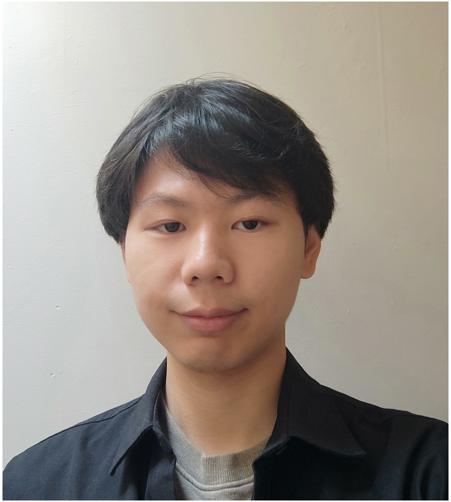
A Framework for Statistical Inference via Randomized Algorithms

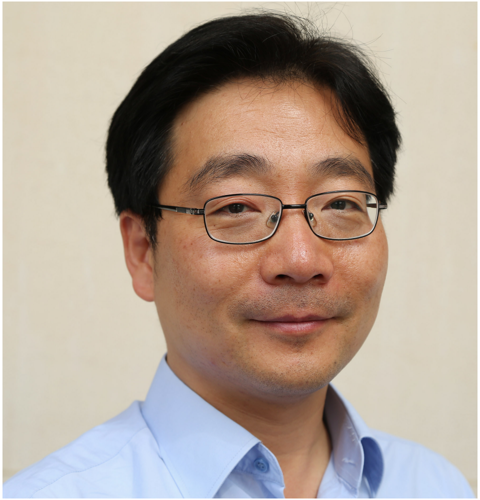
Edgar Dobriban, University of Pennsylvania
joint work with Zhixiang Zhang and Sokbae Lee

July 15, 2023

Collaborators



Zhixiang Zhang



Sokbae Lee

Overview

Overview

General Framework

Sketched Least Squares

Overview

Randomized Algorithms

- Analyzing large datasets: randomized algorithms are a key technique (with distributed, online, etc. methods)

Randomized Algorithms

- Analyzing large datasets: randomized algorithms are a key technique (with distributed, online, etc. methods)
- A randomized algorithm uses external randomness in its steps

Randomized Algorithms

- Analyzing large datasets: randomized algorithms are a key technique (with distributed, online, etc. methods)
- A randomized algorithm uses external randomness in its steps
  - Stochastic optimization
  - Monte Carlo methods (MCMC, ...)
  - Random projection/sketching methods

How Randomness Improves Algorithms

Quanta magazine

Unpredictability can help computer scientists solve otherwise intractable problems.

Random Projections in Science

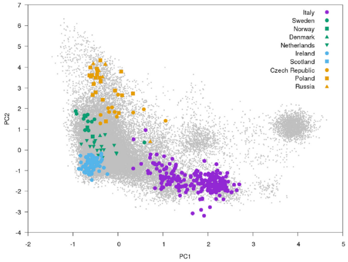
ARTICLE

Fast Principal-Component Analysis Reveals Convergent Evolution of *ADH1B* in Europe and East Asia

Kevin J. Galinsky,[1,2,*] Gaurav Bhatia,[2,3] Po-Ru Loh,[2,3] Stoyan Georgiev,[4] Sayan Mukherjee,[5] Nick J. Patterson,[2,6] and Alkes L. Price[1,2,3,6,*]

Searching for genetic variants with unusual differentiation between subpopulations is an established approach for identifying signals of natural selection. However, existing methods generally require discrete subpopulations. We introduce a method that infers selection using principal components (PCs) by identifying variants whose differentiation along top PCs is significantly greater than the null distribution of genetic drift. To enable the application of this method to large datasets, we developed the FastPCA software, which employs recent advances in random matrix theory to accurately approximate top PCs while reducing time and memory cost from quadratic to linear in the number of individuals, a computational improvement of many orders of magnitude. We apply FastPCA to a cohort of 54,734 European Americans, identifying 5 distinct subpopulations spanning the top 4 PCs.

456 The American Journal of Human Genetics *98*, 456–472, March 3, 2016



[Galinsky et al., 2016]: FastPCA, $n = 54,734$, $p = 162,335$

978-1-108-41498-2 — Advances in Economics and Econometrics

CHAPTER 1

Opportunities and Challenges: Lessons from Analyzing Terabytes of Scanner Data

Serena Ng

This paper seeks to better understand what makes big data analysis different, what we can and cannot do with existing econometric tools, and what issues need to be dealt with in order to work with the data efficiently. As a case study, I set out to extract any business cycle information that might exist in four terabytes of weekly scanner data. The main challenge is to handle the volume, variety, and characteristics of the data within the constraints of our computing environment. Scalable and efficient algorithms are available to ease the computation burden, but they often have unknown statistical properties and are not designed for the purpose of efficient estimation or optimal inference. As well, economic data have unique characteristics that generic algorithms may not accommodate. There is a need for computationally efficient econometric methods as big data is likely here to stay.

[Ng, 2017]: Nielsen scanner data: Four terabytes.

Randomized Algorithms

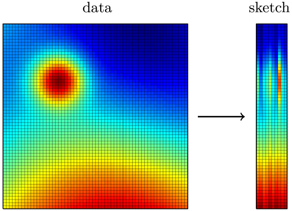
- Wish to analyze a large dataset $\mathcal{D}$

Randomized Algorithms

- Wish to analyze a large dataset $\mathcal{D}$
- Due to its size, cannot process it directly

Randomized Algorithms

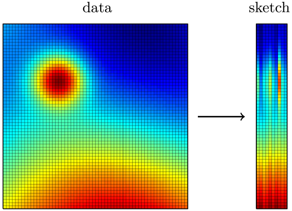
- Wish to analyze a large dataset $\mathcal{D}$
- Due to its size, cannot process it directly
- Instead, can observe the output $Z = \mathcal{A}(\mathcal{D}, S)$ of a randomized algorithm $\mathcal{A}$; where S is a source of randomness



From U.S. Department of Energy Randomized Algorithms Workshop

Randomized Algorithms

- Wish to analyze a large dataset $\mathcal{D}$
- Due to its size, cannot process it directly
- Instead, can observe the output $Z = \mathcal{A}(\mathcal{D}, S)$ of a randomized algorithm $\mathcal{A}$; where S is a source of randomness

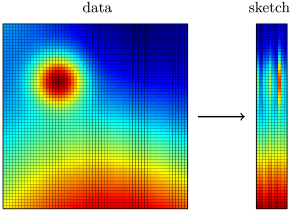


From U.S. Department of Energy Randomized Algorithms Workshop

- Advantage: can save computation time

Randomized Algorithms

- Wish to analyze a large dataset $\mathcal{D}$
- Due to its size, cannot process it directly
- Instead, can observe the output $Z = \mathcal{A}(\mathcal{D}, S)$ of a randomized algorithm $\mathcal{A}$; where S is a source of randomness

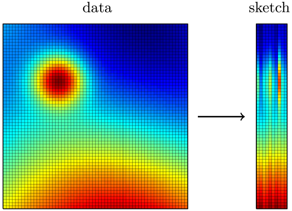


From U.S. Department of Energy Randomized Algorithms Workshop

- Advantage: can save computation time
- Drawback: result is random, unreliable?

Randomized Algorithms

- Wish to analyze a large dataset $\mathcal{D}$
- Due to its size, cannot process it directly
- Instead, can observe the output $Z = \mathcal{A}(\mathcal{D}, S)$ of a randomized algorithm $\mathcal{A}$; where S is a source of randomness



From U.S. Department of Energy Randomized Algorithms Workshop

- Advantage: can save computation time
- Drawback: result is random, unreliable?
- Challenge: How to be "responsibly reckless"—Jack Dongarra, Turing Award Lecture '21

Randomized Algorithms with Reliable Uncertainty

- An emerging idea: the randomized algorithm induces a statistical model; the unknown quantity of interest is a parameter;

Randomized Algorithms with Reliable Uncertainty

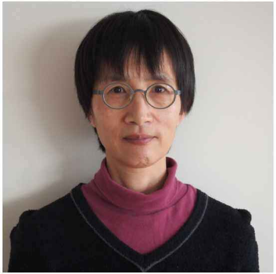
- An emerging idea: the randomized algorithm induces a statistical model; the unknown quantity of interest is a parameter;
- Aim to do statistical inference

Randomized Algorithms with Reliable Uncertainty

- An emerging idea: the randomized algorithm induces a statistical model; the unknown quantity of interest is a parameter;
- Aim to do statistical inference
- Sketched least squares in fixed dimension:
  - CLTs for some sketches [Ahfock et al., 2021, Bartan and Pilanci, 2022];
  - bootstrap for some sketches under conditions [Lopes et al., 2018];
  - heteroskedastic linear models [Lee and Ng, 2022].

