\right) \wedge 1$$

still holds even without the Regularity Conditions 1-3 (only under the assumption that $\|\Sigma_\xi\| \lesssim 1$).
- The asymptotic normality of estimator $f_k(\bar{X}_n)$ also holds, but its asymptotic variance $\langle \Sigma_\xi f'(\theta), f'(\theta)\rangle$ is sub-optimal under the regularity assumptions (so, estimator $f_k(\bar{X}_n)$ is not asymptotically efficient).

Remarks

- One can also construct estimators $f_k(\hat{\theta}_n)$ based on $\hat{\theta}_n := \bar{X}_n$ rather than the MLE. In this case, the upper bound

$$\sup_{\|f\|_{C^s}\leq 1} \inf_{\hat{T}_n} \sup_{\theta\in\mathbb{R}^d} \|f_k(\bar{X}_n) - f(\theta)\|_{L_2(\mathbb{P}_\theta)} \lesssim \left(\frac{1}{\sqrt{n}} \vee \left(\sqrt{\frac{d}{n}}\right)^s\right) \wedge 1$$

  still holds even without the Regularity Conditions 1-3 (only under the assumption that $\|\Sigma_\xi\| \lesssim 1$).
- The asymptotic normality of estimator $f_k(\bar{X}_n)$ also holds, but its asymptotic variance $\langle \Sigma_\xi f'(\theta), f'(\theta)\rangle$ is sub-optimal under the regularity assumptions (so, estimator $f_k(\bar{X}_n)$ is not asymptotically efficient).

Remarks

- For irregular log-concave location families, faster rates of estimation of θ than $\sqrt{\frac{d}{n}}$ become possible.
- An interesting example is the case of i.i.d. points $X_1, \dots, X_n$ uniformly distributed in $\theta + C$, where $C \subset \mathbb{R}^d$ is a symmetric compact convex body. In this case, MLE are the points of the set

$$\hat{\Theta} := (C + X_1) \cap \dots \cap (C + X_n).$$

In particular, it includes Pitman's estimator

$$\hat{\theta}_{Pit} := \frac{\int_{\hat{\Theta}} x dx}{\mathrm{vol}(\hat{\Theta})}.$$

Remarks

- For irregular log-concave location families, faster rates of estimation of θ than $\sqrt{\frac{d}{n}}$ become possible.
- An interesting example is the case of i.i.d. points $X_1, \dots, X_n$ uniformly distributed in $\theta + C$, where $C \subset \mathbb{R}^d$ is a symmetric compact convex body. In this case, MLE are the points of the set

$$\hat{\Theta} := (C + X_1) \cap \cdots \cap (C + X_n).$$

In particular, it includes Pitman's estimator

$$\hat{\theta}_{Pit} := \frac{\int_{\hat{\Theta}} x dx}{\mathrm{vol}(\hat{\Theta})}.$$

Main results for high-dimensional normal model (K (2017, 2018), K& Zhilova (2021))

- $X_1, \dots, X_n$ i.i.d. $N(\mu, \Sigma)$ in $\mathbb{R}^d$ (equipped with the Euclidean norm)
- $\theta := (\mu, \Sigma)$ unknown parameter
- $\Theta := \mathbb{R}^d \times \mathcal{C}_+^d$ the parameter space
- $\mathcal{S}^d$ the space of symmetric $d \times d$ matrices (equipped with the operator norm)
- $\|(w, W)\| := \|w\| + \|W\|, (w, W) \in \mathbb{R}^d \times \mathcal{S}^d$
- $\mathcal{C}_+^d \subset \mathcal{S}^d$ the cone of covariance matrices in $\mathbb{R}^d$
- $\hat{\theta} := (\hat{\mu}, \hat{\Sigma})$,

$$\hat{\mu} = \bar{X} = \frac{X_1 + \dots + X_n}{n}, \quad \hat{\Sigma} = \frac{1}{n-1} \sum_{j=1}^{n} (X_j - \bar{X}) \otimes (X_j - \bar{X})$$

-

$$\mathbb{E}\|\hat{\theta} - \theta\| \lesssim \|\Sigma\|^{1/2} \sqrt{\frac{d}{n}} + \|\Sigma\| \Big(\sqrt{\frac{d}{n}} \vee \frac{d}{n} \Big).$$

Main results for high-dimensional normal model (K (2017, 2018), K& Zhilova (2021))

- $X_1, \ldots, X_n$ i.i.d. $N(\mu, \Sigma)$ in $\mathbb{R}^d$ (equipped with the Euclidean norm)
- $\theta := (\mu, \Sigma)$ unknown parameter
- $\Theta := \mathbb{R}^d \times \mathcal{C}_+^d$ the parameter space
- $\mathcal{S}^d$ the space of symmetric $d \times d$ matrices (equipped with the operator norm)
- $\|(w, W)\| := \|w\| + \|W\|, (w, W) \in \mathbb{R}^d \times \mathcal{S}^d$
- $\mathcal{C}_+^d \subset \mathcal{S}^d$ the cone of covariance matrices in $\mathbb{R}^d$
- $\hat{\theta} := (\hat{\mu}, \hat{\Sigma})$,

$$\hat{\mu} = \bar{X} = \frac{X_1 + \cdots + X_n}{n}, \quad \hat{\Sigma} = \frac{1}{n-1} \sum_{j=1}^n (X_j - \bar{X}) \otimes (X_j - \bar{X})$$

-

$$\mathbb{E}\|\hat{\theta} - \theta\| \lesssim \|\Sigma\|^{1/2} \sqrt{\frac{d}{n}} + \|\Sigma\| \Big(\sqrt{\frac{d}{n}} \vee \frac{d}{n} \Big).$$

Main results for high-dimensional normal model (K (2017, 2018), K& Zhilova (2021))

- $X_1, \dots, X_n$ i.i.d. $N(\mu, \Sigma)$ in $\mathbb{R}^d$ (equipped with the Euclidean norm)
- $\theta := (\mu, \Sigma)$ unknown parameter
- $\Theta := \mathbb{R}^d \times \mathcal{C}_+^d$ the parameter space
- $\mathcal{S}^d$ the space of symmetric $d \times d$ matrices (equipped with the operator norm)
- $\|(w, W)\| := \|w\| + \|W\|, (w, W) \in \mathbb{R}^d \times \mathcal{S}^d$
- $\mathcal{C}_+^d \subset \mathcal{S}^d$ the cone of covariance matrices in $\mathbb{R}^d$
- $\hat{\theta} := (\hat{\mu}, \hat{\Sigma})$,

$$\hat{\mu} = \bar{X} = \frac{X_1 + \cdots + X_n}{n}, \quad \hat{\Sigma} = \frac{1}{n-1} \sum_{j=1}^{n} (X_j - \bar{X}) \otimes (X_j - \bar{X})$$

- $$\mathbb{E}\|\hat{\theta} - \theta\| \lesssim \|\Sigma\|^{1/2} \sqrt{\frac{d}{n}} + \|\Sigma\| \left(\sqrt{\frac{d}{n}} \vee \frac{d}{n} \right).$$

Risk Bounds

Let

$$\Theta(a; d) := \mathbb{R}^d \times \left\{\Sigma \in \mathcal{C}_+^d : \sigma(\Sigma) \subset [1/a, a]\right\}, a \geq 1,$$

$\sigma(\Sigma)$ being the spectrum of Σ

Theorem

Suppose that $f \in C^s(\Theta)$ for some $s = k + 1 + \rho, k \geq 0, \rho \in (0, 1]$. Then, for all $\psi \prec \psi_1$,

$$\sup_{\theta \in \Theta(a;d)} \|f_k(\hat{\theta}) - f(\theta)\|_{L_\psi(\mathbb{P}_\theta)} \lesssim_{s,\psi,a} \|f\|_{C^s} \left[\left(\frac{1}{\sqrt{n}} \bigvee \left(\sqrt{\frac{d}{n}}\right)^s\right) \wedge 1\right].$$

Random homotopy used in the proof: given i.i.d. $Z_1, \dots, Z_n \sim N(0; I_d)$,

$$H(\theta; t) := (1 - t)(\mu, \Sigma) + t(\mu + \Sigma^{1/2}\hat{\mu}_Z, \Sigma^{1/2}\hat{\Sigma}_Z\Sigma^{1/2}), \theta = (\mu, \Sigma) \in \Theta.$$

Risk Bounds

Let
$$\Theta(a; d) := \mathbb{R}^d \times \Big\{\Sigma \in \mathcal{C}_+^d : \sigma(\Sigma) \subset [1/a, a]\Big\}, a \geq 1,$$
$\sigma(\Sigma)$ being the spectrum of Σ

