An Econometrics View of Algorithmic SubSampling

July 2019

Sokbae (Simon) Lee, Columbia University

Serena Ng, Columbia University and NBER

Motivation: Data $A = (y, X) \in \mathbb{R}^{n \times d}$.

- Tasks: inference, estimate moments, density, eigenvalues.
- Problem: too much data! taxing memory, storage, time.
- Want a sketch $\tilde{A} = \Pi A$ that preserves features of A
  - $\Pi \in \mathbb{R}^{m \times n}$, $m \ll n$ rows, faster debugging.
  - H_0 costly to test, try out whether it is worth testing.
- Literature
  - statistics: data squashing (likelihood based), binning
  - machine learning: pasting bites (Breiman)
  - computer science: algorithms for approximation.

This Paper

- Review principles of linear sketching: rows not columns
- Algorithmic Sampling $\neq$ bootstrap or subsampling.
- New results for prediction and inference .
  - Algorithmic goal: fewer rows, fast and efficient.
  - Statistical efficiency: more rows, more efficient.
  - Propose inference conscious guides for sketch size.

Main References

- Sarlos, T. (2006): Improved Approximation Algorithms for Large Matrices via Random Projections, Proceedings of 47th IEEE Symposium, FOCS.
- Drineas, P., et al (2011): Faster Least Squares Approximation, Numerische Mathematik.
- Mahoney, M. W. (2011): Randomized Algorithms for Matrices and Data, *Foundations and Trends in Theoretical Computer Science*.
- Woodruff, D. (2014): Sketching as a Tool for Numerical Linear Algebra, *Foundations and Trends in Theoretical Computer Science*.
- Kane, D. and J. Nelson (2012): Sparser Johnson-Lindenstrauss Transforms, ACM 61:1.
- J. Nelson, J. and H. L. Nguyên (2013): OSNAP: Faster Numerical Linear Algebra Algorithms via Sparser Subspace Embeddings, Proceedings of IEEE 54th Annual Symposium, FOCS.
- Wang, S. et al (2018): Sketched Ridge Regression, Optimization Perspective, Statistical Perspective, and Model Averaging, Journal of Machine Learning Research.

Toy Example

$$A = \begin{pmatrix} 1 & 0 & -.25 & .25 & 0 \\ 0 & 1 & .5 & -.5 & 0 \end{pmatrix}^T .$$

Consider now three 2×2 $\tilde{A}$ matrices :

$$\Pi_1 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix}, \qquad \tilde{A}_1 = \Pi_1 A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\Pi_2 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \qquad \tilde{A}_2 = \Pi_2 A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

$$\Pi_3 = \begin{pmatrix} 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & -1 & 1 & 1 \end{pmatrix}, \qquad \tilde{A}_3 = \Pi_3 A = \begin{pmatrix} 0 & 0 \\ .5 & 0 \end{pmatrix}.$$

- Only Π_1 preserves the rank of A.
- Is rank problem empirically relevant?

An Empirical Example: Mincer Equation

- IPUMS US 1940 (preliminary) complete count data
- White men 16-64. Full sample. $n = 24,640,959$.
- Mincer equations with different covariates:

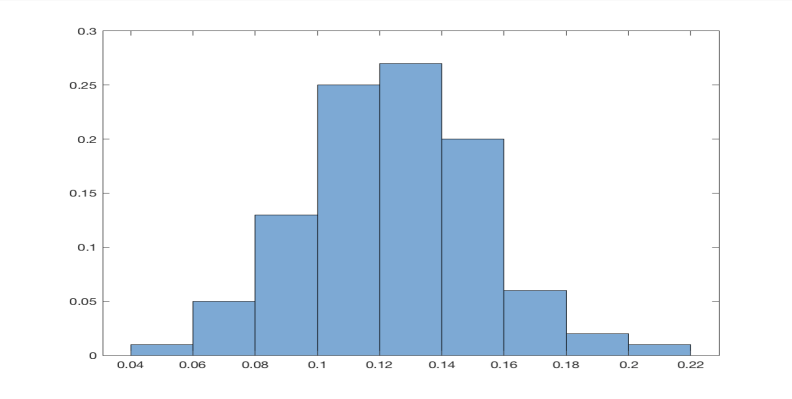
$$\text{log wage} = \beta_0 + \beta_1 \text{edu} + \beta_2 \text{exp} + \beta_3 \text{exp}^2 + \text{error} \quad (1a)$$