Miles Lopes

Michael Mahoney



Serena Ng

Mert Pilanci

Our Work

- Aim to develop a general framework for statistical inference when using randomized methods.
  - Require minimal assumptions: general data/algorithms

Our Work

- Aim to develop a general framework for statistical inference when using randomized methods.
  - Require minimal assumptions: general data/algorithms
- Illustrate with examples (Today, least squares).

Overview

Problem

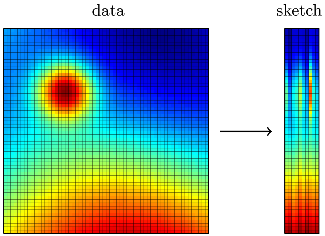
- Dataset $\mathcal{D} = \mathcal{D}_n$, not directly accessible (indexed by "size" n).

Problem

- Dataset $\mathcal{D} = \mathcal{D}_n$, not directly accessible (indexed by "size" n).
- We are interested in a parameter $\theta = \theta(\mathcal{D}) \in \mathbb{R}^p$.

Problem

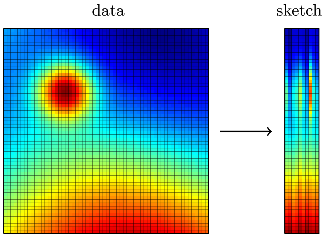
- Dataset $\mathcal{D} = \mathcal{D}_n$, not directly accessible (indexed by "size" n).
- We are interested in a parameter $\theta = \theta(\mathcal{D}) \in \mathbb{R}^p$.
- Observe $Z_m = \mathcal{A}(\mathcal{D}, S_m)$, where $\mathcal{A}$ is known; unobserved S_m has known distribution. Here m represents size of observed data.



From U.S. Department of Energy Randomized Algorithms Workshop

Problem

- Dataset $\mathcal{D} = \mathcal{D}_n$, not directly accessible (indexed by "size" n).
- We are interested in a parameter $\theta = \theta(\mathcal{D}) \in \mathbb{R}^p$.
- Observe $Z_m = \mathcal{A}(\mathcal{D}, S_m)$, where $\mathcal{A}$ is known; unobserved S_m has known distribution. Here m represents size of observed data.

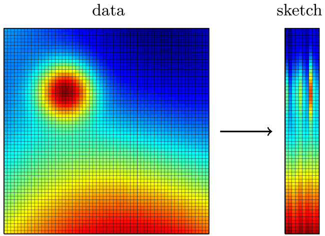


From U.S. Department of Energy Randomized Algorithms Workshop

- Example: In a least squares problem,
  - $\mathcal{D} = (X, y)$, $\theta = (X^\top X)^{-1} X^\top y$,

Problem

- Dataset $\mathcal{D} = \mathcal{D}_n$, not directly accessible (indexed by "size" n).
- We are interested in a parameter $\theta = \theta(\mathcal{D}) \in \mathbb{R}^p$.
- Observe $Z_m = \mathcal{A}(\mathcal{D}, S_m)$, where $\mathcal{A}$ is known; unobserved S_m has known distribution. Here m represents size of observed data.

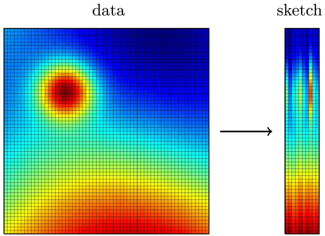


From U.S. Department of Energy Randomized Algorithms Workshop

- Example: In a least squares problem,
  - $\mathcal{D} = (X, y)$, $\theta = (X^\top X)^{-1} X^\top y$,
  - S_m: random sketching/projection matrix,

Problem

- Dataset $\mathcal{D} = \mathcal{D}_n$, not directly accessible (indexed by "size" n).
- We are interested in a parameter $\theta = \theta(\mathcal{D}) \in \mathbb{R}^p$.
- Observe $Z_m = \mathcal{A}(\mathcal{D}, S_m)$, where $\mathcal{A}$ is known; unobserved S_m has known distribution. Here m represents size of observed data.



From U.S. Department of Energy Randomized Algorithms Workshop

- Example: In a least squares problem,
  - $\mathcal{D} = (X, y)$, $\theta = (X^\top X)^{-1} X^\top y$,
  - S_m: random sketching/projection matrix,
  - Observed data $Z_m : (\tilde{X}_m, \tilde{y}_m) = (S_m X, S_m y)$,

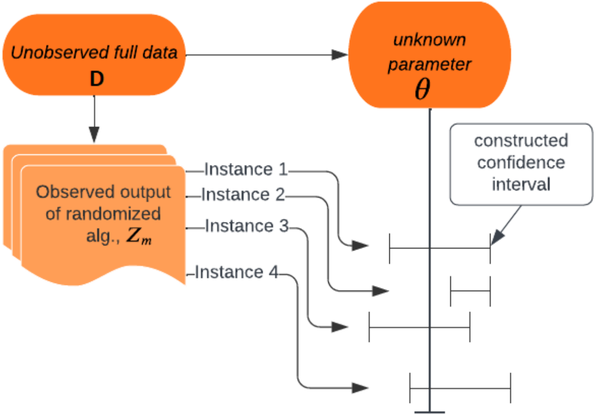
Goal: Constructing a Confidence Region

- Goal: construct confidence region $C_m = C_m(Z_m)$ for θ such that

$$P_{S_m}(\theta \in C_m) \geqslant 1 - \alpha.$$

Goal: Constructing a Confidence Region

- Goal: construct confidence region $C_m = C_m(Z_m)$ for θ such that
$$P_{S_m}(\theta \in C_m) \geqslant 1 - \alpha.$$
- The randomness is from S_m only.



Classical Pivotal Inference

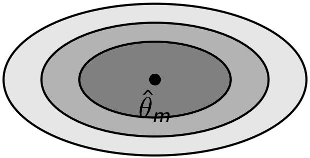
- To construct a confidence region, start from an estimator $\hat{\theta}_m = \hat{\theta}_m(Z_m)$ of θ

Classical Pivotal Inference

- To construct a confidence region, start from an estimator $\hat{\theta}_m = \hat{\theta}_m(Z_m)$ of θ
- Recall classical pivotal inference:
  - If pivot $\hat{T}_m(\hat{\theta}_m - \theta)$ has a known distribution J for a known matrix $\hat{T}_m = \hat{T}_m(Z_m)$,
  - Then for a set Γ such that $J(\Gamma) \geqslant 1 - \alpha$, can take

$$C_m = \hat{\theta}_m - \hat{T}_m^{-1}\Gamma.$$

Clearly $P_{S_m}(\theta \in C_m) \geqslant 1 - \alpha$.



C_m: Linear transform of Γ, centered at $\hat{\theta}_m$.

Classical Asymptotic Pivotal Inference

- More generally, asymptotics: Establish $\hat{T}_m(\hat{\theta}_m - \theta) \Rightarrow J$,
- Meaning: asymptotically, estimator makes errors in a predictable way

Classical Asymptotic Pivotal Inference

- More generally, asymptotics: Establish $\hat{T}_m(\hat{\theta}_m - \theta) \Rightarrow J$,
- Meaning: asymptotically, estimator makes errors in a predictable way
- If J is known, pivotal method has asymptotic coverage;

More General Methods for Inference

- What if J is not known?

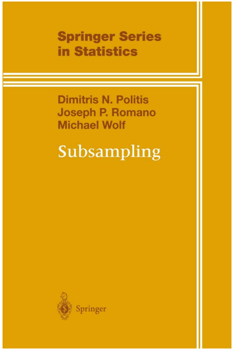
More General Methods for Inference

- What if J is not known?
- Taking inspiration from subsampling [Politis and Romano, 1994, Politis et al., 1999], propose "sub-randomization"

Dimitris Politis

Joseph Romano

Michael Wolf



Sub-randomization

- Generate $K+1$ i.i.d. smaller datasets $Z_{b,i} = \mathcal{A}(\mathcal{D}, S_{b,i})$, for $i = 0, \ldots, K$, with i.i.d. $S_{b,i}$ and $b < m$, by running the randomized algorithm.

