AN ECONOMETRIC VIEW OF ALGORITHMIC SUBSAMPLING

Sokbae Lee[*] Serena Ng [†]

July 3, 2019

Abstract

Datasets that are terabytes in size are increasingly common, but computer bottlenecks often frustrate a complete analysis of the data. While more data are better than less, diminishing returns suggest that we may not need terabytes of data to estimate a parameter or test a hypothesis. But which rows of data should we analyze, and might an arbitrary subset of rows preserve the features of the original data? This paper reviews a line of work that is grounded in theoretical computer science and numerical linear algebra, and which finds that an algorithmically desirable *sketch* of the data must have a *subspace embedding* property. Building on this work, we study how prediction and inference is affected by data sketching within a linear regression setup. The sketching error is small compared to the sample size effect which is within the control of the researcher. As a sketch size that is algorithmically optimal may not be suitable for prediction and inference, we use statistical arguments to provide 'inference conscious' guides to the sketch size. When appropriately implemented, an estimator that pools over different sketches can be nearly as efficient as the infeasible one using the full sample.

Keywords: sketching, coresets, subspace embedding, countsketch, uniform sampling.
JEL Classification: C2, C3.

[*] Department of Economics, Columbia University and Institute for Fiscal Studies

[†] Department of Economics, Columbia University and NBER

The authors would like to thank David Woodruff and Shusen Wang for helpful discussions. The first author would like to thank the European Research Council for financial support (ERC-2014-CoG- 646917-ROMIA) and the UK Economic and Social Research Council for research grant (ES/P008909/1) to the CeMMAP. The second author would like to thank the National Science Foundation for financial support (SES: 1558623).

1 Introduction

The availability of terabytes of data for economic analysis is increasingly common. But analyzing large datasets is time consuming and sometimes beyond the limits of our computers. The need to work around the data bottlenecks was no smaller decades ago when the data were in megabytes than it is today when data are in terabytes and petabytes. One way to alleviate the bottleneck is to work with a *sketch* of the data.[1] These are data sets of smaller dimensions and yet representative of the original data. We study how the design of linear sketches affects estimation and inference in the context of the linear regression model. Our formal statistical analysis complements those in the theoretical computer science and numerical linear algebra derived using different notions of accuracy and whose focus is computation efficiency.

There are several motivations for forming sketches of the data from the full sample. If the data are too expensive to store and/or too large to fit into computer memory, the data would be of limited practical use. It might be cost effective in some cases to get a sense from a smaller sample whether an expensive test based on the full sample is worth proceeding. Debugging is certainly faster with fewer observations. A smaller dataset can be adequate while a researcher is learning how to specify the regression model, as loading a gigabyte of data is much faster than a terabyte even we have enough computer memory to do so. With confidentiality reasons, one might only want to circulate a subset rather than the full set of data.

For a sketch of the data to be useful, the sketch must preserve the characteristics of the original data. Early work in the statistics literature used a sketching method known as 'data squashing'. The idea is to approximate the likelihood function by merging data points with similar likelihood profiles, such as by taking their mean. There are two different ways to squash the data. One approach is to construct subsamples randomly. Deaton and Ng (1998) uses 'binning methods' and uniform sampling to speed up estimation of non-parametric average derivatives. While these methods work well for the application under investigation, its general properties are not well understood. An alternative is to take the data structure into account. Du Mouchel et al. (1999) also forms multivariate bins of the data, but they match low order moments within the bin by non-linear optimization. Owen (1990) reweighs a random sample of X to fit the moments using empirical likelihood estimation. Madigan et al. (1999) uses likelihood-based clustering to select data points that match the target distribution. While theoretically appealing, modeling the likelihood profiles can itself be time consuming and not easily scalable.

Data sketching is of also interest to computer scientists because they are frequently required to

[1] The term 'synopsis' and 'coresets' have also been used. See Comrode et al. (2011), and Agarwal and Varadarajan (2004). We generically refer to these as sketches.

provide summaries (such as frequency, mean, and maximum) of data that stream by continously.[2] Instead of an exact answer which would be costly to compute, *pass-efficient* randomized algorithms are designed to run fast, requires little storage, and guarantee the correct answer with a certain probability. But this is precisely the underlying premise of data sketching in statistical analysis.[3]

Though randomized algorithms are increasingly used for sketching in a wide range of applications, the concept remains largely unknown to economists except for a brief exposition in Ng (2017). This paper provides a gentle introduction to these algorithms in Sections 2 to 4. To our knowledge, this is the first review on sketching in the econometrics literature. We will use the term *algorithmic subsampling* to refer to randomized algorithms designed for the purpose of sketching, to distinguish them from bootstrap and subsampling methods developed for frequentist inference. In repeated sampling, we only observe one sample drawn from the population. Here, the complete data can be thought of as the population which we observe, but we can only use a subsample. Algorithmic subsampling does not make distributional assumptions, and balancing between fast computation and favorable worst case approximation error often leads to algorithms that are oblivious to the properties the data. In contrast, exploiting the probabilistic structure is often an important aspect of econometric modeling.

Perhaps the most important tension between the algorithmic and the statistical view is that while fast and efficient computation tend to favor sketches with few rows, efficient estimation and inference inevitably favor using as many rows as possible. Sampling schemes that are optimal from an algorithmic perspective may not be desirable from an econometric perspective, but it is entirely possible for schemes that are not algorithmically optimal to be statistically desirable. As there are many open questions about the usefulness of these sampling schemes for statistical modeling, there is an increased interest in these methods within the statistics community. Recent surveys on sketching with a regression focus include Ahfock et al. (2017) and Geppert, Ickstadt, Munteanu, Qudedenfeld and Sohler (2017), among others. Each paper offers distinct insights, and the present paper is no exception.

Our focus is on efficiency of the estimates for prediction and inference within the context of the linear regression model. Analytical and practical considerations confine our focus eventually back to uniform sampling, and to a smaller extent, an algorithm known as the countsketch. The results in Sections 5 and 6 are new. It will be shown that data sketching has two effects on estimation, one due to sample size, and one due to approximation error, with the former dominating the later in all quantities of empirical interest. Though the sample size effect has direct implications for the power

[2] The seminal paper on frequency moments is Alon et al. (1999). For a review of the literature, see Comrode et al. (2011) and Cormode (2011).

[3] Pass-efficient algorithms read in data at most a constant number of times. A computational method is referred to as a streaming model if only one pass is needed.

of any statistical test, it is at the discretion of a researcher. We show that moment restrictions can be used to guide the sketch size, with fewer rows being needed when more moments exist. By targeting the power of a test at a prespecified alternative, the size of the sketch can also be tuned so as not to incur excessive power loss in hypothesis testing. We refer to this as the 'inference conscious' sketch size.

There is an inevitable trade-off between computation cost and statistical efficiency, but the statistical loss from using fewer rows of data can be alleviated by combining estimates from different sketches. By the principle of 'divide and conquer', running several estimators in parallel can be statistically efficient and still computationally inexpensive. Both uniform sampling and the countsketch are amenable to parallel processing which facilitates averaging of quantities computed from different sketches. We assess two ways of combining estimators: one that averages the parameter estimates, and one that averages test statistics. Regardless of how information from the different sketches are combined, pooling over subsamples always provides more efficient estimates and more powerful tests. It is in fact possible to bring the power of a test arbitrarily close to the one using the full sample, as will be illustrated in Section 6.

1.1 Motivating Examples

The sketching problem can be summarized as follows. Given an original matrix $A \in \mathbb{R}^{n\times d}$, we are interested in $\widetilde{A} \in \mathbb{R}^{m\times d}$ constructed as

$$\widetilde{A} = \Pi A$$

where $\Pi \in \mathbb{R}^{m\times n}$, $m < n$. In a linear regression setting, $A = [y\ X]$ where y is the dependent variable, and X are the regressors. Computation of the least squares estimator takes $O(nd^2)$ time which becomes costly when the number of rows, n is large. Non-parametric regressions fit into this setup if X is a matrix of sieve basis. Interest therefore arises to use fewer rows of A without sacrificing too much information.

To motivate why the choice of the sampling scheme (ie. Π) matters, consider as an example a 5×2 matrix

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

The rows have different information content as the row norm is $(1, 1, 0.559, 0.559, 0)^T$. Consider now three 2×2 $\widetilde{A}$ matrices constructed as follows:

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

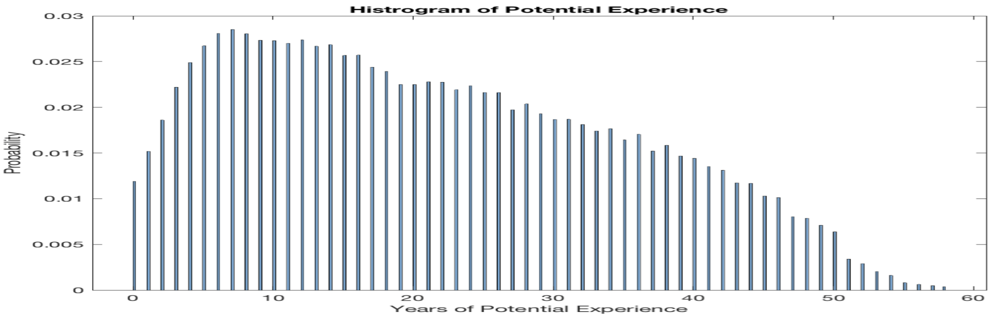
$$\Pi_2 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \qquad \widetilde{A}_2 = \Pi_2 A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

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

Of the three sketches, only Π_1 preserves the rank of A. The sketch defined by Π_2 fails because it chooses row 5 which has no information. The third sketch is obtained by taking linear combination of rows that do not have independent information. The point is that unless Π is chosen appropriately, $\widetilde{A}$ may not have the same rank as A.

Of course, when m is large, changing rank is much less likely and one may also wonder if this pen and pencil problem can ever arise in practice. Consider now estimation of a Mincer equation which has the logarithm of WAGE as the dependent variable, estimated using the IPMUS (2019) dataset which provides a preliminary but complete count data for the 1940 U.S. Census. This data is of interest because it was the first census with information on wages and salary income. For illustration, we use a sample of n =24 million white men between the age of 16 and 64 as the 'full sample'. The predictors that can be considered are years of education, denoted (EDU), and potential experience, denoted (EXP).

Figure 1: Distribution of Potential Experience



Two Mincer equations with different covariates are considered:

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

$$\log \text{wage} \quad = \quad \beta_0 + \beta_1 \text{edu} + \sum_{j=0}^{11} \beta_{2+j} 1\{j \leq \text{exp} < (j+5)\} + \text{error}. \tag{1b}$$

Model (1a) uses EXP and EXP^2 as control variables. Model (1b) replaces potential experience with indicators of experience in five year intervals. Even though there are three predictors including the intercept, the number of covariates K is four in the first model and thirteen in the second. In both cases, the parameter of interest is the coefficient for years of education (β_1). The full sample estimate of β_1 is 0.12145 in specification (1a) and 0.12401 in specification (1b).

Figure 1 shows the histogram of EXP. The values of EXP range from 0 to 58. The problem in this example arises because there are few observations with over 50 years of experience. Hence there is no guarantee that an arbitrary subsample will include observations with EXP > 50. Without such observations, the subsampled covariate matrix may not have full rank. Specification (1b) is more vulnerable to this problem especially when m is small.

We verify that rank failure is empirically plausible in a small experiment with sketches of size $m = 100$ extracted using two sampling schemes. The first method is uniform sampling without replacement which is commonly used in economic applications. The second is the countsketch which will be further explained below. Figure 2 and Figure 3 show the histograms of subsample estimates for uniform sampling and the countsketch, respectively. The left panel is for specification (1a) and the right panel is for specification (1b). In our experiments, singular matrices never occurred with specification (1a); the OLS estimates can be computed using both sampling algorithms and both performed pretty well. However, uniform sampling without replacement produced singular matrices for specification (1b) 77% of the time. The estimates seem quite different from the full sample estimates, suggesting not only bias in the estimates, but also that the bias might not be random. In contrast, the countsketch failed only once out of 100 replications. The estimates are shown in the right panel of Figure 3 excluding the singular case.

This phenomenon can be replicated in a Monte Carlo experiment with $K = 3$ normally distributed predictors. Instead of X_3, it is assumed that we only observe a value of one if its latent value is three standard deviation from the mean. Together with an intercept, there are four regressors. As in the Mincer equation, the regressor matrix has a reduced rank of 3 with probability of 0.58, 0.25, 0.076 when $m = 200, 500, 1000$ rows are sampled uniformly but is always full rank only when $m = 2000$. In contrast, the countsketch never encounters this problem even with $m = 100$. The simple example underscores the point that the choice of sampling scheme matters. As will be seen below, the issue remains in a more elaborate regression with several hundred covariates. This motivates the need to better understand how to form sketches for estimation and inference.

2 Matrix Sketching

This section presents the key concepts in algorithmic sampling. The material is based heavily on the monographs by Mahoney (2011) and Woodruff (2014), as well as the seminal work of Drineas, Mahoney and Muthukrishnan (2006), and subsequent refinements developed in Drineas et

Figure 2: Distributions of Estimates from Uniform Sampling without Replacement

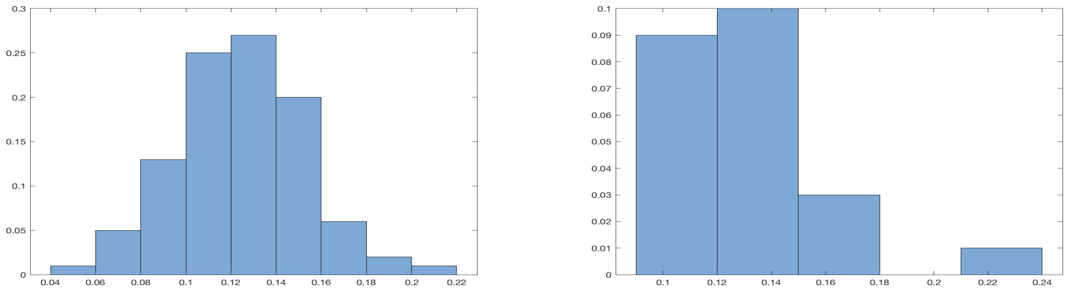
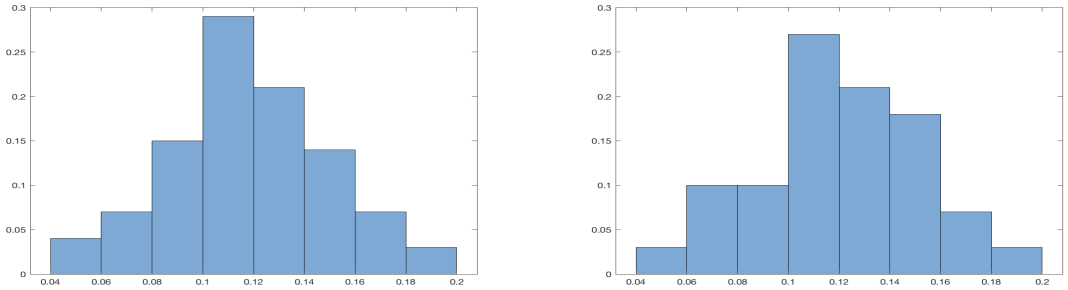


Figure 3: Distributions of Estimates from CountSketch Sampling



al. (2011), Nelson and Nguyen (2013a), Nelson and Nguyen (2014), Cohen, Nelson and Woodruff (2015), Wang, Gittens and Mahoney (2018), among many others.

