Efficient Estimation of Smooth Functionals of High-Dimensional Covariance

Vladimir Koltchinskii

School of Mathematics
Georgia Institute of Technology

Luminy, December 2017

Lepski and Tsybakov

Problems

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

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

Some Related Results

- Efficient estimation of smooth functionals in nonparametric models: Levit (1975, 1978), Ibragimov and Khasminskii (1981);
- In particular, in Gaussian shift model: Ibragimov, Nemirovski and Khasminskii (1987), Nemirovski (1990, 2000)
- Girko (1987–): asymptotically normal estimators of a number of special functionals (such as $\log\det(\Sigma) = \mathrm{tr}(\log\Sigma)$, Stieltjes transform of spectral function of $\Sigma : \mathrm{tr}((I + t\Sigma)^{-1}))$, ... Based on martingale CLT
- Asymptotic normality of log-determinant $\log\det(\hat{\Sigma})$ has been studied by many authors (see, e.g., Cai, Liang and Zhou (2015) for a recent result)

Some Related Results

- Asymptotic normality of $\mathrm{tr}(f(\hat{\Sigma}))$ for a smooth function $f : \mathbb{R} \mapsto \mathbb{R}$: (linear spectral statistic). Common topic in random matrix theory (both for Wigner and for Wishart matrices): Bai and Silverstein (2004), Lytova and Pastur (2009), Sosoe and Wong (2015)
- Estimation of functionals of covariance matrices under sparsity: Fan, Rigollet and Wang (2015)
- Bernstein–von Mises theorems for functionals of covariance: Gao and Zhou (2016)
- Efficient estimation of linear functionals of principal components: Koltchinskii, Löffler and Nickl (2017)

Sample Covariance Operator

- Let
$$\hat{\Sigma} := n^{-1}\sum_{j=1}^{n} X_j \otimes X_j$$
be the sample covariance based on $(X_1, \dots, X_n)$.
- Let
$$\mathcal{S}_{a,d} := \left\{\Sigma \in \mathcal{C}_+^d : a^{-1}I_d \preceq \Sigma \preceq aI_d\right\}, a > 1.$$
- If $\Sigma \in \mathcal{S}_{a,d}$, then
$$\mathbb{E}\|\hat{\Sigma} - \Sigma\| \asymp_a \|\Sigma\|\left(\sqrt{\frac{d}{n}} \bigvee \frac{d}{n}\right)$$
and, for all $t \geq 1$ with probability at least $1 - e^{-t}$,
$$\|\hat{\Sigma} - \Sigma\| \lesssim_a \|\Sigma\|\left(\sqrt{\frac{d}{n}} \bigvee \frac{d}{n} \bigvee \sqrt{\frac{t}{n}} \bigvee \frac{t}{n}\right).$$

Operator Differentiability

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

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

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

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

Bounds on the Remainder of Differentiation

Lemma

Let $S_f(A;H) = f(A+H) - f(A) - (Df)(A;H)$ *be the remainder of differentiation. If, for some* $s \in [1,2]$, $f \in B^s_{\infty,1}(\mathbb{R})$, *then the following bounds hold:*

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

and

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

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

Assumptions and Notations

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

- Loss functions. Let $\mathcal{L}$ be the class of functions $\ell : \mathbb{R} \mapsto \mathbb{R}_+$ such that
  - $\ell(0) = 0$
  - $\ell(-t) = \ell(t), t \in \mathbb{R}$
  - ℓ is convex and nondecreasing on $\mathbb{R}_+$
  - For some $c > 0$, $\ell(t) = O(e^{ct})$ as $t \to \infty$
- Suppose that
  - A.1. $d_n \geq 3 \log n$
  - A.2. for some $\alpha \in (0, 1)$, $d_n \leq n^\alpha$
  - A.3. For some $s > \frac{1}{1-\alpha}$, $f \in B^s_{\infty,1}(\mathbb{R})$.