Unobserved data $\mathcal{D}$

Observed data from randomized alg. $Z_m = \mathcal{A}(\mathcal{D}, S_m)$

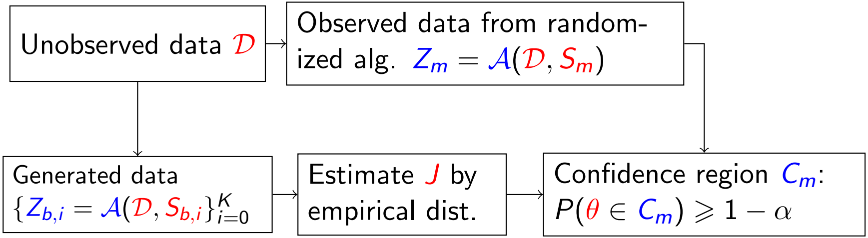
Generated data $\{Z_{b,i} = \mathcal{A}(\mathcal{D}, S_{b,i}\}_{i=0}^{K}$

Estimate J by empirical dist.

Confidence region C_m: $P(\theta \in C_m) \geqslant 1 - \alpha$

Sub-randomization

- Generate $K+1$ i.i.d. smaller datasets $Z_{b,i} = \mathcal{A}(\mathcal{D}, S_{b,i})$, for $i = 0, \ldots, K$, with i.i.d. $S_{b,i}$ and $b < m$, by running the randomized algorithm.



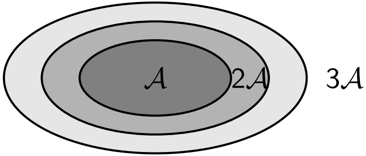
- Estimate J (where $\hat{T}_m(\hat{\theta}_m - \theta) \approx J$) by the empirical distribution of $\hat{\theta}_b(Z_{b,i})$, $i = 1, \ldots, K$, letting

$$\hat{J}(\cdot) = \frac{1}{K} \sum_{i=1}^{K} I\left(\widehat{T}_b(Z_{b,0}) \left[\hat{\theta}_b(Z_{b,i}) - \hat{\theta}_m(Z_m) \right] \in \cdot \right).$$

Scaling a set

- **Definition.** For a closed convex set $\mathcal{B}$, and probability dist. P, s.t. $P(x \cdot \mathcal{B}) \to 1$ as $x \to \infty$, define the scaling

$$\Gamma_P := \inf\{x \geqslant 0 : P(x \cdot \mathcal{B}) \geqslant 1 - \alpha\} \cdot \mathcal{B}$$



Sub-randomization: General result

- Let $\tilde{J}_m$ be the distribution of $\hat{T}_m(Z_m)(\hat{\theta}_m - \theta)$
- **Condition**: asymptotically, estimator makes errors in a predictable way (for both m and b)

$$\tilde{J}_m \Rightarrow J \text{ and } \tilde{J}_b \Rightarrow J.$$

  Do not need to know J; only that it exists.

Sub-randomization: General result

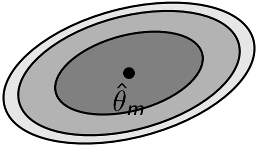
- Let $\tilde{J}_m$ be the distribution of $\hat{T}_m(Z_m)(\hat{\theta}_m - \theta)$
- **Condition**: asymptotically, estimator makes errors in a predictable way (for both m and b)

$$\tilde{J}_m \Rightarrow J \text{ and } \tilde{J}_b \Rightarrow J.$$

  Do not need to know J; only that it exists.
- Recall problem index n, large output size m, small output size b, number of runs K
- **Theorem.** When $n, m, b, K \to \infty$, if $(\widehat{T}_m(Z_m) - \widehat{T}_b(Z_b))^{-1}$ exists w.p. $\to 1$, and Γ_J is a continuity set of J,

$$C_m = \hat{\theta}_m - \left(\widehat{T}_m(Z_m) - \widehat{T}_b(Z_b)\right)^{-1} \Gamma_{\hat{J}}$$

  is an asymptotically valid confidence set: $\liminf_{n\to\infty} P(\theta \in C_m) \geqslant 1 - \alpha$.



C_m: Linear transform of $\Gamma_{\hat{J}}$, centered at $\hat{\theta}_m$.

Plug-in: simplification for a normal limit

- If for some T_m, $T_m(\hat{\theta}_m - \theta_n) \Rightarrow \mathcal{N}(0, I_p)$, we can develop a simpler *plug-in method*

Plug-in: simplification for a normal limit

- If for some T_m, $T_m(\hat{\theta}_m - \theta_n) \Rightarrow \mathcal{N}(0, I_p)$, we can develop a simpler *plug-in method*
- Draw $K > 0$ observations $Z_{m,i} = \mathcal{A}(\mathcal{D}, S_{m,i})$, where $S_{m,i}$ are i.i.d. for $i \in [K]$, from the same process as Z_m.

Plug-in: simplification for a normal limit

- If for some T_m, $T_m(\hat{\theta}_m - \theta_n) \Rightarrow \mathcal{N}(0, I_p)$, we can develop a simpler *plug-in method*
- Draw $K > 0$ observations $Z_{m,i} = \mathcal{A}(\mathcal{D}, S_{m,i})$, where $S_{m,i}$ are i.i.d. for $i \in [K]$, from the same process as Z_m.
- Compute $\hat{\theta}_{m,i} = \hat{\theta}_m(Z_{m,i})$ for $i \in [K]$ and let $\hat{\theta}^*_{K,m} = K^{-1} \sum_{i=1}^K \hat{\theta}_{m,i}$. Let

$$\widehat{\Sigma}_K = K^{-1} \sum_{i=1}^{K} (\hat{\theta}_{m,i} - \hat{\theta}^*_{K,m})(\hat{\theta}_{m,i} - \hat{\theta}^*_{K,m})^\top, \text{ and } \widehat{T}_K = \widehat{\Sigma}_K^{1/2}.$$

Plug-in: simplification for a normal limit

- If for some T_m, $T_m(\hat{\theta}_m - \theta_n) \Rightarrow \mathcal{N}(0, I_p)$, we can develop a simpler *plug-in method*
- Draw $K > 0$ observations $Z_{m,i} = \mathcal{A}(\mathcal{D}, S_{m,i})$, where $S_{m,i}$ are i.i.d. for $i \in [K]$, from the same process as Z_m.
- Compute $\hat{\theta}_{m,i} = \hat{\theta}_m(Z_{m,i})$ for $i \in [K]$ and let $\hat{\theta}^*_{K,m} = K^{-1} \sum_{i=1}^K \hat{\theta}_{m,i}$. Let

$$\widehat{\Sigma}_K = K^{-1} \sum_{i=1}^{K} (\hat{\theta}_{m,i} - \hat{\theta}^*_{K,m})(\hat{\theta}_{m,i} - \hat{\theta}^*_{K,m})^\top, \text{ and } \widehat{T}_K = \widehat{\Sigma}_K^{1/2}.$$

- **Theorem.** Let $m, n, K \to \infty$, s.t. for $A_{m,n} = T_m(\hat{\theta}_m - \theta_n)$, $\mathbb{E}A_{m,n} \to 0$, $\mathbb{E}A_{m,n}A_{m,n}^\top \to I_p$, and $\mathrm{Var}\left[(v^\top A_{m,n})^2\right]$ is bounded over m, n and $v \in \mathbb{R}^p$ with $\|v\| = 1$. Then, for an $1 - \alpha$-probability set Γ under $\mathcal{N}(0, I_p)$,

$$P_{Z_m}\left(\theta_n \in \hat{\theta}_m - \widehat{T}_K^{-1}\Gamma\right) \to_P 1 - \alpha.$$

Multi-run aggregation for small-bias estimators

- If $\hat{\theta}_m$ have small bias, we can develop a more accurate *multi-run aggregation method*

Multi-run aggregation for small-bias estimators

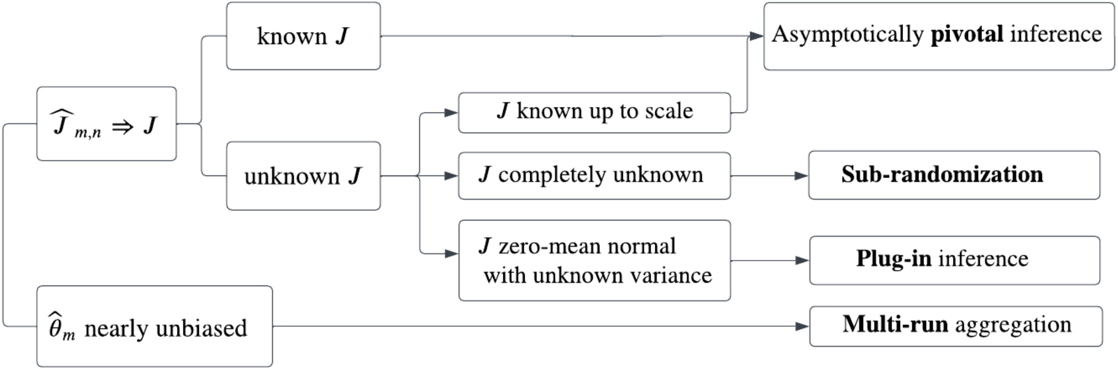
- If $\hat{\theta}_m$ have small bias, we can develop a more accurate *multi-run aggregation method*
- Construct $\hat{\theta}_{b,i}$ as before, but with b instead of m (more flexible)