We begin by setting up the notation. Consider an $n \times d$ matrix positive definite A. Let $A^{(j)}$ denote its j-th column of A and $A_{(i)}$ be its i-th row. Then

$$A = \begin{pmatrix} A_{(1)} \\ \vdots \\ A_{(n)} \end{pmatrix} = \begin{pmatrix} A^{(1)} & \dots & A^{(n)} \end{pmatrix}$$

and $A^T A = \sum_{i=1}^n A_{(i)}^T A_{(i)}$. The singular value decomposition of A is $A = U\Sigma V^T$ where U and V are the left and right eigenvectors of dimensions $(n \times d)$ and $(d \times d)$ respectively. The matrix Σ is $d \times d$ is diagonal with entries containing the singular values of A denoted $\sigma_1, \dots, \sigma_d$, which are ordered such that σ_1 is the largest. Since $A^T A$ is positive definite, its k-th eigenvalue $\omega_k(A^T A)$ equals $\sigma_k(A^T A) = \sigma_k^2(A)$, for $k = 1, \dots d$. The best rank k approximation of A is given by

$$A_k = U_k U_k^T A \equiv P_{U_k} A$$

where U_k is an $n \times k$ orthonormal matrix of left singular vectors corresponding to the k largest singular values of A, and $P_{U_k} = U_k U_k^T$ is the projection matrix.

The Frobenius norm (an average type criterion) is $\|A\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^d |A_{ij}|^2} = \sqrt{\sum_{i=1}^k \sigma_i^2}$. The spectral norm (a worse-case type criterion) is $\|A\|_2 = \sup_{\|x\|_2=1} \|Ax\|_2 = \sqrt{\sigma_1^2}$, where $\|x\|_2$ is the Euclidean norm of a vector x. The spectral norm is bounded above by the Frobenius norm since $\|A\|_2^2 = |\sigma_1|^2 \leq \sum_{i=1}^n \sum_{j=1}^d |A_{ij}|^2 = \|A\|_F^2 = \sum_{i=1}^d \sigma_i^2$.

Let f and g be real valued functions defined on some unbounded subset of real positive numbers n. We say that $g(n) = O(f(n))$ if $|g(n)| \leq k|f(n)|$ for some constant k. This means that $g(n)$ is at most a constant multiple of $f(n)$ for sufficiently large values of n. We say that $g(n) = \Omega(f(n))$ if $g(n) \geq kf(n)$ for all $n \geq n_0$. This means that $g(n)$ is at least $kf(n)$ for some constant k. We say that $g(n) = \Theta(f(n))$ if $k_1 f(n) \leq g(n) \leq k_2 f(n)$ for all $n \geq n_0$. This means that $g(n)$ is at least $k_1 f(n)$ and at most $k_2 f(n)$.

2.1 Approximate Matrix Multiplication

Suppose we are given two matrices, $A \in \mathbb{R}^{n\times d}$ and $B \in \mathbb{R}^{n\times p}$ and are interested in the $d \times p$ matrix $C = A^T B$. The textbook approach is to compute each element of C by summing over dot products:

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

Equivalently, each element is the inner product of two vectors $A_{(i)}$ and $B^{(j)}$. Computing the entire C entails three loops through $i \in [1, d]$, $j \in [1, p]$, and $k \in [1, n]$. An algorithmically more efficient approach is to form C from outer products:

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

making C a sum of n matrices each of rank-1. Viewing C as a sum of n terms suggests to approximate it by summing $m < n$ terms only. But which m amongst the $\frac{n!}{m!(n-m)!}$ possible terms to sum? Consider the following Approximate Matrix Multiplication algorithm (AMM). Let p_j be the probability that row j will be sampled.

Algorithm AMM:

```
Input: $A \in \mathbb{R}^{n\times d}$, $B \in \mathbb{R}^{n\times p}$, $m > 0$, $p = (p_1, \ldots, p_n)$.
for $s = 1 : m$ do
    sample $k_s \in [1, \ldots n]$ with probability $p_{k_s}$ independently with replacement;
    set $\widetilde{A}_{(s)} = \frac{1}{\sqrt{mp_{k_s}}} A_{(k_s)}$ and $\widetilde{B}_{(s)} = \frac{1}{\sqrt{mp_{k_s}}} B_{(k_s)}$
Output: $\widetilde{C} = \widetilde{A}^T \widetilde{B}$.
```

The algorithm essentially produces

$$\widetilde{C} \;\; = \;\; (\Pi A)^T \Pi B = \frac{1}{m} \sum_{s=1}^{m} \frac{1}{p_{k_s}} A^T_{(k_s)} B_{(k_s)} \tag{2}$$

where k_s denotes the index for the non-zero entry in the s row of the matrix

$$\Pi = \frac{1}{\sqrt{m}} \begin{pmatrix} 0 & \frac{1}{\sqrt{p_{k_1}}} & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \frac{1}{\sqrt{p_{k_m}}} & 0 \end{pmatrix}.$$

The Π matrix only has only one non-zero element per row, and the (i,j)-th entry $\Pi_{ij} = \frac{1}{\sqrt{mp_j}}$ with probability p_j. In the case of uniform sampling with $p_k = \frac{1}{n}$ for all i, Π reduces to a sampling matrix scaled by $\frac{\sqrt{n}}{\sqrt{m}}$.

While $\widetilde{C}$ defined by (2) is recognized in econometrics as the estimator of Horvitz and Thompson (1952) which uses inverse probability weighting to account for different proportions of observations in stratified sampling, $\widetilde{C}$ is a sketch of C produced by the *Monte Carlo* algorithm AMM in the theoretical science literature literature.[4] The Monte-Carlo aspect is easily understood if we take A and B to be $n \times 1$ vectors. Then $A^T B = \sum_{i=1}^n A^T_{(i)} B_{(i)} = \sum_{i=1}^n f(i) \approx \int_0^n f(x)dx = f(a)n$ where the last step follows from mean-value theorem for $0 < a < n$. Approximating $f(a)$ by $\frac{1}{m}\sum_{s=1}^m f(k_s)$ gives $\frac{n}{m}\sum_{s=1}^m f(k_s)$ as the Monte Carlo estimate of $\int_0^n f(x)dx$.

Two properties of $\widetilde{C}$ produced by AMM are noteworthy. Under independent sampling,

$$\mathbb{E}\left[\frac{A^T_{(k_s)} B_{(k_s)}}{mp_{(k_s)}}\right] = \sum_{k=1}^{m} p_k \frac{A^T_{(k)} B_{(k)}}{mp_k} = [A^T B]_{ij}.$$

Hence regardless of the sampling distribution, $\widetilde{C}$ is unbiased. The variance of $\widetilde{C}$ defined in terms of the Frobenius norm is

$$\mathbb{E}\Big[\|\widetilde{C} - C\|_F^2\Big] \;\; = \;\; \frac{1}{m} \sum_{k=1}^{n} \frac{1}{p_k} \|A^T_{(k)}\|_2^2 \|B_{(k)}\|_2^2 - \frac{1}{m}\|C\|_F^2$$

which depends on the sampling distribution p. Drineas, Kannan and Mahoney (2006, Theorem 1) shows that minimizing $\sum_{k=1}^n \frac{1}{p_k}\|A^T_{(k)}\|_2^2 \|B_{(k)}\|_2^2$ with respect to p subject to the constraint $\sum_{k=1}^n p_k = 1$ gives[5]

$$p_k = \frac{\|A_{(k)}\|_2 \|B_{(k)}\|_2}{\sum_{s=1}^n \|A_{(s)}\|_2 \|B_{(s)}\|_2}.$$

[4] In Mitzenmacher and Upfal (2006), a Monte Carlo algorithm is a randomized algorithm that may fail or return an incorrect answer but whose time complexity is deterministic and does not depend on the particular sampling. This contrasts with a Las Vegas algorithm which always returns the correct answer but whose time complexity is random. See also Eriksson-Bique et al. (2011).

[5] The first order condition satifies $0 = -\frac{1}{p_k^2}\|A^T_{(k)}\|_2^2 \|B_{(k)}\|_2^2 + \lambda$. Solving for $\sqrt{\lambda}$ and imposing the constraint gives the result stated. Eriksson-Bique et al. (2011) derives probabilities that minimize expected variance for given distribution of the matrix elements.

This optimal p yields a variance of

$$\mathbb{E}\Big[\|\widetilde{C}-C\|_F^2\Big] \leq \frac{1}{m}\Big[\sum_{k=1}^{n}\|A_{(k)}\|_2\|B_{(k)}\|_2\Big]^2 \leq \frac{1}{m}\|A\|_F^2\|B\|_F^2.$$

It follows from Markov's inequality that for given error of size ε and failure probability $\delta > 0$,

$$P\bigg(\|\widetilde{C}-C\|_F^2 > \varepsilon^2\|A\|_F^2\|B\|_F^2\bigg) < \frac{\mathbb{E}\Big[\|\widetilde{C}-C\|_F^2\Big]}{\varepsilon^2\|A\|_F^2\|B\|_F^2} < \frac{1}{m\varepsilon^2},$$

implying that to have an approximation error no larger than ε with probability $1-\delta$, the number of rows used in the approximation must satisfy $m = \Omega(\frac{1}{\delta\varepsilon^2})$.

The approximate matrix multiplication result $\|A^TB - \widetilde{A}^T\widetilde{B}\|_F \leq \varepsilon\|A\|_F\|B\|_F$ is the building block of many of the theoretical results to follow. The result also holds under the spectral norm since it is upper bounded by the Frobenius norm. Since $A_{(i)}^TB_{(i)}$ is a rank one matrix, $\|A_{(i)}^TB_{(i)}\|_2 = \|A_{(i)}^T\|_2\|B_{(i)}\|_2$. Many of the results to follow are in spectral norm because it is simpler to work with a product of two Euclidean vector norms. Furthermore, putting $A = B$, we have

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

One may think of the goal of AMM as preserving the second moment properties of A. The challenge in practice is to understand the conditions that validate the approximation. For example, even though uniform sampling is the simplest of sampling schemes, it cannot be used blindly. Intuitively, uniform sampling treats all data points equally, and when information in the rows are not uniformly dispersed, the influential rows will likely be omitted. From the above derivations, we see that $\text{var}(\widetilde{C}) = O(\frac{n}{m})$ when $p_k = \frac{1}{n}$, which can be prohibitively large. The Mincer equation in the Introduction illustrates the pitfall with uniform sampling when m is too small, but that the problem can by and large be alleviated when $m > 2000$. Hence, care must be taken in using the algorithmic sampling schemes. We will provide some guides below.

2.2 Subspace Embedding

To study the properties of the least squares estimates using sketched data, we first need to make clear what features of A need to be preserved in $\widetilde{A}$. Formally, the requirement is that Π has a 'subspace embedding' property. An embedding is a linear transformation of the data that has the Johnson-Lindenstrauss (JL) property, and a subspace embedding is a matrix generalization of an embedding. Hence it is useful to start with the celebrated JL Lemma.

The JL Lemma, due to Johnson and Lindenstauss (1994), is usually written for linear maps that reduce the number of columns from an $n \times d$ matrix d to k. Given that our interest is ultimately in reducing the number of rows from n to m while keeping d fixed, we state the JL Lemma as follows:

Lemma 1 (JL Lemma) *Let $0 < \epsilon < 1$ and $\{a_1, \dots, a_d\}$ be a set of d points in $\mathbb{R}^n$ with $n > d$. Let $m \geq 8 \log d/\epsilon^2$. There exists a linear map $\Pi : \mathbb{R}^n \to \mathbb{R}^m$ such that $\forall a_i, a_j$*

$$(1-\epsilon)||a_i - a_j||_2^2 \leq ||\Pi a_i - \Pi a_j||_2^2 \leq (1+\epsilon)||a_i - a_j||_2^2.$$

In words, the Lemma states that every set of d points in Euclidean space of dimension n can be represented by a Euclidean space of dimension $m = \Omega(\log d/\epsilon^2)$ with all pairwise distance preserved up to a $1 \pm \epsilon$ factor. Notice that m is logarithmic in d and does not depend on n. A sketch of the proof is given in the online Appendix.

The JL Lemma establishes that d vectors in $\mathbb{R}^n$ can be embedded into $m = \Omega(\log d/\epsilon^2)$ dimensions. But there are situations when we need to preserve the information in the d columns jointly. This leads to the notion of 'subspace embedding' which requires that the norm of vectors in the column space of A be approximately preserved by Π with high probability.

Definition 1 (Subspace-Embedding) *Let A be an $n \times d$ matrix. An L_2 subspace embedding for the column space of A is an $m(\epsilon, \delta, d) \times n$ matrix Π such that $\forall x \in \mathbb{R}^d$,*

$$(1-\epsilon)\|Ax\|_2^2 \leq \|\Pi Ax\|_2^2 \leq (1+\epsilon)\|Ax\|_2^2. \tag{3}$$

Subspace embedding is an important concept and it is useful to understand it from different perspectives. Given that $\|Ax\|_2^2 = x^T A^T Ax$, preserving the column space of A means preserving the information in $A^T A$. The result can analogously be written as

$$\|\Pi Ax\|_2^2 \in \bigg[(1-\epsilon)\|Ax\|_2^2, (1+\epsilon)\|Ax\|_2^2\bigg].$$

Since $Ax = U\Sigma V^T x = Uz$ where $z = \Sigma V^T x \in \mathbb{R}^d$ and U is orthonormal, a change of basis gives:

$$\begin{aligned}
\|\Pi Uz\|_2^2 &\in \bigg[(1-\epsilon)\|Uz\|_2^2, (1+\epsilon)\|Uz\|_2^2\bigg] \\
&= \bigg[(1-\epsilon)\|z\|_2^2, (1+\epsilon)\|z\|_2^2\bigg] \\
&\Leftrightarrow \|(\Pi U)^T(\Pi U) - U^T U\|_2 \leq \epsilon \\
&\Leftrightarrow z^T\bigg((\Pi U)^T(\Pi U) - I_d\bigg)z \leq \epsilon.
\end{aligned}$$

The following Lemma defines subspace embedding in terms of singular value distortions.

Lemma 2 *Let $U \in \mathbb{R}^{n\times d}$ be a unitary matrix and Π be a subspace embedding for the column space of A. Let σ_k is the k-th singular value of A. Then (3) is equivalent to*

$$\sigma_k^2(\Pi U) \in [1-\epsilon, 1+\epsilon] \quad \forall k \in [1, d].$$

To understand Lemma 2, consider the Rayleigh quotient form of ΠU:[6]

$$\omega_k((\Pi U)^T(\Pi U)) = \frac{v_k^T(\Pi U)^T(\Pi U)v_k}{v_k^T v_k}$$

for some vector $v_k \neq 0$. As $\omega_k(A^TA) = \sigma_k^2(A)$,

$$\begin{aligned}\omega_k((\Pi U)^T(\Pi U)) &= \frac{v_k^T v_k - v_k^T\Big(I_d - (\Pi U)^T(\Pi U)\Big)v_k}{v_k^T v_k} \\ &= 1 - \omega_k\Big(I_d - (\Pi U)^T(\Pi U)\Big).\end{aligned}$$

This implies that $|1-\sigma_k^2(\Pi U)| = |\omega_k(I_d - (\Pi U)^T(\Pi U))| = \sigma_k(I_d - (\Pi U)^T(\Pi U))$. It follows that

$$\begin{aligned}|1-\sigma_k^2(\Pi U)| &= \Big|\sigma_k\Big(U^TU - (\Pi U)^T(\Pi U)\Big)\Big| \\ &\leq \sigma_{max}(U^TU - (\Pi U)^T(\Pi^T U)) \\ &= \|U^TU - (\Pi U)^T(\Pi U)\|_2 \leq \epsilon \\ \Leftrightarrow \sigma_k^2(\Pi U) &\in [1-\epsilon, 1+\epsilon] \quad \forall k \in [1,d].\end{aligned}$$

Hence the condition $\|(\Pi U)^T\Pi U - I_d\|_2 \leq \epsilon$ is equivalent to Π generating small singular value distortions. Nelson and Nguyen (2013a) relates this condition to similar results in random matrix theory.[7]

But where to find these embedding matrices? We can look for data dependent or data oblivious ones. We say that Π is a data oblivious embedding if it can be designed without knowledge of the input matrix. The idea of oblivious subspace-embedding first appeared in Sarlos (2006) in which it is suggested that Π can be drawn from a distribution with the JL properties.

