Biometrika (2021), **108**, 2, *pp.* 283–297
Printed in Great Britain
doi: 10.1093/biomet/asaa062
Advance Access publication 11 October 2020

Statistical properties of sketching algorithms

By D. C. Ahfock, W. J. Astle and S. Richardson

MRC Biostatistics Unit, University of Cambridge, Robinson Way, Cambridge CB2 0SR, U.K.
d.ahfock@uq.edu.au wja24@cam.ac.uk sylvia.richardson@mrc-bsu.cam.ac.uk

Summary

Sketching is a probabilistic data compression technique that has been largely developed by the computer science community. Numerical operations on big datasets can be intolerably slow; sketching algorithms address this issue by generating a smaller surrogate dataset. Typically, inference proceeds on the compressed dataset. Sketching algorithms generally use random projections to compress the original dataset, and this stochastic generation process makes them amenable to statistical analysis. We argue that the sketched data can be modelled as a random sample, thus placing this family of data compression methods firmly within an inferential framework. In particular, we focus on the Gaussian, Hadamard and Clarkson–Woodruff sketches and their use in single-pass sketching algorithms for linear regression with huge samples. We explore the statistical properties of sketched regression algorithms and derive new distributional results for a large class of sketching estimators. A key result is a conditional central limit theorem for data-oblivious sketches. An important finding is that the best choice of sketching algorithm in terms of mean squared error is related to the signal-to-noise ratio in the source dataset. Finally, we demonstrate the theory and the limits of its applicability on two datasets.

Some key words: Computational efficiency; Random projection; Randomized numerical linear algebra; Sketching.

1. Introduction

Sketching is a general probabilistic data compression technique involving random projections (Cormode, 2011). Even routine calculations can be prohibitively computationally expensive if performed on massive datasets. Computational time can be reduced to an acceptable level by allowing some approximation error in the results. Sketching algorithms simplify the computational task by generating a compressed version of the original dataset that then serves as a surrogate for calculations. The compressed dataset is referred to as a sketch, because it acts as a compact representation of the full dataset. Sketching algorithms use a randomized compression stage, which makes them interesting from a statistical viewpoint. Sketching algorithms for linear regression have attracted significant attention in the numerical linear algebra and theoretical computer science communities (Mahoney, 2011; Woodruff, 2014).

To describe sketched regression in more detail, we first assume that the data consist of a length-n response vector y and an $n \times p$ matrix of covariates, X, which is of full rank. It is assumed throughout that $n > p$. The objective is to find the least squares coefficients. Given sufficient computational resources, these can be computed exactly as

$$\beta_{\mathrm{F}} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y,$$

where the subscript F indicates the connection to the full dataset. Only two quantities are needed to determine β_{F}: the Gram matrix $X^{\mathrm{T}}X$ and the marginal associations $X^{\mathrm{T}}y$. Calculation of $X^{\mathrm{T}}X$ requires $O(np^2)$ operations, while computation of $X^{\mathrm{T}}y$ needs only $O(np)$ calculations. There are two broad methods for sketched regression, namely complete sketching and partial sketching. Complete sketching is based on approximating both $X^{\mathrm{T}}X$ and $X^{\mathrm{T}}y$, whereas partial sketching approximates only the Gram matrix. Drineas et al. (2006) established many important results for complete sketching, and Dhillon et al. (2013) and Pilanci & Wainwright (2016) derived foundational results for partial sketching.

Sketching algorithms use random linear mappings to reduce the size of the dataset from n to k observations. The random linear mapping can be represented as a $k \times n$ sketching matrix S. Complete sketching generates a length-k sketched response vector $\tilde{y}$ and a $k \times p$ matrix of sketched predictors $\tilde{X}$. The sketched data are computed through the linear mappings $\tilde{y} = Sy$ and $\tilde{X} = SX$. Assuming that $\tilde{X}$ is of rank p, the complete sketching estimator β_{S} is defined to be the set of least squares coefficients using the sketched responses and predictors,

$$\beta_{\mathrm{S}} = (\tilde{X}^{\mathrm{T}}\tilde{X})^{-1}\tilde{X}^{\mathrm{T}}\tilde{y}. \tag{1}$$

The partial sketching estimator, β_{P}, is defined as

$$\beta_{\mathrm{P}} = (\tilde{X}^{\mathrm{T}}\tilde{X})^{-1}X^{\mathrm{T}}y. \tag{2}$$

The key difference between (1) and (2) is that the partial sketching estimator β_{P} is constructed using the exact marginal associations $X^{\mathrm{T}}y$. Given the sketched data, computation of β_{S} or β_{P} requires only $O(kp^2)$ operations, compared with the $O(np^2)$ operations required for β_{F}.