Multi-run aggregation for small-bias estimators

- If $\hat{\theta}_m$ have small bias, we can develop a more accurate *multi-run aggregation method*
- Construct $\hat{\theta}_{b,i}$ as before, but with b instead of m (more flexible)
- **Theorem.** Let $b, n, K \to \infty$, and suppose there is $a > 0$ such that $\mathbb{E}|v^\top \hat{\theta}_b|^{2+a}$ is uniformly bounded over b, n and all $v \in \mathbb{R}^p$ with $\|v\| = 1$. Let $\lambda_b = \lambda_{\min}\left(\mathrm{Cov}\left[\hat{\theta}_b\right]\right)$, and suppose that $\|\mathbb{E}\hat{\theta}_b - \theta_n\| = o(K^{-1/2}{\lambda_b}^{1/2})$. Then,

$$P\left(\theta_n \in \frac{1}{K}\sum_{i=1}^{K}\hat{\theta}_{b,i} - \frac{1}{K^{1/2}}\widehat{T}_{K,b}^{-1}\Gamma\right) \to 1 - \alpha.$$

Summary of inference methods via randomized algorithms



Overview

Overview

General Framework

Sketched Least Squares

Sketched Least Squares

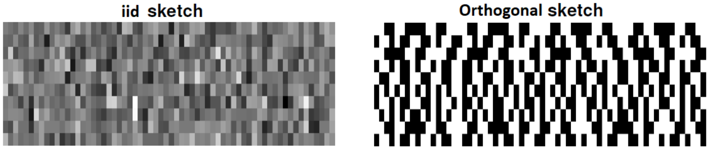
- Unobserved full data (X, y), X is $n \times p$

Sketched Least Squares

- Unobserved full data (X, y), X is $n \times p$
- Parameter of interest $\theta = \beta$ (fixed p); $\theta = c^\top \beta$ (growing p)
  - Least squares parameter $\beta = (X^\top X)^{-1} X^\top y$,

Sketched Least Squares

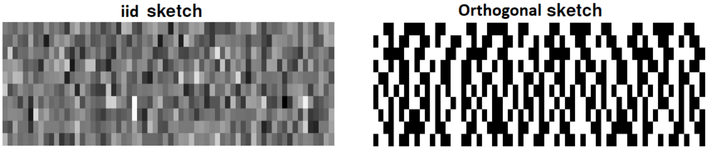
- Unobserved full data (X, y), X is $n \times p$
- Parameter of interest $\theta = \beta$ (fixed p); $\theta = c^\top \beta$ (growing p)
  - Least squares parameter $\beta = (X^\top X)^{-1} X^\top y$,
- Observed data $Z_m : (\tilde{X}_m, \tilde{y}_m) = (S_m X, S_m y)$ where S_m is random sketching/projection matrix



From Pilanci: Information-theoretic bounds on sketching

Sketched Least Squares

- Unobserved full data (X, y), X is $n \times p$
- Parameter of interest $\theta = \beta$ (fixed p); $\theta = c^\top \beta$ (growing p)
  - Least squares parameter $\beta = (X^\top X)^{-1} X^\top y$,
- Observed data $Z_m : (\tilde{X}_m, \tilde{y}_m) = (S_m X, S_m y)$ where S_m is random sketching/projection matrix



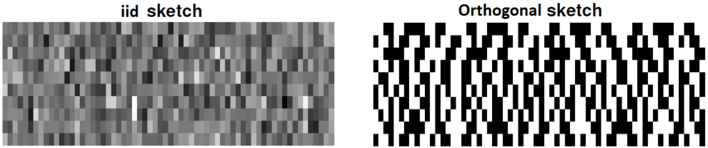
From Pilanci: Information-theoretic bounds on sketching

- Sketch-and-solve estimator:

$$\hat{\theta}_m = (\tilde{X}_m{}^\top \tilde{X}_m)^{-1} \tilde{X}_m{}^\top \tilde{y}_m.$$

Sketched Least Squares

- Unobserved full data (X, y), X is $n \times p$
- Parameter of interest $\theta = \beta$ (fixed p); $\theta = c^\top \beta$ (growing p)
  - Least squares parameter $\beta = (X^\top X)^{-1} X^\top y$,
- Observed data $Z_m : (\tilde{X}_m, \tilde{y}_m) = (S_m X, S_m y)$ where S_m is random sketching/projection matrix



From Pilanci: Information-theoretic bounds on sketching

- Sketch-and-solve estimator:

$$\hat{\theta}_m = (\tilde{X}_m{}^\top \tilde{X}_m)^{-1} \tilde{X}_m{}^\top \tilde{y}_m.$$

- Later: partial sketching, iterative sketching: sketch-and-project, ...

Growing p

Consider $c^\top \hat{\beta}_m$ for some fixed sequence of $c \in \mathbb{R}^p$. We aim to show

$$\tau_m \sigma^{-1}(c^\top \hat{\beta}_m - c^\top \beta) \Rightarrow \mathcal{N}(0, 1).$$

Growing p

Consider $c^\top \hat{\beta}_m$ for some fixed sequence of $c \in \mathbb{R}^p$. We aim to show

$$\tau_m \sigma^{-1}(c^\top \hat{\beta}_m - c^\top \beta) \Rightarrow \mathcal{N}(0, 1).$$

- As $p, m, n \to \infty$, one of the following asymptotic regimes holds:
  1. Proportional limit: $0 < \liminf p/n \leq \limsup p/n < 1$, and $0 < \liminf p/m \leq \limsup p/m < 1$.
  2. Non-proportional limit: $p/n \to 0$, and $p/m \to 0$.

Growing p

Consider $c^\top \hat{\beta}_m$ for some fixed sequence of $c \in \mathbb{R}^p$. We aim to show

$$\tau_m \sigma^{-1}(c^\top \hat{\beta}_m - c^\top \beta) \Rightarrow \mathcal{N}(0, 1).$$

- As $p, m, n \to \infty$, one of the following asymptotic regimes holds:
  1. Proportional limit: $0 < \liminf p/n \leq \limsup p/n < 1$, and $0 < \liminf p/m \leq \limsup p/m < 1$.
  2. Non-proportional limit: $p/n \to 0$, and $p/m \to 0$.
- **Theorem.** For i.i.d. sketching, when all moments of the entries of the sketching matrices are bounded, $\mathbb{E}|S_{1,1}|^l \leqslant C_l$ for any $(C_l)_{l \geqslant 1}$ with $C_l < \infty$, we have asymptotic normality with $\tau_m = (m-p)^{1/2}$, and

$$\sigma^2 = \frac{m-p}{m}(\kappa_4 - 3)\sum_{k=1}^{n}[c^\top (X^\top X)^{-1} x_k \varepsilon_k]^2 + c^\top (X^\top X)^{-1} c \|\varepsilon\|^2.$$

For Gaussian sketching, consistently estimated by $\widehat{\sigma^2} := \frac{m}{m-p} c^\top (\widetilde{X}_m^\top \widetilde{X}_m)^{-1} c \|\tilde{\varepsilon}_n\|^2$;

Growing p

Consider $c^\top \hat{\beta}_m$ for some fixed sequence of $c \in \mathbb{R}^p$. We aim to show

$$\tau_m \sigma^{-1}(c^\top \hat{\beta}_m - c^\top \beta) \Rightarrow \mathcal{N}(0,1).$$

- As $p, m, n \to \infty$, one of the following asymptotic regimes holds:
  1. Proportional limit: $0 < \liminf p/n \le \limsup p/n < 1$, and $0 < \liminf p/m \le \limsup p/m < 1$.
  2. Non-proportional limit: $p/n \to 0$, and $p/m \to 0$.
- **Theorem.** For i.i.d. sketching, when all moments of the entries of the sketching matrices are bounded, $\mathbb{E}|S_{1,1}|^l \leqslant C_l$ for any $(C_l)_{l\geqslant 1}$ with $C_l < \infty$, we have asymptotic normality with $\tau_m = (m-p)^{1/2}$, and

$$\sigma^2 = \frac{m-p}{m}(\kappa_4 - 3)\sum_{k=1}^{n}[c^\top (X^\top X)^{-1} x_k \varepsilon_k]^2 + c^\top (X^\top X)^{-1} c \|\varepsilon\|^2.$$

  For Gaussian sketching, consistently estimated by $\widehat{\sigma^2} := \frac{m}{m-p} c^\top (\widetilde{X}_m^\top \widetilde{X}_m)^{-1} c \|\tilde{\varepsilon}_n\|^2$;
- Sub-randomization, plug-in, aggregation, pivotal: ✓

Growing p: Orthogonal sketching

Consider $c^\top \hat{\beta}_m$ for some fixed sequence of $c \in \mathbb{R}^p$. We aim to show

$$\tau_m \sigma^{-1}(c^\top \hat{\beta}_m - c^\top \beta) \Rightarrow \mathcal{N}(0, 1).$$

- Hadamard sketching?