Definition 2 *A random matrix $\Pi \in \mathbb{R}^{m\times n}$ drawn from a distribution F forms a JL transform with parameters ϵ, δ, d if there exists a function f such that for any $0 \leq \epsilon, \delta \leq 1$ and $m = \Omega(\log(\frac{d}{\epsilon^2}f(\delta)))$, $(1-\epsilon)\|x\|_2^2 \leq \|\Pi x\|_2^2 \leq (1+\epsilon)\|x\|_2^2$ holds with probability at least $1-\delta$ for all d-vector $x \subset \mathbb{R}^n$.*

A JL transform is often written JLT(ϵ, δ, d) for short. Embedding matrices Π that are JLT guarantee good approximation to matrix products in terms of Frobenius norm. This means that for such Πs with a suitable choice of m, it holds that for conformable matrices A, B having n rows:

$$P\Big(\|(\Pi A)^T(\Pi B) - A^TB\|_F \leq \epsilon\|A\|_F\|B\|_F\Big) \geq 1-\delta. \tag{4}$$

[6] For a Hermitian matrix M, the Rayleigh quotient is $\frac{c^TMc}{c^Tc}$ for a nonzero vector c. By Rayleigh-Ritz Theorem, $\min(\sigma(M)) \leq \frac{c^TMc}{c^Tc} \leq \max(\sigma(M))$ with equalities when c is the eigenvector corresponding to the smallest and largest eigenvalues of M, respectively. See, e.g. Hogben (2007, Section 8.2).

[7] Consider a $T \times N$ matrix of random variables with mean zero and unit variance with $c = \lim_{N,T\to\infty}\frac{N}{T}$. In random matrix theory, the largest and smallest eigenvalues of the sample covariance matrix have been shown to converge to $(1+\sqrt{c})^2, (1-\sqrt{c})^2$, respectively. See, e.g., Yin et al. (1988) and Bai and Yin (1993).

The Frobenius norm bound has many uses. If $A = B$, then $\|\Pi A\|_F^2 = (1 \pm \epsilon)\|A\|_F^2$ with high probability. The result also holds in the spectral norm, Sarlos (2006, Corollary 11).

3 Random Sampling, Random Projections, and the Countsketch

There are two classes of Πs with the JL property: random sampling which reduces the row dimension by randomly picking rows of A, and random projections which form linear combinations from the rows of A. A scheme known as a countsketch that is not a JL transform can also achieve subspace embedding efficiently. We will use a pen and pencil example with $m = 3$ and $n = 9$ to provide a better understanding of the three types of Πs. In this example, A has 9 rows given by $A_1, \dots, A_9$.

3.1 Random Sampling (RS)

Let D be a diagonal *rescaling* matrix with $\frac{1}{\sqrt{mp_i}}$ in the i-th diagonal and p_i is the probability that row i is chosen. Under random sampling,

$$\Pi \quad = \quad DS,$$

where $S_{jk} = 1$ if row k is selected in the j-th draw and zero otherwise so that the j-th row of the selection matrix S is the jth-row of an n dimensional indentity matrix. Examples of sampling schemes are:

RS1. Uniform sampling without replacement: $\Pi \in \mathbb{R}^{m\times n}$, $D \in \mathbb{R}^{m\times m}$, $p_i = \frac{1}{n}$ for all i. Each row is sampled at most once.

RS2. Uniform sampling with replacement: $\Pi \in \mathbb{R}^{m\times n}$, $D \in \mathbb{R}^{m\times m}$, $p_i = \frac{1}{n}$ for all i. Each row can be sampled more than once.

RS3. Bernoulli sampling uses an $n \times n$ matrix $\Pi = DS$, where $D = \sqrt{\frac{n}{m}} I_n$, S is initialized to be $0_{n\times n}$ and the j-th diagonal entry is updated by

$$S_{jj} = \begin{cases} 1 & \text{with probability } \frac{m}{n} \\ 0 & \text{with probability } 1 - \frac{m}{n} \end{cases}$$

Each row is sampled at most once, and m is the expected number of sampled rows.

RS4. Leverage score sampling: the sampling probabilities are taken from *importance sampling distribution*

$$p_i = \frac{\ell_i}{\sum_{i=1}^n \ell_i} = \frac{\ell_i}{d}, \tag{5}$$

where for A with $\textsc{svd}(A) = UDV^T$,

$$\ell_i = \|U_{(i)}\|_2^2 = \|e_i^T U\|_2^2,$$

is the leverage score for row i, $\sum_i \ell_i = \|U\|_F^2 = d$, and e_i is a standard basis vector.

Notably, the rows of the sketch produced by random sampling are the rows of the original matrix A. For example, If rows 9,5,1 are randomly chosen by uniform sampling, RS1 would give

$$\widetilde{A} = D \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} A = \frac{\sqrt{9}}{\sqrt{3}} \begin{pmatrix} A_9 \\ A_5 \\ A_1 \end{pmatrix}.$$

Ipsen and Wentworth (2014, Section 3.3) shows that sampling schemes RS1-RS3 are similar in terms of the condition number and rank deficiency in the matrices that are being subsampled. Unlike these three sampling schemes, leverage score sampling is not data oblivious and warrants further explanation.

As noted above, uniform sampling may not efficient. More precisely, uniform sampling does not work well when the data have high coherence, where coherence refers to the maximum of the row leverage scores ℓ_i defined above. Early work suggests to use sampling weights that depend on the Euclidean norm, $p_i = \frac{\|A_i\|_2^2}{\|A\|_F^2}$. See, e.g., Drineas, Kannan and Mahoney (2006) and Drineas and Mahoney (2005). Subsequent work finds that a better approach is to sample according to the leverage scores which measure the correlation between the left singular vectors of A with the standard basis, and thus indicates whether or not information is spread out. The idea of leverage-sampling, first used in Jolliffe (1972), is to sampling a row more frequently if it has more information.[8] Of course, ℓ_i is simply the i-th diagonal element of the hat matrix $A(A^TA)^{-1}A^T$, known to contain information about influential observations. In practice, computation of the leverage scores requires an eigen decomposition which is itself expensive. Drineas et al. (2012) and Cohen, Lee, Musco, Musco, Peng and Sidford (2015) suggest fast approximation of leverage scores.

3.2 Random Projections (RP)

Some examples of random projections are:

RP1. Sub-Gaussian random projections:[9]

i. Gaussian random projection, $\Pi \in \mathbb{R}^{m \times n}$ where $\Pi_{ij} = \frac{1}{\sqrt{m}} N(0,1)$.

[8] There are other variations of leverage score sampling. McWilliams et al. (2014) considers subsampling in linear regression models when the observations of the covariances may be corrupted by an additive noise. The influence of observation i is defined by $d_i = \frac{e_i^2 \ell_i}{(1-\ell_i)^2}$, where e_i is the OLS residual and ℓ_i is the leverage score. Unlike leverage scores, d_i takes into account the relation between the predictor variables and the y.

[9] A mean-zero vector $s \in \mathbb{R}^n$ is 1-sub-Gaussian if for any $u \in \mathbb{R}^n$ and for all $\epsilon > 0$, $P(s,u) \geq \epsilon \|u\|_2) \leq e^{-\epsilon^2/2}$.

ii A random matrix with entries of $\{+1,-1\}$, Sarlos (2006), Achiloptas (2003).

RP2. Randomized Orhthogonal Systems: $\Pi = \sqrt{\frac{n}{m}} PHD$ where D is an $n \times n$ is diagonal Rademacher matrix with entries of ± 1, P is a sparse matrix, and H is an orthonormal matrix.

RP3. Sparse Random Projections (SRP)

$$\Pi = DS$$

where $D \in \mathbb{R}^{m \times m}$ is a diagonal matrix of $\sqrt{\frac{s}{m}}$ and $S \in \mathbb{R}^{m \times n}$

$$S_{ij} = \begin{cases} -1 & \text{with probability } \frac{1}{2s} \\ 0 & \text{with probability } 1 - \frac{1}{s} \\ 1 & \text{with probability } \frac{1}{2s} \end{cases}$$

The rows of the sketch produced by random projections are linear combinations of the rows of the original matrix. For example, RP3 with $s = 2$ could give

$$\widetilde{A} = D \begin{pmatrix} 0 & 0 & 1 & 1 & 0 & -1 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & -1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \end{pmatrix} A = \frac{\sqrt{3}}{\sqrt{9}} \begin{pmatrix} A_3 + A_4 - A_6 \\ A_1 - A_3 - A_5 + A_9 \\ A_2 + A_7 - A_9 \end{pmatrix}$$

Early work on random projections such as Dasgupta et al. (2010) uses Πs that are dense, an example being RP1. Subsequent work favors sparser Πs, an example being RP3. Achiloptas (2003) initially considers $s = 3$. Li et al. (2006) suggests to increase s to $\sqrt{n}$. Given that uniform sampling is algorithmically inefficient when information is concentrated, the idea of randomized orthogonal systems is to first randomize the data by the matrix H to destroy uniformity, so that sampling in a data oblivious manner using P and rescaling by D remains appropriate. The randomization step is sometimes referred to as 'preconditioning'. Common choices of H are the Hadamard matrix as in the SRHT of Ailon and Chazelle (2009)[10] and the discrete Fourier transform as in FJLT of Woolfe et al. (2008).

3.3 Countsketch

While sparse Πs reduce computation cost, Kane and Nelson (2014, Theorem 2.3) shows that each column of Π must have $\Theta(d/\epsilon)$ non-zero entries to create a L_2 subspace embedding. This would seem to suggest that Π cannot be too sparse. However, Clarkson and Woodruff (2013) argues that if the non-zero entries of Π are carefully chosen, Π need not be a JLT and a very sparse subspace embedding is actually possible. Their insight is that Π need not preserve the norms of an arbitrary

[10] The Hadamard matrix is defined recursively by $H_n = \begin{pmatrix} H_{n/2} & H_{n/2} \\ H_{n/2} & -H_{n/2} \end{pmatrix}$, $\quad H_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$. A constraint is that n must be in powers of two.

subset of vectors in $\mathbb{R}^n$, but only those that sit in the d-dimensional subspace of $\mathbb{R}^n$. The sparse embedding matrix considered in Clarkson and Woodruff (2013) is the countsketch.[11]

A countsketch of sketching dimension m is a random linear map $\Pi = PD : \mathbb{R}^n \to \mathbb{R}^m$ where D is an $n \times n$ random diagonal matrix with entries chosen independently to be $+1$ or -1 with equal probability. Furthermore, $P \in \{0,1\}$ is an $m \times n$ binary matrix such that $P_{h(i),i} = 1$ and 0 otherwise, and $h : [n] \to [m]$ is a random map such that for each $i \in [n]$, $h(i) = m'$ for $m' \in [m]$ with probability $\frac{1}{m}$. As an example, a countsketch might be

$$\widetilde{A} = \begin{pmatrix} 0 & 0 & 1 & 0 & 1 & -1 & 0 & 0 & 1 \\ -1 & 0 & 0 & -1 & 0 & 0 & 0 & -1 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} A = \begin{pmatrix} A_3 + A_5 - A_6 + A_9 \\ A_1 - A_4 - A_8 \\ A_2 + A_7 \end{pmatrix}$$

Like random projections, the rows of a countsketch are also a linear combinations of the rows of A.

Though the countsketch is not a JLT, Nelson and Nguyen (2013b) and Meng and Mahoney (2013) show that the following Frobenius norm bound holds for the countsketch with appropriate choice of m:

$$P\Bigg(\|(\Pi U)^T(\Pi U) - I_d\|_2 > 3\epsilon \Bigg) \leq \delta \tag{6}$$

which implies that countsketch provides a $1+\varepsilon$ subspace embedding for the column space of A in spite of not being a JLT, see Woodruff (2014, Theorem 2.6).

The main appeal of the countsketch is that the run time needed to compute ΠA can be reduced to $O(\text{NNZ}(A))$, where NNZ(A) denotes the number of non-zero entries of A. The efficiency gain is due to extreme sparsity of a countsketch Π which only has one non-zero element per column. Still, the Π matrix can be costly to store when n is large. Fortunately, it is possible to compute the sketch without constructing Π.

The streaming version of the countsketch is a variant of the frequent-items algorithm where we recall that having to compute summaries such as the most frequent item in the data that stream by was instrumental to the development of sketching algorithms. The streaming algorithm proceeds by initializing $\widetilde{A}$ to an $m \times n$ matrix of zeros. Each row $A_{(i)}$ of A is updated as

$$\widetilde{A}_{h(i)} = \widetilde{A}_{h(i)} + g(i)A_{(i)}$$

where $h(i)$ sampled uniformly at random from $[1, 2, \ldots m]$ and g_i sampled from $\{+1, -1\}$ are independent. Computation can be done one row at a time.[12] The online Appendix provides the streaming implementation of the example above.

[11] The definition is taken from Dahiya et al. (2018). Given input j, a count-sketch matrix can also be characterized by a hash function $h(j)$ such that $\forall j, j', j \neq j' \to h(j) \neq h(j')$. Then $\Pi_{h(j),j} = \pm 1$ with equal probability 1/2.

[12] See Ghashami et al. (2016). Similar schemes have been proposed in Charika et al. (2002); Conmode and Muthukrishnan (2005).

3.4 Properties of the Πs

To assess the actual performance of the different Πs, we conduct a small Monte Carlo experiment with 1000 replications. For each replication b, we simulate an $n \times d$ matrix A and construct the seven JL embeddings considered above. For each embedding, we count the number of times that $|||\Pi(a_i - a_j)||_2^2$ is within $(1 \pm \epsilon)$ of $||a_i - a_j||_2^2$ for all $d(d+1)/2$ pairs of distinct (i,j). The success rate for the replication is the total count divided by $d((d+1)/2$. We also record $||\frac{\sigma(\Pi A)}{\sigma(A)} - 1||_2$ where $\sigma(\Pi A)$ is a vector of d singular values of ΠA. According to theory, the pairwise distortion of the vectors should be small if $m \geq C \log d/\epsilon^2$. We set $(n,d) = (20,000)$ and $\epsilon = 0.1$. Four values of $C = \{1,2,3,4,5,6,8,16\}$ are considered. We draw A from the (i) normal distribution, and (ii) the exponential distribution. In MATLAB, these are generated as X=RANDN(N,D) and X=EXPRND(D,[N D]). Results for the Pearson distribution using X=PEARSRND(0,1,1,5,N,D) are similar and not reported.

Table 1 reports the results averaged over 1000 simulations. With probability around 0.975, the pairwise distance between columns with 1000 rows is close to the pairwise distance between columns with 20000 rows. The average singular value distortion also levels off with about 1000 rows of data. Hence, information in n rows can be well summarized by a smaller matrix with m rows. The takeaway from Table 1 is that the performance of the different Πs are quite similar, making computation cost and analytical tractability two important factor in deciding which ones to use.

Choosing a Π is akin to choosing a kernel and many will work well, but there are analytical differences. Any $\Pi^T\Pi$ can be written as $I_n + R_{11} + R_{12}$ where R_{11} is a generic diagonal and R_{12} is a generic $n \times n$ matrix with zeros in each diagonal entry. Two features will be particularly useful.

$$\Pi^T\Pi = I_n + R_{11} \tag{7a}$$

$$\Pi\Pi^T = \frac{n}{m} I_m \tag{7b}$$

Property (7a) imposes that $\Pi^T\Pi$ is a diagonal matrix, allowing $R_{11} \neq 0$ but restricting $R_{12} = 0$. Each $\Pi\Pi^T$ can also be written as $\Pi\Pi^T = \frac{n}{m} I_m + R_{21} + R_{22}$ where R_{21} is a generic diagonal and R_{22} is a generic $m \times m$ matrix with zeros in each diagonal entry. Property (7b) requires that $\Pi\Pi^T$ is proportional to an identity matrix, and hence that $R_{21} = R_{22} = 0_{m\times m}$.

For the Πs previously considered, we summarize their properties as follows:

	(7a)	(7b)
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
RP4 (Countsketch)	no	no

Property (7a) holds for all three random sampling methods but of all the random projection methods considered, the property only holds for SRHT. This is because SRHT effectively performs uniform sampling of the randomized data. For property (7b), it is easy to see that $R_{21} = 0$ and $R_{22} = 0_{m\times m}$ when uniform sampling is done without replacement and $\Pi\Pi^T = \frac{n}{m} I_m$. By implication, the condition also holds for SRHT if sampling is done without replacement since H and D are orthonormal matrices. But uniform sampling is computationally much cheaper than SRHT and has the distinct advantange over the SRHT that the rows of the sketch are those of the original matrix and hence interpretable. For this reason, we will subsequently focus on uniform sampling and use its special structure to obtain precise statistical results. Though neither condition holds for the countsketch, its extreme sparse structure permits useful statements to be made. This is useful because of the computational advantages associted with the countsketch.

