A new theory for sketching in linear regression

Edgar Dobriban[1] Sifan Liu[2]

[1]Wharton Statistics Department
University of Pennsylvania

[2]Department of Mathematical Sciences
Tsinghua University

March 27, 2019

Overview

Overview

The Age of Data

- We live in the *Age of Data*
- An enormous variety of digital data is generated and recorded every day
- Examples: Web (pages, links, ads, reviews), Science (health records, genetics, high-energy physics), ...

Big data

- It is expensive to analyze large datasets
  - Storage: Hard disk
  - Memory: RAM
  - Computation: CPU, GPU, TPU
  - Communication
- Fast algorithms are important, because they end up saving real resources (money, time, energy)

Big data

- Many approaches to efficiently processing big data (from computer science, statistics, optimization, machine learning, data mining, signal processing and information theory, etc)
  - Efficient data structures
  - Distributed and parallel computing
  - Data reduction and compression
  - Online and streaming algorithms
  - Randomized algorithms
  - ....

Randomized algorithms for processing matrices and data

- Many of data problems can be modeled via matrices
- E.g., Collect data on n patients. Measure their features (demographics, lifestyle, genetics, preferences, medical tests, health outcomes)
- Arrange features of patient i into $p \times 1$ vector x_i, and let $x_i^\top$ be the rows of data matrix $X \in \mathbb{R}^{n \times p}$.
- Data matrix X is big $\rightarrow$ reduce it to smaller matrix
- Sample size reduction: $S \in \mathbb{R}^{r \times n}$ $(r < n)$

$$X \in \mathbb{R}^{n \times p} \rightarrow SX \in \mathbb{R}^{r \times p}$$

- Dimension reduction: $T \in \mathbb{R}^{p \times t}$ $(p < t)$

$$X \in \mathbb{R}^{n \times p} \rightarrow XT \in \mathbb{R}^{n \times t}$$

- How to choose S, T? Randomize, leading to random projections

Uses of Random Projections

- Randomized numerical linear algebra [Drineas and Mahoney, 2016]: matrix multiplication, SVD, low rank approximation, etc.
- Statistics, Machine learning and Data mining: nonparametric regression [Yang et al., 2017], ridge regression [Lu et al., 2013], two sample testing [Lopes et al., 2011], classification [Cannings and Samworth, 2017], clustering [Fern and Brodley, 2003], PCA [Rokhlin et al., 2009], deep learning [Abadi et al., 2016], etc.
- Convex optimization [Pilanci and Wainwright, 2015, 2016, 2017]
- Econometrics [Ng, 2017]
- Genomics [Galinsky et al., 2016]

Johnson-Lindenstrauss (JL) Lemma

- JL lemma is possibly the most fundamental result on RP for big data
- Informally, RP can preserve pairwise distances if the target dimension is not too small
- Formally (Johnson-Lindenstrauss Lemma): any set V of n points in p-dim Euclidean space can be projected down to a randomly chosen subspace of $k = c\varepsilon^{-2} \log n$ dimensions s.t. $\forall x, y \in V$

$$\|x - y\|(1 - \varepsilon) \leq \sqrt{n/k}\|Px - Py\| \leq \|x - y\|(1 + \varepsilon)$$

holds with constant probability (w.r.t. P). [Johnson and Lindenstrauss, 1984]

- Can be translated into guarantees for algorithms. Sometimes too loose to be useful.

Comments

- RP/randomized algorithms:
- Strengths: General-purpose, works for arbitrary/adversarial data, fast,
- Limitations: May fail with some probability, results depend on random draw (large variance), hard to tune, unknown/mysterious when and how well they work
- We shed light on the mystery, giving concrete tools for practitioners to be able to decide how and when to use random projections - "instruction manual" for RP/sketching

Overview

Linear Regression

- Fundamental statistical method to study dependence of outcome of interest (e.g., blood pressure) and predictor variables (e.g., lifestyle, age, weight, demographics, genetic variables)
- Data:
  1. n samples/datapoints. Arrange features of sample i into $p \times 1$ vector x_i, and let $x_i^\top$ be the rows of data matrix $X \in \mathbb{R}^{n \times p}$, $n > p$.
  2. Outcomes y_i arranged into $n \times 1$ vector Y.
- Goal: understand effect of features on outcome
- Linear regression model: $Y = X\beta + \varepsilon$
  1. Unknown regression coefficients $\beta \in \mathbb{R}^p$,
  2. Noise $\varepsilon \in \mathbb{R}^n$. The ε_i's are uncorrelated, with $\mathbb{E}[\varepsilon_i] = 0$, $\mathbb{E}\left[\varepsilon_i^2\right] = \sigma^2$
- Ordinary Least Squares (OLS) estimator (most widely used multivariate statistical method)