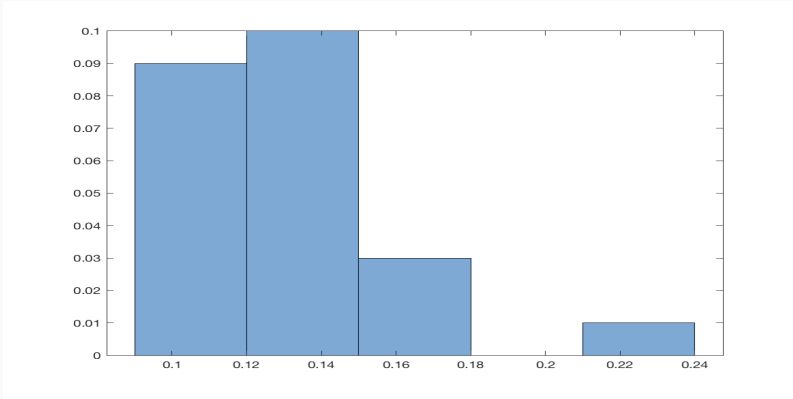
$$\text{log wage} = \beta_0 + \beta_1 \text{edu} + \sum_{j=0}^{11} \beta_{2+j} 1_{\text{exp} \in [j, j+5]} + \text{error}. \quad (1b)$$

- Full sample estimates of β_1: 0.12145 and 0.12401.
- Potential problem: $\text{EXP} \in [0, 58]$ but few $\text{EXP} > 50$.

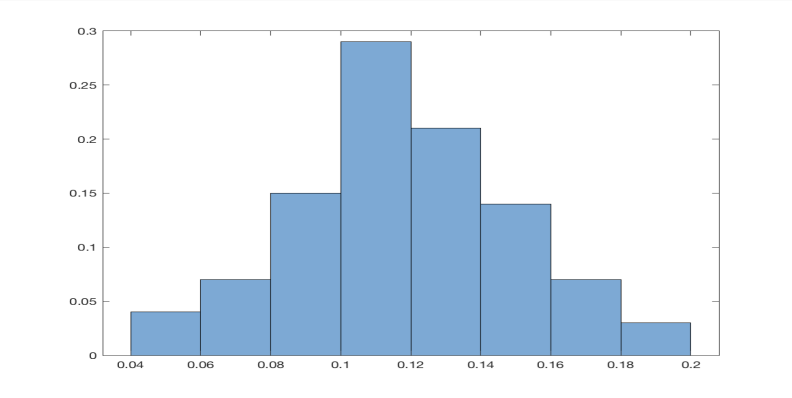
Distribution of Estimates

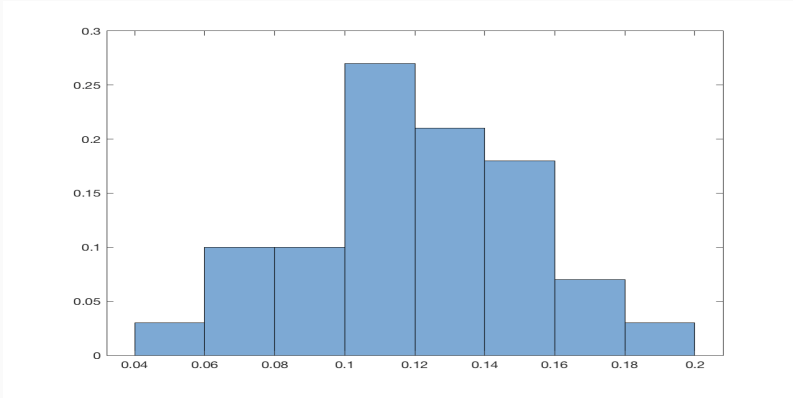
Uniform Sampling w/o replacement: singular in 77 of 100 sketches





CountSketch: singular in 1 of 100 sketches





Design of Π matters.

Matrix Multiplication: $A \in \mathbb{R}^{n\times d}$, $B \in \mathbb{R}^{n\times p}$.

Goal: compute $A^T B$.

- Standard: each element is a sum of inner product

$$C_{ij} = [A^T B]_{ij} = \sum_{k=1}^{n} A_{ik}^T B_{kj}.$$

.