4 Algorithmic Results for the Linear Regression Model

The linear regression model with K regressors is $y = X\beta + e$. The least squares estimator minimizes $\|y - Xb\|_2$ with respect to b and is defined by

$$\hat{\beta} = (X^T X)^{-1} X^T y = V\Sigma^{-1} U^T y.$$

We are familiar with the statistical properties of $\hat{\beta}$ under assumptions about e and X. But even without specifying a probabilistic structure, $\|y - Xb\|_2$ with $n > K$ is an over-determined system of equations and can be solved by algebraically. The SVD solution gives $\beta^* = X^- y$ where $\text{SVD}(X) = U\Sigma V^T$, the pseudoinverse is $X^- = V\Sigma^{-1}U^T$. The 'Choleski' solution starts with the normal equations $X^T X\beta = X^T y$ and factorizes $X^T X$. The algebraic properties of these solutions are well studied in the matrix computations literature when all data are used.

Given an embedding matrix Π and sketched data $(\Pi y, \Pi X)$, minimizing $= \|\Pi(y - Xb)\|_2^2$ with respect to b gives the sketched estimator

$$\widetilde{\beta} = \Big((\Pi X)^T \Pi X \Big)^{-1} (\Pi X)^T \Pi y.$$

Let $\widehat{\text{SSR}} = \|y - X\hat{\beta}\|_2^2$ be the full sample sum of squared residuals. For an embedding matrix $\Pi \in \mathbb{R}^{m\times n}$, let $\widetilde{\text{SSR}} = \|\widetilde{y} - \widetilde{X}\widetilde{\beta}\|_2^2$ be the sum of squared residuals from using the sketched data. Assume that the following two conditions hold with probability $1-\delta$ for $0 < \epsilon < 1$:

$$|1 - \sigma_k^2(\Pi U)| \;\leq\; \frac{1}{\sqrt{2}} \quad \forall k = 1, \ldots, K; \tag{8a}$$

$$\|(\Pi U)^T \Pi(y - X\hat{\beta})\|_2^2 \;\leq\; \epsilon\, \widehat{\text{SSR}}^2/2. \tag{8b}$$

Condition (8a) is is equivalent to $\|(\Pi U)^T(\Pi U) - U^T U\|_2 \leq \frac{1}{\sqrt{2}}$ as discussed above. Since $\sigma_i(U) = 1$ for all $k \in [1, K]$, the condition requires the smallest singular value, $\sigma_K(\Pi U)$, to be positive so that ΠX has the same rank as X. A property of the least squares estimator is for the least squares residuals to be orthogonal to X, ie. $U^T(y - X\hat{\beta}) = 0$. Condition (8b) requires near orthogonality when both quantities are multiplied by Π. The two algorithmic features of sketched least squares estimation are summarized below.

Lemma 3 *Let the sketched data be* $(\Pi y, \Pi X) = (\widetilde{y}, \widetilde{X})$ *where* $\Pi \in \mathbb{R}^{m\times n}$ *is a subspace embedding matrix. Let* $\sigma_{min}(X)$ *be the smallest singular value of* X*. Suppose that conditions (8a) and (8b) hold. Then with probability at least* $(1-\delta)$ *and for suitable choice of* m*,*

(i). $\widetilde{\text{SSR}} \leq (1+\epsilon)\widehat{\text{SSR}}$*;*

(ii). $\|\widetilde{\beta} - \hat{\beta}\|_2 \leq \epsilon \widehat{\text{SSR}}/\sigma_{min}$.

Sarlos (2006) provides the proof for random projections, while Drineas, Mahoney and Muthukrishnan (2006) analyzes the case of leverage score sampling. The desired m depends on the result and the sampling scheme.

Part (i) is based on subspace embedding arguments. By optimality of $\widetilde{\beta}$ and JL Lemma,

$$\begin{aligned}
\widetilde{\text{SSR}} &= \|\Pi(y - X\widetilde{\beta})|_2 \\
&\leq \|\Pi(y - X\hat{\beta})\|_2 \quad \text{by optimality of } \widetilde{\beta} \\
&\leq (1+\epsilon)\|y - X\hat{\beta}\|_2 \quad \text{by subspace embedding} \\
&= (1+\epsilon)\widehat{\text{SSR}}.
\end{aligned}$$

Part (ii) shows that the sketching error is data dependent. Consider a reparameterization of $X\hat{\beta} = U\Sigma V^T\hat{\beta} = U\hat{\theta}$ and $X\widetilde{\beta} = U\Sigma V^T\widetilde{\beta} = U\widetilde{\theta}$. As shown in the online Appendix, $\|\widetilde{\theta} - \hat{\theta}\|_2 \leq \sqrt{\epsilon}\, \widehat{\text{SSR}}$. Taking norms on both sides of $X(\widetilde{\beta} - \hat{\beta}) = U(\widetilde{\theta} - \hat{\theta})$ and since U is orthonormal,

$$\|\widetilde{\beta} - \hat{\beta}\|_2 \;\leq\; \frac{\|U\widetilde{\theta}\|_2}{\sigma_{\min}}.$$

Notably, difference between $\hat{\beta}$ and $\widetilde{\beta}$ depends on the minimum singular value of X. Recall that for consistent estimation, we also require that the minimum eigenvalue to diverge.

The non-asymptotic worse case error bounds in Lemma 3 are valid for any subspace embedding matrix Π, though more precise statements are available for certain Πs. For leverage score sampling, see Drineas, Mahoney and Muthukrishnan (2006), for uniform sampling and SRHT, see Drineas et al. (2011); and for the countsketch, Woodruff (2014, Theorem 2.16), Meng and Mahoney (2013, Theorem 1), Nelson and Nguyen (2013a). These algorithmic results are derived without reference to the probabilistic structure of the data. Hence the results do not convey information such as bias and sampling uncertainty. An interesting question is whether optimality from an algorithmic perspective implies optimality from a statistical perspective. Using Taylor series expansion, Ma et al. (2014) shows that leverage-based sampling does not dominate uniform sampling in terms of bias and variance, while Raskutti and Mahoney (2016) finds that prediction efficiency requires m to be quite large. Pilanci and Wainwright (2015) shows that the solutions from sketched least squares regressions have larger variance than the oracle solution that uses the full sample. Pilanci and Wainwright (2015) provides a result that relates m to the rank of the matrix. Wang, Gittens and Mahoney (2018) studies four sketching methods in the context of ridge regressions that nests least squares as a special case. It is reported that sketching schemes with near optimal algorithmic properties may have features that not statistically optimal. Chi and Ipsen (2018) decomposes the variance of $\widetilde{\beta}$ into a model induced component and an algorithm induced component.

5 Statistical Properties of $\widetilde{\beta}$

We consider the linear regression model with K regressors:

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

where y is the dependent variable, X is the $n \times K$ matrix of regressors, β is the $K \times 1$ vector of regression coefficients whose true value is β_0. It should be noted that K is the number of predictors which is generally larger than d, which is the number of covariates available since the predictors may include transformation of the d covariates. In the Mincer example, we have data for EDU, EXP collected into A with $d = 2$ columns. But the regressor matrix X is $n \times K$ where $K = 4$ in regression (1a) and $K = 13$ in regression (1b).

The full sample estimator using data (y, X) is $\hat{\beta} = (X^TX)^{-1}X^Ty$. For a given Π, the estimator using sketched data $(\widetilde{y}, \widetilde{X}) = (\Pi y, \Pi X)$ is

$$\widetilde{\beta} = (\widetilde{X}^T\widetilde{X})^{-1}\widetilde{X}^T\widetilde{y}.$$

Assumption OLS:

(i) the regressors X are non-random, has SVD $X = U\Sigma V^T$, and $X^T X$ is non-singular;

(ii) $\mathbb{E}[e_i] = 0$ and $\mathbb{E}[ee^T] = \Omega_e$ is a diagonal positive definite matrix.

Assumption PI:

(i) Π is independent of e;

(ii) for given singular value distortion parameter $\varepsilon_\sigma \in (0,1)$, there exists failure parameter $\delta_\sigma \in (0,1)$ such that for all $k = 1, \dots K$, $P\Big(|1 - \sigma_k^2(\Pi U)| \le \varepsilon_\sigma\Big) \ge 1 - \delta_\sigma$.

(iii) $\Pi^T\Pi$ is an $n \times n$ diagonal matrix and $\Pi\Pi^T = \frac{n}{m} I_m$.

Assumption OLS is standard in regression analyses. The errors are allowed to be possibly heteroskedastic but not cross-correlated. Under Assumption OLS, $\hat{\beta}$ is unbiased, i.e. $\mathbb{E}[\hat{\beta}] = \beta_0$ with a sandwich variance

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

Assumption PI.(i) is needed for $\widetilde{\beta}$ to be unbiased. Assumption PI.(ii) restricts attention to Π matrices that have subspace embedding property. As previously noted, the condition is equivalent to $\|I_d - (\Pi U)^T(\Pi U)\|_2 \le \varepsilon_\sigma$ holding with probability $1 - \delta_\sigma$. We use PI2 to refer Assumptions PI (i) and (ii) holding jointly. Results under PI2 are not specific to any Π.

Assumption PI.(iii) simplifies the expression for $\mathbb{V}(\widetilde{\beta}|\Pi)$ and we will use PI3 to denote Assumptions PI (i)-(iii) holding jointly. PI3 effectively narrows the analysis to uniform sampling and SRHT without replacement. We will focus on uniform sampling without replacement for a number of reasons. It is simple to implement, and unlike the SRHT, the rows have meaningful intepretation. In a regression context, uniform sampling has an added advantage that each time we add or drop a variable in the X matrix, most Π would likely require $(\widetilde{y}, \widetilde{X})$ to be reconstructed. This can be cumbersome when variable selection is part of the empirical exercise. Uniform sampling without replacement is an exception since the columns are unaffected once the rows are randomly chosen.

For regressions, we need to know not only the error in approximating $X^T X$, but also the error in approximating $(X^T X)^{-1}$. This is made precise in the next Lemma.

Lemma 4 *Suppose that PI2 is satisfied. For given non-random matrix $X \in \mathbb{R}^{n \times K}$ of full rank with* SVD$(X) = U\Sigma V^T$, *consider any non-zero $K \times 1$ vector c. It holds that*

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

The Lemma follows from the fact that

$$\begin{aligned}(X^TX)^{-1} &= V\Sigma^{-2}V^T \equiv PP^T \\ \Big((\Pi X)^T(\Pi X)\Big)^{-1} &= PQP^T\end{aligned}$$

where $P = V\Sigma^{-1}$ and $Q^{-1} = (\Pi U)^T(\Pi U)$. By the property of Rayleigh quotient, the smallest eigenvalue of $(U^T\Pi^T\Pi U)$ is bounded below by $(1-\varepsilon_\sigma)$. Hence

$$\left|\frac{c^T(PQP^T - PP^T)c}{c^TPP^Tc}\right| = \left|\frac{c^TP(I_d - Q^{-1})QP^Tc}{c^TPP^Tc}\right| \le \|Q\|_2\|I_d - Q^{-1}\|_2 \le \frac{\varepsilon_\sigma}{(1-\varepsilon_\sigma)}.$$

The approximation error $(X^TX)^{-1}$ is thus larger than that for X^TX, which equals ε_σ.

Under Assumptions OLS and PI2, $\widetilde{\beta}$ is unbiased and has sandwich variance

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

since $\Pi\Pi^T = \frac{n}{m}I_m$. The variance of $\widetilde{\beta}$ is inflated over that of $\hat{\beta}$ through the sketching error on the 'bread' $(\widetilde{X}^T\widetilde{X})^{-1}$, as well as on the 'meat' because $X^T\Omega_eX$ is now approximated by $\widetilde{X}^T\Pi\Omega_e\Pi^T\widetilde{X}$. If $e \sim (0, \sigma_e^2 I_n)$ is homoskedastic, then $\widetilde{\beta}$ has variance

$$\mathbb{V}(\widetilde{\beta}|\Pi) = \sigma_e^2\frac{n}{m}(\widetilde{X}^T\widetilde{X})^{-1}. \tag{9}$$

Though $\hat{\beta}$ is best linear unbiased under homoskedasticity, $\widetilde{\beta}$ may not be best in the class of linear estimators using sketched data.

5.1 Efficiency of $\widetilde{\beta}$ Under Uniform Sampling

Suppose we are interested in predicting y at some x^0. According to the model, $\mathbb{E}[y|x = x^0] = \beta^Tx^0$. Feasible predictions are obtained upon replacing β with $\hat{\beta}$ and $\widetilde{\beta}$. Since both estimators are unbiased, their respective variance is also the mean-squared prediction error.

Theorem 1 *Suppose that $e_i \sim (0, \sigma_e^2)$ and Assumptions OLS and PI3 hold. Let $\textsc{mse}(x_0^T\hat{\beta})$ and $\textsc{mse}(x_0^T\widetilde{\beta}|\Pi)$ be the mean-squared prediction error of y at x_0 using $\hat{\beta}$ and $\widetilde{\beta}$ conditional on Π, respectively. Then with probability at least $1-\delta_\sigma$, it holds that*

$$\frac{\textsc{mse}(x_0^T\widetilde{\beta}|\Pi)}{\textsc{mse}(x_0^T\hat{\beta})} \le \underbrace{\frac{n}{m}}_{\textit{sample size}}\underbrace{\left(\frac{1}{1-\varepsilon_\sigma}\right)}_{\textit{sketching error}}.$$

The prediction error has has two components: a sample size effect given by $\frac{n}{m} > 1$, and a sketching effect given by $\frac{1}{1-\varepsilon_\sigma} > 1$. The result arises because under homoskedasticity,

$$x_0^T(\widetilde{X}^T\widetilde{X})^{-1}x_0 - x_0^T(X^TX)^{-1}x_0 = \frac{n}{m}x_0^T\left[\left(X^T\Pi^T\Pi X\right)^{-1} - (X^TX)^{-1}\right]x_0 + \frac{n-m}{m}x_0^T(X^TX)^{-1}x_0.$$

It follows that

$$\left|\frac{x_0^T\mathbb{V}(\widetilde{\beta}|\Pi)x_0 - x_0^T\mathbb{V}(\hat{\beta})x_0}{x_0^T\mathbb{V}(\hat{\beta})x_0}\right| = \left|\frac{n}{m}\frac{x_0^T[\big((\Pi X)^T\Pi X\big)^{-1} - (X^TX)^{-1}]x_0}{x_0^T(X^TX)^{-1}x_0} + \frac{n-m}{m}\right|$$
$$\leq \frac{n}{m}\frac{\varepsilon_\sigma}{1-\varepsilon_\sigma} + \frac{n-m}{m}.$$

We will subsequently be interested in the effect of sketching for testing linear restrictions as given by a $K \times 1$ vector c. The estimated linear combination $c^T\widetilde{\beta}$ has variance $\mathbb{V}(c^T\widetilde{\beta}|\Pi) = c^T\mathbb{V}(\widetilde{\beta}|\Pi)c$. When c is a vector of zeros except in the k-th entry, $\text{var}(c^T\widetilde{\beta}|\Pi)$ is the variance of $\widetilde{\beta}_k$. When c is a vector of ones, $\text{var}(c^T\widetilde{\beta}|\Pi)$ is the variance of the sum of estimates. A straightforward generalization of Theorem 1 leads to the following.

Corollary 1 *Let c be a known $K \times 1$ vector. Under the Assumptions of Theorem 1, it holds with probability $1-\delta_\sigma$ that*

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

The relative error is thus primarily determined by the relative sample size. As m is expected to be much smaller than n, the efficiency loss is undeniable.