Growing p: Orthogonal sketching

Consider $c^\top \hat{\beta}_m$ for some fixed sequence of $c \in \mathbb{R}^p$. We aim to show

$$\tau_m \sigma^{-1}(c^\top \hat{\beta}_m - c^\top \beta) \Rightarrow \mathcal{N}(0,1).$$

- Hadamard sketching?
- Haar sketching: S_m uniform over $m \times n$ partial orthogonal matrices.

Growing p: Orthogonal sketching

Consider $c^\top \hat{\beta}_m$ for some fixed sequence of $c \in \mathbb{R}^p$. We aim to show

$$\tau_m \sigma^{-1}(c^\top \hat{\beta}_m - c^\top \beta) \Rightarrow \mathcal{N}(0,1).$$

- Hadamard sketching?
- Haar sketching: S_m uniform over $m \times n$ partial orthogonal matrices.
- As $p, m, n \to \infty$, one of the following holds:
  1. Proportional limit: $\limsup p/n < 1$, $\limsup p/m < 1$, and $0 < \liminf m/n \le \limsup m/n < 1$.
  2. Non-proportional limit: $\lim p/n = 0$, $\limsup p/m < 1$, and $\lim m/n = 0$.

Growing p: Orthogonal sketching

Consider $c^\top \hat{\beta}_m$ for some fixed sequence of $c \in \mathbb{R}^p$. We aim to show

$$\tau_m \sigma^{-1}(c^\top \hat{\beta}_m - c^\top \beta) \Rightarrow \mathcal{N}(0, 1).$$

- Hadamard sketching?
- Haar sketching: S_m uniform over $m \times n$ partial orthogonal matrices.
- As $p, m, n \to \infty$, one of the following holds:
  1. Proportional limit: $\limsup p/n < 1$, $\limsup p/m < 1$, and $0 < \liminf m/n \leq \limsup m/n < 1$.
  2. Non-proportional limit: $\lim p/n = 0$, $\limsup p/m < 1$, and $\lim m/n = 0$.
- **Theorem.** For Haar sketching, $\tau_m = \left(\frac{(m-p)(n-p)}{n-m}\right)^{1/2}$, and $\sigma^2 = c^\top (X^\top X)^{-1} c \|\varepsilon\|^2$.

Growing p: Orthogonal sketching

Consider $c^\top \hat{\beta}_m$ for some fixed sequence of $c \in \mathbb{R}^p$. We aim to show

$$\tau_m \sigma^{-1}(c^\top \hat{\beta}_m - c^\top \beta) \Rightarrow \mathcal{N}(0, 1).$$

- Hadamard sketching?
- Haar sketching: S_m uniform over $m \times n$ partial orthogonal matrices.
- As $p, m, n \to \infty$, one of the following holds:
  1. Proportional limit: $\limsup p/n < 1$, $\limsup p/m < 1$, and $0 < \liminf m/n \le \limsup m/n < 1$.
  2. Non-proportional limit: $\lim p/n = 0$, $\limsup p/m < 1$, and $\lim m/n = 0$.
- **Theorem.** For Haar sketching, $\tau_m = \left(\frac{(m-p)(n-p)}{n-m}\right)^{1/2}$, and $\sigma^2 = c^\top (X^\top X)^{-1} c \|\varepsilon\|^2$.
- Consistently estimated by

$$\widehat{\sigma^2} := \frac{m(n-m)}{(m-p)(n-p)} c^\top (\widetilde{X}_m^\top \widetilde{X}_m)^{-1} c \|\tilde{\varepsilon}_n\|^2;$$

Growing p: Orthogonal sketching

Consider $c^\top \hat{\beta}_m$ for some fixed sequence of $c \in \mathbb{R}^p$. We aim to show

$$\tau_m \sigma^{-1}(c^\top \hat{\beta}_m - c^\top \beta) \Rightarrow \mathcal{N}(0,1).$$

- Hadamard sketching?
- Haar sketching: S_m uniform over $m \times n$ partial orthogonal matrices.
- As $p, m, n \to \infty$, one of the following holds:
  1. Proportional limit: $\limsup p/n < 1$, $\limsup p/m < 1$, and $0 < \liminf m/n \le \limsup m/n < 1$.
  2. Non-proportional limit: $\lim p/n = 0$, $\limsup p/m < 1$, and $\lim m/n = 0$.
- **Theorem.** For Haar sketching, $\tau_m = \left(\frac{(m-p)(n-p)}{n-m}\right)^{1/2}$, and $\sigma^2 = c^\top (X^\top X)^{-1} c \|\varepsilon\|^2$.
- Consistently estimated by

$$\widehat{\sigma^2} := \frac{m(n-m)}{(m-p)(n-p)} c^\top (\widetilde{X}_m^\top \widetilde{X}_m)^{-1} c \|\tilde{\varepsilon}_n\|^2;$$

- Sub-randomization, plug-in, aggregation, pivotal: ✓

Proof idea for i.i.d. case, growing dimension

Goal: find asy distribution of $m^{1/2}(c^\top \hat{\beta}_m - c^\top \beta)$.

Proof idea for i.i.d. case, growing dimension

Goal: find asy distribution of $m^{1/2}(c^\top \hat{\beta}_m - c^\top \beta)$. We use a trigonometric interpolation strategy inspired by Götze et al. [2017], Baik et al. [2018]; dating back to Bentkus [2003]

Proof idea for i.i.d. case, growing dimension

Goal: find asy distribution of $m^{1/2}(c^\top \hat{\beta}_m - c^\top \beta)$. We use a trigonometric interpolation strategy inspired by Götze et al. [2017], Baik et al. [2018]; dating back to Bentkus [2003]

- Gaussian case: Recall that $X = U \Lambda V^\top$, and $U_\perp$ is an orthogonal complement of U; then

$$c^\top \hat{\beta}_m = c^\top \beta + c^\top V \Lambda^{-1} (U^\top S^\top S U)^{-1} (SU)^\top \underbrace{S U_\perp U_\perp^\top y}_{\text{independent of } SU}.$$

Proof idea for i.i.d. case, growing dimension

Goal: find asy distribution of $m^{1/2}(c^\top \hat{\beta}_m - c^\top \beta)$. We use a trigonometric interpolation strategy inspired by Götze et al. [2017], Baik et al. [2018]; dating back to Bentkus [2003]

- Gaussian case: Recall that $X = U\Lambda V^\top$, and $U_\perp$ is an orthogonal complement of U; then

$$c^\top \hat{\beta}_m = c^\top \beta + c^\top V \Lambda^{-1} (U^\top S^\top S U)^{-1} (SU)^\top \underbrace{S U_\perp U_\perp^\top y}_{\text{independent of } SU}.$$

- General distribution: Consider the following trigonometric interpolation matrix for $\theta \in [0, \pi/2]$:

$$Y_m(\theta) = S_m \sin\theta + W_m \cos\theta,$$

where W_m consists of Gaussian variables.

Proof idea for i.i.d. case, growing dimension

Goal: find asy distribution of $m^{1/2}(c^\top \hat{\beta}_m - c^\top \beta)$. We use a trigonometric interpolation strategy inspired by Götze et al. [2017], Baik et al. [2018]; dating back to Bentkus [2003]

- Gaussian case: Recall that $X = U\Lambda V^\top$, and $U_\perp$ is an orthogonal complement of U; then

$$c^\top \hat{\beta}_m = c^\top \beta + c^\top V \Lambda^{-1}(U^\top S^\top S U)^{-1}(SU)^\top \underbrace{S U_\perp U_\perp^\top y}_{\text{independent of } SU}.$$

- General distribution: Consider the following trigonometric interpolation matrix for $\theta \in [0, \pi/2]$:

$$Y_m(\theta) = S_m \sin\theta + W_m \cos\theta,$$

  where W_m consists of Gaussian variables.
- Let $q(\theta) = m^{1/2}(c^\top \hat{\beta}_m(\theta) - c^\top \beta)$. Find the characteristic function $\mathbb{E}e^{itq(\pi/2)}$.