- Algorithmic: outer product, sum n rank-1 matrices

$$C = \underbrace{A^T B}_{d\times p} = \sum_{i=1}^{n} \underbrace{A_{(i)}^T B_{(i)}}_{(d\times 1)\times(1\times p)} ,$$

- AMM: Approximate Matrix Multiplication:

$$\tilde{C} \;=\; (\Pi A)^T \Pi B = \frac{1}{m}\sum_{s=1}^{m} \frac{1}{p_{k_s}} A_{(k_s)}^T B_{(k_s)}$$

Properties of AMM

- $\mathbb{E}[\tilde{C}] = C$ (unbiased).
- $\mathbb{V}[\tilde{C}] \equiv \mathbb{E}[\|\tilde{C} - C\|_F^2] \leq \frac{1}{m} \sum_{s=1}^{m} \frac{1}{p_s} \|A_{(s)}\|_F^2 \|B_{(s)}\|_F^2$.
- Using variance minimizing $p_k = \frac{\|A_{(k)}\|_2 \|B_{(k)}\|_2}{\sum_{s=1}^{n} \|A_{(s)}\|_2 \|B_{(s)}\|_2}$. gives

$$P(\|(\Pi A)^T(\Pi B) - A^T B)\|_2 \geq \epsilon \|A\|_2 \|B\|_2) < \delta.$$

- Put $A = B$,

$$P(\|(\Pi A)^T(\Pi A) - A^T A)\|_2 \geq \epsilon \|A\|_2^2) < \delta.$$

- Goal algorithmic sampling can be understood as findings conditions to preserve second moments of A.
- Conditions are problem specific

Why not do Uniform Sampling?

- When the row norms are not evenly distributed (ie. some rows have more informative than others), uniform sampling is not efficient. More rows needed.
    - Deaton and Ng (1998).
- Alternative: sample informative rows more frequently.
- Ideal: use leverage score sampling probabilities, but computationally costly.
- Look for easy to compute $\tilde{A}$s that preserve features of A.

JL Lemma

Let $x_1, \ldots, x_d \in \mathbb{R}^n$ be arbitrary points and $\epsilon \in (0, 1/2)$.
There exists $\Pi : \mathbb{R}^n \to \mathbb{R}^m$ with $m = O(\frac{log(d)}{\epsilon^2})$, s.t. $\forall i, j$,

$$\|\Pi(x_i - x_j)\|_2^2 = (1 \pm \epsilon)\|x_i - x_j\|_2^2.$$

- m logarithmic in d but does not depend on n.
- Every set of d points in n-dim. Euclidean space can be represented by a m dim. with all pairwise distance preserved up to a $1 \pm \epsilon$ factor.
- For regressions, we need to preserve d-dimensional subspace, not just d individual vectors.

Subspace Embedding

- An L_2 subspace embedding for the column space of $A \in \mathbb{R}^{n \times d}$ with distortion ϵ is a Π such that for all $x \in \mathbb{R}^d$,

$$\|\Pi Ax\|_2^2 = (1 \pm \epsilon)\|Ax\|_2^2.$$

- Change basis: $Ax = U\Sigma V^T x \equiv Uz$ where $z = \Sigma V^T x$

- $\Pi \in \mathbb{R}^{m \times n}$ is a subspace embedding for U if

$$\|(\Pi U)^T(\Pi U) - \underbrace{U^T U}_{I_d}\|_2 \leq \epsilon.$$

- Subspace embedding implies small eigenvalue distortions:

.

$$\sigma_i^2(\Pi U) \in [1 - \epsilon, 1 + \epsilon], \quad \forall i.$$

Construction of Π

Where to get Π?

- take those with JL properties.
  - Random Sampling:
    - rows in $\tilde{A}$ are rows of A.
    - eg. uniform sampling w or w/o replacement
    - eg. leverage score sampling.
  - Random Projections:
    - rows in $\tilde{A}$ are linear combinations of rows of A.
    - eg. Gaussian, sparse Π.
    - eg. SRHT (uniformize, then random sample).
- countsketch:
  - Very sparse Π: one non-zero entry per column.
  - Streaming version available.

- Uniform sampling

$$\tilde{A} = D \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} A = \frac{\sqrt{10}}{\sqrt{3}} \begin{pmatrix} A_{10} \\ A_5 \\ A_1 \end{pmatrix}.$$

- Sparse random projections

$$\tilde{A} = D \begin{pmatrix} 0 & 0 & 1 & 1 & 0 & -1 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & -1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \end{pmatrix} A = \frac{\sqrt{3}}{\sqrt{9}} \begin{pmatrix} A_3 + A_4 - \\ A_1 - A_3 - A_5 \\ A_2 + A_7 - \end{pmatrix}$$