There is a large literature concerned with designing appropriate distributions for the random sketching matrix S. Our focus is on data-oblivious random projections, such that the distribution of the sketching matrix is not a function of the source data (y, X). An example is the Gaussian sketch, where each element is independently distributed as an $N(0, 1/k)$ variate. We also consider the Hadamard sketch and the Clarkson–Woodruff sketch, random projections that exploit structure and sparsity for computational efficiency. A motivation for this work is that there are no clear ties between data-oblivious random projections and classical subsampling techniques.

Most existing results on the accuracy of sketching are universal worst-case bounds (Woodruff, 2014; Mahoney & Drineas, 2016). This is typical for randomized algorithms; however, a more detailed error analysis can provide important insights (Halko et al., 2011). We investigate the statistical properties of β_{P} and β_{S} when data-oblivious sketches are used. An important finding is that the signal-to-noise ratio in the source dataset strongly influences the relative efficiency of complete to partial sketching. The statistical analysis also allows the construction of exact confidence intervals for the Gaussian sketch and asymptotic confidence intervals for other random projections, paving the way for their wider use in the statistical community.

At its core, sketched regression is a randomized algorithm for approximate computation of β_{F}. Repeated application of the sketching algorithm to the same dataset will produce different results. The first stage in our analysis is to establish the distributional properties of the sketching estimators with the source dataset held fixed. An important result is a conditional central limit theorem for the sketched dataset that connects the Hadamard and Clarkson–Woodruff projections to the Gaussian sketch. The conditional analysis of the randomized algorithms is then extended to cover situations where sketching is used for approximate statistical inference. Given a statistical model for the response $y = X\beta_0 + \epsilon$, with population parameter β_0 and error term ϵ, distributional properties of β_{P} and β_{S} can be determined by integrating over the conditional distributions of the sketching estimators that take y and X as fixed.

2. Background and related work

2.1. *Preliminaries*

We define a number of quantities related to the full dataset before moving on. The total, residual and model sum of squares are given by $\text{TSS}_\text{F} = y^\text{T}y$, $\text{RSS}_\text{F} = \|y - X\beta_\text{F}\|_2^2$ and $\text{MSS}_\text{F} = \|X\beta_\text{F}\|_2^2$, respectively, with $\text{TSS}_\text{F} = \text{MSS}_\text{F} + \text{RSS}_\text{F}$. The proportion of variance explained by the model is $R_\text{F}^2 = \text{MSS}_\text{F}/\text{TSS}_\text{F}$. These values will be important in characterizing the behaviour of β_S and β_P. The source data are generically represented by the $n \times d$ matrix $A = (y, X)$.

There are two general categories of distributions for the random matrix S: data-aware random projections and data-oblivious random projections. A data-aware random projection uses information in the source data (y, X) to generate S. In contrast, a data-oblivious random projection can be sampled without knowledge of y or X. Data-aware random projections are closely connected to finite population sampling methods in the statistics literature (Ma & Sun, 2015). Our main focus is on data-oblivious random projections, as their mechanism for data compression is not obviously tied to subsampling. Data-oblivious random projections generate a dataset of k pseudo-observations using the source dataset as a component in the generative process.

2.2. *Data-oblivious sketches*

The Gaussian sketch was one of the first projections proposed for sketched regression (Sarlos, 2006). Recall that a Gaussian sketch is formed by independently sampling each element of S from an $N(0, 1/k)$ distribution. A drawback of the Gaussian sketch is that computation of the sketched data is quite demanding, taking $O(ndk)$ operations. Therefore, work has been done on designing more computationally efficient random projections. Woodruff (2014) gives an excellent survey of work in this area.