Proof idea for i.i.d. case continued

- We show that for any bounded complex-valued function f with bounded derivatives up to the fifth order, we have the ODE

$$\frac{d\mathbb{E}f(q(\theta))}{d\theta} - 2(\kappa_4 - 3)\sin^3\theta\cos\theta \cdot \Psi(X, y) \cdot \mathbb{E}f^{(2)}(q(\theta)) = o(1)$$

uniformly for $\theta \in [0, \pi/2]$ (ODE for expected functions of characteristic function: Tikhomirov [1981]).

Proof idea for i.i.d. case continued

- We show that for any bounded complex-valued function f with bounded derivatives up to the fifth order, we have the ODE

$$\frac{d\mathbb{E}f(q(\theta))}{d\theta} - 2(\kappa_4 - 3)\sin^3\theta\cos\theta \cdot \Psi(X, y) \cdot \mathbb{E}f^{(2)}(q(\theta)) = o(1)$$

  uniformly for $\theta \in [0, \pi/2]$ (ODE for expected functions of characteristic function: Tikhomirov [1981]).
- Requires Taylor expansion up to fifth order, leave-one and -two-out calculations, controlling numerous sums of products, categorizing them into classes (cca. 20 pages)

Proof idea for i.i.d. case continued

- We show that for any bounded complex-valued function f with bounded derivatives up to the fifth order, we have the ODE

$$\frac{d\mathbb{E}f(q(\theta))}{d\theta} - 2(\kappa_4 - 3)\sin^3\theta\cos\theta \cdot \Psi(X, y) \cdot \mathbb{E}f^{(2)}(q(\theta)) = o(1)$$

  uniformly for $\theta \in [0, \pi/2]$ (ODE for expected functions of characteristic function: Tikhomirov [1981]).
- Requires Taylor expansion up to fifth order, leave-one and -two-out calculations, controlling numerous sums of products, categorizing them into classes (cca. 20 pages)
- Integrate the ODE: Let $\psi(\theta) = e^{itq(\theta)}$, and $g(\theta) = \psi(\theta)\exp\left\{t^2(\kappa_4 - 3)\sin^4\theta\Psi/2\right\}$

Proof idea for i.i.d. case continued

- We show that for any bounded complex-valued function f with bounded derivatives up to the fifth order, we have the ODE

$$\frac{d\mathbb{E}f(q(\theta))}{d\theta} - 2(\kappa_4 - 3)\sin^3\theta\cos\theta \cdot \Psi(X, y) \cdot \mathbb{E}f^{(2)}(q(\theta)) = o(1)$$

  uniformly for $\theta \in [0, \pi/2]$ (ODE for expected functions of characteristic function: Tikhomirov [1981]).
- Requires Taylor expansion up to fifth order, leave-one and -two-out calculations, controlling numerous sums of products, categorizing them into classes (cca. 20 pages)
- Integrate the ODE: Let $\psi(\theta) = e^{itq(\theta)}$, and $g(\theta) = \psi(\theta)\exp\left\{t^2(\kappa_4 - 3)\sin^4\theta\Psi/2\right\}$
- We have

$$\frac{d\mathbb{E}g(\theta)}{d\theta} = \left(\frac{d\mathbb{E}\psi}{d\theta} + (\mathbb{E}\psi)2t^2(\kappa_4 - 3)\sin^3\theta\cos\theta\Psi\right) \times \exp\left\{t^2(\kappa_4 - 3)\sin^4\theta\Psi/2\right\} = o(1),$$

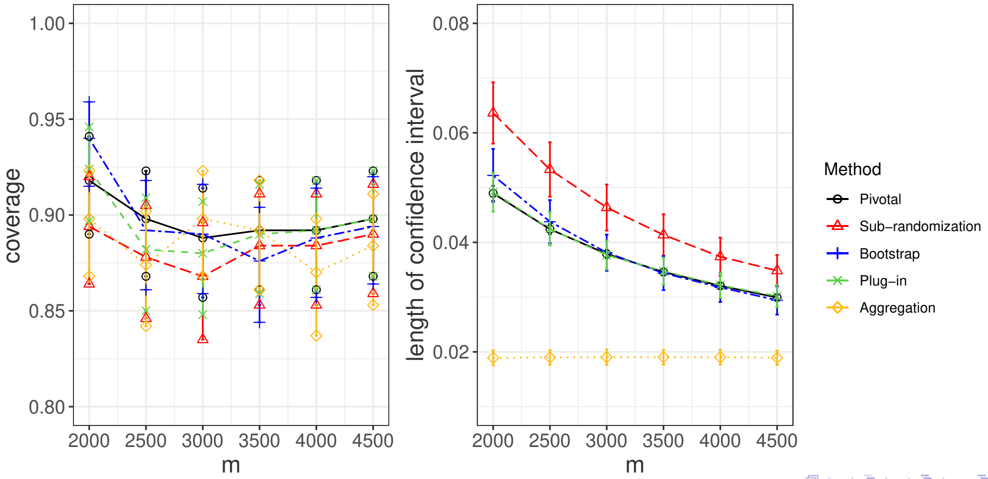
  thus $\mathbb{E}g(\pi/2) = \mathbb{E}g(0)$ (matching Gaussian case).

Numerical Experiments: Coverage

- Generate X_n with i.i.d. $\mathcal{N}(0,1)$ entries, and y_n with i.i.d. $\mathrm{Unif}(0,1)$ entries

Numerical Experiments: Coverage

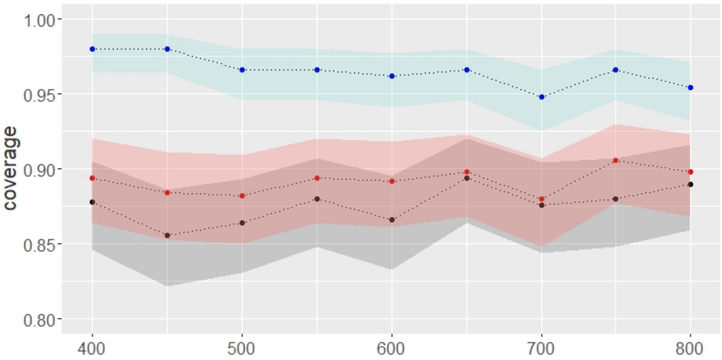
- Generate X_n with i.i.d. $\mathcal{N}(0,1)$ entries, and y_n with i.i.d. $\mathrm{Unif}(0,1)$ entries
- With $n = 8,000$, $p = 500$, $b = 600$ and $K = 100$, find coverage of Hadamard sketching 90% intervals for the first coordinate of β; and 95% Clopper-Pearson interval for coverage



Empirical Data Example

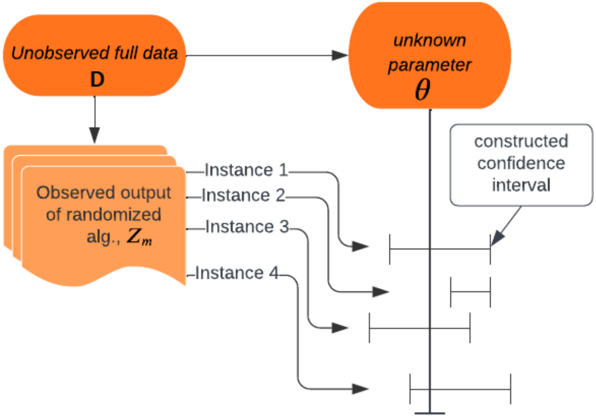
- Use subset of HGDP data, with $n = 1043$ indiv., $p = 200$ SNPs, $b = 300$, $K = 100$; Hadamard sketching.
- Show coverage of 90% intervals for the first coordinate of β; and 95% Clopper-Pearson interval for coverage

Pivotal method (black); sub-randomization (red); bootstrap (blue).



Summary

- A framework for statistical inference when using randomized algorithms
  - Requires showing $\widehat{T}_m(\hat{\theta}_m - \theta)$ has a limit distribution (sometimes more: normal limit, small bias)
- Example: Sketching in least squares. Aggregation is the "best" of the methods considered.
  - Allow broad classes of projections, algorithms
  - Handle growing data dimension



Partial Sketching

The partial sketching estimator: $\hat{\beta}_m^{(\mathrm{pa})} := (\widetilde{X}_m^\top \widetilde{X}_m)^{-1} X^\top y$. Advantageous when $\|X\beta\|/\|\varepsilon\|$ is "small".

Partial Sketching

The partial sketching estimator: $\hat{\beta}_m^{(\mathrm{pa})} := (\widetilde{X}_m^\top \widetilde{X}_m)^{-1} X^\top y$. Advantageous when $\|X\beta\|/\|\varepsilon\|$ is "small".