Theorem

Suppose that $f \in C^s(\Theta)$ for some $s = k + 1 + \rho, k \geq 0, \rho \in (0, 1]$. Then, for all $\psi \prec \psi_1$,

$$\sup_{\theta \in \Theta(a;d)} \|f_k(\hat{\theta}) - f(\theta)\|_{L_\psi(\mathbb{P}_\theta)} \lesssim_{s,\psi,a} \|f\|_{C^s} \Bigg[\Bigg(\frac{1}{\sqrt{n}} \bigvee \Bigg(\sqrt{\frac{d}{n}}\Bigg)^s\Bigg) \wedge 1\Bigg].$$

Random homotopy used in the proof: given i.i.d. $Z_1, \dots, Z_n \sim N(0; I_d)$,

$$H(\theta; t) := (1 - t)(\mu, \Sigma) + t(\mu + \Sigma^{1/2}\hat{\mu}_Z, \Sigma^{1/2}\hat{\Sigma}_Z\Sigma^{1/2}), \theta = (\mu, \Sigma) \in \Theta.$$

Risk Bounds

Proposition

The following minimax bound holds:

$$\sup_{\|f\|_{C^s}\leq 1} \inf_{T} \sup_{\theta\in\Theta(a;d)} \|T - f(\theta)\|_{L_2(\mathbb{P}_\theta)} \gtrsim_{a,s} \left(\left(\frac{1}{\sqrt{n}} \bigvee \left(\sqrt{\frac{d}{n}}\right)^s\right) \wedge 1\right),$$

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

Phase transition in convergence rates

- If $d = d_n \leq n^{\alpha}$ for some $\alpha \in (0,1)$ and $s \geq \frac{1}{1-\alpha}$, then
$$\sup_{\theta \in \Theta(a;d)} \|f_k(\hat{\theta}) - f(\theta)\|_{L_{\psi}(\mathbb{P}_{\theta})} = O(n^{-1/2}).$$
- For $d = d_n \geq n^{\alpha}, \alpha \in (0,1)$ and $s < \frac{1}{1-\alpha}$, there are functionals f with $\|f\|_{C^s} \leq 1$ that could not be estimated with a rate better than $n^{-s(1-\alpha)/2}$, which is slower than $n^{-1/2}$.

Efficiency

Let

$$\sigma_f^2(\theta) := \|\Sigma^{1/2} f'_\mu(\mu, \Sigma)\|^2 + 2\|\Sigma^{1/2} f'_\Sigma(\mu, \Sigma)\Sigma^{1/2}\|_2^2, \theta = (\mu, \Sigma) \in \Theta.$$

Theorem

Suppose $d = d_n \leq n^\alpha$ *for some* $\alpha \in (0, 1)$. *Then, for all* $s = k + 1 + \rho > \frac{1}{1-\alpha}$, $k \geq 0$, $\rho \in (0, 1]$,

$$\sup_{\|f\|_{C^s} \leq 1} \sup_{\theta \in \Theta(a; d_n)} \left| n\mathbb{E}_\theta (f_k(\hat{\theta}) - f(\theta))^2 - \sigma_f^2(\theta) \right| \to 0, \ n \to \infty$$

and, for all $\sigma_0 > 0$ *and for* $Z \sim N(0, 1)$,

$$\sup_{\|f\|_{C^s} \leq 1} \sup_{\theta \in \Theta(a; d_n), \sigma_f(\theta) \geq \sigma_0} W_{2, \mathbb{P}_\theta}\left(\frac{\sqrt{n}(f_k(\hat{\theta}) - f(\theta))}{\sigma_f(\theta)}; Z \right) \to 0, n \to \infty.$$

Functional estimation via linear aggregation of plug-in estimators

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

Given a “base estimator" $\hat{\theta}_n$ and a smooth functional f, choose sample sizes $1 \le n_1 < n_2 \cdots < n_k \le n$ and coefficients $C_1, \dots, C_k$ such that the biases of plug-in estimators $f(\hat{\theta}_{n_j}), j = 1, \dots, k$ almost cancel out and the bias of estimator

$$T_f(X_1, \dots, X_n) := \sum_{j=1}^{k} C_j f(\hat{\theta}_{n_j})$$

is small. Note that $f(\hat{\theta}_{n_j})$ are often replaced by the corresponding U-statistics yielding a “jackknife" estimator.

Bias reduction via linear aggregation of plug-in estimators

- $Y \sim P$ a r.v. in a Banach space E with unknown mean $\mathbb{E}Y$
- $f : E \mapsto \mathbb{R}$
- Goal: estimate $f(\mathbb{E}Y)$ based on i.i.d. observations $Y_1, \dots, Y_n$ of Y

An expansion of the bias of plug-in estimator

Proposition

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

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

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

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

Moreover, if f is a polynomial of degree k, then (1) holds with $R = 0$.

Bias reduction via linear aggregation of plug-in estimators

- Let $k \geq 2$ and let $n/c \leq n_1 < \cdots < n_k \leq n$ for some $c > 1$.
- Let

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

where $C_1, \ldots, C_k$ satisfy
  - $\sum_{j=1}^{k} C_j = 1$
  - $\sum_{j=1}^{k} \frac{C_j}{n_j^l} = 0, l = 1, \ldots, k-1.$

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

It is easy to check that

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

Assumption

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

Clearly, for this assumption to hold, one needs $n_{j+1} - n_j \asymp n$ (for instance, $n_j = q^{k-j} n, q \in (0, 1)$).

Bounding the bias

Proposition

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

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

Bounding the bias: proof

It follows from the expansion that

$$\mathbb{E}\hat{T}_{f,k}(Y_1,\dots,Y_n) - f(\mathbb{E}Y) = \sum_{j=1}^{k} C_j(\mathbb{E}f(\bar{Y}_{n_j}) - f(\mathbb{E}Y))$$
$$= \sum_{l=1}^{k-1} \beta_{l,k}(P) \underbrace{\sum_{j=1}^{k} \frac{C_j}{n_j^l}}_{=0} + \sum_{j=1}^{k} C_j R_{n_j} = \sum_{j=1}^{k} C_j R_{n_j}$$

and

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

Bias and concentration

- Note also that

$$\|\hat{T}_{f,k}(Y_1,\dots,Y_n) - \mathbb{E}\hat{T}_{f,k}(Y_1,\dots,Y_n)\|_{L_p}$$
$$\leq \sum_{j=1}^{k} |C_j| \|f(\bar{Y}_{n_j}) - \mathbb{E}f(\bar{Y}_{n_j})\|_{L_p}$$
$$\lesssim_k \max_{1\leq j\leq k} \|f(\bar{Y}_{n_j}) - \mathbb{E}f(\bar{Y}_{n_j})\|_{L_p}.$$

- Conclusion: To bound

$$\|\hat{T}_{f,k}(Y_1,\dots,Y_n) - f(\mathbb{E}Y)\|_{L_p},$$

it is enough to bound

$$\mathbb{E}\|\bar{Y}_n - \mathbb{E}Y\|^{k+\rho} \text{ and } \|f(\bar{Y}_n) - \mathbb{E}f(\bar{Y}_n)\|_{L_p}.$$

Bias and concentration

- Note also that

$$\|\hat{T}_{f,k}(Y_1,\dots,Y_n) - \mathbb{E}\hat{T}_{f,k}(Y_1,\dots,Y_n)\|_{L_p}$$
$$\le \sum_{j=1}^{k} |C_j| \|f(\bar{Y}_{n_j}) - \mathbb{E}f(\bar{Y}_{n_j})\|_{L_p}$$
$$\lesssim_k \max_{1\le j\le k} \|f(\bar{Y}_{n_j}) - \mathbb{E}f(\bar{Y}_{n_j})\|_{L_p}.$$

- Conclusion: To bound

$$\|\hat{T}_{f,k}(Y_1,\dots,Y_n) - f(\mathbb{E}Y)\|_{L_p},$$

it is enough to bound

$$\mathbb{E}\|\bar{Y}_n - \mathbb{E}Y\|^{k+\rho} \text{ and } \|f(\bar{Y}_n) - \mathbb{E}f(\bar{Y}_n)\|_{L_p}.$$

Symmetrized estimators (jackknife)

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

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

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

Symmetrized estimators (jackknife)

Note that

- $\mathbb{E}\hat{T}^{\mathrm{sym}}_{f,k}(Y_1,\dots,Y_n) = \mathbb{E}\hat{T}_{f,k}(Y_1,\dots,Y_n)$
- for all $p \geq 1$,

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

Symmetrized estimators (jackknife)

Moreover, let

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

be the remainder of the first order Taylor expansion.