The Hadamard sketch is a structured random matrix (Ailon & Chazelle, 2009). The sketching matrix is formed as $S = \Phi HD/\sqrt{k}$, where Φ is a $k \times n$ matrix and H and D are both $n \times n$ matrices. The fixed matrix H is a Hadamard matrix of order n. A Hadamard matrix is a square matrix with elements that are either $+1$ or -1 and orthogonal rows. Although Hadamard matrices do not exist for all integers n, the source dataset can be padded with zeros so that a conformable Hadamard matrix is available. The matrix D is a diagonal matrix whose n diagonal entries are independent Rademacher random variables. The random matrix Φ subsamples k rows of H with replacement. The structure of the Hadamard sketch allows for fast matrix multiplication, reducing calculation of the sketched dataset to $O(nd \log k)$ operations.

The Clarkson–Woodruff sketch is a sparse random matrix (Clarkson & Woodruff, 2013). The projection can be represented as the product of two independent random matrices, $S = \Gamma D$, where Γ is a random $k \times n$ matrix and D is a random $n \times n$ matrix. The matrix Γ is initialized as a matrix of zeros. Independently in each column, one element is selected and set to $+1$. The matrix D is a diagonal matrix whose n diagonal entries are independent Rademacher random variables. The sparsity of the Clarkson–Woodruff sketch speeds up matrix multiplication, decreasing the complexity of generating the sketched dataset to $O(nd)$.

3. Gaussian sketching

3.1. *Complete sketching*

The Gaussian sketch is mathematically tractable, and it is possible to establish a number of exact finite-sample results regarding the performance of the sketching estimators. In this section we derive the distribution of β_S in the case where a Gaussian sketch is used. As mentioned

previously, all results treat y and X as fixed. The variability in β_S is solely due to the use of the random sketching matrix S. Let $(\tilde{y}_j, \tilde{x}_j^T)$ $(j = 1, \ldots, k)$ refer to the jth row of the sketched data matrix $\tilde{A} = (\tilde{y}, \tilde{X})$. Similarly, let s_j^T denote the jth row of the sketching matrix S. The sketched dataset consists of k random units $(\tilde{y}_j, \tilde{x}_j^T)$ $(j = 1, \ldots, k)$. The jth sketched response is given by $\tilde{y}_j = s_j^T y$, and the jth sketched predictor is calculated as $\tilde{x}_j^T = s_j^T X$ $(j = 1, \ldots, k)$. The k sketched instances are independently distributed, because rows of the sketching matrix are independent.

It can be shown that the joint distribution of the sketched data, $p(\tilde{y} \mid \tilde{X}, y, X)\, p(\tilde{X} \mid y, X)$, has the structure of a hierarchical Gaussian linear model. The sketched dataset has a multivariate normal distribution, conditional on the source dataset. This is because the sketched dataset can be expressed as a linear combination of Gaussian random variables. Specifically, row j in the sketched dataset is $(\tilde{y}_j, \tilde{x}_j^T) = s_j^T A$. Given the source dataset $A = (y, X)$, $A^T s_j$ is a linear combination of independent Gaussians as $s_j \sim N(0, I_d/k)$, and so $(\tilde{y}_j, \tilde{x}_j^T)$ must be jointly normally distributed, conditional on the source data $A = (y, X)$. It is easily shown that the conditional joint distribution of the sketched responses and predictors is then

$$\begin{pmatrix} \tilde{y}_j \\ \tilde{x}_j \end{pmatrix} \,\Big|\, y, X \sim N\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \frac{1}{k} \begin{pmatrix} y^T y & y^T X \\ X^T y & X^T X \end{pmatrix} \right\} \quad (j = 1, \ldots, k).$$

From standard results on the multivariate normal distribution, it follows that the conditional distribution of $\tilde{y}_j$ given $\tilde{x}_j$ is also normal with conditional mean $E_S(\tilde{y}_j \mid \tilde{x}_j, y, X) = \tilde{x}_j^T \beta_F$. The subscript S is used with the expectation operator to emphasize that the only random quantity is the sketching matrix. The conditional distribution of $\tilde{y}_j$ given the sketched predictors $\tilde{x}_j$ and the source dataset (y, X) is

$$\tilde{y}_j \mid \tilde{x}_j, y, X \sim N\left(\tilde{x}_j^T \beta_F, \frac{\mathrm{RSS}_F}{k}\right) \quad (j = 1, \ldots, k).$$