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

Theorem

Under assumptions A.1-A.3, there exists an estimator $h(\hat{\Sigma})$ such that for all $\sigma_0 > 0$

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

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

$$\sup_{\Sigma \in \mathcal{S}_{a,d_n}, \sigma_f(\Sigma;B) \geq \sigma_0} \left| \mathbb{E}_\Sigma \ell \left(\frac{n^{1/2}\Big(\langle h(\hat{\Sigma}), B\rangle - \langle f(\Sigma), B\rangle\Big)}{\sigma_f(\Sigma, B)} \right) - \mathbb{E}\ell(Z) \right| \to 0.$$

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

Theorem

Let $f \in B^1_{\infty,1}(\mathbb{R})$. *Suppose* $d_n \geq 1, a > 1, \sigma_0 > 0$ *are such that, for some* $1 < a' < a$ *and* $\sigma_0' > \sigma_0$ *and for all large enough* n,

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

Then, the following bound holds:

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

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

Perturbation Theory: Application to Functions of Sample Covariance (The Delta Method)

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

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

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

and, similarly,

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

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

Perturbation Theory: Application to Functions of Sample Covariance (The Delta Method)

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

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

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

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

iff $\alpha < 1/2$.

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

Normal Approximation for Smooth Functions of Sample Covariance

Theorem

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

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

Then

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

Normal Approximation Bounds for Smooth Functions of Sample Covariance

- Let $\Sigma := \sum_{\lambda \in \sigma(\Sigma)} \lambda P_\lambda$ be the spectral decomposition of Σ
- $f \in B^s_{\infty,1}(\mathbb{R})$
- $\|B\|_1 < \infty$
- $\sigma_f(\Sigma; B) := \sqrt{2}\|\Sigma^{1/2} Df(\Sigma; B)\Sigma^{1/2}\|_2$
- $\mu_f(\Sigma; B) := \|\Sigma^{1/2} Df(\Sigma; B)\Sigma^{1/2}\|_3$
- $\gamma_s(f; \Sigma) := \frac{\|f\|_{B^s_{\infty,1}} \|B\|_1 \|\Sigma\|^s}{\sigma_f(\Sigma; B)}$
- $t_n(\Sigma) := t_{n,s}(f; \Sigma) := \left[-\log \gamma_s(f; \Sigma) + \frac{s-1}{2} \log\left(\frac{n}{d}\right)\right] \bigvee 1.$

Normal Approximation Bounds for Smooth Functions of Sample Covariance

Theorem

Let $f \in B^{s}_{\infty,1}(\mathbb{R})$ *for some* $s \in (1,2]$. *Then*

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left\{\frac{n^{1/2}\left\langle f(\hat{\Sigma}) - \mathbb{E}f(\hat{\Sigma}), B\right\rangle}{\sigma_f(\Sigma, B)} \le x\right\} - \Phi(x)\right| \lesssim_s \left(\frac{\mu_f(\Sigma; B)}{\sigma_f(\Sigma; B)}\right)^3 \frac{1}{\sqrt{n}}$$

$$+\gamma_s(f;\Sigma)\left(\left(\frac{d}{n}\right)^{(s-1)/2} \bigvee \left(\frac{t_n(\Sigma)}{n}\right)^{(s-1)/2} \bigvee \left(\frac{t_n(\Sigma)}{n}\right)^{s-1/2}\right)\sqrt{t_n(\Sigma)}.$$

Perturbation Theory for Functions of Sample Covariance

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

  implies that

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

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

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

Perturbation Theory for Functions of Sample Covariance

- The linear term

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

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

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

- The centered remainder

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

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

Concentration of the Remainder

Theorem

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

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

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

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

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

Wishart Operators and Bias Reduction

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

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

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

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

Wishart Operators and Bias Reduction

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

- To find an estimator of $f(\Sigma)$ with a small bias, one needs to solve (approximately) the following integral equation ("the Wishart equation")
$$\mathcal{T}g(\Sigma) = f(\Sigma), \Sigma \in \mathcal{C}_+^d.$$