- Generally
$$\tau_m(\Sigma')^{-1/2}(\hat{\beta}_m^{(\mathrm{pa})} - \beta) \Rightarrow \mathcal{N}(0, I_p).$$

Partial Sketching

The partial sketching estimator: $\hat{\beta}_m^{(\mathrm{pa})} := (\widetilde{X}_m^\top \widetilde{X}_m)^{-1} X^\top y$. Advantageous when $\|X\beta\|/\|\varepsilon\|$ is "small".

- Generally
$$\tau_m (\Sigma')^{-1/2} (\hat{\beta}_m^{(\mathrm{pa})} - \beta) \Rightarrow \mathcal{N}(0, I_p).$$
- **Theorem.** For i.i.d. sketching: $\tau_m = m^{1/2}$, $\Sigma' = \beta\beta^\top + \Sigma_0'$,
$$\Sigma_0' = (X^\top X)^{-1} \left\{ \sum_{i=1}^{n} (y_i - \varepsilon_i)^2 \left[X^\top X + (\kappa_4 - 3) x_i x_i^\top \right] \right\} (X^\top X)^{-1}$$

Partial Sketching

The partial sketching estimator: $\hat{\beta}_m^{(\mathrm{pa})} := (\widetilde{X}_m^\top \widetilde{X}_m)^{-1} X^\top y$. Advantageous when $\|X\beta\|/\|\varepsilon\|$ is "small".

- Generally
$$\tau_m (\Sigma')^{-1/2} (\hat{\beta}_m^{(\mathrm{pa})} - \beta) \Rightarrow \mathcal{N}(0, I_p).$$
- **Theorem.** For i.i.d. sketching: $\tau_m = m^{1/2}$, $\Sigma' = \beta\beta^\top + \Sigma_0'$,
$$\Sigma_0' = (X^\top X)^{-1} \left\{ \sum_{i=1}^{n} (y_i - \varepsilon_i)^2 \left[X^\top X + (\kappa_4 - 3) x_i x_i^\top \right] \right\} (X^\top X)^{-1}$$
- Sub-randomization: ✓

Partial Sketching

The partial sketching estimator: $\hat{\beta}_m^{(\mathrm{pa})} := (\widetilde{X}_m^\top \widetilde{X}_m)^{-1} X^\top y$. Advantageous when $\|X\beta\|/\|\varepsilon\|$ is "small".

- Generally
$$\tau_m (\Sigma')^{-1/2} (\hat{\beta}_m^{(\mathrm{pa})} - \beta) \Rightarrow \mathcal{N}(0, I_p).$$
- **Theorem.** For i.i.d. sketching: $\tau_m = m^{1/2}$, $\Sigma' = \beta\beta^\top + \Sigma_0'$,
$$\Sigma_0' = (X^\top X)^{-1} \left\{ \sum_{i=1}^{n} (y_i - \varepsilon_i)^2 \left[X^\top X + (\kappa_4 - 3) x_i x_i^\top \right] \right\} (X^\top X)^{-1}$$
- Sub-randomization: ✓
- Pivotal: ✓

Partial Sketching

- **Theorem.** Asymptotic normality holds for Haar/Hadamard partial sketching. The plug-in estimators estimate Σ' (below) consistently.

τ_m, Σ for asymptotic distribution of $\hat{\beta}_m^{(\mathrm{pa})}$ when p is fixed

	τ_m	Σ'
i.i.d. ($\kappa_4 = 3$)	$m^{1/2}$	$(X^\top X)^{-1}\|X\beta\|^2 + \beta\beta^\top$
Haar	$\left(\frac{mn}{n-m}\right)^{1/2}$	$(X^\top X)^{-1}\|X\beta\|^2 + \beta\beta^\top$
Hadamard	$\left(\frac{mn}{n-m}\right)^{1/2}$	$(X^\top X)^{-1}\|X\beta\|^2 + 2\beta\beta^\top$

The asymptotic variances of $\hat{\beta}_m^{(\mathrm{pa})}$ differ slightly for Haar and Hadamard.

Iterative Sketching

- Iterative methods can converge to the truth even with a fixed sketch dimension.

Iterative Sketching

- Iterative methods can converge to the truth even with a fixed sketch dimension.
- Examples: Iterative Hessian sketching [Pilanci and Wainwright, 2016], Sketch-and-project: randomized Kaczmarz/Newton, etc. [Gower and Richtárik, 2015, Gower et al., 2019]

Iterative Sketching

- Iterative methods can converge to the truth even with a fixed sketch dimension.
- Examples: Iterative Hessian sketching [Pilanci and Wainwright, 2016], Sketch-and-project: randomized Kaczmarz/Newton, etc. [Gower and Richtárik, 2015, Gower et al., 2019]
- Iterative Hessian sketching: start with $\hat{\beta}_{0,m,n} = 0$; update for $t \geqslant 1$:

$$\hat{\beta}_{t,m,n} = (\widetilde{X}_{t,m}^\top \widetilde{X}_{t,m})^{-1} X^\top (y - X\hat{\beta}_{t-1,m,n}) + \hat{\beta}_{t-1,m,n}.$$

Iterative Sketching

- Iterative methods can converge to the truth even with a fixed sketch dimension.
- Examples: Iterative Hessian sketching [Pilanci and Wainwright, 2016], Sketch-and-project: randomized Kaczmarz/Newton, etc. [Gower and Richtárik, 2015, Gower et al., 2019]
- Iterative Hessian sketching: start with $\hat{\beta}_{0,m,n} = 0$; update for $t \geqslant 1$:

$$\hat{\beta}_{t,m,n} = (\widetilde{X}_{t,m}^\top \widetilde{X}_{t,m})^{-1} X^\top (y - X\hat{\beta}_{t-1,m,n}) + \hat{\beta}_{t-1,m,n}.$$

- **Theorem.** For i.i.d. sketching,

$$m^{T/2}\left(\hat{\beta}_{T,m,n} - \beta\right) \Rightarrow V\Lambda^{-1}\left(\prod_{i=T}^{1} \mathcal{G}_i\right) U^\top y,$$

  where $\mathcal{G}_i$ are i.i.d. symmetric zero-mean Gaussian, and for $j_1 \leq k_1$, $j_2 \leq k_2$

$$\mathrm{Cov}\left[(\mathcal{G}_i)_{j_1 k_1}, (\mathcal{G}_i)_{j_2 k_2}\right] = \delta_{j_1 j_2}\delta_{k_1 k_2} + \delta_{j_1 k_2}\delta_{k_1 j_2} + (\kappa_4 - 3)\lim_{n\to\infty}\sum_{\ell=1}^{n} U_{\ell,j_1} U_{\ell,k_1} U_{\ell,j_2} U_{\ell,k_2}.$$

- Sub-randomization & pivotal can be used.

Growing p, Partial Sketching

We aim to show

$$\tau_m \sigma'^{-1} \left(\alpha_m c^\top \hat{\beta}_m^{(\mathrm{pa})} - c^\top \beta \right) \Rightarrow \mathcal{N}(0, 1).$$

Growing p, Partial Sketching

We aim to show

$$\tau_m \sigma'^{-1} \left(\alpha_m c^\top \hat{\beta}_m^{(\mathrm{pa})} - c^\top \beta \right) \Rightarrow \mathcal{N}(0, 1).$$

- **Theorem.** For i.i.d. sketching, $\alpha_m = \frac{m-p}{m}$, $\tau_m = (m-p)^{1/2}$,

$$\sigma'^2 = \frac{m-p}{m}(\kappa_4 - 3) \sum_{k=1}^{n} \left[c^\top (X^\top X)^{-1} x_k (y_k - \varepsilon_k) \right]^2 + \left[\|X\beta\|^2 c^\top (X^\top X)^{-1} c + (c^\top \beta)^2 \right].$$

Growing p, Partial Sketching

We aim to show

$$\tau_m \sigma'^{-1} \left(\alpha_m c^\top \hat{\beta}_m^{(\mathrm{pa})} - c^\top \beta \right) \Rightarrow \mathcal{N}(0,1).$$

- **Theorem.** For i.i.d. sketching, $\alpha_m = \frac{m-p}{m}$, $\tau_m = (m-p)^{1/2}$,

$$\sigma'^2 = \frac{m-p}{m}(\kappa_4 - 3) \sum_{k=1}^{n} \left[c^\top (X^\top X)^{-1} x_k (y_k - \varepsilon_k) \right]^2 + \left[\|X\beta\|^2 c^\top (X^\top X)^{-1} c + (c^\top \beta)^2 \right].$$

- **Theorem.** For Haar sketching, $\alpha_m = \frac{n(m-p)}{m(n-p)}$, $\tau_m = \left(\frac{(m-p)(n-p)}{n-m} \right)^{1/2}$,

$$\sigma'^2 = \left[\|X\beta\|^2 c^\top (X^\top X)^{-1} c + (c^\top \beta)^2 \right].$$

- Sub-randomization & pivotal can be used.