Proposition

For all $p \geq 1$,

$$\Big\| \hat{T}^{\rm sym}_{f,k}(Y_1,\dots,Y_n) - \mathbb{E}\hat{T}^{\rm sym}_{f,k}(Y_1,\dots,Y_n) - \langle \bar{Y}_n - \mathbb{E}Y, f'(\mathbb{E}Y)\rangle \Big\|_{L_p}$$

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

Symmetrized estimators (proof)

Note that

$$\hat{T}_{f,k}(Y_1,\dots,Y_n) - \mathbb{E}\hat{T}_{f,k}(Y_1,\dots,Y_n)$$
$$= \sum_{j=1}^{k} C_j(f(\bar{Y}_{n_j}) - \mathbb{E}f(\bar{Y}_{n_j})) = \sum_{j=1}^{k} C_j\langle \bar{Y}_{n_j} - \mathbb{E}Y, f'(\mathbb{E}Y)\rangle$$
$$+ \sum_{j=1}^{k} C_j(S_f(\mathbb{E}Y; \bar{Y}_{n_j} - \mathbb{E}Y) - \mathbb{E}S_f(\mathbb{E}Y; \bar{Y}_{n_j} - \mathbb{E}Y))$$

and, since $\mathbb{E}(\bar{Y}_{n_j}|\mathcal{F}_{\rm sym}) = \bar{Y}_n$,

$$\mathbb{E}\Big(\sum_{j=1}^{k} C_j\langle \bar{Y}_{n_j} - \mathbb{E}Y, f'(\mathbb{E}Y)\rangle|\mathcal{F}_{\rm sym}\Big) = \langle \bar{Y}_n - \mathbb{E}Y, f'(\mathbb{E}Y)\rangle.$$

Symmetrized estimators (proof)

It follows that

$$\hat{T}^{\mathrm{sym}}_{f,k}(Y_1,\dots,Y_n) - \mathbb{E}\hat{T}^{\mathrm{sym}}_{f,k}(Y_1,\dots,Y_n) - \langle \bar{Y}_n - \mathbb{E}Y, f'(\mathbb{E}Y)\rangle =$$

$$= \sum_{j=1}^{k} C_j \mathbb{E}(S_f(\mathbb{E}Y; \bar{Y}_{n_j} - \mathbb{E}Y) - \mathbb{E}S_f(\mathbb{E}Y; \bar{Y}_{n_j} - \mathbb{E}Y))|\mathcal{F}_{\mathrm{sym}}),$$

implying the claim. □

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

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

Estimation of Hölder smooth functionals of covariance via linear aggregation of plug-in estimators

Covariance operators in Banach spaces

- X a r.v. in a separable Banach space E with the dual space E^*
  $\mathbb{E}\langle X, u\rangle^2 < \infty, u \in E^*$
  $\Sigma : E^* \mapsto E$ is the covariance operator of X :

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

- In other words,
  $\langle \Sigma u, v\rangle = \mathbb{E}\langle X, u\rangle\langle X, v\rangle = \mathrm{cov}(\langle X, u\rangle, \langle X, v\rangle), u, v \in E^*$
  $\Sigma = \mathbb{E}(X \otimes X)$
  Σ is a symmetric positively semidefinite operator
- X a centered Gaussian r.v. in E iff $\forall u \in E^*$ $\langle X, u\rangle$ is a normal r.v.
  Distribution of X is characterized by its covariance operator Σ :
  $X \sim N(0, \Sigma)$

Covariance operators in Banach spaces

- X a r.v. in a separable Banach space E with the dual space E^*
  $\mathbb{E}\langle X, u\rangle^2 < \infty, u \in E^*$
  $\Sigma : E^* \mapsto E$ is the covariance operator of X :

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

- In other words,
  $\langle \Sigma u, v\rangle = \mathbb{E}\langle X, u\rangle\langle X, v\rangle = \mathrm{cov}(\langle X, u\rangle, \langle X, v\rangle), u, v \in E^*$
  $\Sigma = \mathbb{E}(X \otimes X)$
  Σ is a symmetric positively semidefinite operator
- X a centered Gaussian r.v. in E iff $\forall u \in E^*$ $\langle X, u\rangle$ is a normal r.v.
  Distribution of X is characterized by its covariance operator Σ :
  $X \sim N(0, \Sigma)$

Covariance operators in Banach spaces

- X a r.v. in a separable Banach space E with the dual space E^*
  $\mathbb{E}\langle X, u\rangle^2 < \infty, u \in E^*$
  $\Sigma : E^* \mapsto E$ is the covariance operator of X :

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

- In other words,
  $\langle \Sigma u, v\rangle = \mathbb{E}\langle X, u\rangle\langle X, v\rangle = \mathrm{cov}(\langle X, u\rangle, \langle X, v\rangle), u, v \in E^*$
  $\Sigma = \mathbb{E}(X \otimes X)$
  Σ is a symmetric positively semidefinite operator
- X a centered Gaussian r.v. in E iff $\forall u \in E^*$ $\langle X, u\rangle$ is a normal r.v.
  Distribution of X is characterized by its covariance operator Σ :
  $X \sim N(0, \Sigma)$

Hölder smooth functionals of covariance

- Let $L(E^*, E)$ be the space of symmetric bounded linear operators $A : E^* \mapsto E$ equipped with the operator norm:

$$\|A\| := \sup_{\|u\|\leq 1} \|Au\| = \sup_{\|u\|,\|v\|\leq 1} |\langle Au, v\rangle|.$$

- Goal: estimate $f(\Sigma)$ for a functional $f : L(E^*, E) \mapsto \mathbb{R}$ of Hölder smoothness $s > 0$ based on i.i.d. observations $X_1, \dots, X_n \sim N(0, \Sigma)$.

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

- $E = E^* = \mathbb{H}$, $\mathbb{H}$ 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: spectral projections

- $\Sigma_0 : \mathbb{H} \mapsto \mathbb{H}$ a covariance operator with eigenvalues $\|\Sigma_0\| = \lambda_1 = \cdots = \lambda_l > \lambda_{l+1} \geq \ldots$.
  $g_l := \lambda_l - \lambda_{l+1}$ 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 $f(A) := \langle P(A), B \rangle$,

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

Example of Hölder smoothness: spectral projections

- $\Sigma_0 : \mathbb{H} \mapsto \mathbb{H}$ a covariance operator with eigenvalues $\|\Sigma_0\| = \lambda_1 = \cdots = \lambda_l > \lambda_{l+1} \geq \ldots$.
  $g_l := \lambda_l - \lambda_{l+1}$ 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 $f(A) := \langle P(A), B \rangle$,

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

Sample covariance

- $X_1, \dots, X_n$ i.i.d. copies of $X \sim N(0, \Sigma)$
- $\hat{\Sigma}_n : E^* \mapsto E$ is the sample covariance operator based on $X_1, \dots, X_n$

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

- In other words,

$$\hat{\Sigma}_n = n^{-1} \sum_{j=1}^{n} X_j \otimes X_j$$

Dimension dependent bounds on sample covariance

- Let $E := \mathbb{R}^d$. Then

$$\mathbb{E}\|\hat{\Sigma}_n - \Sigma\| \lesssim \|\Sigma\|\Big(\sqrt{\frac{d}{n}} \vee \frac{d}{n}\Big).$$

- Moreover, for all $t > 0$ with probability at least $1 - e^{-t}$,

$$\|\hat{\Sigma}_n - \Sigma\| \lesssim \|\Sigma\|\Big(\sqrt{\frac{d}{n}} \vee \frac{d}{n} \vee \sqrt{\frac{t}{n}} \vee \frac{t}{n}\Big),$$

which implies that for all $p \geq 1$

$$\mathbb{E}^{1/p}\|\hat{\Sigma}_n - \Sigma\|^p \lesssim \|\Sigma\|\Big(\sqrt{\frac{d}{n}} \vee \frac{d}{n} \vee \sqrt{\frac{p}{n}} \vee \frac{p}{n}\Big).$$

Effective rank of covariance operator

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

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

- $\mathbf{r}(\lambda\Sigma) = \mathbf{r}(\Sigma), \lambda > 0$
- $\mathbf{r}(\Sigma) \leq \mathrm{rank}(\Sigma) \leq \dim(E)$
- 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 $\mathbf{r}(\Sigma) \asymp d$.

Effective rank of covariance operator

- 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 $\mathbf{r}(\Sigma) \asymp d$.

Effective rank of covariance operator

- 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 $m \sim \eta(\varepsilon) r$ which is $(1 \pm \varepsilon)$-isomorphic to ℓ_2^m.

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

Bounds on 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

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

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

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

- 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).

Bounds on 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

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

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

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

- 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).