- Countsketch (sparsest, nnz(A) time)

$$\tilde{A} = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \end{pmatrix} A = \begin{pmatrix} A_4 - A_6 \\ A_1 - A_3 - A_5 + A_8 \\ A_2 + A_7 - A_{10} \end{pmatrix}$$

Monte Carlo Experiments

- A: $\mathrm{N}(n, d)$ or $\textsc{exprnd}(d, [n, d])$, $(n, d) = (20,000, 5)$
- Each simulation, evaluate
  - (Norm approx.) count total times in $\frac{d(d+1)}{2}$ pairs s.t. $|||\Pi(a_i - a_j)||_2^2 \in (1 \pm \epsilon)||a_i - a_j||_2^2$.
  - (Eigenvalue distortion) $||\frac{\sigma(\Pi A)}{\sigma(A)} - 1||_2$.
- Results are averaged over 100 simulations.

m	Random Sampling			Random Projections					
	u-wo	u-w	bern	rndn	± 1	srht	ssrp	count	lev
rndn	Norm approximation								
161	0.627	0.624	0.538	0.628	0.633	0.631	0.640	0.642	0.766
322	0.801	0.792	0.700	0.790	0.795	0.795	0.800	0.793	0.906
483	0.883	0.879	0.800	0.877	0.873	0.878	0.882	0.881	0.961
644	0.931	0.931	0.871	0.926	0.929	0.927	0.931	0.928	0.980
805	0.961	0.956	0.898	0.956	0.952	0.957	0.956	0.957	0.991
966	0.978	0.972	0.932	0.971	0.974	0.974	0.975	0.972	0.997
1288	0.990	0.987	0.973	0.990	0.991	0.989	0.990	0.991	0.999
2576	1.000	1.000	0.998	1.000	1.000	1.000	1.000	1.000	1.000
	Eigenvalue distortion								
161	0.189	0.191	0.191	0.189	0.187	0.188	0.189	0.188	0.155
322	0.126	0.128	0.127	0.127	0.127	0.128	0.126	0.129	0.105
483	0.099	0.100	0.098	0.101	0.102	0.102	0.099	0.100	0.082
644	0.082	0.085	0.084	0.086	0.084	0.085	0.085	0.086	0.070
805	0.073	0.074	0.073	0.075	0.075	0.074	0.074	0.075	0.061
966	0.065	0.067	0.065	0.067	0.066	0.067	0.067	0.068	0.054
1288	0.055	0.056	0.055	0.056	0.056	0.056	0.055	0.055	0.045
2576	0.033	0.036	0.033	0.036	0.037	0.036	0.035	0.037	0.030

m	Random Sampling			Random Projections					
	u-wo	u-w	bern	rndn	± 1	srht	ssrp	count	lev
Exp	Norm approximation								
161	0.432	0.429	0.402	0.627	0.624	0.636	0.628	0.637	0.735
322	0.580	0.578	0.548	0.796	0.795	0.794	0.800	0.791	0.890
483	0.697	0.681	0.656	0.885	0.880	0.876	0.882	0.885	0.943
644	0.747	0.738	0.717	0.925	0.930	0.929	0.930	0.928	0.977
805	0.799	0.790	0.767	0.951	0.954	0.955	0.953	0.959	0.986
966	0.851	0.840	0.812	0.971	0.968	0.973	0.969	0.972	0.994
1288	0.899	0.894	0.866	0.990	0.988	0.989	0.991	0.989	0.997
2576	0.986	0.974	0.975	1.000	1.000	1.000	1.000	1.000	1.000
	Eigenvalue distortion								
161	0.263	0.257	0.259	0.188	0.193	0.188	0.190	0.188	0.157
322	0.176	0.177	0.175	0.126	0.128	0.127	0.127	0.127	0.103
483	0.136	0.141	0.138	0.100	0.098	0.100	0.099	0.099	0.082
644	0.116	0.118	0.116	0.084	0.083	0.083	0.082	0.085	0.067
805	0.101	0.107	0.103	0.074	0.073	0.073	0.073	0.073	0.060
966	0.090	0.094	0.090	0.066	0.067	0.066	0.065	0.065	0.053
1288	0.076	0.079	0.075	0.055	0.055	0.055	0.054	0.055	0.045
2576	0.048	0.052	0.048	0.036	0.036	0.037	0.035	0.036	0.029

Algorthmic (worse case) results for overdetermined systems

$\min_\beta \|y - X\beta\|_2^2$.