This is the exact form of a standard Gaussian linear model, where the regression coefficient is β_F and the conditional variance is RSS_F/k. The distribution $p(\tilde{X} \mid y, X)$ is easily obtained as the marginal distribution of $\tilde{x}_j$ is also multivariate normal,

$$\tilde{x}_j \mid y, X \sim N(0, X^T X/k) \quad (j = 1, \ldots, k).$$

A Gaussian sketch effectively simulates a series of observations from a Gaussian linear model parameterized in terms of β_F and RSS_F, where the design matrix has a matrix normal distribution. The distribution of β_S conditional on the sketched predictors $\tilde{X}$ follows immediately from standard results on linear models (Searle, 1997, Ch. 3). To obtain the marginal distribution of β_S, it is necessary to integrate over the random sketched design matrix $\tilde{X}$. Using properties of the normal distribution (Eaton, 2007), it is possible to show that $(\tilde{X}^T \tilde{X}) \mid y, X \sim \mathrm{Wis}(k, X^T X/k)$. Hence,

$$(\tilde{X}^T \tilde{X})^{-1} \mid y, X \sim \mathrm{IW}\{k, k(X^T X)^{-1}\},$$

where IW denotes the inverse Wishart distribution. The marginal distribution of β_S can then be described using the normal inverse Wishart distribution (Gelman et al., 2014, p. 73). The following theorem characterizes the distribution of β_S under the Gaussian sketch.

THEOREM 1. *Suppose β_S is computed using a Gaussian sketch and that $k \geqslant p$. Then:*

(i) *the conditional distribution of β_S is*

$$\beta_S \mid \tilde{X}, y, X \sim N\left\{\beta_F, \frac{\mathrm{RSS}_F}{k}(\tilde{X}^T \tilde{X})^{-1}\right\};$$

(ii) *the marginal distribution of* β_S *is*

$$\beta_S \mid y, X \sim \text{Student}\left\{\beta_F,\ \frac{\text{RSS}_F}{k-p+1}(X^TX)^{-1},\ k-p+1\right\}.$$

For a proof see the Supplementary Material.

An immediate consequence of (i) is the ability to generate exact confidence intervals for the elements of β_S, an approach that does not seem to have been considered in the existing literature. The variance of β_S,

$$\text{var}(\beta_S \mid y, X) = \frac{\text{RSS}_F}{(k-p-1)}(X^TX)^{-1}, \tag{3}$$

is not dependent on the compression ratio k/n. Although RSS_F can be expected to grow linearly with n, this will generally be counterbalanced by $(X^TX)^{-1}$ decreasing linearly with n.

3.2. *Partial sketching*

Partial sketching was first proposed by Dhillon et al. (2013) using uniform subsampling, and was later studied for general sketches by Pilanci & Wainwright (2016). Existing results on partial sketching highlight that the model sum of squares influences the approximation error of the partial sketching estimator β_P. It is easy to see that the variance of the partial sketching estimator will not be a function of the residual sum of squares. From the normal equations it follows that $X^Ty = X^TX\beta_F$. Using this property, we see that conditional on y and X, the variance of the random linear combination $\beta_P = (X^TS^TSX)^{-1}X^Ty = (X^TS^TSX)^{-1}X^TX\beta_F$ will be a function of the covariates X and the fitted values $X\beta_F$. The residual vector has no influence on the variance of the partial sketching estimator, and as such the variance of β_P will not be related to the residual sum of squares. This suggests that when the noise level is high, partial sketching may become preferable to complete sketching (Dhillon et al., 2013; Becker et al., 2015).