Bounds on sample covariance: Schatten p-norms

- $E = \mathbb{H}$ a Hilbert space, $A : \mathbb{H} \mapsto \mathbb{H}$ is a compact self-adjoint operator
- $|A| := \sqrt{A^2}$
- for $p \geq 1$, $\|A\|_p := (\mathrm{tr}(|A|^p))^{1/p}$ and $\mathcal{S}_p := \{A : \|A\|_p < \infty\}$
- if $\dim(\mathbb{H}) = d < \infty$, then $\|A\|_p \leq d^{1/p}\|A\|$
- Therefore,

$$\mathbb{E}^{1/p}\|\hat{\Sigma}_n - \Sigma\|_p^p \lesssim \|\Sigma\| d^{1/p}\Big(\sqrt{\frac{d}{n}} \vee \frac{d}{n} \vee \sqrt{\frac{p}{n}} \vee \frac{p}{n}\Big).$$

Bounds on sample covariance: Schatten p-norms

Proposition

Suppose that $\mathbf{r}(\Sigma) \lesssim n$. Then, for all $p \geq 2$,

$$\mathbb{E}^{1/p}\|\hat{\Sigma}_n - \Sigma\|_p^p$$

$$\lesssim \|\Sigma\|\mathbf{r}(\Sigma)^{1/p}\Big(\sqrt{\frac{\mathbf{r}(\Sigma)p}{n}} \vee p\sqrt{\frac{\log n}{n}}\Big)\Big(1 \vee \Big(\frac{p}{n}\Big)^{1/4-1/2p} \vee \Big(\frac{p}{n}\Big)^{1/2-1/p}\Big)$$

$$\lesssim_p \|\Sigma\|\mathbf{r}(\Sigma)^{1/p}\Big(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \vee \sqrt{\frac{\log n}{n}}\Big).$$

Functionals of covariance, 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

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

- $f(\Sigma), \Sigma \in U$ 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

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

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

Estimators

- Let $k \geq 2$ and let $n/c \leq n_1 < \cdots < n_k \leq n$ for some $c > 1$ and $n_{j+1} - n_j \asymp_k n, j = 1, \dots, k$. Denote

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

- Linear aggregation

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

- Symmetrized (jackknife) estimator

$$T^{\mathrm{sym}}_{f,k}(X_1, \dots, X_n) = \hat{T}^{\mathrm{sym}}_{f,k}(X_1 \otimes X_1, \dots, X_n \otimes X_n)$$

$$:= \mathbb{E}(\hat{T}_{f,k}(X_1 \otimes X_1, \dots, X_n \otimes X_n) | \mathcal{F}_{\mathrm{sym}}) = \sum_{j=1}^{k} C_j U_n f(\hat{\Sigma}_{n_j}).$$

Estimators

- Let $k \geq 2$ and let $n/c \leq n_1 < \cdots < n_k \leq n$ for some $c > 1$ and $n_{j+1} - n_j \asymp_k n, j = 1, \dots, k$. Denote

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

- Linear aggregation

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

- Symmetrized (jackknife) estimator

$$T_{f,k}^{\mathrm{sym}}(X_1, \dots, X_n) = \hat{T}_{f,k}^{\mathrm{sym}}(X_1 \otimes X_1, \dots, X_n \otimes X_n)$$

$$:= \mathbb{E}(\hat{T}_{f,k}(X_1 \otimes X_1, \dots, X_n \otimes X_n) | \mathcal{F}_{\mathrm{sym}}) = \sum_{j=1}^{k} C_j U_n f(\hat{\Sigma}_{n_j}).$$

Estimators

- Let $k \geq 2$ and let $n/c \leq n_1 < \cdots < n_k \leq n$ for some $c > 1$ and $n_{j+1} - n_j \asymp_k n, j = 1, \dots, k$. Denote
$$C_j := \prod_{i \neq j} \frac{n_j}{n_j - n_i}, j = 1, \dots, k.$$
- Linear aggregation
$$T_{f,k}(X_1, \dots, X_n) := \hat{T}_{f,k}(X_1 \otimes X_1, \dots, X_n \otimes X_n) = \sum_{j=1}^{k} C_j f(\hat{\Sigma}_{n_j})$$
- Symmetrized (jackknife) estimator
$$T^{\mathrm{sym}}_{f,k}(X_1, \dots, X_n) = \hat{T}^{\mathrm{sym}}_{f,k}(X_1 \otimes X_1, \dots, X_n \otimes X_n)$$
$$:= \mathbb{E}(\hat{T}_{f,k}(X_1 \otimes X_1, \dots, X_n \otimes X_n) | \mathcal{F}_{\mathrm{sym}}) = \sum_{j=1}^{k} C_j U_n f(\hat{\Sigma}_{n_j}).$$

A global risk bound

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 $\rho \in (0, 1]$. Let $s := k + \rho$.

Theorem

For all $p \geq 1$,

$$\Big\| T_{f,k}(X_1, \dots, X_n) - f(\Sigma) \Big\|_{L_p}$$

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

The same bound holds for estimator $T^{\mathrm{sym}}_{f,k}(X_1, \dots, X_n)$.

Phase transition in the error rates

Suppose that $\|f\|_{\rm Lip} \lesssim 1$, $\|f^{(k)}\|_{{\rm Lip}_\rho} \lesssim 1$, $\|\Sigma\| \lesssim 1$ and $\mathbf{r}(\Sigma) \lesssim n$. Then

$$\|T_{f,k}(X_1,\dots,X_n) - \tau_f(\Sigma)\|_{L_p} \lesssim_{s,p} \frac{1}{\sqrt{n}} + \biggl(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}}\biggr)^s$$

$$\asymp \begin{cases} \frac{1}{\sqrt{n}} & \text{if } \mathbf{r}(\Sigma) \lesssim n^\alpha, \alpha \in (0,1) \text{ and } s \geq \frac{1}{1-\alpha} \\ n^{-s(1-\alpha)/2} \gg \frac{1}{\sqrt{n}} & \text{if } \mathbf{r}(\Sigma) \asymp n^\alpha, \alpha \in (0,1) \text{ and } s < \frac{1}{1-\alpha}. \end{cases}$$

A local risk bound

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 $\rho \in (0, 1]$. Let $s := k + \rho$.

Theorem

Suppose that $\mathbf{r}(\Sigma) \lesssim n$ *and, for a sufficiently large* $C > 0$, $C\|\Sigma\|\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} < \delta$. *Then, for all* $p \geq 1$,

$$\|T_{f,k}(X_1, \dots, X_n) - f(\Sigma)\|_{L_p}$$
$$\lesssim_{s,p} \|f\|_{\mathrm{Lip}} \frac{\|\Sigma\|}{\sqrt{n}} + \|f^{(k)}\|_{\mathrm{Lip}_\rho(U)} \|\Sigma\|^s \left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}}\right)^s$$
$$+ \max_{2 \leq j \leq k} \|f^{(j)}(\Sigma)\| \left(\frac{\|\Sigma\|}{\sqrt{n}}\right)^j \exp\left\{-cn\left(\frac{\delta^2}{\|\Sigma\|^2} \wedge \frac{\delta}{\|\Sigma\|}\right)\right\}$$

with some constant $c > 0$.

Normal approximation and efficiency

Theorem

Suppose that $\mathbf{r}(\Sigma) \lesssim n$. *Let* $f : L(E^*, E) \mapsto \mathbb{R}$ *be* k *times Fréchet differentiable for some* $k \geq 2$ *with* $\|f'\|_{C^1} < \infty$ *and with* $\|f^{(k)}\|_{\mathrm{Lip}_\rho} < \infty$ *for some* $\rho \in (0, 1]$. *Let* $s := k + \rho$. *Then*

$$\left\| T^{\mathrm{sym}}_{f,k}(X_1, \dots, X_n) - f(\Sigma) - \langle \hat{\Sigma}_n - \Sigma, f'(\Sigma) \rangle \right\|_{L_2}$$
$$\lesssim_s \|f'\|_{C^1} \frac{\|\Sigma\|^2}{\sqrt{n}} \sqrt{\frac{\mathbf{r}(\Sigma)}{n}} + \|f^{(k)}\|_{\mathrm{Lip}_\rho} \|\Sigma\|^s \left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \right)^s.$$

Normal approximation and efficiency

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

- $$\left|\sqrt{n}\left\| T^{\mathrm{sym}}_{f,k}(X_1,\dots,X_n) - f(\Sigma)\right\|_{L_2} - \sigma_f(\Sigma)\right| \lesssim_s \|f'\|_{C^1}\|\Sigma\|^2\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} + \|f^{(k)}\|_{\mathrm{Lip}_\rho}\|\Sigma\|^s\sqrt{n}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}}\right)^s.$$

- $$W_2\left(\frac{\sqrt{n}(T^{\mathrm{sym}}_{f,k}(X_1,\dots,X_n) - f(\Sigma))}{\sigma_f(\Sigma)}, N(0,1)\right) \lesssim_s \frac{\|\Sigma\|^2\|f'(\Sigma)\|^2}{\sigma_f^2(\Sigma)}\frac{1}{\sqrt{n}} + \frac{\|f'\|_{C^1}\|\Sigma\|^2}{\sigma_f(\Sigma)}\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} + \frac{\|f^{(k)}\|_{\mathrm{Lip}_\rho}\|\Sigma\|^s}{\sigma_f(\Sigma)}\sqrt{n}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}}\right)^s.$$