Put $A = [y\ , X] \in \mathbb{R}^{n \times d}$, $x = \begin{pmatrix} 1 & -\beta \end{pmatrix}^T \in \mathbb{R}^d$, $d = p + 1$.

$$\begin{aligned} \frac{1}{(1+\epsilon)} \|y - X\tilde{\beta}\|_2^2 &\leq \|\Pi(y - X\tilde{\beta})\|_2^2 \quad \text{by subspace embedding} \\ &\leq \|\Pi(y - X\hat{\beta})\|_2^2 \quad \text{since } \tilde{\beta} \text{ is optimal} \\ &\leq (1+\epsilon)\|y - X\hat{\beta}\|_2^2 \text{ by subspace embedding} \end{aligned}$$

$$\underbrace{\|y - X\tilde{\beta}\|_2^2}_{=\text{approx obj } L(\tilde{\beta})} \leq (1+\epsilon)^2 \underbrace{\|y - X\hat{\beta}\|_2^2}_{=\text{full sample obj} L(\hat{\beta})} .$$

Two Useful Features

Choosing a Π is like choosing a kernel.

$$\Pi^T \Pi = I_n + R_{11} \tag{2a}$$

$$\Pi \Pi^T = \frac{n}{m} I_m \tag{2b}$$

	(2a)	(2b)
RS1 (Uniform,w/o)	yes	yes
RS2 (Uniform,w)	yes	no
RS3 (Bernoulli)	yes	no
RS4 (Leverage)	yes	no
RP1 (Gaussian)	no	no
RP2 (SRHT)	yes	yes
RP3 (SRP)	no	no
CS (countsketch)	no	no

Will focus on uniform sampling w/o replacement.

Setup

Linear regression model with K regressors:

$$y = X^T\beta + e, \qquad e_i \sim (0, \Omega_e).$$

Assumption OLS:

(i) $X = U\Sigma V^T$ are non-random and X^TX is non-singular;
(ii) $\mathbb{E}[e_i] = 0$ and $\mathbb{E}[ee^T] = \Omega_e$ is diagonal, positive definite.

Assumption PI:

(i) Π is independent of e;
(ii) for given singular value distortion parameter $\varepsilon_\sigma \in (0, 1)$, there exists failure parameter $\delta_\sigma \in (0, 1)$ s.t. $\forall k \in [1, \dots K]$, $P\Big(|1 - \sigma_k^2(\Pi U)| \leq \varepsilon_\sigma \Big) \geq 1 - \delta_\sigma$.
(iii) $\Pi^T\Pi$ is an $n \times n$ diagonal matrix and $\Pi\Pi^T = \frac{n}{m} I_m$.

- For any non-zero $K \times 1$ vector c,

$$\left| \frac{c^T[(X^TX)^{-1} - (\tilde{X}^T\tilde{X})^{-1}]c}{c^T(X^TX)^{-1}c} \right| \leq \frac{\varepsilon_\sigma}{1-\varepsilon_\sigma}.$$

- Variance of $\tilde{\beta}$ under OLS and PI2:

$$\mathbb{V}(\hat{\beta}) = (\tilde{X}^T\tilde{X})^{-1}(\tilde{X}^T\Pi\Omega_e\Pi^T\tilde{X})(\tilde{X}^T\tilde{X})^{-1}.$$

  - Under PI3, $\Pi\Pi^T = \frac{n}{m}I_m$,

$$\mathbb{V}(\tilde{\beta}) = \frac{n}{m}(\tilde{X}^T\tilde{X})^{-1}(\tilde{X}^T\Omega_e\tilde{X})^{-1}(\tilde{X}^T\tilde{X})^{-1}$$

  - If, in addition, $\Omega_e = \sigma_e^2 I_n$

$$\mathbb{V}(\tilde{\beta}|\Pi) = \sigma_e^2\frac{n}{m}(\tilde{X}^T\tilde{X})^{-1}.$$

Prediction at x_0

Under OLS and PI3:

$$\frac{\mathrm{MSE}(x_0^T\tilde{\beta}|\Pi)}{\mathrm{MSE}(x_0^T\hat{\beta})} \leq \underbrace{\frac{n}{m}}_{\text{sample size}} \underbrace{\left(\frac{1}{1-\varepsilon_\sigma}\right)}_{\text{sketching error}}.$$

- Corollary: for a $K \times 1$ vector c.

$$\frac{c^T\mathbb{V}(\tilde{\beta}|\Pi)c}{c^T\mathbb{V}(\hat{\beta})c} \leq \frac{n}{m}\left(\frac{1}{1-\varepsilon_\sigma}\right).$$

- Upper and Lower bound:

$$\frac{n}{m}\left(\frac{1}{1+\varepsilon_\sigma}\right) \leq \frac{c^T\mathbb{V}(\tilde{\beta}|\Pi)c}{c^T\mathbb{V}(\hat{\beta})c} \leq \frac{n}{m}\left(\frac{1}{1-\varepsilon_\sigma}\right).$$

Results hold with prob $1-\delta_\sigma$.