The hierarchical model for complete sketching provides an intuitive statistical perspective on the mechanics of the algorithm. Partial sketching seems to lack a similar conceptual device. The least squares coefficients can be represented as the solution to the linear system of equations $X^TXb = X^Ty$. Partial sketching simply returns the solution, b, to the approximate linear system $\tilde{X}^T\tilde{X}b = X^Ty$. Lacking a convenient representation for the estimator, we must proceed in a more pedestrian manner. The mean squared error of the estimator β_P can be determined using only mean and variance information, and this will be the goal for now. The key observation is that $(\tilde{X}^T\tilde{X})^{-1} \mid y, X \sim \text{IW}\{k,\ k(X^TX)^{-1}\}$. Conditional on y and X, the estimator $\beta_P = (\tilde{X}^T\tilde{X})^{-1}X^Ty$ is a linear combination of the elements of an inverse Wishart random matrix. However, this is a nonstandard distribution, and it is difficult to express directly the distribution function of β_P. Despite this obstacle, it is straightforward to determine the mean and variance of β_P. From properties of the inverse Wishart distribution, it can be seen that the partial sketching estimator is biased, with mean

$$E_S(\beta_P \mid y, X) = \frac{k}{(k-p-1)}\beta_F,$$

where it is assumed that $k > p+3$. This motivates an alternative unbiased estimator

$$\beta_P^* = \frac{(k-p-1)}{k}(\tilde{X}^T\tilde{X})^{-1}X^Ty = \frac{(k-p-1)}{k}\beta_P.$$

Determining the variance of β_{P} and the unbiased β_{P}^* is a more lengthy computation, which is given in the Supplementary Material. The variance of the unbiased estimator β_{P}^* is

$$\mathrm{var}(\beta_{\mathrm{P}}^* \mid y, X) = \frac{(k-p-1)}{(k-p)(k-p-3)} \left\{ \mathrm{MSS}_{\mathrm{F}} (X^{\mathrm{T}} X)^{-1} + \frac{k-p+1}{k-p-1} \beta_{\mathrm{F}} \beta_{\mathrm{F}}^{\mathrm{T}} \right\}. \tag{4}$$

By making a connection with method-of-moments estimation it is possible to establish asymptotic normality of both β_{P} and β_{P}^* as k tends to infinity. This motivates the construction of approximate confidence intervals. As the exact variance is unknown, we propose the following estimator of $\mathrm{var}(\beta_{\mathrm{P}}^* \mid y, X)$ using the sketched model sum of squares $\mathrm{MSS}_{\mathrm{S}}$:

$$\frac{(k-p-1)}{(k-p)(k-p-3)} \left\{ \left(\frac{k-p-1}{k} \right) \mathrm{MSS}_{\mathrm{S}} (\tilde{X}^{\mathrm{T}} \tilde{X})^{-1} + \beta_{\mathrm{P}}^* \beta_{\mathrm{P}}^{*\mathrm{T}} \right\}.$$

3.3. *Relative efficiency*

The relative efficiencies of complete and partial sketching are also of interest. As the plug-in estimator β_{P} has a greater mean squared error than β_{P}^*, it will not be considered in this subsection. The performance of the complete sketching estimator β_{S} and the unbiased partial sketching estimator β_{P}^* will be compared in terms of mean squared error. As both β_{S} and β_{P}^* are unbiased, the mean squared errors can be computed using $\mathrm{var}(\beta_{\mathrm{S}} \mid y, X)$ and $\mathrm{var}(\beta_{\mathrm{P}}^* \mid y, X)$. Comparing (3) and (4), it can be seen that the variance of β_{P}^* is dependent on $\mathrm{MSS}_{\mathrm{F}}$ whereas the variance of β_{S} is dependent on $\mathrm{RSS}_{\mathrm{F}}$. This suggests that the signal-to-noise ratio in the source dataset will be an influential factor in determining which estimator is more efficient. In the Supplementary Material it is shown that for $k > p + 3$ the relative efficiency can be bounded in terms of the signal-to-noise ratio

$$\frac{R_{\mathrm{F}}^2}{1 - R_{\mathrm{F}}^2} \leqslant \frac{E_{\mathrm{S}}(\|\beta_{\mathrm{P}}^* - \beta_{\mathrm{F}}\|_2^2 \mid y, X)}{E_{\mathrm{S}}(\|\beta_{\mathrm{S}} - \beta_{\mathrm{F}}\|_2^2 \mid y, X)} \leqslant \frac{2(k-p-1)}{(k-p-3)} \frac{R_{\mathrm{F}}^2}{1 - R_{\mathrm{F}}^2}.$$

When R_{F}^2 is close to 1, complete sketching can be orders of magnitude more efficient than partial sketching; and when R_{F}^2 is close to 0, partial sketching can be orders of magnitude more efficient than complete sketching.