Normal approximation and efficiency

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

- $$\left|\sqrt{n}\left\| T_{f,k}^{\rm sym}(X_1,\dots,X_n) - f(\Sigma)\right\|_{L_2} - \sigma_f(\Sigma)\right| \lesssim_s \|f'\|_{C^1}\|\Sigma\|^2\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} + \|f^{(k)}\|_{\mathrm{Lip}_\rho}\|\Sigma\|^s\sqrt{n}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}}\right)^s.$$

- $$W_2\left(\frac{\sqrt{n}(T_{f,k}^{\rm sym}(X_1,\dots,X_n) - f(\Sigma))}{\sigma_f(\Sigma)}, N(0,1)\right) \lesssim_s \frac{\|\Sigma\|^2\|f'(\Sigma)\|^2}{\sigma_f^2(\Sigma)}\frac{1}{\sqrt{n}} + \frac{\|f'\|_{C^1}\|\Sigma\|^2}{\sigma_f(\Sigma)}\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} + \frac{\|f^{(k)}\|_{\mathrm{Lip}_\rho}\|\Sigma\|^s}{\sigma_f(\Sigma)}\sqrt{n}\left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}}\right)^s.$$

Normal approximation and efficiency

In particular, if $\mathbf{r}(\Sigma) \lesssim n^{\alpha}$ for some $\alpha \in (0,1)$ and $s = k + \rho > \frac{1}{1-\alpha}$, then

$$\frac{\sqrt{n}\left\| T_{f,k}^{\mathrm{sym}}(X_1, \dots, X_n) - f(\Sigma) \right\|_{L_2}}{\sigma_f(\Sigma)} \to 1$$

and

$$\frac{\sqrt{n}(T_{f,k}^{\mathrm{sym}}(X_1, \dots, X_n) - f(\Sigma))}{\sigma_f(\Sigma)} \xrightarrow{d} N(0,1).$$

Example: an application to spectral projections

- $\Sigma_0 : \mathbb{H} \mapsto \mathbb{H}$ 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
- $f(A) := \langle P(A), B\rangle, A \in U, \|B\|_1 < \infty$
  $f \in C^\infty(U)$

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
- $f(A) := \langle P(A), B\rangle, A \in U, \|B\|_1 < \infty$
  $f \in C^\infty(U)$

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 all* $k \geq 1$ *and for all* $p \geq 1$,

$$\sup_{\|\Sigma-\Sigma_0\|<\delta,\ \mathbf{r}(\Sigma)\leq r} \Big\| T_{f,k}(X_1,\dots,X_n) - f(\Sigma) \Big\|_{L_p}$$

$$\lesssim_{k,p} \|B\|_1 \biggl(\frac{\gamma}{\sqrt{n}} + \Bigl(\gamma\sqrt{\frac{r}{n}}\Bigr)^{k+1} \biggr).$$

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).

Proof ingredients: bounds on the bias

- For $s = k + \rho, k \geq 2, \rho \in (0, 1]$,

$$|\mathbb{E}_\Sigma T_{f,k}(X_1, \dots, X_n) - f(\Sigma)| \lesssim_k \|f^{(k)}\|_{\mathrm{Lip}_\rho} \mathbb{E}\|\hat{\Sigma}_n - \Sigma\|^s.$$

- Since

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

it follows that

$$|\mathbb{E}_\Sigma T_{f,k}(X_1, \dots, X_n) - f(\Sigma)| \lesssim_s \|f^{(k)}\|_{\mathrm{Lip}_\rho} \|\Sigma\|^s \Big(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \vee \frac{\mathbf{r}(\Sigma)}{n}\Big)^s.$$

Proof ingredients: bounds on the bias

- For $s = k + \rho$, $k \geq 2$, $\rho \in (0, 1]$,

$$|\mathbb{E}_\Sigma T_{f,k}(X_1, \dots, X_n) - f(\Sigma)| \lesssim_k \|f^{(k)}\|_{\mathrm{Lip}_\rho} \mathbb{E}\|\hat{\Sigma}_n - \Sigma\|^s.$$

- Since

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

it follows that

$$|\mathbb{E}_\Sigma T_{f,k}(X_1, \dots, X_n) - f(\Sigma)| \lesssim_s \|f^{(k)}\|_{\mathrm{Lip}_\rho} \|\Sigma\|^s \Big(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \vee \frac{\mathbf{r}(\Sigma)}{n}\Big)^s.$$

Proof ingredients: concentration bound for $f(\hat{\Sigma}_n)$

Proposition

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

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

Proof of concentration bound for $f(\hat{\Sigma}_n)$

- $E^n := E \times \cdots \times E := \{x^{(n)} = (x_1, \dots, x_n) : x_1, \dots, x_n \in E\}$
- $\|x^{(n)}\| := (\|x_1\|^2 + \cdots + \|x_n\|^2)^{1/2}, x^{(n)} \in E^n$
- $(E^n)^* = (E^*)^n = E^* \times \cdots \times E^*$
- $X^{(n)} = (X_1, \dots, X_n) \sim N(0, \Sigma^{(n)})$, where

$$\langle \Sigma^{(n)} u^{(n)}, v^{(n)} \rangle = \sum_{j=1}^{n} \langle \Sigma u_j, v_j \rangle, u^{(n)}, v^{(n)} \in (E^*)^n$$

- $\|\Sigma^{(n)}\| = \|\Sigma\|$
- $\hat{\Sigma}_n = \hat{\Sigma}_n(X^{(n)})$

Proof of concentration bound for $f(\hat{\Sigma}_n)$

- Bounds on local Lipschitz constants
  - $(L\hat{\Sigma}_n)(X^{(n)}) \leq \frac{2\|\hat{\Sigma}_n\|^{1/2}}{\sqrt{n}}$
  - $(L\|\hat{\Sigma}_n\|^{1/2})(X^{(n)}) \leq \frac{1}{\sqrt{n}}$
  - $(Lf(\hat{\Sigma}_n))(X^{(n)}) \leq \frac{2\|f\|_{\rm Lip}\|\hat{\Sigma}_n\|^{1/2}}{\sqrt{n}}$
- Gaussian concentration

$$\|f(\hat{\Sigma}_n) - \mathbb{E}f(\hat{\Sigma}_n)\|_{L_p} \lesssim \|\Sigma\|^{1/2}\sqrt{p}\|(Lf(\hat{\Sigma}_n))(X^{(n)})\|_{L_p}$$

$$\lesssim \|f\|_{\rm Lip}\|\Sigma\|^{1/2}\sqrt{\frac{p}{n}}\|\|\hat{\Sigma}_n\|^{1/2}\|_{L_p} \lesssim \|f\|_{\rm Lip}\|\Sigma\|^{1/2}\sqrt{\frac{p}{n}}\mathbb{E}\|\hat{\Sigma}_n\|^{1/2}$$

$$+ \|f\|_{\rm Lip}\|\Sigma\|^{1/2}\sqrt{\frac{p}{n}}\|\|\hat{\Sigma}_n\|^{1/2} - \mathbb{E}\|\hat{\Sigma}_n\|^{1/2}\|_{L_p}$$

Proof of concentration bound for $f(\hat{\Sigma}_n)$

- Bounds on local Lipschitz constants
  - $(L\hat{\Sigma}_n)(X^{(n)}) \leq \frac{2\|\hat{\Sigma}_n\|^{1/2}}{\sqrt{n}}$
  - $(L\|\hat{\Sigma}_n\|^{1/2})(X^{(n)}) \leq \frac{1}{\sqrt{n}}$
  - $(Lf(\hat{\Sigma}_n))(X^{(n)}) \leq \frac{2\|f\|_{\mathrm{Lip}}\|\hat{\Sigma}_n\|^{1/2}}{\sqrt{n}}$
- Gaussian concentration

$$\|f(\hat{\Sigma}_n) - \mathbb{E}f(\hat{\Sigma}_n)\|_{L_p} \lesssim \|\Sigma\|^{1/2}\sqrt{p}\|(Lf(\hat{\Sigma}_n))(X^{(n)})\|_{L_p}$$
$$\lesssim \|f\|_{\mathrm{Lip}}\|\Sigma\|^{1/2}\sqrt{\frac{p}{n}}\|\|\hat{\Sigma}_n\|^{1/2}\|_{L_p} \lesssim \|f\|_{\mathrm{Lip}}\|\Sigma\|^{1/2}\sqrt{\frac{p}{n}}\mathbb{E}\|\hat{\Sigma}_n\|^{1/2}$$
$$+ \|f\|_{\mathrm{Lip}}\|\Sigma\|^{1/2}\sqrt{\frac{p}{n}}\|\|\hat{\Sigma}_n\|^{1/2} - \mathbb{E}\|\hat{\Sigma}_n\|^{1/2}\|_{L_p}$$