Test Linear Restrictions

Goal: test q linear restrictions, $H_0 : R\beta = r$. Under normality,

$$F(\hat{\beta}) \;=\; R(\hat{\beta}-\beta_0)^T\left(\hat{\mathbb{V}}(R\hat{\beta})\right)^{-1}R(\hat{\beta}-\beta_0) \sim F_{q,n-d}$$

- Power against fixed alternative $R\beta_0 = r$ depends on $\mathbb{V}(\hat{\beta})$ through non-centrality parameter λ. For $q = 1$,

$$\hat{\lambda} \;=\; \frac{(R\beta_0 - r)^2}{\mathbb{V}(R\hat{\beta})} > \frac{(R\beta_0 - r)^2}{\mathbb{V}(R\tilde{\beta})} = \tilde{\lambda}.$$

- Relative non-centrality:

$$\frac{\hat{\lambda}}{\tilde{\lambda}} = \frac{\mathbb{V}(R\tilde{\beta})}{\mathbb{V}(R\hat{\beta})} \leq \frac{n}{m}\left(\frac{1}{1-\varepsilon_\sigma}\right).$$

Back of Envelope Calculations: $E[F] = \frac{(n-d)(q+\lambda)}{q(n-d-2)}$

	Δ	$\phi_n = \frac{n\Delta^2}{2}$	$\mathbb{E}[F]$	power
n =1e6	0	0	1.000	0.050
m =2000			1.001	0.050
m =1000			1.002	0.050
n =1e6	0.1	5000	5001	1.0
m =2000		9.09	10.10	0.885
m =1000		4.54	5.55	0.607
n =1e6	0.2	20000	20001	1.0
m =2000		36.36	37.40	0.999
m =1000		18.18	19.22	0.993
n =1e6	0.3	45000	45001	1.0
m =2000		81.81	82.90	1.0
m =1000		40.90	41.88	1.0

$$m^* = \frac{m}{1+\epsilon}$$

Unless small data analysis, $\frac{n-d}{n-d-2} \approx \frac{m-d}{m-d-2}$ in big data.

Local Power:

- For $q = 1$, $R\beta_0 - r = \frac{\Delta}{\sqrt{n}}$
- How informative is the test at testing $\Delta_m = \frac{\Delta}{\sqrt{m}}$

Δ	$n = 1e6$	$m = 100$	$m = 200$	$m = 1000$	$m = 2000$
2	0.292	0.266	0.268	0.270	0.270
4	0.807	0.761	0.765	0.768	0.769
6	0.988	0.979	0.980	0.981	0.981

Local alternative changes slowly when m is not small.

Divide and Conquer

Create J sketches $(\tilde{y}_j, \tilde{X}_j) = (\Pi_j y, \Pi_j X)$, without replacement:

$$\begin{aligned}
\mathbb{V}(\bar{\beta}) &= \frac{1}{J^2}\sum_{j=1}^{J}\sum_{k=1}^{J}(\tilde{X}_j^T\tilde{X}_j)^{-1}\Big(\tilde{X}_j\Pi_j\mathbb{E}[ee^T|\Pi_1,\ldots,\Pi_J]\Pi_k\tilde{X}_k\Big)(\tilde{X}_k^T\tilde{X}_k)^{-1} \\
&= \sigma^2\frac{1}{J^2}\sum_{j=1}^{J}(\tilde{X}_j^T\tilde{X}_j)^{-1}(\tilde{X}_j\Pi_j\Pi_j^T\tilde{X}_j)^{-1}(\tilde{X}_j^T\tilde{X}_j) \\
&= \frac{n}{mJ}\left(\frac{1}{J}\sum_{j=1}^{J}\mathbb{V}(\tilde{\beta}_j)\right).
\end{aligned}$$