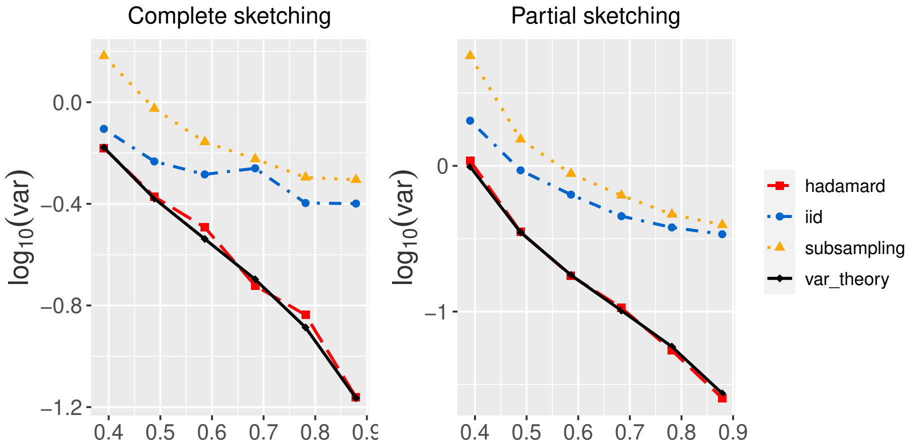
Numerical Experiments: Variance prediction

- Take $n = 2,048$, $p = 500$; m ranging from 800 to 1,800

Numerical Experiments: Variance prediction

- Take $n = 2,048$, $p = 500$; m ranging from 800 to 1,800
- Generate (then fix) the data X with i.i.d. $\mathcal{N}(0,1)$ entries; and y from Unif(0,1)

Log of variance of $\sqrt{m}c^\top\hat{\beta}_m$.



Numerical Experiments: Coverage

- Generate synthetic data, with $n = 8,000$ and $p = 500$ (as in Lopes et al. [2018]).

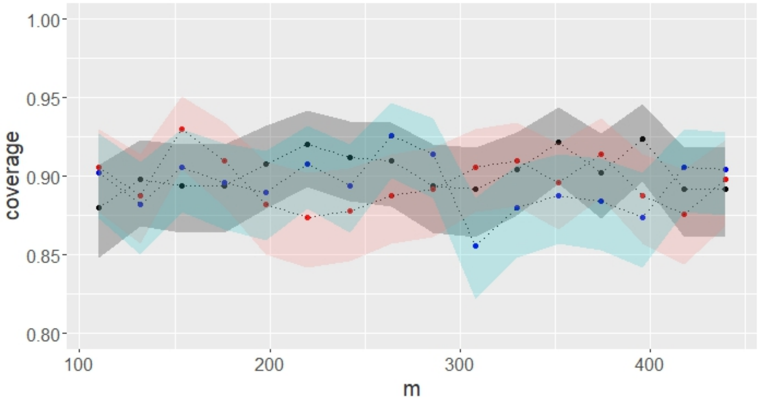
Numerical Experiments: Coverage

- Generate synthetic data, with $n = 8,000$ and $p = 500$ (as in Lopes et al. [2018]).
- $X = U\Lambda V^\top$, where
  - U consists of the left singular vectors of a matrix with i.i.d. $t_2(0, \Sigma)$, $\Sigma_{i,j} = 2 \times 0.5^{|i-j|}$ rows,
  - Λ is a diagonal matrix with entries uniformly spaced between 0.1 and 1,
  - V is the right singular matrix of a $p \times p$ standard Gaussian matrix.
- $y = Xb + \delta$, where $b = (\mathbf{1}_{0.2p}, t\mathbf{1}_{0.6p}, \mathbf{1}_{0.2p})$ with $t = 0.1$, and $\delta \sim \mathcal{N}(0, 0.01^2 I)$.

Empirical Data, fixed dimension

- Use "cpusmall" dataset, with $n = 8293$, $p = 11$, $b = 50$, $K = 100$, S_m (and $S_{b,i}$) be Hadamard sketching matrices, find coverage of 90% intervals for the first coordinate of β; and 95% Clopper-Pearson interval for coverage

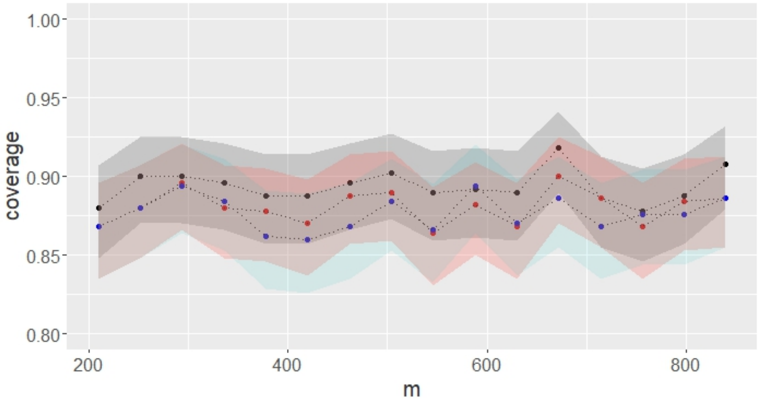
Pivotal method (black); sub-randomization (red); bootstrap (blue).



Empirical Data, fixed dimension

- Use "nycflight13" dataset, with $n = 60448, p = 21, b = 100, K = 100, S_m$ (and $S_{b,i}$) be Hadamard sketching matrices, find coverage of 90% intervals for the first coordinate of β; and 95% Clopper-Pearson interval for coverage

Pivotal method (black); sub-randomization (red); bootstrap (blue).



References I

Daniel C Ahfock, William J Astle, and Sylvia Richardson. Statistical properties of sketching algorithms. *Biometrika*, 108(2):283–297, 2021.

Jinho Baik, Ji Oon Lee, and Hao Wu. Ferromagnetic to paramagnetic transition in spherical spin glass. *Journal of Statistical Physics*, 173:1484–1522, 2018.

Burak Bartan and Mert Pilanci. Distributed sketching for randomized optimization: Exact characterization, concentration and lower bounds. *arXiv preprint arXiv:2203.09755*, 2022.

V Bentkus. A new method for approximations in probability and operator theories. *Lithuanian Mathematical Journal*, 43:367–388, 2003.

Kevin J Galinsky, Gaurav Bhatia, Po-Ru Loh, Stoyan Georgiev, Sayan Mukherjee, Nick J Patterson, and Alkes L Price. Fast principal-component analysis reveals convergent evolution of adh1b in europe and east asia. *The American Journal of Human Genetics*, 98(3):456–472, 2016.

Friedrich Götze, Alexey Naumov, and Alexander Tikhomirov. Distribution of linear statistics of singular values of the product of random matrices. *Bernoulli*, 23(4B):3067–3113, 2017.

References II

Robert Gower, Dmitry Kovalev, Felix Lieder, and Peter Richtárik. Rsn: Randomized subspace newton. *Advances in Neural Information Processing Systems*, 32, 2019.

Robert M Gower and Peter Richtárik. Randomized iterative methods for linear systems. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1660–1690, 2015.

Sokbae Lee and Serena Ng. Least squares estimation using sketched data with heteroskedastic errors. In *International Conference on Machine Learning*, pages 12498–12520. PMLR, 2022.

Miles E Lopes, Shusen Wang, and Michael W Mahoney. Error estimation for randomized least-squares algorithms via the bootstrap. *arXiv preprint arXiv:1803.08021*, 2018.

Serena Ng. Opportunities and challenges: Lessons from analyzing terabytes of scanner data. Technical report, National Bureau of Economic Research, 2017.

Mert Pilanci and Martin J Wainwright. Iterative hessian sketch: Fast and accurate solution approximation for constrained least-squares. *The Journal of Machine Learning Research*, 17(1): 1842–1879, 2016.

References III

Dimitris N Politis and Joseph P Romano. Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, pages 2031–2050, 1994.

Dimitris N Politis, Joseph P Romano, and Michael Wolf. *Subsampling*. Springer Science & Business Media, 1999.

Alexander N Tikhomirov. On the convergence rate in the central limit theorem for weakly dependent random variables. *Theory of Probability & Its Applications*, 25(4):790–809, 1981.