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y$$

Sketched Linear Regression

- Complexity of full OLS: $O(np^2)$ floating point operations (flops).
- n, p are large, so this is too expensive.
- e.g., $n = 10^8$, $p = 10^6$, get $\approx 10^{22}$ flops. Typical CPU - 10^{11} flops/second. Takes 3,000 years!
- Approximation is necessary
- Apply random projection/sketching matrix $S \in \mathbb{R}^{r \times n}$ of the samples to get $(\tilde{X}, \tilde{Y}) = (SX, SY)$, then do least-squares on $(\tilde{X}, \tilde{Y})$, which gives

$$\hat{\beta}_s = (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top \tilde{Y}.$$

Questions

- Faster but less accurate... By how much?
- How does this work from a statistical point of view? What is the loss? How many dimensions do we need?
- How to choose sketching matrix S? for instance
  - Uniform sampling of rows of X (i.e., subsampling)
  - iid entries (Gaussian or $-1, 0, 1$),
  - Haar
  - randomized Hadamard/Fourier transform
  - leverage sampling
- Under what conditions on the data X?

Statistical efficiency

- How to measure the statistical efficiency?
- Standard measures for evaluating estimator $\hat{\beta}$ of parameter β
  - Mean Squared Error: $\mathbb{E}\left[\|\hat{\beta} - \beta\|^2\right]$
  - In regression context, Residual Error: $\mathbb{E}\left[\|Y - X\hat{\beta}\|^2\right]$
  - In regression context, Test Error: Let x_t, y_t be a test datapoint. We use $\hat{y}_t = x_t^\top \hat{\beta}$ to predict the unobserved y_t: $\mathbb{E}\left[(y_t - x_t^\top \hat{\beta})^2\right]$

Relative Efficiencies

- How to compare the statistical efficiency of two estimators?
- Relative efficiency: ratio of errors

$$\text{Variance efficiency: } VE(\hat{\beta}_s, \hat{\beta}) := \frac{\mathbb{E}\left[\|\beta - \hat{\beta}_s\|^2\right]}{\mathbb{E}\left[\|\beta - \hat{\beta}\|^2\right]},$$

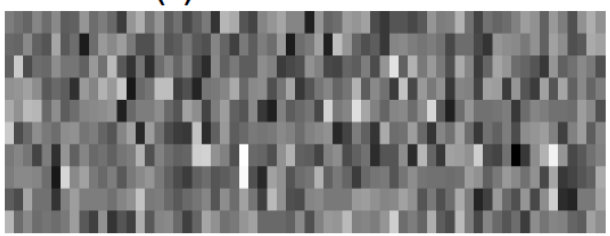
$$\text{Residual efficiency: } RE(\hat{\beta}_s, \hat{\beta}) := \frac{\mathbb{E}\left[\|Y - X\hat{\beta}_s\|^2\right]}{\mathbb{E}\left[\|Y - X\hat{\beta}\|^2\right]},$$

$$\text{The out-of-sample efficiency: } OE(\hat{\beta}_s, \hat{\beta}) := \frac{\mathbb{E}\left[(y_t - x_t^\top \hat{\beta}_s)^2\right]}{\mathbb{E}\left[(y_t - x_t^\top \hat{\beta})^2\right]}.$$

Sketching Methods

- Three broad categories
  1. Sampling methods: sample rows of X independently, either iid or according to "importance" (fast/non-robust)
  2. iid entries: entries of S are iid (popular/easy to understand/inaccurate/slow)
  3. Structured orthogonal: S consists of random rows from a structured orthonormal/unitary matrix such as Fourier/Hadamard transform (fast/accurate/robust/hard to analyze and understand)