Variance Reduction by Averaging

Inference using $\tilde{\beta}$

$$\bar{\beta} = \frac{1}{J}\sum_{j=1}^{J}\tilde{\beta}_j, \qquad \text{var}(\bar{\beta}) = \frac{1}{J(J-1)}\sum_{j=1}^{J}\left[\text{se}(\tilde{\beta}_j)\right]^2,$$

$$\bar{t}_2 = J^{-1}\sum_{j=1}^{J}\hat{t}_j, \qquad \text{var}(\bar{t}_2) = \frac{1}{J-1}\sum_{j=1}^{J}(\tilde{t}_j - \bar{t}_2)^2.$$

Consider two pooled t statistics:

$$\begin{aligned} \bar{T}_1 &= \frac{\bar{\beta} - \beta_0}{se(\bar{\beta})} \\ \bar{T}_2 &= \sqrt{J}\frac{\bar{t}_2}{se(\bar{t}_2)}. \end{aligned}$$

Assumption PI-Avg

(i) $(\Pi_1, \ldots, \Pi_J)$ is independent of e;

(ii) for all j, k such that $j \neq k$, $\Pi_j \Pi_j^T = \frac{n}{m} I_m$ and $\Pi_j \Pi_k^T = O_m$.

(iii) for given $\varepsilon_\sigma \in (0,1)$, there exists $\delta_\sigma \in (0,1)$ st

$$P\Big(|1 - \sigma_k^2(\Pi_j U)| \leq \varepsilon_\sigma\Big) \geq 1 - \delta_\sigma.$$

Theorem: Under OLS and PI-Avg hold,

$$\frac{c^T \mathbb{V}(\bar{\beta}|\Pi_1, \ldots, \Pi_J) c}{c^T \mathbb{V}(\hat{\beta}) c} \leq \frac{n}{mJ} \frac{1}{(1 - \varepsilon_\sigma)}.$$

- By choice of J, can we make $\mathbb{V}(\bar{\beta})$ close to $\mathbb{V}(\hat{\beta})$!

$K = 3$

m	J	RS1	SRHT	CS	LEV	RS1	SRHT	CS	LEV
		$\hat{\beta}_3$ with $\beta_3 = 1.0$				se($\hat{\beta}_3$)			
500	1	0.999	1.002	0.999	1.000	0.046	0.044	0.045	0.039
500	5	0.999	1.000	1.000	1.001	0.021	0.020	0.020	0.018
500	10	1.000	1.000	1.000	1.000	0.014	0.015	0.015	0.012
1000	1	1.000	0.999	0.998	1.000	0.032	0.032	0.031	0.026
1000	5	1.000	1.000	1.000	1.000	0.014	0.015	0.015	0.012
1000	10	1.000	0.999	1.000	1.000	0.010	0.010	0.010	0.008
2000	1	1.001	1.000	1.001	1.000	0.021	0.022	0.022	0.019
2000	5	1.000	1.000	1.000	1.000	0.010	0.010	0.010	0.008
2000	10	1.000	1.000	1.000	1.000	0.007	0.007	0.007	0.006
5000	1	1.000	1.001	0.999	1.000	0.014	0.014	0.014	0.012
5000	5	1.000	1.000	0.999	1.000	0.006	0.006	0.006	0.005
5000	10	1.000	1.000	1.000	1.000	0.005	0.005	0.004	0.004

m	J	RS1	SRHT	CS	LEV	RS1	SRHT	CS	LEV
		Size				Power, $\beta_3 = 0.98$			
500	1	0.050	0.040	0.062	0.063	0.081	0.069	0.071	0.104
500	5	0.035	0.029	0.021	0.045	0.114	0.115	0.123	0.185
500	10	0.039	0.051	0.053	0.037	0.276	0.258	0.265	0.345
1000	1	0.048	0.044	0.050	0.052	0.101	0.101	0.085	0.113
1000	5	0.024	0.046	0.032	0.023	0.221	0.218	0.233	0.320
1000	10	0.041	0.035	0.044	0.042	0.461	0.454	0.452	0.617
2000	1	0.045	0.052	0.058	0.055	0.136	0.142	0.147	0.189
2000	5	0.034	0.022	0.035	0.025	0.436	0.432	0.451	0.545
2000	10	0.040	0.043	0.038	0.053	0.763	0.761	0.767	0.902
5000	1	0.053	0.046	0.040	0.047	0.298	0.322	0.275	0.399
5000	5	0.026	0.018	0.026	0.019	0.835	0.832	0.829	0.930
5000	10	0.045	0.046	0.036	0.054	0.987	0.993	0.989	0.999

Algorithmic vs Statistical Optimality

- computation efficiency: favors smaller m
  - run faster, lower storage cost.
- statistical efficiency: favors bigger m
  - smaller variance, more powerful tests.

