Title: An Econometric View of Algorithmic Subsampling

Submitted by Simon Lee[1] and Serena Ng[2]

to Big Data and High-Performance Computing for Financial Economics Workshop

1 Motivation

The availability of terabytes of data for economic analysis is increasingly common. A case in point is weekly the Nielsen scanner data analyzed in Ng (2017)[3] whose size has now grown to six terabytes. The data have what are referred to as the "three Vs" characteristics: volume, variety, and velocity. Econometricians may think of them as short panels of big, often unbalanced, high-frequency, highly heterogeneous data in the product, spatial, and time dimensions. Financial data share many similar characteristics.

Large datasets are costly to store and to analyze, and memory constraints often make a complete analysis beyond the limits of our computers. Suppose that we have a data matrix A of dimension $n \times d$, and n is very big. Do we really need all n rows to compute the eigenvalues of A, or to compute the projection matrix $A(A^TA)^{-1}A^T$? Can we approximate A^TA using only $m << n$ rows of A. Suppose we are interested in testing a hypothesis. Can we get a glimpse of the result using just $m << n$ rows?

A dataset constructed from just m rows of the original data is known as a sketch. A sketch is useful only if it preserves the structure of A, and computer scientists have shown that a sketch with small eigenvalue distortions can be obtained by *random projections*. But their focus is on run time and certain worst case error bounds. Without specifying a probability model, not much can be said about the effect of sketching on statistical analysis.

We are interested in methods that reduce storage demands, speeding up calculations, while allowing for inference to be almost as accurate as using all the data. This paper considers how sketches affect estimation and inference in the context of the normal linear model with p predictors, $y_i = X_i'\beta_0 + e_i$, $e_i \sim N(0, \sigma^2)$. We observe a $n \times d$ data matrix $A = (y, X_i)$, $d = p + 1$. The singular value decomposition of A is UDV^T. The full sample estimator is $\hat{\beta} = \min_\beta ||y - X\beta||_2^2$. The subsampled estimator is $\tilde{\beta} = \min_\beta ||\tilde{y} - \tilde{X}\beta||_2^2$, where the $m \times d$ coreset $(\tilde{y}, \tilde{X}) = (\Pi y, \Pi X)$ is constructed via an embedding matrix $\Pi \in \mathbb{R}^{m \times n}$. Our emphasis is on the choice and construction of Π from an econometric perspective. Some highlights of the results are as follows:

A. Results On Choice of Π

- Π must preserve the column space of A, a result also known as subspace-embedding. As Π is random, this requires that for appropriately chosen m, $P_{\Pi \in \mathcal{D}}(||(\Pi U)^T(\Pi U) - I_d||_2 > \epsilon) < \delta$. The condition implies that the eigenvalues of ΠA are within $(1 \pm \epsilon)$ of the eigenvalues of A.

[1]Columbia University

[2]Corresponding author: Columbia University and NBER, 420 W. 118 St. New York, NY 10027. (212)-854-5488

[3]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.

- Uniform sampling uses a selection matrix as Π that will, in general, require m to be large (same order as n) unless the data are uniformly dispersed. In other words, uniform sampling will not, in general, be a desirable way to reduce the dimension of the problem.
- Sketching matrices Π satisfying $\Pi\Pi^T = \frac{n}{m} I_d$ will produce more efficient $\tilde{\beta}$. Such a Π can be constructed by Subsampled Random Hadamard transform (SRHT), ie. apply uniform sampling without replacement to data that are first uniformized and randomized.
- If A is known to be sparse, such as when there are lots of dummy variable predictors, a very sparse Π can be obtained by COUNTSKETCH. The sketch can be obtained without storing the whole data set in memory.

B. Implications for Inference: $\Pi\Pi^T = \frac{n}{m} I_m$

- For prediction at arbitrary x_0, the relative prediction variance is $\frac{\mathbb{V}(x_0^T \tilde{\beta})}{\mathbb{V}(x_0^T \hat{\beta})} \leq \frac{n}{m}(\frac{1}{1-\epsilon})$ where $\frac{n}{m}$ is the sample size effect, and $\frac{1}{1-\epsilon}$ is the sketching error effect.
- For testing against fixed alternative $R\beta_0 = r$, the relative power of an F test depends on relative non-centrality: $\frac{\lambda_m}{\lambda_n} = \frac{n(1+\epsilon)}{m}$. In spite of an inevitable loss in relative efficiency, the loss in absolute power is small because in a big data setting, m is not small. Hence $\frac{n-d}{n-d-2} \approx \frac{m-d}{m-d-2}$ for large m and n, as is the case.
- Local power: Suppose we use $\hat{\beta}$ to test $H_1^n = \frac{c}{\sqrt{n}}$ but $\tilde{\beta}$ to test $H_1^m = \frac{c}{\sqrt{m}}$. We find that local power changes slowly when m is large (even though it may be small relative to n).

  The above results only change slightly when $||\Pi\Pi^T - \frac{n}{m} I_m||_2 = o(1)$.

C. Divide and Conquer/Distributed Computing

- Construct $\tilde{\beta}_j$ for each $j = 1, \ldots, J$ sketch. Form $\bar{\beta} = \frac{1}{J}\sum_{j=1}^J \tilde{\beta}_j$ and $\bar{t} = \frac{1}{J}\sum_{j=1}^J t_j$.
- For sampling without replacement, letting $J = \frac{n}{m}$ would use all information in the data. We study the efficiency implications. We also consider a t test based on $\bar{\beta}$, as well as an average of the individual t tests, denoted $\bar{t}$. These have better power than a single t test.

2 Broader Impact

In summary, we consider ΠA that preserves the subspace spanned by A and make precise the econometric implications of specific Πs. We show that if Π preserves the subspace of A, there will be little effect on inference. Our asymptotic analysis complements the time cost and finite sample error bounds derived in the theoretical computer science literature. Our results are relevant to all 'Big Data' analysis that requires high performance computing. A obvious application is financial data. A draft of the paper (in preparation for the Annals of Economic Review) is available on request.