Proof of concentration bound for $f(\hat{\Sigma}_n)$

It remains to use the bounds

$$\mathbb{E}\|\hat{\Sigma}_n\|^{1/2} \leq \mathbb{E}^{1/2}\|\hat{\Sigma}_n\| \leq \|\Sigma\|^{1/2} + \mathbb{E}^{1/2}\|\hat{\Sigma}_n - \Sigma\|,$$

$$\mathbb{E}\|\hat{\Sigma}_n - \Sigma\| \lesssim \|\Sigma\| \left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \vee \frac{\mathbf{r}(\Sigma)}{n} \right)$$

and

$$\|\|\hat{\Sigma}_n\|^{1/2} - \mathbb{E}\|\hat{\Sigma}_n\|^{1/2}\|_{L_p} \lesssim \|\Sigma\|^{1/2} \sqrt{\frac{p}{n}}$$

to complete the proof. □

Proof ingredients: concentration bound for the remainder of Taylor expansion

Recall that

$$S_f(\Sigma, \hat{\Sigma}_n - \Sigma) := f(\hat{\Sigma}_n) - f(\Sigma) - \langle \hat{\Sigma}_n - \Sigma, f'(\Sigma)\rangle.$$

Proposition

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

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

Upper bound on the L_p-risk in Hólder classes of functionals

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

Theorem

Let $a > 0, r \ge 1$ *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* $s := k + \rho$, $k \ge 0, \rho \in (0,1]$. *Then, for all* $p \ge 1$,

$$\sup_{\|f\|_{C^{s,a}(U)}\le 1} \sup_{\Sigma\in S(a,r), \|\Sigma-\Sigma_0\|<\delta} \Big\| T_{f,k}(X_1,\dots,X_n) - f(\Sigma)\Big\|_{L_p(\mathbb{P}_\Sigma)}$$

$$\lesssim_{s,p} \frac{1}{\sqrt{n}} \vee \Big(\sqrt{\frac{r}{n}}\Big)^s.$$

The same bound also holds for estimator $T^{\mathrm{sym}}_{f,k}(X_1,\dots,X_n)$.

Lower bounds

A local minimax lower bound

- $E = \mathbb{H}$ 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

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

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

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

$$\max_{1 \leq k \leq d} \inf_{T_k} \sup_{\Sigma \in S(a,r), \|\Sigma - \Sigma_0\| < \delta} \left\| T_k(X_1, \dots, X_n) - f_k(\Sigma) \right\|_{L_2(\mathbb{P}_\Sigma)}$$
$$\gtrsim_{s,\gamma_1,\gamma_2} \left(\sqrt{\frac{r}{n}} \right)^s.$$

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

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

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

$$\max_{1 \le k \le d} \max_{\omega \in B} \left\| \hat{T}_k(X_1, \dots, X_n) - f_k(\Sigma_{\theta_\omega}) \right\|_{L_2(\mathbb{P}_{\Sigma_{\theta_\omega}})} < \delta,$$

  then

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

- Based on $\hat{T}_k$, it is not hard to construct estimator $\hat{\theta} := \theta_{\hat{\omega}}, \hat{\omega} \in B$ such that

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

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

Another local minimax lower bound

Let $E = \mathbb{H}$ be a Hilbert space. Let $a > 1, r \geq 1$ and let Σ_0 be a spiked covariance operator of rank $[r]$ with non-zero eigenvalues in $[a^{-1} + \bar{\delta}, a - \bar{\delta}]$ for some $\bar{\delta} > 0$. Let $U := B(\Sigma_0, 2\bar{\delta})$ and suppose $f \in C^1(U)$. Let

$$\omega_{f'}(\Sigma_0, \delta) := \sup_{\|\Sigma - \Sigma_0\| < \delta} \|f'(\Sigma) - f'(\Sigma_0)\|, \delta \leq \bar{\delta}.$$

Theorem

For all $\beta > 2$, there exists a constant $D_\beta > 0$ such that, for all $\delta < \bar{\delta}$,

$$\inf_{T_n} \sup_{\Sigma \in \mathcal{S}(a,r), \|\Sigma - \Sigma_0\| < \delta} \frac{\sqrt{n}\|T_n(X_1, \dots, X_n) - f(\Sigma)\|_{L_2(\mathbb{P}_\Sigma)}}{\sigma_f(\Sigma)}$$

$$\geq 1 - D_\beta \left[\frac{a\omega_{f'}(\Sigma_0, \delta)}{\sigma_f(\Sigma_0)} + a^\beta \delta + \frac{a^2}{\delta^2 n}\right].$$

Hájek-LeCam local asymptotic minimaxity

If $\sigma_f(\Sigma_0)$ is bounded away from zero and $\omega_{f'}(\Sigma_0, \delta) \to 0$ as $\delta \to 0$, then

$$\lim_{c\to\infty} \liminf_{n\to\infty} \inf_{T_n} \sup_{\Sigma\in\mathcal{S}(a,r),\|\Sigma-\Sigma_0\|<\frac{c}{\sqrt{n}}} \frac{\sqrt{n}\|T_n(X_1,\dots,X_n) - f(\Sigma)\|_{L_2(\mathbb{P}_\Sigma)}}{\sigma_f(\Sigma)} \geq 1.$$

Estimation of trace functionals of covariance via linear aggregation of plug-in estimators

Gaussian model in a Hilbert space

- $\mathbb{H}$ is a separable Hilbert space
- X is a centered Gaussian r.v. in $\mathbb{H}$ iff $\forall u \in \mathbb{H}$ $\langle X, u\rangle$ is a mean zero normal r.v.
- $\Sigma = \mathbb{E}(X \otimes X) : \mathbb{H} \mapsto \mathbb{H}$ is the covariance operator of X :

$$\langle \Sigma u, v\rangle = \mathbb{E}\langle X, u\rangle\langle X, v\rangle, u, v \in \mathbb{H}$$

- Σ is a self-adjoint, positively semi-definite and nuclear operator
- $\lambda_1(\Sigma) \geq \lambda_2(\Sigma) \geq \cdots \geq 0$ are the eigenvalues of Σ (repeated with their multiplicities)

Trace functionals

- $f : \mathbb{R}_+ \mapsto \mathbb{R}$ a "test function"
- a trace functional:

$$\tau_f(\Sigma) := \mathrm{tr}(f(\Sigma)) = \sum_{j\geq 1} f(\lambda_j(\Sigma)) = \int_{\mathbb{R}_+} f d\mu_\Sigma$$

where μ_Σ is the spectral measure of Σ :

$$\mu_\Sigma(B) := \sum_{j\geq 1} I_B(\lambda_j(\Sigma)), B \subset \mathbb{R}_+$$

- Goal: given f, estimate $\tau_f(\Sigma)$ based on i.i.d. observations $X_1, \dots, X_n \sim N(0, \Sigma)$
- Warning: functional $\tau_f(\Sigma)$ is not Hölder smooth!

Sample covariance operator and linear spectral statistics

- sample covariance operator: $\hat{\Sigma}_n : \mathbb{H} \mapsto \mathbb{H}$,

$$\hat{\Sigma}_n := n^{-1} \sum_{j=1}^{n} X_j \otimes X_j$$

- linear spectral statistic (plug-in estimator)

$$\tau_f(\hat{\Sigma}_n) = \mathrm{tr}(f(\hat{\Sigma}_n)) = \sum_{j \geq 1} f(\lambda_j(\hat{\Sigma}_n))$$

Asymptotic results in random matrix theory

- $\mathbb{H} = \mathbb{R}^d$, $d = d_n \to \infty$, $\gamma_n := \frac{d_n}{n} \to \gamma \in (0, +\infty)$
- $\hat{\mu}_n := d_n^{-1} \mu_{\hat{\Sigma}_n}$ normalized spectral measure of $\hat{\Sigma}_n$
- Marchenko and Pastur, 60s: for $\Sigma = I_d$,

$$d_n^{-1} \tau_f(\hat{\Sigma}_n) = \int_{\mathbb{R}_+} f d\hat{\mu}_n \to \int_{\mathbb{R}_+} f d\nu_\gamma \text{ as } n \to \infty,$$

ν_γ is Marchenko-Pastur Law:

$$\nu_\gamma(A) := \begin{cases} \tilde{\nu}_\gamma(A) & \text{for } \gamma \in (0, 1] \\ (1 - \gamma^{-1}) I_A(0) + \tilde{\nu}_\gamma(A) & \text{for } \gamma > 1, \end{cases}$$

$$\tilde{\nu}_\gamma(dx) := \frac{1}{2\pi\gamma} \frac{\sqrt{(x-a)(b-x)}}{x} I_{[a,b]}(x) dx,$$