A lower bound in estimation error can be obtained for embedding matrices $\Phi \in \mathbb{R}^{m\times n}$ satisfying $\|\Phi U\|_2^2 \leq 1+\varepsilon_\sigma$. For a sketch size of m rows, define a class of OLS estimators as follows:

$$\mathcal{B}(m,n,\varepsilon_\sigma) := \Big\{\breve{\beta} := ((X\Phi)^T\Phi X)^{-1}(X\Phi)^T\Phi y\ \Big\}.$$

For such $\breve{\beta}$, let $\mathbb{V}(\breve{\beta}|\Phi)$ denote its mse for given Φ. Assuming that $e_i \sim (0,\sigma_e^2)$,

$$\begin{aligned}\frac{c^T\mathbb{V}(\breve{\beta}|\Phi)c}{c^T\mathbb{V}(\hat{\beta})c} &= \frac{m^{-1}\sigma_e^2c^T((\Phi X)^T\Phi X)^{-1}c}{n^{-1}\sigma_e^2c^T(X^TX)^{-1}c} = \frac{n}{m}\frac{c^T((\Phi X)^T\Phi X)^{-1}c}{c^T(X^TX)^{-1}c}\\ &= \frac{n}{m}\frac{c^T(V\Sigma U^T\Phi^T\Phi U\Sigma V^T)^{-1}c}{c^T(V\Sigma^2V^T)^{-1}c} = \frac{n}{m}\frac{c^TV\Sigma^{-1}(U^T\Phi^T\Phi U)^{-1}\Sigma^{-1}V^Tc}{c^TV\Sigma^{-2}V^Tc}\\ &\geq \frac{n}{m}\sigma_{\min}[(U^T\Phi^T\Phi U)^{-1}].\end{aligned}$$

But by the definition of spectral norm, $\|\Phi U\|_2^2 = \sigma^2_{\max}(\Phi U)$ for any Φ. Thus the subspace embedding condition $\|\Phi U\|_2^2 \leq 1+\varepsilon_\sigma$ implies $\sigma^2_{\max}(\Phi U) = \sigma_{\max}((\Phi U)^T\Phi U) \leq 1+\varepsilon_\sigma$, and hence

$$\sigma_{\min}\bigg((U^T\Phi^T\Phi U)^{-1}\bigg) \geq \frac{1}{1+\varepsilon_\sigma}.$$

This leads to the following lower bound for $\breve{\beta}$:

$$\frac{c^T\mathbb{V}(\breve{\beta}|\Phi)c}{c^T\mathbb{V}(\hat{\beta})c} \geq \frac{n}{m}\bigg(\frac{1}{1+\varepsilon_\sigma}\bigg).$$

Combining the upper and lower bounds leads to the following:

Theorem 2 *Under OLS and PI3, the estimator $\widetilde{\beta}$ with $e_i \sim (0, \sigma_e^2)$ has mean-squared error relative to the full sample estimator $\hat{\beta}$ bounded by*

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

These are the upper and lower bounds for uniform sampling when implemented by sampling without replacement.

It is also of interest to know how heteroskedasticity affects the sketching error. Let $\Omega_{e,ii}$ denote the ith diagonal element of Ω_e. Under OLS and PI3, it holds with probability at least $1-\delta_\sigma$ that

$$\frac{\text{MSE}(x_0^T\widetilde{\beta}|\Pi)}{\text{MSE}(x_0^T\hat{\beta})} \leq \left(\frac{\max_i \Omega_{e,ii}}{\min_i \Omega_{e,ii}}\right)\left(\frac{n}{m}\right)\frac{(1+\varepsilon_\sigma)}{(1-\varepsilon_\sigma)^2}.$$

Hence heteroskedasticity independently interacts with the structure of Π to inflate the mean-squared prediction error. The upper and lower bound for $\mathbb{V}(c\widetilde{\beta}|\Phi)$ are larger than under homoskedasticity by a magnitude that depends on the extent of dispersion in $\Omega_{e,ii}$. A formal result is given in the online appendix.

5.2 Efficiency of $\widetilde{\beta}$ under Countsketch

Condition PI.iiii puts restrictions on $\Pi^T\Pi$ and holds for uniform sampling. But the condition does not hold for the countsketch. In its place, we assume the following to obtain a different embedding result for the countsketch:

Assumption CS: For given $\varepsilon_\Pi > 0$ and for all $U \in \mathbb{R}^{n\times K}$ satisfying $U^TU = I_K$, there exists an $n \times n$ matrix $A(\Omega_e, m, n)$, which may depend on (Ω_e, m, n), and a constant $\delta_\Pi \in (0,1)$ such that

$$P\Big(\|U^T\Pi^T\Pi\Omega_e\Pi^T\Pi U - U^TA(\Omega_e,m,n)U\|_2 \leq \frac{n}{m}\varepsilon_\Pi\Big) \geq 1-\delta_\Pi.$$

The conditions for Assumption CS are verified in online Appendix B. The Assumption is enough to provide a subspace embedding result for $\Pi^T\Pi\Omega_e\Pi^T\Pi$ because as shown in the online Appendix, the following holds for Countsketch,

$$\begin{aligned}
&\Big\|U^T\Pi^T\Pi\Omega_e\Pi^T\Pi U - \frac{n}{m}U^T\Omega_eU\Big\|_2 \\
&\leq \big\|U^T\Pi^T\Pi\Omega_e\Pi^T\Pi U - U^TA(\Omega_e,m,n)U\big\|_2 + \Big\|U^TA(\Omega_e,m,n)U - \frac{n}{m}U^T\Omega_eU\Big\|_2 \\
&\leq \frac{n}{m}\Big[\varepsilon_\Pi + \Big\|\frac{m}{n}A(\Omega_e,m,n) - \Omega_e\Big\|_2\Big]
\end{aligned} \tag{10}$$

where

$$A(\Omega_e,m,n) = \Omega_e + \frac{1}{m}\Big(\text{tr}(\Omega_e)I_n - \Omega_e\Big).$$

Hence under OLS, PI2, and CS it holds with probability at least $1-\delta_\Pi-\delta_\sigma$ that

$$\frac{\text{MSE}(x_0^T\widetilde{\beta}|\Pi)}{\text{MSE}(x_0^T\hat{\beta})}\leq\left(\frac{\max_i\Omega_{e,ii}}{\min_i\Omega_{e,ii}}\right)\left(\frac{n}{m}\right)\left(\frac{1}{1-\varepsilon_\sigma}\right)\frac{\left[1+\varepsilon_\Pi+\|\frac{m}{n}A(\Omega_e,m,n)-\Omega_e\|_2\right]}{(1-\varepsilon_\sigma)}.$$

The prediction error of the countsketch depends on the quantity $A(\Omega_e,m,n)$. But if $\Omega_e=\sigma_e^2I_n$, $\left\|\frac{m}{n}A(\Omega_e,m,n)-\Omega_e\right\|_2=\sigma_e^2(\frac{m-1}{n})$, which will be negligible if $m/n=o(1)$. Hence to a first approximation, Theorem 1 also holds under the countsketch. This result is of interest since the countsketch is computationally inexpensive.

5.3 Hypothesis Testing

The statistical implications of sketching in a regression setting have largely focused on properties of the point estimates. The implications for inference are largely unknown. We analyze the problem from view point of hypothesis testing.

Consider the goal of testing q linear restrictions formulated as $H_0:R\beta=r$ where R is a $q\times K$ matrix of restrictions with no unknowns. In this subsection, we further assume that $e_i\sim N(0,\sigma_e^2)$. Under normality, the F test is exact and has the property that at the true value of $\beta=\beta^0$,

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

Under the null hypothesis that β_0 is the true value of β, F_n has a Fisher distribution with q and $n-d$ degrees of freedom. The power of a test given a data of size n against a fixed alternative $\beta_1\neq\beta_0$ depends on $\mathbb{V}(\hat{\beta})$ only through non-centrality parameter ϕ_n,[13] defined in Wallace (1972) as

$$\phi_n=\frac{(R\beta_0-r)^T\mathbb{V}(R\hat{\beta})^{-1}(R\beta-r)}{2}.$$

The non-centrality parameter is increasing in $|R\beta_0-r|$ and the sample size through the variance, but decreasing in σ^2. In the case of one restriction ($q=1$),

$$\phi_n=\frac{(R\beta_0-r)^2}{\mathbb{V}(R\hat{\beta})}>\frac{(R\beta_0-r)^2}{\mathbb{V}(R\widetilde{\beta})}=\phi_m.$$

This leads to the relative non-centrality

$$\frac{\phi_n}{\phi_m}=\frac{\mathbb{V}(R\widetilde{\beta})}{\mathbb{V}(R\hat{\beta})}\leq\frac{n}{m}\frac{1}{(1-\varepsilon_\phi)}.$$

which also has a sample size effect and an effect due to sketching error. The effective size of the subsample from the viewpoint of power can be thought of as $m(1-\varepsilon_\phi)$.

[13] The definition of non-centrality is not universal, sometimes the factor of two is omitted. See, for example, Cramer (1987) and Rudd (2000).

A loss in power is to be expected when $\widetilde{\beta}$ is used. But by how much? Insights can be performed from some back of the envelope calculations. Recall that if U and V are independent χ^2 variables with ν_1 and ν_2 degrees of freedom, V is central and U has non-centrality parameter ϕ,

$$\mathbb{E}[F] = \mathbb{E}\left[\frac{(U/\nu_1)}{(V/\nu_2)}\right] = \frac{\nu_2(\nu_1 + \phi_n)}{\nu_1(\nu_2 - 2)}.$$

In the full sample case, $\nu_1 = q$ and $\nu_2 = n - d$ and hence

$$\mathbb{E}[F_n] \quad \approx \quad \frac{(n-d)(q+\phi_n)}{q(n-d-2)}.$$

For the subsampled estimator, $\nu_1 = q$ and $\nu_2 = m - d$, giving

$$\mathbb{E}[F_m] \quad \approx \quad \frac{(m-d)(q+\phi_m)}{q(m-d-2)}.$$

While q and ϕ_m affect absolute power, the relative power of testing a hypothesis against a fixed alternative is mainly driven by the relative sample size, $\frac{m}{n}$.

The top panel of Table 2 presents some power calculations with $\phi_n = n\Delta^2$, $\phi_m = m(1-\varepsilon_\phi)\Delta^2$ for $\Delta = (0, 5, 10)$. The result that stands out is that the power loss from using $\widetilde{\beta}$ to test hypothesis can be made negligible. The reason is that in a big data setting, we have the luxury of allowing m to be as large as we wish. This result does not depend on q.

It is also of interest to consider the local power of the F test. Using the Pitman formulation, the full sample of size n allows us to test the local alternative by putting $\Delta_n = \frac{\Delta}{\sqrt{n}}$. Similarly, a subsample of size m allows us to test the local alternative $\Delta_m = \frac{\Delta}{\sqrt{m}}$. The power difference is primarily due to sample size effect. As seen from the lower panel of Table 2, The effect of sketching is small.

6 Econometrically Motivated Refinements

As more and more data are being collected, sketching continues to be an active area of research. For any sketching scheme, a solution of higher accuracy can be obtained by iteration. The idea is to approximate the deviation from an initial estimate $\Delta = \hat{\beta} - \widetilde{\beta}^{(1)}$ by solving, for example, $\hat{\Delta}^{(1)} = \text{argmin}_\Delta \|y - (X(\widetilde{\beta}^{(1)} + \Delta)\|_2^2$ and update $\widetilde{\beta}^{(2)} = \widetilde{\beta}^{(1)} + \hat{\Delta}^{(1)}$. Pilanci and Wainwright (2016) starts with the observation that the least squares objective function $\|y - X\beta\|_2^2 = \|y\|_2^2 + \|X\beta\|_2^2 - 2y^T X\beta$ and suggests to sketch the quadratic term $\|X\beta\|_2^2$ but not the linear term $y^T X\beta$. The result is a Hessian sketch of β, defined as $((\Pi X)^T(\Pi X))^{-1}X^T y$. Wang et al. (2017) suggests that this can be seen as a type of Newton updating with the true Hessian replaced by the sketched Hessian, and the iterative Hessian sketch is also a form of iterative random projection.

In the rest of this section, we consider statistically motivated ways to improve upon $\widetilde{\beta}$. Subsection 1 considers pooling estimates from multiple sketches. Subsection 2 suggests an m that is motivated by hypothesis testing.

6.1 Combining Sketches

The main result of the previous section is that the least squares estimates using sketched data has two errors, one due to a smaller sample size, and one due to sketching. The efficiency loss is hardly surprising, and there are different ways to improve upon it. Dhillon et al. (2013) proposes a two-stage algorithm that uses m rows of (y, X) to obtain an initial estimate of Σ_{XX} and Σ_{Xy}. In the second stage, the remaining rows are used to estimate the bias of the first stage estimator. The final estimate is a weighted average of the two estimates. An error bound of $O(\frac{\sqrt{K}}{\sqrt{n}})$ is obtained. This bound is independent of the amount of subsampling provided $m > O(\sqrt{K/n})$. Chen et al. (2016) suggests to choose sample indices from an importance sampling distribution that is proportional to a sampling score computed from the data. They show that the optimal p_i depends on whether minimizing mean-squared error of $\widetilde{\beta}$ or of $X\widetilde{\beta}$ is the goal, though $\mathbb{E}[e_i^2]$ plays a role in both objectives.

The sample size effect is to be expected, and is the cost we pay for not being able to use the full data. But if it is computationally simple to create a sample of size m, the possibility arises that we can better exploit information in the data without hitting the computation bottleneck by generating many subsamples and subsequently pool estimates constructed from the subsamples. Breiman (1999) explored an idea known as *pasting bites* that, when applied to regressions, would repeatedly form training samples of size m by random sampling from the original data, then make prediction by fitting the model to the training data. The final prediction is the average of the predictions. Similar ideas are considered in Chawla et al. (2004) and Christmann et al. (2007). Also related is distributed computing which takes advantage of many nodes in the computing cluster. Typically, each machine only sees a subsample of the full data and the parameter of interest is updated. Heince et al. (2016) studies a situation when the data are distrubuted across workers according features of X rather than the sample and show that their DUAL LOCAL algorithm has bounded approximation error that depends only weakly on the number of workers.

Consider $\widetilde{\beta}_1, \ldots, \widetilde{\beta}_J$ computed from J subsamples each of size K. As mentioned above, uniform sampling with too few rows is potentially vulnerable to omitting influential observations. Computing multiple sketches also provides the user with an opportunity to check the rank across sketches. Let $\widetilde{t}_j = \frac{\widetilde{\beta}_j - \beta_0}{se(\widetilde{\beta}_j)}$ be the t test from sketch j. For a given m, we consider sampling without replacement and J is then at most n/m. Define the average quantities

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

$$\bar{t}_2 = J^{-1}\sum_{j=1}^{J}\hat{t}_j, \qquad \text{se}(\bar{t}_2) = \sqrt{\frac{1}{J-1}\sum_{j=1}^{J}\left(\widetilde{t}_j - \bar{t}_2\right)^2}.$$

Strictly speaking, pooling requires that subsamples are non-overlapping and observations are independent across different subsamples. Assuming independence across j, the pooled estimator $\bar{\beta}$ has $\text{var}(\bar{\beta}) = \frac{1}{J^2}\sum_{j=1}^{J}\text{var}(\widetilde{\beta}_j)$. Thus, $\text{se}(\bar{\beta}) \approx \sqrt{\frac{1}{J^2}\sum_{j=1}^{J}\left[\text{se}(\widetilde{\beta}_j)\right]^2} \geq \frac{1}{\sqrt{J}}\left[\frac{1}{J}\sum_{j=1}^{J}\text{se}(\widetilde{\beta}_j)\right]$ because of Jansen's inequality. Our estimator for the standard error $\bar{\beta}$ uses $J-1$ term in the denominator to allow for a correction when J is relatively small.

Consider two pooled t statistics:

$$\bar{T}_1 = \frac{\bar{\beta}-\beta_0}{se(\bar{\beta})} \tag{11a}$$

$$\bar{T}_2 = \sqrt{J}\frac{\bar{t}_2}{se(\bar{t}_2)}. \tag{11b}$$

Critical values from the standard normal distribution can be used for $\bar{T}_1$. For example, for the 5%-level test, we reject H_0 if $|\bar{T}_1| > 1.96$. For $\bar{T}_2$, we recommend using critical values from the t distribution with $J-1$ degrees of freedom. For example, for the 5%-level test, we reject H_0 if $|\bar{T}_2| > 2.776$ for $J = 5$.