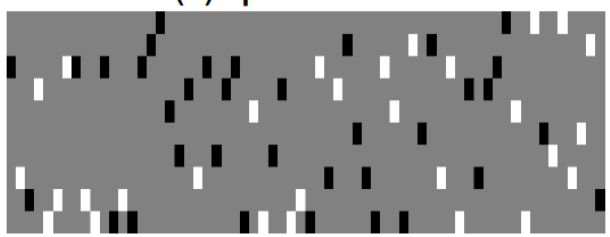
(a) Gaussian sketch



(b) ± 1 random sign sketch

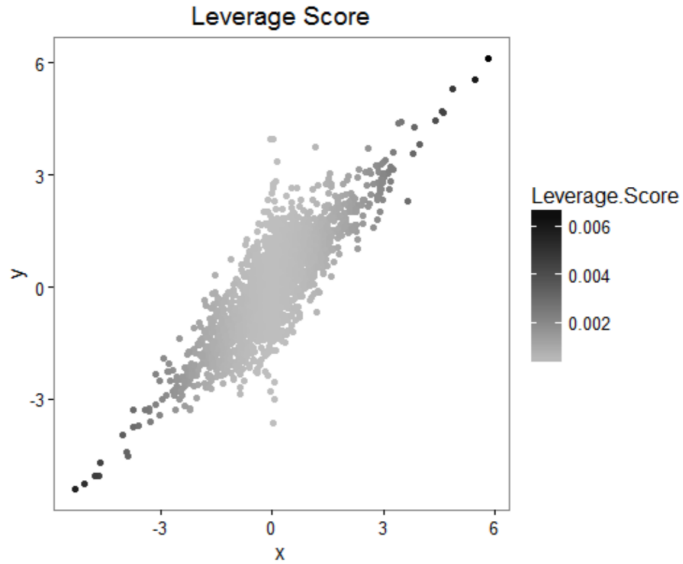
(c) ROS sketch

(d) sparse sketch



Sampling Methods

- Uniform sampling: S samples each row of X with equal probability
- Leverage score sampling: S samples each row of X w.r.t. its leverage scores $H_{ii} = x_i^\top (X^\top X)^{-1} x_i$, where $H = X(X^\top X)^{-1} X^\top$ is the "hat" matrix.



- Performing sampling: $O(n)$ flops
- Computing leverage scores: $O(np^2)$
- Estimating leverage scores by Hadamard transform: $O(pn \log n)$ [Drineas et al., 2012] (better to use Hadamard transform directly)

iid entries

- Gaussian projection: S has iid $\mathcal{N}(0,1)$ entries
- Sparse projection: S has iid random entries with $\mathbb{P}(S_{ij} = \pm 1) = 1/(2s)$, $\mathbb{P}(S_{ij} = 0) = 1 - 1/s$, where s is some fixed constant greater than 1.
- Generally: S_{ij} are iid standardized random variables
- Computing the projection SX: $O(rnp)$

Structured orthogonal

- Fourier transform:
  - Start with Discrete Fourier Transform matrix F, with entries $F_{uv} = n^{-1/2} e^{-2\pi i \frac{(u-1)(v-1)}{n}}$.
  - Let $S = DFQP$, where Q has iid ± 1 diagonals, D selects random rows, P is uniform permutation matrix
- Hadamard transform: Same construction, starting with recursively defined Hadamard matrix, $H_1 = 1$,

$$H_n = \begin{pmatrix} H_{n/2} & H_{n/2} \\ H_{n/2} & -H_{n/2} \end{pmatrix},$$

- Computing SX via Fast Fourier transform $O(pn \log n)$

Overview

Theoretical results

Table: Approximate efficiency of sketching. Original linear model: $n \times p$, with n samples and p dimensions ($n > p$). Sketched linear model: $r \times p$ ($n > r > p$). The loss functions are VE (variance efficiency) and OE (out-of-sample prediction efficiency).

X	S	VE$=\frac{\mathbb{E}\left[\|\beta-\hat{\beta}_s\|^2\right]}{\mathbb{E}\left[\|\beta-\hat{\beta}\|^2\right]}$	OE$=\frac{\mathbb{E}\left[(y_t-x_t^\top \hat{\beta}_s)^2\right]}{\mathbb{E}\left[(y_t-x_t^\top \hat{\beta})^2\right]}$
Arbitrary	iid entries	$1+\frac{n-p}{r-p}$	$\frac{nr-p^2}{n(r-p)}$
	Fourier/Hadamard	$\frac{n-p}{r-p}$	$\frac{r(n-p)}{n(r-p)}$
Ortho-invariant	Uniform sampling		
Elliptical: $WZ\Sigma^{\frac{1}{2}}$	Leverage sampling	$\frac{\eta_{sw^2}^{-1}(1-p/n)}{\eta_{w^2}^{-1}(1-p/n)}$	$\frac{1+\mathbb{E}\left[w^2\right]\eta_{sw^2}^{-1}(1-\gamma)}{1+\mathbb{E}[w^2]\eta_{w^2}^{-1}(1-\gamma)}$