We will use inference considerations to tune algorithmic guides.

Sketch of ideas

- Algorithmic guide: $m \geq 6\varepsilon_\sigma^{-2} n \ell_{\max} \log(2J \cdot K/\delta_\sigma)$
- $p_i = \frac{\ell_i}{K} = \|U_{(i)}\|^2$, ℓ_i is leverage score of row i.
- For $J = 1$, the algorithmic guide says

$$m = \Omega\left(nK \log(K) \cdot p_{\max} \right)$$

- $\ell_i \leq \sigma_K^{-1}(S_X) \frac{1}{n} \|X_{(i)}\|_2^2$. But $\left\|X_{(i)}\right\|_2^2 \leq K \cdot X_{\max}^2$.
- Hence $p_{\max} \leq \frac{\sigma_K^{-1}(S_X) X_{\max}^2}{n}$.

A New Deterministic Rule

Assumptions for $X_{\max} = o_p((nK)^{1/r})$.

a. $\sigma_K(S_X)$ is bounded below by c_X w.p.1 as $n \to \infty$;

b. $\mathbb{E}[|X_{(i,j)}|^r] \leq C_X$ for some C_X and some $r \geq 2$.

Proposition A deterministic rule:

$$\begin{cases} m_1 & = \Omega\left((nK)^{1+2/r} \log K/n\right) \quad \text{if} \quad r < \infty \\ m_1 & = \Omega(K \log(nK)) \qquad\qquad \text{if in addition } \mathbb{E}[\exp(tX_{(i,j)})] \leq C_X. \end{cases}$$

Two Inference Conscious Rules

Proposition Suppose that $e_i \sim N(0, \sigma_e^2)$ and the Assumptions of Theorem 1 hold. Let $\bar{\gamma}$ be the target power of a one-sided test τ_1 at β^0 and $\bar{\alpha}$ be the tolerated Type 1 error. Define

$$S(\bar{\alpha}, \bar{\gamma}) = \Phi_{\bar{\gamma}}^{-1} + \Phi_{1-\bar{\alpha}}^{-1}, \quad \text{and} \quad \tau_2(m) = \frac{c^T(\beta^0 - \beta_0)}{\mathrm{SE}(c^T\tilde{\beta})}.$$

- data dependent rule: Given $\tilde{\beta}(m_1)$ and effect size $\beta^0 - \beta_0$,

$$m_2(m_1) = m_1 \frac{S^2(\bar{\alpha}, \bar{\gamma})}{\tau_2^2(m_1)}.$$

- data oblivious rule: given a pre- specified $\tau_2(\infty)$,

$$m_3 = n \frac{S^2(\bar{\alpha}, \bar{\gamma})}{\tau_2^2(\infty)}.$$

Values of $S(\alpha, \gamma)$

	γ				
α	0.500	0.600	0.700	0.800	0.900
0.010	2.326	2.580	2.851	3.168	3.608
0.050	1.645	1.898	2.169	2.486	2.926
0.100	1.282	1.535	1.806	2.123	2.563

- Since $\frac{m_2}{m_1}$, any $\tau_2 < S$ will adjust m_1 up.
- $\frac{n}{m_3} = \frac{\tau_2^2(\infty)}{S^2(\bar{\alpha},\gamma)}$, sample size effect is tied to power.

m_2: Trade-off

$n = 1e7, r = 10, K = 10, m_0 = 1000, \bar{\alpha} = 0.05$

		$(\beta_1^0 - \beta_{10})$				
$\bar{\gamma}$	σ_e	.005	.01	.015	.02	.025
0.50	0.50	29686	7421	3298	1855	1187
0.80	0.50	67837	16959	7537	4240	2713
0.90	0.50	93965	23491	10441	5873	3759
0.50	1.00	98296	24574	10922	6143	3932
0.80	1.00	224620	56155	24958	14039	8985
0.90	1.00	311136	77784	34571	19446	12445
0.50	3.00	981128	245282	109014	61321	39245
0.80	3.00	2242020	560505	249113	140126	89681
0.90	3.00	3105562	776391	345062	194098	124222

Summary of Main Findings

1. Π should have subspace embedding property.
   - For speed and simplicity: CS and uniform sampling.
   - For easy interpretation: uniform sampling.
   - For improved estimates: form multiple sketches, average.
2. Sketching error= sample size effect+ approximation error.
3. For inference: need bigger m than algorithmically optimal.
4. Propose inference conscious m.