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 $\Pi_j\Pi_k^T = O_m$ where O_m is an $m \times m$ matrix of zeros.

(iii) for given singular value distortion parameter $\varepsilon_\sigma \in (0,1)$, there exists failure parameter $\delta_\sigma \in (0,1)$ such that for all $k = 1,\ldots K$ and for all $j = 1,\ldots,J$, $P\Big(|1-\sigma_k^2(\Pi_jU)| \leq \varepsilon_\sigma\Big) \geq 1-\delta_\sigma$.

Condition (ii) is crucial; it is satisfied, for example, if $\Pi_1,\ldots,\Pi_J$ are non-overlapping subsamples and each of them is sampled uniformly without replacement. Assumptions OLS and PI-Avg (i) and (ii) along with the homoskedastic error assumption ensure that

$$c^T\mathbb{V}(\bar{\beta}|\Pi_1,\ldots,\Pi_J)c = \sigma_e^2\frac{n}{mJ^2}\sum_{j=1}^{J}c^T\left(X^T\Pi_j^T\Pi_jX\right)^{-1}c.$$

Condition (iii) is equivalent to the statement that $1-\varepsilon_\sigma \leq \sigma_k(U^T\Pi_j^T\Pi_jU) \leq 1+\varepsilon_\sigma \quad \forall\ k = 1,\ldots,K,\ \ \forall\, j = 1,\ldots,J.$.

Theorem 3 *Consider J independent sketches obtained by uniform sampling. Suppose that $e_i \sim (0,\sigma_e^2)$, Assumptions OLS and PI-Avg hold. Then with probability at leat $1-\delta_\sigma$, the mean squared error of $c^T\bar{\beta}$ conditional on $(\Pi_1,\ldots,\Pi_J)$ satisifies*

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

The significance of the Theorem is that by choice of m and J, the pooled estimator can be almost as efficient as the full sample estimator. If we set $J = 1$, the theorem reduces to Theorem 1.

We use a small Monte Carlo experiment to assess the effectiveness of combining statistics computed from different sketches. The data are generated as $y = X\beta + e$ where e is normally distributed, and X is drawn from a non-normal distribution. In MATLAB, X=PEARSRND(0,1,1,5,N,K). With $n = 1e6$, we consider different values of m and J. Most of our results above were derived for uniform sampling, so it is also of interest to evaluate the properties of the $\widetilde{\beta}$ using data sketched by Πs that do not satisfy PI.(iii). Four sketching schemes are considered: uniform sampling without replacement labeled as RS1, SRHT, the countsketch labeled CS, and leverage score sampling, labeled LEV. It should be mentioned that results for SHRT and LEV took significantly longer time to compute than RS1 and CS.

The top panel of Table 3 reports results for $\hat{\beta}_3$ when $K = 3$. Both sampling schemes precisely estimate β_3 whose true value is one. The standard error is larger the smaller is m, which is the sample size effect. But averaging $\widetilde{\beta}_j$ over j reduces variability. The standard error of LEV is slghtly less efficient. The size of the t test for $\beta_3 = 1.0$ is accurate, and the power of the test against $\beta_3 < 1$ when $\beta_3 = 0.98$ is increasing in the amount of total information used. Combining J sketches of size m generally gives a powerful test than a test based on a sketch size of mJ. The bottom panel of Table 3 reports for $K = 9$, focusing on uniform sampling without replacement and the countsketch. The results are similar to those for $K = 3$. The main point to highlight is that while there is a sample size effect from sketching, it can be alleviated by pooling across sketches.

6.2 The Choice of m

The JL Lemma shows that $m = O(\log d\epsilon^{-2})$ rows are needed for d vectors from $\mathbb{R}^n$ to be embedded into an m dimensional subspace. A rough and ready guide for embedding a d dimensional subspace is $m = \Omega(d \log d\epsilon^{-2})$. This is indeed the generic condition given in, for example, Sarlos (2006), though more can be said for certain Πs.[14] Notably, these desired m for random projections depends only on d but not on n.

As shown in Boutidis and Gittens (2013, Lemma 4.3), subspace embedding by uniform sampling without replacement requires that

$$m \geq 6\varepsilon_\sigma^{-2} n \ell_{\max} \log(2J \cdot K/\delta_\sigma). \tag{12}$$

where $\ell_{\max} = \max_i \ell_i$ is the maximum leverage score, also known as coherence. When coherence is large, the information in the data is not well spread out, and more rows are required for uniform

[14] The result for SRHT is proved in Lemma 4.1 of Boutidis and Gittens (2013). The result for count sketch is from Theorem 2 of Nelson and Nguyen (2013a).

sampling to provide subspace embedding. Hence unlike random projections, the desired m for uniform sampling is not data oblivious.

But while this choice of m is algorithmically desirable, statistical analysis often cares about the variability of the estimates in repeated sampling, and a larger m is always desirable for $\mathbb{V}(\widetilde{\beta})$. The question arises as to whether m can be designed to take both algorithmic and statistical considerations into account. We suggest two ways to to fine tune the algorithmic condition. Now

$$\begin{aligned}\ell_i = \|U_{(i)}\|^2 = X_{(i)}^T (X^T X)^{-1} X_{(i)} &= \frac{1}{n} X_{(i)}^T S_X^{-1} X_{(i)} \\ &\leq (\sigma_K^{-1}(S_X)) \frac{1}{n} \|X_{(i)}\|_2^2 = \sigma_K^{-1}(S_X) \frac{1}{n} \|X_{(i)}\|_2^2.\end{aligned}$$

where $\sigma_K(S_X)$ is the minimum eigenvalue of $S_X = n^{-1} X^T X$, $X_{(i)}^T$ denotes the i-th row of X, and $X_{(i,j)}$ the (i,j) element of X. But

$$\left\|X_{(i)}\right\|_2^2 = \sum_{j=1}^{K} \left[X_{(i,j)}\right]^2 \leq K \cdot \max_{j=1,\ldots,K} \left[X_{(i,j)}\right]^2 = K \cdot X_{\max}^2,$$

where $X_{\max} = \max_{i=1,\ldots,n} \max_{j=1,\ldots,d} \left|X_{(i,j)}\right|$. This implies $n \cdot \ell_{\max} \leq \sigma_K^{-1}(S_X) \cdot K \cdot X_{\max}^2$. Recalling that $p_i = \frac{\ell_i}{K}$ defines the importance sampling distribution, we can now restate the algorithmic condition for m when $J = 1$ as

$$m = \Omega\Big(nK \log(K) \cdot p_{\max}\Big) \quad \text{where} \quad p_{\max} \leq \frac{\sigma_K^{-1}(S_X) X_{\max}^2}{n}. \tag{13}$$

It remains to relate $X_{\max}$ with quantities of statistical interest.

Assumption M:

a. $\sigma_K(S_X)$ is bounded below by c_X with probability approaching one as $n \to \infty$;

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

Condition (a) is a standard identification condition for S_X to be positive definite so that it will converge in probability to $\mathbb{E}\left[X_{(i)} X_{(i)}^T\right]$. Condition (b) requires the existence of r moments so as to bound extreme values.[15] If the condition holds,

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

Proposition 1 *Suppose that Assumption M holds. A deterministic rule for sketched linear regressions by uniform sampling is*

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

[15] Similar conditions are used to obtain results for the hat matrix. See, for example, Section 6.23 of Hansen (2019)'s online textbook.

Proposition 1 can be understood as follows. Suppose that $r = 6$ moments are known to exist. Proposition 1 suggests a sketch of $m_1 = \Omega(\log K(nK)^{4/3}/n)$ rows which will generally be larger than $\Omega(K \log K)$, which is the sketch size suggested for data with thin tails. For such data, the moment generating function is uniformly bounded and $\mathbb{E}[\exp(tX_{(i,j)})] \leq C_X$ for some constant C_X and some $t > 0$ so that $X_{\max} = o_p(\log(nK))$. In both cases, the desired m increases with both n and K, as well as the number of data points in the regressor matrix, $n \cdot K$. This contrasts with the algorithmic condition for m which does not depend on n.

To use Proposition 1, we can either (i) fix r to determine m_1 or (ii) target 'observations-per-regressor ratio'. As an example, suppose $n = 1e7$ and $K = 10$. If $r = 6$, Proposition 1 suggests to sample $m_1 = 10,687$ rows, implying $\frac{m_1}{K} \approx 1000$. If instead we fix $\frac{m}{K}$ at 100 and uniformly sample $m_1 = 1000$ rows, we must be ready to defend the existence of $r = \frac{2\log(nK)}{\log(\frac{m}{K}) - \log(\log K)} \approx 10$ moments. There is a clear trade-off between m_1 and r.

Though m_1 depends on n, it is still a deterministic rule. To obtain a rule that is data dependent, consider again $c^T\widetilde{\beta}$, where c is a $K \times 1$ vector, and assume that $e_i \sim N(0, \sigma^2)$ so that $\text{var}(\widetilde{\beta}) = \frac{n}{m}\sigma_e^2(\widetilde{X}^T\widetilde{X})^{-1}$ where β_0 is the true (unknown) value of β. Define

$$\tau_0(m) = \frac{c^T(\widetilde{\beta} - \beta_0)}{\text{SE}(c^T\widetilde{\beta})} = \frac{c^T(\widetilde{\beta} - \beta^0) + c^T(\beta^0 - \beta_0)}{\text{SE}(c^T\widetilde{\beta})}.$$

It holds that $P_{\beta_0}(\tau_0 > z|\Pi, X_1, \ldots, X_n) = \Phi(-z)$ for some z, where $\Phi(\cdot)$ is the cdf of the standard normal distribution. Now consider a one-sided test τ_1 based on $\widetilde{\beta}$ against an alternative, say, β^0. The test τ_1 is related to τ_0 by

$$\tau_1(m) = \frac{c^T(\widetilde{\beta} - \beta^0)}{\text{SE}(c^T\widetilde{\beta})} + \frac{c^T(\beta^0 - \beta_0)}{\text{SE}(c^T\widetilde{\beta})} = \tau_0(m) + \tau_2(m).$$

The power of τ_1 at nominal size α is then

$$P_{\beta_0}(\tau_0 + \tau_2 > \Phi^{-1}_{(1-\alpha)}|\Pi, X_1, \ldots, X_n) = \Phi\left(-\Phi^{-1}_{(1-\alpha)} + \tau_2\right) \equiv \gamma.$$

Define

$$S(\alpha, \gamma) = \Phi^{-1}_{\gamma} + \Phi^{-1}_{1-\alpha}.$$

Common values of α and γ give the following:

Selected 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

Proposition 2 *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 and $\bar{\alpha}$ be the nominal size of the test.*

- *Let $\widetilde{\beta}$ be obtained from a sketch of size m_1. For a given effect size of $\beta^0 - \beta_0$, a data dependent 'inference conscious' sketch size is*

$$m_2(m_1) = S^2(\bar{\alpha}, \bar{\gamma}) \frac{m_1 var(c^T \widetilde{\beta})}{[c^T(\beta^0 - \beta_0)]^2} = m_1 \frac{S^2(\bar{\alpha}, \bar{\gamma})}{\tau_2^2(m_1)}. \tag{14}$$

- *For a pre- specified $\tau_2(\infty)$, a data oblivious 'inference conscious' sketch size is*

$$m_3 = n \frac{S^2(\bar{\alpha}, \bar{\gamma})}{\tau_2^2(\infty)}. \tag{15}$$

Inference considerations suggests to adjust m_1 by a factor that depends on $S(\bar{\alpha}, \bar{\gamma})$ and τ_2. For these values of $\bar{\gamma}$ and $\bar{\alpha}$, m_1 will be adjusted upwards when the τ_2 is less than two. The precise adjustment depends on the choice of τ_2.

The proposed m_2 in Part (i) requires an estimate of $\text{VAR}(c^T \widetilde{\beta})$ from a preliminary sketch. Table 4 provides an illustration for one draw of simulated data with $n = 1e7$, $K = 10$, and three values of σ_e, $\bar{\gamma}$ and effect size $\beta_1^0 - \beta_{10}$. Assuming $r = 10$ gives m_1 of roughly 1000. The sketched data are used to obtain an estimate of τ_2. If σ_e is small, m_1 almost achieves the target power of 0.5, but a target of 0.9 would require almost four times more rows, since m_2 is 3759. The larger is σ_e, the less precise is $\widetilde{\beta}$ for a given m, and more rows are needed. It is then up to the user how to trade of computation cost and power of the test.

The proposed m_3 in Part (ii) is motivatead by the fact that setting m_1 to n gives $m_2(n) = n\frac{S^2(\bar{\alpha},\bar{\gamma})}{\tau_2^2(n)}$. Though computation of $\tau_2(n)$ is infeasible, $\tau_2(n)$ is asymptotically normal as $n \to \infty$. Now if the full sample t-statistic cannot reject the null hypothesis, a test based on sketched data will unlikely reject the null hypothesis. But when full sample t statistic is expected to be relatively large (say, 5), the result can be used in conjunction with $S(\bar{\alpha}, \bar{\gamma})$ to give m_3. Say if $S(\bar{\alpha}, \bar{\gamma})$ is 2, $m_3 = (2/5)^2 n$. This allows us to gauge the sample size effect since $\frac{n}{m} = \frac{\tau_2^2(\infty)}{S^2(\bar{\alpha},\bar{\gamma})}$. Note that m_3 only requires the choice of $\bar{\alpha}, \bar{\gamma}$, and $\tau_2(\infty)$ which, unlike m_2, can be computed without a preliminary sketch.

Though Propositions 1 and 2 were derived for uniform sampling, they can still be used for other choice of Π. The one exception in which some caution is warranted is the countsketch. The rule given for the countsketch in Nelson and Nguyen (2013a, Theorem 5) of $m \geq \varepsilon^{-2} K(K+1)\delta^{-1}$ is data oblivious, which it is generally larger than the rule given in Boutidis and Gittens (2013) for uniform sampling especially if K is large, because the cost of sparsity of Π is a larger m. Thus, one might want to first use a small r to obtain a conservative m_1 for the countsketch. One can then use Proposition 2 to obtain an 'inference conscious' guide.

To illustrate how to use m_1, m_2 and m_3, we consider Belenzon et al. (2017) which studies firms' performance from naming the company after its owners, a phenomenon known as 'eponymy'. The parameter of interest is α_1 in a 'return on assets' regression

$$\text{roa}_{it} = \alpha_0 + \alpha_1 \text{eponymous}_{it} + Z_{it}^T \beta + \eta_i + \tau_t + c_i + \epsilon_{it}.$$

The coefficient gives the effect of the eponymous dummy after controlling for time varying firm specific variables Z_{it}, SIC dummies η_i, country dummies c_i, and year dummies τ_t. The panel of data includes 1.8 million companies from 2002-2012, but we only use data for one year. An interesting aspect of this regression is that even in the full sample with $n = 562160$, some dummies are sparse while others are collinear, giving an effective number of $K = 423$ regressors. We will focus on the four covariates: the indicator variable for being eponymous, the log of assets, the log number of shareholders, and equity dispersion.

Given the values of (n, K) for this data, any assumed value of r less than 8 would give an m_1 larger than n which is not sensible.[16] This immediately restrict us to $r \geq 6$. As point of reference, $(r, m_1) = (8, 317657)$ and $(r, m_1) = (15, 33476)$. The smallest m_1 is obtained by assuming that the data have thin tails, resulting in $m_1 = K \log(nK) = 8158$.

Table 5 presents the estimation results for several values of m. The top panel presents results for uniform sampling. Note that more than 50 covariates for uniform sampling are omitted due to colinearity, even with a relatively large sketch size. The first column shows the full sample estimates for comparison. Column (2) shows that the point estimates given by the smallest sketch with $m_1 = 8158$ are not too different from those in column (1), but the precision estimates are much worse. To solve this problem, we compute $m_2(m_1)$ by plugging in the t statistic for equity dispersion (ie. $\hat{\tau}_0 = 0.67$) as τ_2. This gives an m_2 of 112358. A similar sketch size can be obtained by assuming $r = 10$ for m_1, or by plugging in $\tau_2(\infty) = 5$ for m_3. As seen from Table 5, the point estimates of all sketches are similar, but the inference conscious sketches are larger in size and give larger test statistics.