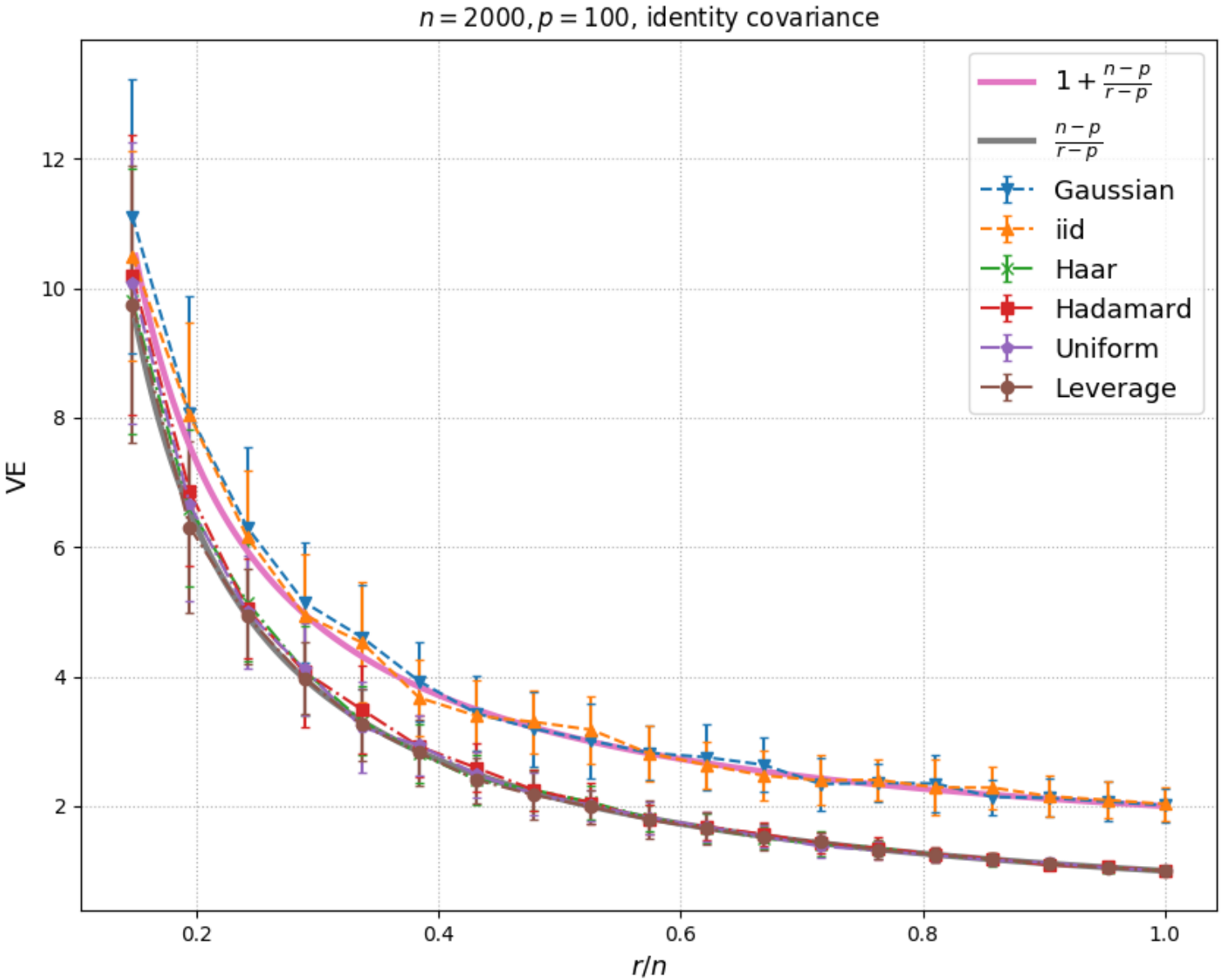
Comments on Our Results

- Accurate in numerical simulations and data examples
- Can be used as "sample size calculations"