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

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

Wishart Operators and Bias Reduction

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

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

- Then

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

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

Bootstrap Chain

- $$\mathcal{T}g(\Sigma) := \mathbb{E}_{\Sigma} g(\hat{\Sigma}) = \int_{\mathcal{C}_+^d} g(V) P(\Sigma; dV), \Sigma \in \mathcal{C}_+^d$$
- $$\mathcal{T}^k g(\Sigma) = \mathbb{E}_{\Sigma} g(\hat{\Sigma}^{(k)}), \Sigma \in \mathcal{C}_+^d,$$
  where
  $$\hat{\Sigma}^{(0)} = \Sigma \to \hat{\Sigma}^{(1)} = \hat{\Sigma} \to \hat{\Sigma}^{(2)} \to \dots$$
  is a Markov chain in $\mathcal{C}_+^d$ with transition probability kernel P.
- Note that $\hat{\Sigma}^{(j+1)}$ is the sample covariance based on n i.i.d. observations $\sim N(0; \hat{\Sigma}^{(j)})$ (conditionally on $\hat{\Sigma}^{(j)}$)
- Conditionally on $\hat{\Sigma}^{(j)}$, with a “high probability",
  $$\|\hat{\Sigma}^{(j+1)} - \hat{\Sigma}^{(j)}\| \lesssim \|\hat{\Sigma}^{(j)}\| \sqrt{\frac{d}{n}}$$

k-th order difference

- k-th order difference along the Markov chain:

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

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

- $\sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} f(\hat{\Sigma}^{(j)})$ is the k-th order difference of f along the trajectory of the Bootstrap Chain. For f of smoothness k, it should be of the order $\left(\sqrt{\frac{d}{n}}\right)^k$.

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

Theorem

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

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

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

Corollary

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

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

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

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

Sketch of the proof: reduction to orthogonally invariant functions

- $f(x) = x\psi'(x)$
- $g(\Sigma) := \mathrm{tr}(\psi(\Sigma))$
- g is orthogonally invariant function on $\mathcal{C}_+^d$: $g(U\Sigma U^{-1}) = g(\Sigma)$

Proposition

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

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

Proposition

$\forall k\ \mathcal{D}\mathcal{T}^k = \mathcal{T}^k\mathcal{D}$ and $\mathcal{D}\mathcal{B}^k = \mathcal{B}^k\mathcal{D}$.

- $Dg(\Sigma) = \psi'(\Sigma),\ \ \mathcal{D}g(\Sigma) = f(\Sigma)$
- $\mathcal{B}^k f(\Sigma) = \mathcal{D}\mathcal{B}^k g(\Sigma)$

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

- g orthogonally invariant, $\Sigma^{1/2}W\Sigma^{1/2} = U(W^{1/2}\Sigma W^{1/2})U^{-1}$ imply

$$\mathcal{T}g(\Sigma) = \mathbb{E}_\Sigma g(\hat{\Sigma}) = \mathbb{E}g(\Sigma^{1/2}W\Sigma^{1/2}) = \mathbb{E}g(W^{1/2}\Sigma W^{1/2}),$$

  where $W := n^{-1}\sum_{j=1}^n Z_j \otimes Z_j$, $Z_1, \dots, Z_n$ i.i.d. standard normal
- By induction,

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

  where $W_1, \dots, W_k$ i.i.d. copies of W
- $$\mathcal{B}^k g(\Sigma) = (\mathcal{T} - \mathcal{I})^k g(\Sigma) = \sum_{j=0}^k (-1)^{k-j}\binom{k}{j}\mathcal{T}^j g(\Sigma)$$

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

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

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

- Denote $V_j(t_j) := I + t_j(W_j^{1/2} - I), t_j \in [0,1]$
- $S(t_1, \dots, t_k) := V_1(t_1) \dots V_k(t_k)$
- $\phi(t_1, \dots, t_k) := g(S(t_1, \dots, t_k)^* \Sigma S(t_1, \dots, t_k)), (t_1, \dots, t_k) \in [0,1]^k$
- $\Delta_i \phi(t_1, \dots, t_k) := \phi(t_1, \dots, 1, \dots, t_k) - \phi(t_1, \dots, 0, \dots, t_k)$
- $$\mathcal{B}^k g(\Sigma) = \mathbb{E}\Delta_1 \dots \Delta_k \phi = \mathbb{E}\int_0^1 \dots \int_0^1 \frac{\partial^k \phi(t_1, \dots, t_k)}{\partial t_1 \dots \partial t_k} dt_1 \dots dt_k$$
- Integral representation:

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

Sketch of the proof: bounding partial derivatives

Lemma

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

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

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

It implies that

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

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

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