The bottom panel of Table 5 presents results for the countsketch. Compared to uniform sampling in the top panel, only one or two covariates are now dropped. Though the estimate of α_1 is almost almost identical to the one for the full sample and for uniform sampling, the estimated coefficient for equity dispersion is somewhat different. This might be due to the fact that uniform sampling drops much more covariates than countsketch.

[16] This is based on $m_1 = (nK)^{1+2/r} \log K/n$. A smaller r is admissible if $m_1 = c_1 (nK)^{1+2/r} \log K/n$ for some constant $0 < c_1 < 1$. We limit our attention to $c_1 = 1$.

7 Concluding Remarks

This paper provides an gentle introduction to sketching and studies its implications for prediction and inference using a linear model. Sample code for constructing the sketches are avaialble in MATLAB, R, and STATA. Our main findings are as follows:

1. Sketches incur an approximation error that is small relative to the sample size effect.
2. For speed and parallelization: use countsketch.
3. For simple implementation: use uniform sampling.
4. For improved estimates: average over multiple sketches.
5. Statistical analysis may require larger sketch size that what is algorithmically desirable. We propose two inference conscious rules for the sketch size.

Sketching has also drawn attention of statisticians in recent years. Ahfock et al. (2017) provides an inferential framework to obtain distributional results for a large class of sketched estimators. Geppert, Ickstadt and Munteanu (2017) considers random projections in Bayesian regressions and provides sufficient conditions for a Gaussian likelihood based on sketched data to have an error of $1 + O(\epsilon)$ in terms of L_2 Wasserstein distance. In the design of experiments literature, the goal is to reveal as much information as possible given a fixed budget.[17] Since sketching is about forming random samples, it is natural to incorporate the principles in design of experiments into sketching. Wang, Zhu and Ma (2018) considers the design subsamples for logistic regressions. A-optimality and practical considerations suggest to use $p_i = \frac{|\hat{e}_i| \|x_i\|}{\sum_i |\hat{e}_i| \|x_i\|}$, which may be understood as score based sampling. Wang, Yang and Stufken (2018) considers the homoskedastic normal linear regression model. The principle of D-optimality suggests to recursively selecting data according to extreme values of covariances. The algorithm is suited for distributed storage and parallel computing.

While using sketches to overcome the computation burden is a step forward, sometimes we need more than a basic sketch. We have been silent about how to deal with data that are dependent over time or across space, such as due to network effects. We may want our sketch to preserve, say, the size distribution of firms in the original data. The sampling algorithms considered in this review must then satisfy additional conditions. When the data have a probabilistic structure, having more data is not always desirable, Boivin and Ng (2006). While discipline-specific problems require discipline-specific input, there is also a lot to learn from what has already been done in other literatures. Cross-disciplinary work is a promising path towards efficient handling of large volumes of data.

[17] A criterion that uses the trace norm for ordering matrices is A-optimality. A criterion that uses the determinant to order matrices is a D-optimal design.

A Appendix

Proof of Lemma 3 By an orthogonal decomposition of the least squares residuals,

$$\begin{aligned}
\|y - X\widetilde{\beta}\|_2^2 &= \|y - X\hat{\beta}\|_2^2 + \|X\widetilde{\beta} - X\hat{\beta}\|_2^2 \\
&= \|y - U\hat{\theta}\|_2^2 + \|U(\widetilde{\theta} - \hat{\theta})\|_2^2 \\
&= \widehat{\text{SSR}}^2 + \|U(\widetilde{\theta} - \hat{\theta})\| \\
&= \widehat{\text{SSR}}^2 + \|\widetilde{\theta} - \hat{\theta}\|_2, \qquad (16)
\end{aligned}$$

where

$$\begin{aligned}
\|\widetilde{\theta} - \hat{\theta}\|_2 &= \|(\Pi U)^T \Pi U(\widetilde{\theta} - \hat{\theta}) + (\widetilde{\theta} - \hat{\theta}) + (\Pi U)^T \Pi U(\hat{\theta} - \widetilde{\theta})\|_2 \\
&\le \|(\Pi U)^T \Pi U(\widetilde{\theta} - \hat{\theta})\|_2 + \|(\Pi U)^T \Pi U(\hat{\theta} - \widetilde{\theta}) - (\hat{\theta} - \widetilde{\theta})\|_2 \\
&\le \|(\Pi U)^T \Pi U(\widetilde{\theta} - \hat{\theta})\|_2 + \|(\Pi U)^T \Pi U - I_d\|_2 \|(\widetilde{\theta} - \hat{\theta})\|_2 \\
&\le \|(\Pi U)^T (\Pi U)(\widetilde{\theta} - \hat{\theta})\|_2 + \frac{1}{\sqrt{2}}\|\widetilde{\theta} - \hat{\theta}\|_2 \\
&\le \sqrt{2}\|(\Pi U)^T \Pi U(\widetilde{\theta} - \hat{\theta})\|_2
\end{aligned}$$

by triangle inequality, Cauchy-Schwarz inequality, condition (8a), and rearranging terms. Now the normal equations implies $(\Pi U)^T(\Pi U)\widetilde{\theta} = (\Pi U)^T \Pi(y - X\widetilde{\beta})$. Hence

$$\begin{aligned}
\|\widetilde{\theta} - \hat{\theta}\|_2 &\le \sqrt{2}\|(\Pi U)^T \Pi(y - U\hat{\theta})\|_2 \\
&\le \sqrt{\epsilon_0}\, \widehat{\text{SSR}}
\end{aligned}$$

by condition (8b) and for some failure probability δ_0. It follows from (16) that $\widetilde{\text{SSR}}^2 \le (1+\epsilon_2)\widehat{\text{SSR}}^2$ holds with probability $1 - \delta_2$ where $\epsilon_2 = \epsilon_0^2$ and $\delta_2 < 2\delta_0$. This probability can be made higher with suitable choice of ϵ_0 and m, which can be controlled by the researcher.

Proof of Theorem 3 Note that

$$c^T(\bar{\beta} - \beta) \;=\; \frac{1}{J}\sum_{j=1}^{J} c^T \big(X^T \Pi_j^T \Pi_j X\big)^{-1} \big(X^T \Pi_j^T \Pi_j e\big).$$

Thus,

$$E[c^T(\bar{\beta} - \beta)|\Pi_1, \ldots, \Pi_J] = \frac{1}{J}\sum_{j=1}^{J} c^T \big(X^T \Pi^T \Pi X\big)^{-1} \big(X^T \Pi^T \Pi E[e|\Pi_1, \ldots, \Pi_J]\big) = 0.$$

Write

$$
\begin{aligned}
&E[\{c^T(\bar{\beta}-\beta)\}^2|\Pi_1,\ldots,\Pi_J] \\
&= \frac{1}{J^2}\sum_{j=1}^{J}\sum_{k=1}^{J} c^T\left(X^T\Pi_j^T\Pi_j X\right)^{-1}\left(X^T\Pi_j^T\Pi_j E[ee^T|\Pi_1,\ldots,\Pi_J]\Pi_k^T\Pi_k X\right)\left(X^T\Pi_k^T\Pi_k X\right)^{-1}c \\
&= \sigma_e^2\frac{1}{J^2}\sum_{j=1}^{J} c^T\left(X^T\Pi_j^T\Pi_j X\right)^{-1}\left(X^T\Pi_j^T\Pi_j\Pi_j^T\Pi_j X\right)\left(X^T\Pi_j^T\Pi_j X\right)^{-1}c \\
&+ \sigma_e^2\frac{1}{J^2}\sum_{j=1}^{J}\sum_{k=1,k\neq J}^{J} c^T\left(X^T\Pi_j^T\Pi_j X\right)^{-1}\left(X^T\Pi_j^T\Pi_j\Pi_k^T\Pi_k X\right)\left(X^T\Pi_k^T\Pi_k X\right)^{-1}c \\
&= \sigma_e^2\frac{n}{mJ^2}\sum_{j=1}^{J} c^T\left(X^T\Pi_j^T\Pi_j X\right)^{-1}c.
\end{aligned}
$$

Then, the desired result is obtained by arguments identical to those used in proving Theorem 1. In particular, we can show that

$$
\frac{n}{m}\frac{c^T\left(X^T\Pi_j^T\Pi_j X\right)^{-1}c}{c^T(X^TX)^{-1}c} \leq \frac{n}{m}\frac{1}{(1-\varepsilon_\sigma)} \tag{17}
$$

jointly for all $j = 1,\ldots,J$ with probability at least $1-\delta_\sigma$. *Q.E.D.*

B Verification of Assumption CS for Countsketch

In this part of the appendix, we verify Assumption CS for the countsketch. Here, we use d to denote the column dimension of Π.

Lemma 5 *Let $\Pi \in \mathbb{R}^{n\times d}$ be a random matrix such that (i) the (i,j) element Π_{ij} of Π is $\Pi_{ij} = \delta_{ij}\sigma_{ij}$, where σ_{ij}'s are i.i.d. ± 1 random variables and δ_{ij} is an indicator random variable for the event $\Pi_{ij} \neq 0$; (ii) $\sum_{i=1}^{m}\delta_{ij} = 1$ for each $j = 1,\ldots,n$; (iii) for any $S \subset [n]$, $\mathbb{E}\left(\Pi_{j\in S}\delta_{ij}\right) = m^{-|S|}$; (iv) the columns of Π are i.i.d. Furthermore, there is a universal constant C_e such that $max_{i=1,\ldots,n}\Omega_{e,ii} \leq C_e$. Suppose that*

$$
\frac{(d^2+1)m}{n} + \frac{(d^2+d)}{m} \leq \frac{\delta_\Pi\varepsilon_\Pi^2}{8C_e^2}. \tag{18}
$$

Let

$$
A(\Omega_e,m,n) = \Omega_e + \frac{1}{m}\left\{tr(\Omega_e)I_n - \Omega_e\right\}. \tag{19}
$$

Then, we have that

$$
\mathbb{P}\left(\left\|U^T\Pi^T\Pi\Omega_e\Pi^T\Pi U - U^TA(\Omega_e,m,n)U\right\|_2 > \frac{n}{m}\varepsilon_\Pi\right) \leq \delta_\Pi.
$$

This lemma states that Condition CS is satisfied for countsketch, provided that all the diagonal elements of Ω_e are bounded by a universal constant, $d^2 m/n = o(\delta_\Pi \varepsilon_\Pi^2)$ and $d^2/m = o(\delta_\Pi \varepsilon_\Pi^2)$. The rate conditions in (18) are not stringent. When n is very large, d is of a moderate size and δ_Π and ε_Π are given, there is a range of m that satisfies (18).

Proof of Lemma 5 Since we assume that each diagonal element of Ω_e is bounded by a universal constant C_e,

$$\mathrm{tr}(\Omega_e^2) \le C_e^2 n, \ \mathrm{tr}(\Omega_e) = C_e n, \ \text{ and } \ \|\Omega_e\|_2 = C_e.$$

Then Lemma F.1, which is given in the online appendix, implies that

$$\begin{aligned}
&\mathbb{P}\left(\left\|U^T \Pi^T \Pi \Omega_e \Pi^T \Pi U - U^T A(\Omega_e, m, n) U\right\|_2 > \epsilon\right) \\
&\le 2\epsilon^{-2} \Bigg\{ \frac{2d^2(m-1)}{m^2} \mathrm{tr}(\Omega_e^2) + \frac{2d^2}{m^2} \|\Omega_e\|_2 \mathrm{tr}(\Omega_e) + \frac{d^2}{m^3} \left\{ [\mathrm{tr}(\Omega_e)]^2 + 2\|\Omega_e\|_2^2 \right\} \\
&\qquad + \frac{2}{m} \mathrm{tr}(\Omega_e^2) + \frac{2}{m^2} \mathrm{tr}(\Omega_e) + \frac{1}{m^3} \left\{ d\,[\mathrm{tr}(\Omega_e)]^2 + 2\mathrm{tr}(\Omega_e^2) \right\} \Bigg\} \\
&\le 2\epsilon^{-2} \Bigg\{ \frac{2d^2(m-1)}{m^2} C_e^2 n + \frac{2d^2}{m^2} C_e^2 n + \frac{d^2}{m^3} \left\{ C_e^2 n^2 + 2C_e^2 \right\} \\
&\qquad + \frac{2}{m} C_e^2 n + \frac{2}{m^2} C_e n + \frac{1}{m^3} \left\{ d C_e^2 n^2 + 2C_e^2 n \right\} \Bigg\} \\
&\le 8C_e^2 \epsilon^{-2} \left(\frac{(d^2+1)n}{m} + \frac{(d^2+d)n^2}{m^3} \right).
\end{aligned}$$

If we take $\epsilon = \frac{n}{m}\varepsilon_\Pi$, then

$$\mathbb{P}\left(\left\|U^T \Pi^T \Pi \Omega_e \Pi^T \Pi U - U^T A(\Omega_e, m, n) U\right\|_2 > \frac{n}{m}\varepsilon_\Pi\right) \le 8C_e^2 \varepsilon_\Pi^{-2} \left(\frac{(d^2+1)m}{n} + \frac{(d^2+d)}{m} \right).$$

To satisfy the probability above is bounded by δ_Π, we need to assume that

$$8C_e^2 \varepsilon_\Pi^{-2} \left(\frac{(d^2+1)m}{n} + \frac{(d^2+d)}{m} \right) \le \delta_\Pi,$$

which is imposed by (18). *Q.E.D.*

Table 1: Assessment of JL Lemma: $n = 20,000$, $d = 5$.

m	Random Sampling			Random Projections					
	RS1	RS2	RS3	RP1	RP2	RP2	RP3	CS	LEV
Normal	Norm approximation								
161	0.627	0.624	0.538	0.628	0.633	0.631	0.640	0.642	0.757
322	0.801	0.792	0.700	0.790	0.795	0.795	0.800	0.793	0.909
644	0.931	0.931	0.871	0.926	0.929	0.927	0.931	0.928	0.982
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	1.000
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	0.158
322	0.126	0.128	0.127	0.127	0.127	0.128	0.126	0.129	0.105
644	0.082	0.085	0.084	0.086	0.084	0.085	0.085	0.086	0.071
966	0.065	0.067	0.065	0.067	0.066	0.067	0.067	0.068	0.055
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	0.029
Exponential	Norm approximation								
161	0.432	0.429	0.402	0.627	0.624	0.636	0.628	0.637	0.717
322	0.580	0.578	0.548	0.796	0.795	0.794	0.800	0.791	0.875
644	0.747	0.738	0.717	0.925	0.930	0.929	0.930	0.928	0.972
966	0.851	0.840	0.812	0.971	0.968	0.973	0.969	0.972	0.992
1288	0.899	0.894	0.866	0.990	0.988	0.989	0.991	0.989	0.998
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	0.158
322	0.176	0.177	0.175	0.126	0.128	0.127	0.127	0.127	0.104
644	0.116	0.118	0.116	0.084	0.083	0.083	0.082	0.085	0.069
966	0.090	0.094	0.090	0.066	0.067	0.066	0.065	0.065	0.055
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	0.030

Table 2: Back of Envelope Power Calculations

	Δ	ϕ_n	$\mathbb{E}[F]$	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
n =1e6	0.4	80000	80001	1.0
m =2000		145.45	146.60	1.0
m =1000		72.72	73.87	1.0
n =1e6	0.5	125000	125001	1.0
m =2000		227.27	227.50	1.0
m =1000		113.63	114.86	1.0

$$\Delta_m = \frac{\Delta}{\sqrt{m}}$$

c	$n = 1e6$	$m = 100$	$m = 200$	$m = 1000$	$m = 2000$
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

Table 3: Monte Carlo Experiments: Properties of $\widetilde{\beta}$ and $t_{\hat{\beta}_3}$, $n = 1e6$

$K - 3$

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
		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