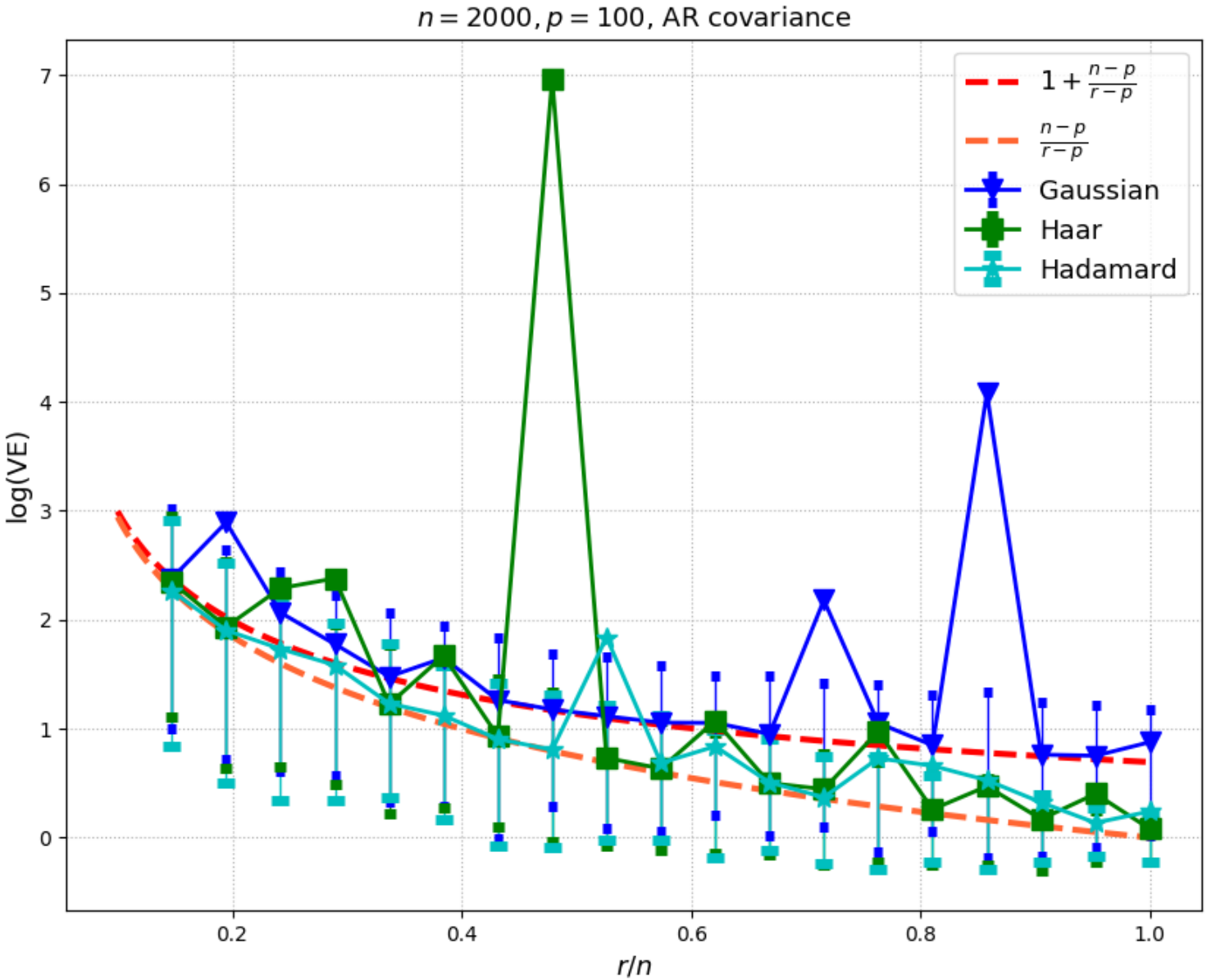
Numerical Experiments

- Take $n = 2000, p = 100$, with r ranging from 300 to 2000.
- Generate and fix the data matrix X from
  - a standard Gaussian distribution
  - multivariate t distribution with AR-1 covariance
- Generate β, ε as standard Gaussian
- At each dimension r, randomly generate 50 sketching matrices S
- Compute $\hat{\beta}_s$ and $\hat{\beta}$, then find VE

Numerical Results



Numerical Results



Empirical Data Analysis Results

- We make empirically testable predictions - how much do residuals increase after sketching?
- These can be tested using data without making any assumptions (very different from most work in high-dimensional statistics...)
- We test our results on the Million Song Year Prediction Dataset (MSD) [Bertin-Mahieux et al., 2011].
- $n = 515,344$ samples and $p = 90$ features
- We take a random test set of size 10,000.
- For each target dimension r, we show the mean, 5% and 95% quantiles over 10 repetitions.