$a := (1 - \sqrt{\gamma})^2, b := (1 + \sqrt{\gamma})^2.$

- general case: $\Sigma = \Sigma^{(n)}$,

$$\mu_n := d_n^{-1}\mu_{\Sigma^{(n)}} \xrightarrow{w} \mu \text{ as } n \to \infty.$$

Then

$$d_n^{-1}\tau_f(\hat{\Sigma}_n) = \int_{\mathbb{R}_+} f d\hat{\mu}_n \to \int_{\mathbb{R}_+} f\, d(\mu \boxtimes \nu_\gamma)$$

- Gaussian fluctuations:

$$\tau_f(\hat{\Sigma}_n) - d_n \int_{\mathbb{R}_+} f\, d(\mu_n \boxtimes \nu_{\gamma_n}) =: \xi_n(f) \xrightarrow{f.d.d.} \xi(f) \text{ as } n \to \infty,$$

$\xi(f)$ being a Gaussian process

Asymptotic results in random matrix theory

- In other words,

$$\tau_f(\hat{\Sigma}_n) \overset{d}{\approx} d_n \int_{\mathbb{R}_+} f \, d(\mu_n \boxtimes \nu_{\gamma_n}) + \xi(f),$$

  whereas we want to estimate

$$\tau_f(\Sigma^{(n)}) = d_n \int_{\mathbb{R}_+} f \, d\mu_n.$$

- Free multiplicative deconvolution: N. El Karoui (2008), F. Benaych-Georges and M. Debbah (2010)

Example: estimation of log-determinant

- $\mathbb{H} = \mathbb{R}^d$, $\log\det(\Sigma) = \mathrm{tr}(\log \Sigma)$, Σ nonsingular
- $\hat{T}_n := \log\det(\hat{\Sigma}_n) - \underbrace{\sum_{k=1}^{d} \log\Big(1 - \frac{k}{n}\Big)}_{\text{bias correction}}$
- T. Cai, T. Liang and H. Zhou, 2013: if $\frac{d}{n} \to \gamma \in [0,1)$, then

$$\frac{\hat{T}_n - \log\det(\Sigma)}{\sqrt{-2\log(1-\frac{d}{n})}} \overset{d}{\to} Z \sim N(0,1) \text{ as } n \to \infty$$

- Note that, in this case,

$$\log\det(\hat{\Sigma}_n) - \log\det(\Sigma) = \log\det(\hat{\Sigma}_n^Z) \overset{d}{=} \sum_{k=1}^{d} \underbrace{\log(\chi^2_{n-k+1})}_{\text{independent}} - d\log n$$

Bounds for estimators of trace functionals (K (2024))

Theorem

Let $f \in C^s(\mathbb{R}_+)$ *for some* $s = m + \rho$, $m \geq 2$, $\rho \in (0, 1]$ *and let* $f(0) = 0$. *Suppose also that* $\|\Sigma\| \lesssim 1$ *and* $\mathbf{r}(\Sigma) \lesssim n$. *Then, for all* $p \geq 1$,

$$\begin{aligned}
&\|T_{\tau_f,m}(X_1, \dots, X_n) - \tau_f(\Sigma)\|_{L_p} \\
&\lesssim_{m,p} \frac{\|\Sigma f'(\Sigma)\|_2}{\sqrt{n}} + \frac{\mathbf{r}(\Sigma)}{n} + \mathbf{r}(\Sigma)\Bigl(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}}\Bigr)^s \\
&\lesssim_{m,p} \sqrt{\frac{\mathbf{r}(\Sigma)}{n}} + \mathbf{r}(\Sigma)\Bigl(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}}\Bigr)^s.
\end{aligned}$$

The same bound holds for estimator $T^{\rm sym}_{\tau_f,m}(X_1, \dots, X_n)$.

Phase transition in the error rates

$$\|T_{\tau_f,m}(X_1,\dots,X_n)-\tau_f(\Sigma)\|_{L_2}\lesssim_s \sqrt{\frac{\mathbf{r}(\Sigma)}{n}}+\mathbf{r}(\Sigma)\Big(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}}\Big)^s$$

$$\asymp \begin{cases}\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} & \text{if } \mathbf{r}(\Sigma)\lesssim n^{\alpha}, \alpha\in(0,1) \text{ and } s\geq \frac{1+\alpha}{1-\alpha}\\ n^{\alpha-s(1-\alpha)/2}\gg\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} & \text{if } \mathbf{r}(\Sigma)\asymp n^{\alpha}, \alpha\in(0,1) \text{ and } s< \frac{1+\alpha}{1-\alpha}.\end{cases}$$

Bounds for estimators of trace functionals

Theorem

Let $f \in C^s(\mathbb{R}_+)$ for some $s = m + \rho$, $m \geq 2$, $\rho \in (0, 1]$ and let $f(0) = 0$. Suppose also that $\|\Sigma\| \lesssim 1$ and $\mathbf{r}(\Sigma) \lesssim n$. Then, for all $p \geq 1$,

$$\| T^{\mathrm{sym}}_{\tau_f, m}(X_1, \dots, X_n) - \tau_f(\Sigma) - \langle f'(\Sigma), \hat{\Sigma}_n - \Sigma \rangle \|_{L_p} \lesssim_{m,p} \frac{\mathbf{r}(\Sigma)}{n} + \mathbf{r}(\Sigma) \Big(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \Big)^s.$$

Normal approximation and efficiency

- Under the assumption that $\mathbf{r}(\Sigma) \lesssim n^{\alpha}$ for some $\alpha \in (0,1)$ and $s > \frac{1+\alpha}{1-\alpha}$ (and some additional assumptions, for instance, that $\|\Sigma f'(\Sigma)\|_2^2 \asymp \mathbf{r}(\Sigma)$),

$$\frac{\sqrt{n}(T^{\rm sym}_{\tau_f,m}(X_1,\dots,X_n) - \tau_f(\Sigma))}{\sqrt{2}\|\Sigma f'(\Sigma)\|_2} \xrightarrow{d} Z \sim N(0,1)$$

and

$$\frac{n\mathbb{E}_{\Sigma}(T^{\rm sym}_{\tau_f,m}(X_1,\dots,X_n) - \tau_f(\Sigma))^2}{2\|\Sigma f'(\Sigma)\|_2^2} \to 1.$$

- Hájek-LeCam local asymptotic minimaxity:

$$\lim_{c\to\infty} \liminf_{n\to\infty} \inf_{T_n} \sup_{\|\Sigma-\Sigma_0\|_2 < \frac{c}{\sqrt{n}}} \frac{n\mathbb{E}_{\Sigma}(T_n(X_1,\dots,X_n) - \tau_f(\Sigma))^2}{2\|\Sigma f'(\Sigma)\|_2^2} \geq 1.$$

Lifshits-Krein spectral shift formula

Theorem (M. Krein (1953))

Let A, H be self-adjoint operators in $\mathbb{H}$ and $\|H\|_1 < \infty$. Then, for all continuously differentiable functions $f : \mathbb{R} \mapsto \mathbb{R}$ such that $\|\tilde{f}\|_{L_1(\mathbb{R})} < \infty$ and $\|\widetilde{f'}\|_{L_1(\mathbb{R})} < \infty$, $\|f(A+H) - f(A)\|_1 < \infty$. Moreover, there exists a unique function $\eta_{A,H} : \mathbb{R} \mapsto \mathbb{R}$ such that

$$\|\eta_{A,H}\|_{L_1(\mathbb{R})} \lesssim \|H\|_1$$

and

$$\mathrm{tr}(f(A+H) - f(A)) = \int_{\mathbb{R}} f'(\lambda)\eta_{A,H}(\lambda)d\lambda.$$

It follows that

$$|\mathrm{tr}(f(A+H) - f(A))| \lesssim \|f'\|_{L_\infty}\|H\|_1.$$

Higher order spectral shift formula

Potapov, Skripka and Sukochev (2013)

- Let A, H be self-adjoint operators in $\mathbb{H}$
- $\|H\|_{m+1} < \infty$ for some $m \geq 1$
- $\exists\ \eta_{m,A,H} : \mathbb{R} \mapsto \mathbb{R}$ such that $\|\eta_{m,A,H}\|_{L_1(\mathbb{R})} \lesssim_m \|H\|_{m+1}^{m+1}$ and, for all $m+1$ times continuously differentiable functions $f : \mathbb{R} \mapsto \mathbb{R}$ such that $\|\widetilde{f^{(j)}}\|_{L_1(\mathbb{R})} < \infty, j = 0, \ldots, m+1,$

$$\mathrm{tr}\Big(f(A+H) - f(A) - \sum_{k=1}^{m} \frac{1}{k!} \frac{d^k}{dt^k} f(A+tH)|_{t=0}\Big)$$
$$= \int_{\mathbb{R}} f^{(m+1)}(\lambda)\eta_{m,A,H}(\lambda)d\lambda.$$