$K = 9$

m	J	RS1	CS	RS1	CS	RS1	CS	RS1	CS
		Size		Power		$\hat{\beta}_9$		se($\hat{\beta}_9$)	
500	1	0.049	0.067	0.076	0.079	0.999	0.998	0.046	0.047
500	5	0.038	0.029	0.129	0.120	1.000	1.000	0.021	0.020
500	10	0.032	0.039	0.241	0.268	0.999	1.000	0.014	0.015
1000	1	0.052	0.041	0.099	0.087	1.000	1.000	0.032	0.031
1000	5	0.036	0.027	0.219	0.214	1.000	1.000	0.014	0.014
1000	10	0.033	0.042	0.461	0.484	1.000	1.000	0.010	0.010
2000	1	0.043	0.050	0.143	0.128	1.000	0.999	0.022	0.022
2000	5	0.025	0.028	0.411	0.400	1.000	1.000	0.010	0.010
2000	10	0.041	0.044	0.782	0.773	1.000	1.000	0.007	0.007
5000	1	0.051	0.057	0.260	0.292	0.999	1.000	0.014	0.015
5000	5	0.021	0.037	0.839	0.813	1.000	1.000	0.006	0.007
5000	10	0.033	0.044	0.988	0.990	1.000	1.000	0.005	0.005

Table 4: Inference Conscious Choice of m

$n = 1e7, r = 10, K = 10, m_0 = 1000$

		$(\beta_1^0 - \beta_{10})$				
$\bar{\gamma}$	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

Table 5: Example of Belenzon et al. (2017)

	(1)	(2)	(3)	(4)
Uniform Sampling	Full Sample	m_1	$m_2(m_1)$	m_3
		$K\log(nK)$	$(\bar{\alpha},\bar{\gamma}) = (.05,.8)$	$\tau_2(\infty) = 5$
Dummy for eponymous	0.031**	0.035**	0.031**	0.031**
	(0.001)	(0.010)	(0.003)	(0.002)
ln(assets)	0.005**	0.000	0.006**	0.007**
	(0.001)	(0.004)	(0.001)	(0.001)
ln(no. shareholders)	-0.032**	-0.025**	-0.030**	-0.031**
	(0.001)	(0.007)	(0.002)	(0.002)
Equity dispersion	-0.012**	-0.007	-0.016**	-0.017**
	(0.001)	(0.011)	(0.003)	(0.003)
Omitted Covariates		132	58	54
Observations	562,170	8,158	112,355	139,022

	(1)	(2)	(3)	(4)
Countsketch	Full Sample	m_1	m_2	m_3
Dummy for eponymous	0.031**	0.035**	0.030**	0.032**
	(0.001)	(0.011)	(0.003)	(0.003)
ln(assets)	0.005**	0.009**	0.004**	0.005**
	(0.001)	(0.003)	(0.001)	(0.001)
ln(no. shareholders)	-0.032**	-0.028**	-0.032**	-0.032**
	(0.001)	(0.008)	(0.002)	(0.002)
Equity dispersion	-0.012**	-0.024	-0.010*	-0.011**
	(0.001)	(0.014)	(0.004)	(0.004)
Omitted Covariates		2	1	1
Observations	562,170	8,147	112,347	139,015

Robust standard errors in parentheses
** $p<0.01$, * $p<0.05$

References

Achiloptas, D. 2003, Database Friendly Random Projections: Johnson-Lindenstrauss with Binary Coins, *Journal of Computer and System Sciences* **66**(4), 671–687.

Agarwal, P. S. H.-P. and Varadarajan, K. 2004, Approximating Extent Measures of Points, *Journal of the ACM* **51**(4), 606–635.

Ahfock, D., Astle, W. and Richardson, S. 2017, Statistical Properties of Sketching Algorithms, arXiv:1706.03665v1.

Ailon, N. and Chazelle, B. 2009, The Fast Johnson-Lindenstrauss Transform and Approximate Nearest Neighbors, *SIAM Journal Computing* **39**(1), 302–322.

Alon, N., Matias, Y. and Onak, K. 1999, The Space Complexity of Approximating the Frequency Moments, *Journal of Computational Systems Science* **58**(1), 137–147.

Bai, Z. and Yin, Y. 1993, Limit of the Smallest Eigenvalue of a Large Dimensional Sample Coveriacne Matrix, *Annals of Probability* **21:3**, 1275–1294.

Belenzon, S., Chatterji, A. and Dailey, B. 2017, Eponymous Enterpreneurs, *American Economic Reivew* **107:6**, 1638–1655.

Boivin, J. and Ng, S. 2006, Are More Data Always Better for Factor Analysis, *Journal of Econometrics* **132**, 169–194.

Boutidis, C. and Gittens, A. 2013, Improved Matrix Algorithms via the Subsampled Randomized Hadamadr Transform, *SIAM Journal on Matrix Analysis* **34:3**, 1301–1340.

Breiman, L. 1999, Pasting Bites Together for Prediction in Large Data Sets and On-Line, *Machine Learning* **36**(2), 85–103. https://www.stat.berkeley.edu/ breiman/pastebite.pdf.

Charika, M., Chen, K. and Farach-Colton, M. 2002, Finding Frequent Items in Data Streams, *Proceedings of the International Colliquium in Automata, Languages and Programming* pp. 283–703.

Chawla, N., Hall, L., Bowyer, K. and Kegelmeyer, P. 2004, Learning Ensembles from Bites: A Scalable and Accurate Approach, *Journal of Machine Learning Research* **5**, 421–451.

Chen, S., Varma, R., Singh, A. and Kovacevic, J. 2016, A Statistical Perspective of Sampling Scores for Linear Regresson, *IEEE International Symposium on Information theory*.

Chi, J. and Ipsen, I. 2018, Randomized Least Squares Regression: Combining Model and Algorithm Induced Uncertainties, xrXiv:1808.05924v1.

Christmann, A., Steinwart, I. and Hubert, M. 2007, Robust Learning from Bites for Data Mining, *Computational Statistics and Data Analysis* **52**, 347–361.

Clarkson, K. and Woodruff, D. 2013, Low Rank Approximation and Regression in Input Sparsity Time, *Proceedings of the 45th ACM Symposium on the Theory of Computing*.

Cohen, M., Lee, Y., Musco, C., Musco, C., Peng, R. and Sidford, A. 2015, Uniform Sampling for Matrix Approximation, *Proceedings of the 46th ACM Symposium on the Theory of Computing* pp. 181–190.

Cohen, M., Nelson, J. and Woodruff, D. 2015, Optimal Approximate Matrix Product in Terms of Stable Rank, arXiv:1507.02268.

Comrode, G., Garofalakis, M., Haas, P. and Jermaine, C. 2011, Synposes for Massive Data: Samples, Histograms, Wavelets,Sketches, *Foundations and Trends in Databases* **4**(1), 1–294.

Conmode, G. and Muthukrishnan, S. 2005, An Improved Data Stream Summary: The Count-Min Sketch and Applications, *Journal of Algorithms* **55**, 29–38.

Cormode, G. 2011, Sletch Techniques for Approximate Query Processing, *Foundations and Trends in Databases*.

Cramer, J. S. 1987, Mean and Variance of R^2 in Small and Moderate Samples, *Journal of Econometrics* **35**, 253–266.

Dahiya, Y., Konomis, D. and Woodruff, D. 2018, An Empirical Evaluation of Sketching for Numerical Linear Algebra, KDD, http://www.cs.cmu.edu/afs/cs/user/dwoodruf/www/dkw18.pdf.

Dasgupta, A., Kumar, R. and Sarlos, T. 2010, A Sparse Lindenstrauss Transform, *STOC* pp. 341–350.

Deaton, A. and Ng, S. 1998, Parametric and Nonparametric Approaches to Tax Reform, *Journal of the American Statistical Association* **93**(443), 900–909.

Dhillon, P., Lu, Y., Foster, D. and Ungar, L. 2013, New Subsampling Algorithms for Faster Least Squares Regression, *Advances in Neural Information Processing Systems (NIPS)* **26**, 360–368.

Drineas, P. and Mahoney, M. 2005, On the Nystrom Method for Approximating a Gram Matrix for Improved Kernel Based Learning, *Journal of Machine Learning Research* **6**, 2152–2125.

Drineas, P., Kannan, R. and Mahoney, M. 2006, Fast Monte Carlo Algorithms for Matrices I: Approximating Matrix Multiplications, *SIAM Journal on Computing* **36**, 132–157.

Drineas, P., Magdon-Ismail, M., Mahoney, M. and Woodruff, D. 2012, Fast Approximation of Matrix Coherence and Statistical Leverage, *Journal of Machine Learning Research* **13**, 3441–3472.

Drineas, P., Mahoney, M. and Muthukrishnan, S. 2006, Sampling Algorithms for L2 Regression and Applications, *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms* pp. 1127–1136.

Drineas, P., Mahoney, M., Muthukrishnan, S. and Sarlos, T. 2011, Faster Least Squares Approximation, *Numerical Mathematics* **117**, 219–249.

Du Mouchel, W., Volinsky, C., Johnson, T., Cortes, C. and Pregibon, D. 1999, Squashing Flat Files Flatter, *Proceedings of the Fifth ACM Conference on Knowledge Discovery and Data Mining* pp. 6–15.

Eriksson-Bique, S., Solberg, M., Stefanelli, M., Warkentin, S., Abbey, R. and Ipsen, I. 2011, Importance Sampling for a Monte Carlo Matrix Multiplication Algorithm with Application to Information Retrieval, *SIAM Journal Computing* **33**(4), 1689–1706.

Geppert, L., Ickstadt, K. and Munteanu, A. 2017, Random Projections for Bayesian Regression, *Statistics and Computing* **27**(1), 79–101.

Geppert, L., Ickstadt, K., Munteanu, A., Qudedenfeld, J. and Sohler, C. 2017, *Statistical Computing* **27**, 79–101.

Ghashami, M., Liberty, E., Phillips, M. and Woodruff, D. 2016, Frequent Directions: Simple and Deterministic Matrix Sketching, *SIAM Journal Computing* **45**(5), 1762–1792.

Heince, C., McWilliams, B. and Meinshausen, N. 2016, Dual Loco: Distributing Statistical Estimation Using Random Projections, *19th International Conference on Artificial Intelligence* **51**, 875–883.

Hogben, L. 2007, *Handbook of Linear Algebra*, Chapman and Hall.

Horvitz, D. and Thompson, D. 1952, A Generalization of Sampling Replacement from a Finite Universe, *Journal of the American Statistical Association* **47**, 663–685.

IPMUS 2019, Ruggles, S. and S. Flood and R. Goeken and J. Grover and E. Meyer and J. Pacas and M. Sobek, Vol. Version 9, Minneapolis, p. `http://doi.org/10.18128/D010.V90`.

Ipsen, I. and Wentworth, T. 2014, The Effect of Coherence on Sampling From Matrices with Orthonormal Columns and Preconditioned Least Squares Problems, *Siam Journal of Matrix Analysis and Applicatons* **35**(4), 1490–1520.

Johnson, W. and Lindenstauss, J. 1994, Extensions of Lipschitz Maps into a Hilbert Space, *Contemporary Mathematics*.

Jolliffe, I. 1972, Discarding Variables in a Principal Component Analysis: Artificial Data, *Applied Statistics* **21**(2), 160–173.

Kane, D. and Nelson, J. 2014, Sparser Johnson-Lindenstrauss Transforms, *Journal of the ACM* **61:1**, 4.1–4.10.

Li, P., Hastie, T. and Church, K. 2006, Very Sparse Random Projections, *KDD* pp. 287–296.

Ma, P., Mahoney, M. W. and Yu, B. 2014, A Statistical Perspective on Algorithmic Leveraging, *Proceedings of the 31st ICML Conference*, Vol. arXiv: 1306.5362.

Madigan, D., Raghavan, N., Dumouchel, W., Nason, M., Posse, C. and Ridgeway, G. 1999, Likelihood-Based Data Squashing: A Modeling Approach to Instance Construction, *Technical report*, AT and T Labs Ressearch.

Mahoney, M. W. 2011, Randomized Algorithms for Matrices and Data, *Foundations and Trends in Machine Learning*, http://dx.doi.org/10.1561/2200000035 edn, Vol. 3:2, NOW, pp. 123–224.

McWilliams, B., Krummenacher, C., Lucic, G. and Buhmann, J. 2014, Fast and Robust Least Squares Estimation in Corrupted Linear Models, *Advances in Neural Information Processing Systems (NIPS)* pp. 415–423.

Meng, X. and Mahoney, M. 2013, Low Distortion Subspace Embeddings in Input-Sparsity time and Applications to Robust Linear Regression, *Proceedings of the 45th ACM Symposium on the Theory of Computing*.

Mitzenmacher, M. and Upfal, E. 2006, *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*, Cambridge University Press.

Nelson, J. and Nguyen, H. 2013a, OSNAP: Faster Numerical Linear Algebra Algorithms via Sparser Subspace Embeddings, *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*.

Nelson, J. and Nguyen, H. 2013b, Sparsity Lower Bounds for Dimensionality Reducing Maps, *STOC* pp. 101–110.

Nelson, J. and Nguyen, H. 2014, Lower Bounds for Oblvious Subspace Embeddings, *Proceedings of the 41st Interational Colluqium on Automata, Langauges and Programming*.

Ng, S. 2017, Opportunities and Challenges: Lessons from Analyzing Terabytes of Scanner Data, *in* B. Honore, A. Pkes, M. Piazzesi and L. Samuelson (eds), *Advances in Economics and Econometrics, Eleventh World Congress of the Econometric Society*, Vol. II, Cambridge University Press, pp. 1–34.

Owen, A. 1990, Empirical Likelihood Ratio Confidence Region, *Annals of Statistics* **18**, 90–120.

Pilanci, M. and Wainwright, M. 2015, Randomized Sketches of Convex Programs with Sharp Guarantees, *IEEE Transactions of Information Theory* **61**(9), 5096–5115.

Pilanci, M. and Wainwright, M. 2016, Iterative Hessian Sketch: and Accurate Solution Approximation for Constrained Least Squares, *Journal of Machine Learning Research* pp. 1–33.

Raskutti, G. and Mahoney, M. 2016, A Statistical Perspective on Randomized Sketching for Ordinary Least Squares, *Journal of Machine Learning Research* pp. 1–31.

Rudd, P. 2000, *An Introduction to Classical Econometric Theory*, Oxford University Press.

Sarlos, T. 2006, Improved Approximation Algorithms for Large Matrices via Random Projections, *Proceedings of the 47 IEEE Symposium on Foundations of Computer Science.*

Wallace, T. 1972, Weaker Criteria and Tests for Linear Restrictions, *Econometrica* **40**(4), 689–698.

Wang, H., Yang, M. and Stufken, J. 2018, Information-Based Optimal Subdata Selection for Big Data Linear Research, *Journal of the American Statistical Association.*

Wang, H., Zhu, R. and Ma, P. 2018, Optimal Subsampling for Large Sample Logistic Regression, *Journal of the American Statistical Association* **113**(522), 849–844.

Wang, J., Lee, J., Mahdav, M., Kolar, M. and Srebo, N. 2017, Sketching Meets Random Projection in the Dual: A Provable Recovery Algorithm for Big and High Dimensional Data, *Electronic Journal of Statistics* **11**, 4896–4944.

Wang, S., Gittens, A. and Mahoney, M. 2018, Sketched Ridge Regression: Optimization Perspective, Statistical Perspective, and Model Averaging, *Proceedings of the 34th International Conference on Machine Learning* **19**, 1–50.

Woodruff, D. 2014, Sketching as a Tool for Numerical Linear Algebra, *Foundations and Trends in Theoretical Computer Science* **10**(1-2), 1–157.

Woolfe, F., Liberty, E., Vladmir, R. and Mark, T. 2008, A Fast Randomized Algorithm for the Approximation of matrices, *Applied and Computational Harmonic Analysis* **25:3**, 335–366.

Yin, Y., Bai, Z. and Krishnaiah, P. 1988, On the Limit of the Largest Eigenvalue of the Largest Dimensional Sample Covariance Mtrix, **78:4**, 509–521.