Million Song Year Prediction Dataset

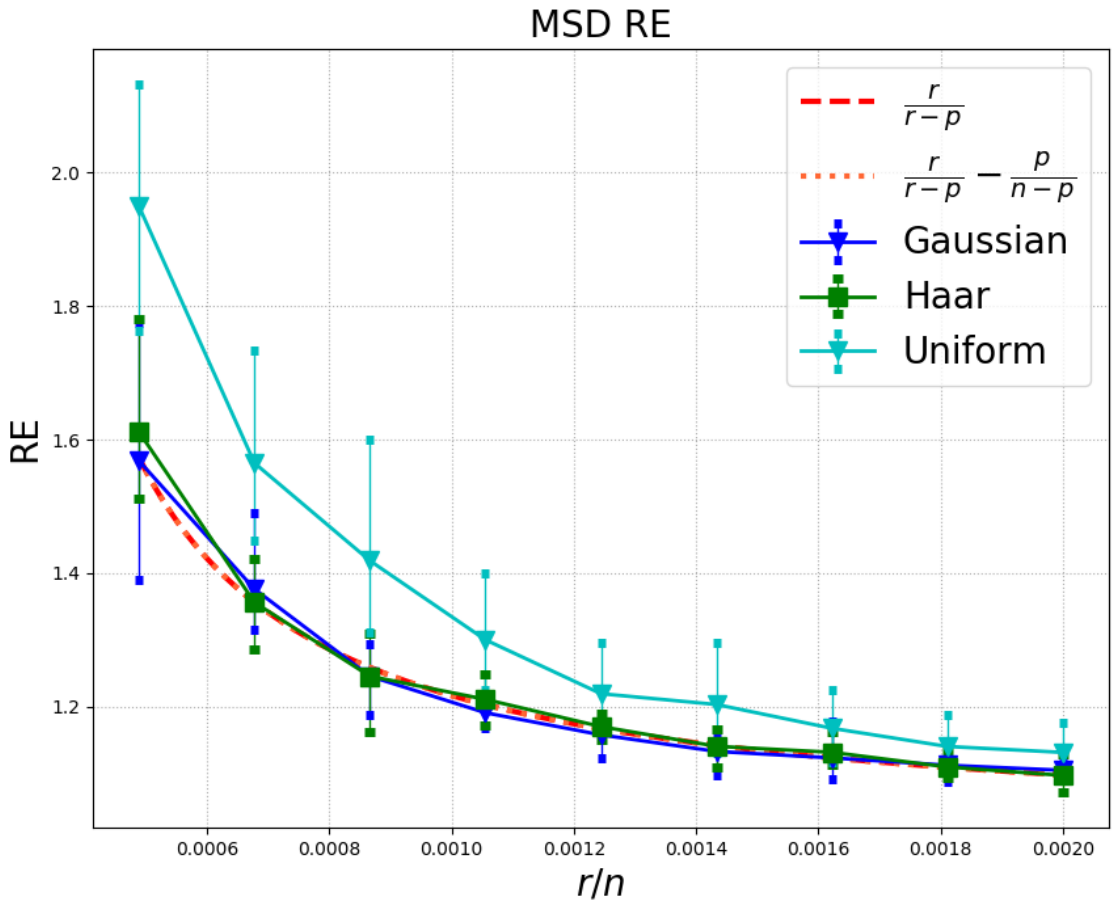


Figure: Good agreement of theory and empirical results for Gaussian and Hadamard/Haar projections. For uniform sampling, our theory requires the data matrix X to be rotationally invariant, which may not hold, leading to less accuracy.

Proof aspects

- Our proofs rely on asymptotic random matrix theory and free probability
- Perfect fit: algorithm random, need weak assumptions
- "Standard" results (such as the Marchenko-Pastur law) are *not* enough.
- To study the subsampled randomized Hadamard transform (SRHT), we discovered that we can use the results of Anderson and Farrell [2014] on *asymptotically liberating sequences.*

Anderson & Farrell's paper

Asymptotically liberating sequences of random unitary matrices

Greg W. Anderson [a,*], Brendan Farrell [b,1]

[a] *School of Mathematics, University of Minnesota, Minneapolis, MN 55455, USA*
[b] *Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA 91125, USA*

ARTICLE INFO

Article history:
Received 22 February 2013
Accepted 21 December 2013
Available online 29 January 2014
Communicated by Dan Voiculescu

MSC:
60B20
42A61
46L54
15B52

Keywords:
Free probability

ABSTRACT

A fundamental result of free probability theory due to Voiculescu and subsequently refined by many authors states that conjugation by independent Haar-distributed random unitary matrices delivers asymptotic freeness. In this paper we exhibit many other systems of random unitary matrices that, when used for conjugation, lead to freeness. We do so by first proving a general result asserting "asymptotic liberation" under quite mild conditions, and then we explain how to specialize these general results in a striking way by exploiting Hadamard matrices. In particular, we recover and generalize results of the second-named author and of Tulino, Caire, Shamai and Verdú.