It follows that

$$\Big|\mathrm{tr}\Big(f(A+H) - f(A) - \sum_{k=1}^{m} \frac{1}{k!} \frac{d^k}{dt^k} f(A+tH)|_{t=0}\Big)\Big|$$
$$\lesssim_m \|f^{(m+1)}\|_{L_\infty} \|H\|_{m+1}^{m+1}.$$

Formulas for higher order derivatives

$$\frac{d^k}{dt^k} f(A+tH)|_{t=0} = (D^k f)(A)[H,\dots,H],$$

where

$$(D^k f)(A)[H_1,\dots,H_k]$$
$$= \sum_{\pi\in S_k} \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f^{[k]}(s_1,\dots,s_{k+1}) E_A(ds_1) H_{\pi(1)} \dots H_{\pi(k)} dE_A(ds_{k+1})$$

$f^{[k]}$ is the k-th order divided difference of f : $f^{[0]} = f$ and for $k \geq 1$

$$f^{[k]}(s_1,\dots,s_{k+1}) := \begin{cases} \frac{f^{[k-1]}(s_1,\dots,s_{k-1},s_k) - f^{[k-1]}(s_1,\dots,s_{k-1},s_{k+1})}{s_k - s_{k+1}} & s_k \neq s_{k+1} \\ \frac{\partial}{\partial t} f^{[k-1]}(s_1,\dots,s_{k-1},t)|_{t=s_k} & s_k = s_{k+1} \end{cases}$$

E_A is the resolution of identity of operator A.

Bounds on higher order derivatives

It follows from PSS (2013) that

$$\left|\operatorname{tr}((D^k f)(A)[H,\dots,H])\right| \lesssim_k \|f^{(k)}\|_{L_\infty}\|H\|_k^k,$$

which implies the boundedness of multilinear forms

$$\operatorname{tr}((D^k f)(A)[H_1,\dots,H_k])$$

on the Schatten space $\mathcal{S}_k$.

Bounding the bias

Theorem

Let $f \in C^s(\mathbb{R}_+)$ for some $s = m + \rho$, $m \geq 2$, $\rho \in (0,1]$ and $f(0) = 0$. Then

$$\Big|\mathbb{E} T_{\tau_f, m}(X_1, \dots, X_n) - \tau_f(\Sigma)\Big| \lesssim_s \|f\|_{C^s} \|\Sigma\|^s \mathbf{r}(\Sigma) \biggl(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \vee \sqrt{\frac{\log n}{n}}\biggr)^s.$$

Expansion of the bias of plug-in estimator

Proposition

Let $m \geq 2$ and let $f : \mathbb{R} \mapsto \mathbb{R}$ be $m+1$ times continuously differentiable functions such that $f(0) = 0$, $\|f^{(m+1)}\|_{L_\infty} < \infty$ and $\|\widetilde{f^{(j)}}\|_{L_1(\mathbb{R})} < \infty, j = 0, \dots, m+1$. Then

$$\mathbb{E}\tau_f(\hat{\Sigma}_n) - \tau_f(\Sigma) = \sum_{l=1}^{m-1} \frac{\beta_{l,m}(\Sigma, f)}{n^l} + R,$$

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

$$|R| \lesssim \|f^{(m+1)}\|_{L_\infty} \|\Sigma\|^{m+1} \mathbf{r}(\Sigma) \Bigl(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \vee \sqrt{\frac{\log n}{n}}\Bigr)^{m+1}.$$

Expansion of the bias of plug-in estimator: comments on the proof

- $$\mathbb{E}\tau_f(\hat{\Sigma}_n) - \tau_f(\Sigma) = \sum_{k=1}^{m} \frac{\mathbb{E}\mathrm{tr}((D^k f)(\Sigma)[\hat{\Sigma}_n - \Sigma, \dots, \hat{\Sigma}_n - \Sigma])}{k!} + \mathbb{E}\rho$$

- $$\sum_{k=1}^{m} \frac{\mathbb{E}\mathrm{tr}((D^k f)(\Sigma)[\hat{\Sigma}_n - \Sigma, \dots, \hat{\Sigma}_n - \Sigma])}{k!} = \sum_{l=1}^{m-1} \frac{\beta_{l,m}(\Sigma, f)}{n^l}$$

- Using bounds on Schatten norms errors of sample covariance,

$$|\mathbb{E}\rho| \lesssim_m \|f^{(m+1)}\|_{L_\infty} \mathbb{E}\|\hat{\Sigma}_n - \Sigma\|_{m+1}^{m+1}$$
$$\lesssim_m \|f^{(m+1)}\|_{L_\infty} \|\Sigma\|^{m+1} \mathbf{r}(\Sigma) \Big(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \vee \sqrt{\frac{\log n}{n}}\Big)^{m+1}.$$

Bounding the bias (sketch of the proof)

It follows from the expansion that, for f satisfying the conditions $f(0) = 0$, $\|f^{(m+1)}\|_{L_\infty} < \infty$ and $\|\widetilde{f^{(j)}}\|_{L_1(\mathbb{R})} < \infty, j = 0, \dots, m+1$,

$$\mathbb{E}\underbrace{T_{\tau_f,m}(X_1,\dots,X_n)}_{=\sum_{j=1}^m C_j \tau_f(\hat{\Sigma}_{n_j})} - \tau_f(\Sigma) = \sum_{l=1}^{k-1} \beta_{l,m}(\Sigma, f) \underbrace{\sum_{j=1}^m \frac{C_j}{n_j^l}}_{=0} + \sum_{j=1}^m C_j R_{n_j}$$

$$= \sum_{j=1}^m C_j R_{n_j}$$

and, as a consequence,

$$\Big|\mathbb{E}T_{\tau_f,m}(X_1,\dots,X_n) - \tau_f(\Sigma)\Big| \leq \sum_{j=1}^m |C_j||R_{n_j}| \lesssim_k \max_{1\leq j\leq m} |R_{n_j}|$$

$$\lesssim_m \|f^{(m+1)}\|_{L_\infty} \|\Sigma\|^{m+1} \mathbf{r}(\Sigma) \left(\sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \vee \sqrt{\frac{\log n}{n}}\right)^{m+1}.$$

Bounding the bias (sketch of the proof)

The rest of the proof is based on extending the bound on the bias to arbitrary functions f such that $f(0) = 0, \|f\|_{C^s} < \infty$ by approximation arguments.

Concentration bounds

- for $f \in C^1$ and $p \geq 1$,

$$\|\langle f'(\Sigma), \hat{\Sigma}_n - \Sigma\rangle\|_{L_p} \lesssim \|\Sigma f'(\Sigma)\|_2 \sqrt{\frac{p}{n}} \vee \|\Sigma f'(\Sigma)\| \frac{p}{n}$$

- Remainder of the first order Taylor expansion:
  - $R_f(A, H) := \tau_f(A + H) - \tau_f(A) - \langle f'(A), H\rangle, A, H \in \mathcal{S}_1$
  - $f : \mathbb{R}_+ \mapsto \mathbb{R}$ continuously differentiable, f' Lipschitz, $f(0) = 0$, $\mathbf{r}(\Sigma) \lesssim n$.
  - For all $p \geq 1$,

$$\left\|R_f(\Sigma, \hat{\Sigma}_n - \Sigma) - \mathbb{E}R_f(\Sigma, \hat{\Sigma}_n - \Sigma)\right\|_{L_p}$$
$$\lesssim \|f'\|_{\mathrm{Lip}} \|\Sigma\|^2 \left(\frac{\mathbf{r}(\Sigma)}{\sqrt{n}} \sqrt{\frac{p}{n}} \vee \left(\frac{\mathbf{r}(\Sigma)}{\sqrt{n}} \vee 1\right) \frac{p}{n} \vee \left(\frac{p}{n}\right)^2\right).$$

References

- V. K., Asymptotically Efficient Estimation of Smooth Functionals of Covariance Operators. *J. European Mathematical Society*, 2021, 23, 3, 765–843
- V. K. & M. Zhilova, Efficient estimation of smooth functionals in Gaussian shift models. *Ann. Inst. H. Poincaré - Probab. et Statist.*, 2021, 57, 1, 351–386.
- V. K. & M. Zhilova, Estimation of Smooth Functionals in Normal Models: Bias Reduction and Asymptotic Efficiency, *Annals of Statistics*, 2021, 49, 5, 2577–2610.
- V.K. & M. Wahl, Functional estimation in log-concave location families. High Dimensional Probability IX. Progress in Probability, vol 80, pp. 393–440, Birkhäuser, 2023.
- V.K., Estimation of smooth functionals of covariance operators: jackknife bias reduction and bounds in terms of effective rank, *Ann. Inst. H. Poincaré - Probab. et Statist.*, 2025, 61, 665-712.
- V.K., Estimation of trace functionals and spectral measures of covariance operators in Gaussian models. *arXiv:2402.11321*