Proof outline for VE

- Expand $\mathbb{E}\|\beta - \hat{\beta}_S\|^2$ in terms of trace:

$$\mathbb{E}\|\beta - \hat{\beta}_S\|^2 = \mathbb{E}\,\mathrm{tr}[QQ^\top] = \|Q\|_{Fr}^2$$
$$Q = (X^\top S^\top S X)^{-1} X^\top S^\top S$$

- Calculate
  - Finite-sample expectation for Gaussian S (using *Wishart properties*)
  - Limiting expectation for iid S (using *Lindeberg swapping* [Chatterjee, 2006])
  - For orthogonal S, reduces to $\mathrm{tr}(X^\top S^\top S X)^{-1}$

Proof outline for VE, Hadamard/Fourier

- Find limit of $\mathrm{tr}(X^\top S^\top S X)^{-1}$, X arbitrary, $S = DHQP$
  - $D = diag(B_i)$, $B_i \sim Bernoulli(r/n)$ selects random rows,
  - H is Hadamard,
  - Q has iid ± 1 diagonals,
  - P is uniform permutation matrix
- Let $W = P^\top QHQP$ be the *signed Hadamard matrix*. Then

$$X^\top S^\top S X = X^\top (P^\top QH) D (HQP) X \stackrel{d}{=} X^\top W D W X.$$

- Anderson & Farrell: $D, n^{-1} WXX^\top W$ are asymptotically freely independent

Related Work

Raskutti and Mahoney [2016] showed that for Gaussian projection,

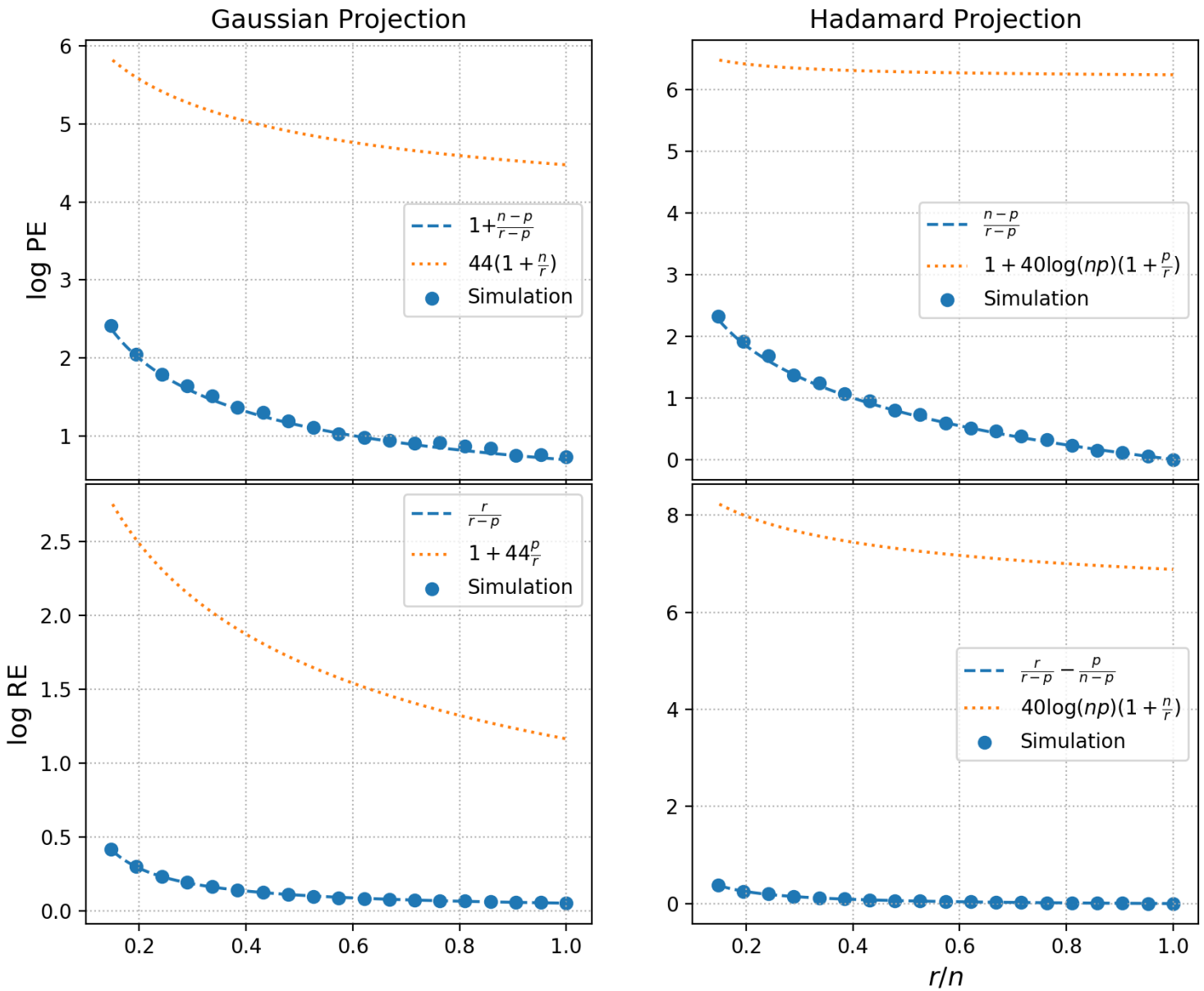
$$PE \leq 44(1 + \frac{n}{r}),$$
$$RE \leq 1 + 44\frac{p}{r},$$

and for Hadamard/Fourier transform,

$$PE \leq 1 + 40\log(np)(1 + \frac{p}{r}),$$
$$RE \leq 40\log(np)(1 + \frac{n}{r}),$$

hold with some constant probability.
Our contribution: We find the asymptotically EXACT answer!

Comparing results



Overview

Highlights

- Sketching, a promising general approach for big data analysis
- Our results are accurate & conveniently usable
- Interesting & important to study randomized algorithms for other problems: PCA, classification, hypothesis testing, etc
- Real opportunity for RMT/FP in randomized algorithms for data science

References I

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318. ACM, 2016.

Greg W Anderson and Brendan Farrell. Asymptotically liberating sequences of random unitary matrices. *Advances in Mathematics*, 255:381–413, 2014.

Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

Timothy I Cannings and Richard J Samworth. Random-projection ensemble classification. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):959–1035, 2017.

Sourav Chatterjee. A generalization of the lindeberg principle. *The Annals of Probability*, 34(6):2061–2076, 2006.

Petros Drineas and Michael W Mahoney. Randnla: randomized numerical linear algebra. *Communications of the ACM*, 59(6):80–90, 2016.

References II

Petros Drineas, Malik Magdon-Ismail, Michael W Mahoney, and David P Woodruff. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13(Dec):3475–3506, 2012.

Xiaoli Z Fern and Carla E Brodley. Random projection for high dimensional data clustering: A cluster ensemble approach. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 186–193, 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.

William B Johnson and Joram Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics*, 26(189-206):1, 1984.

Miles Lopes, Laurent Jacob, and Martin J Wainwright. A more powerful two-sample test in high dimensions using random projection. In *Advances in Neural Information Processing Systems*, pages 1206–1214, 2011.

Yichao Lu, Paramveer Dhillon, Dean P Foster, and Lyle Ungar. Faster ridge regression via the subsampled randomized hadamard transform. In *Advances in neural information processing systems*, pages 369–377, 2013.

References III

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. Randomized sketches of convex programs with sharp guarantees. *IEEE Transactions on Information Theory*, 61(9): 5096–5115, 2015.

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.

Mert Pilanci and Martin J Wainwright. Newton sketch: A near linear-time optimization algorithm with linear-quadratic convergence. *SIAM Journal on Optimization*, 27(1):205–245, 2017.

Garvesh Raskutti and Michael W Mahoney. A statistical perspective on randomized sketching for ordinary least-squares. *The Journal of Machine Learning Research*, 17(1):7508–7538, 2016.

Vladimir Rokhlin, Arthur Szlam, and Mark Tygert. A randomized algorithm for principal component analysis. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1100–1124, 2009.

References IV

Yun Yang, Mert Pilanci, and Martin J Wainwright. Randomized sketches for kernels: Fast and optimal nonparametric regression. *The Annals of Statistics*, 45(3): 991–1023, 2017.