3.4. *Combined estimator*

So far we have assumed that an analyst must choose between one of the two methods; but obtaining both β_{P}^* and β_{S} from a single sketch is computationally cheap and may be an attractive strategy. The most demanding operation with the sketched data is calculating $(\tilde{X}^{\mathrm{T}} \tilde{X})^{-1}$. Given this quantity, it is economical to compute both β_{S} and β_{P}^*. Becker et al. (2015) mentioned that they were investigating such a strategy, but did not give any details. Our development of a combined estimator is motivated by the fact that, even when using a single sketch $(\tilde{y}, \tilde{X})$, the two estimators are uncorrelated, i.e., $\mathrm{cov}(\beta_{\mathrm{P}}^*, \beta_{\mathrm{S}} \mid y, X) = 0$. This is established in the Supplementary Material by taking iterated expectations and using the hierarchical model from § 3.1. A simple strategy is then to take a weighted combination of β_{S} and β_{P}^*. A combined estimator β_{C} can be defined as

$$\beta_{\mathrm{C}} = \phi \beta_{\mathrm{S}} + (1 - \phi) \beta_{\mathrm{P}}^*$$

for some $0 \leqslant \phi \leqslant 1$. The value of ϕ that minimizes the mean squared error is $\phi_{\mathrm{opt}} = \mathrm{tr}\{\mathrm{var}(\beta_{\mathrm{P}}^*) \mid y, X\} / [\mathrm{tr}\{\mathrm{var}(\beta_{\mathrm{P}}^* \mid y, X)\} + \mathrm{tr}\{\mathrm{var}(\beta_{\mathrm{S}} \mid y, X)\}]$. Use of the weighted estimator is expected to be

most beneficial when the signal-to-noise ratio is moderate, i.e., $R_{\mathrm{F}}^2 \approx 0.5$. When the signal-to-noise ratio is either very high or very low, there is little advantage in using the weighted estimator, as either the complete or the partial estimator will dominate.

3.5. *One-step correction*

As noted by a referee, the combined estimator is related to another strategy in the sketching literature for improving β_{S}. Dhillon et al. (2013) and Pilanci & Wainwright (2016) proposed a refinement procedure that uses gradient information from the source dataset. The one-step corrected estimator is defined as

$$\beta_{\mathrm{H}} = \beta_{\mathrm{S}} + (\tilde{X}^{\mathrm{T}}\tilde{X})^{-1}X^{\mathrm{T}}(y - X\beta_{\mathrm{S}}) = \{I - (\tilde{X}^{\mathrm{T}}\tilde{X})^{-1}X^{\mathrm{T}}X\}\beta_{\mathrm{S}} + (\tilde{X}^{\mathrm{T}}\tilde{X})^{-1}X^{\mathrm{T}}y. \tag{5}$$

Now the least squares solution β_{F} satisfies $X^{\mathrm{T}}(y - X\beta_{\mathrm{F}}) = 0$, so

$$\beta_{\mathrm{F}} = \beta_{\mathrm{F}} + (\tilde{X}^{\mathrm{T}}\tilde{X})^{-1}X^{\mathrm{T}}(y - X\beta_{\mathrm{F}}) = \{I - (\tilde{X}^{\mathrm{T}}\tilde{X})^{-1}X^{\mathrm{T}}X\}\beta_{\mathrm{F}} + (\tilde{X}^{\mathrm{T}}\tilde{X})^{-1}X^{\mathrm{T}}y. \tag{6}$$

Subtracting (6) from (5) gives the following expression for the error:

$$\beta_{\mathrm{H}} - \beta_{\mathrm{F}} = \{I - (\tilde{X}^{\mathrm{T}}\tilde{X})^{-1}X^{\mathrm{T}}X\}(\beta_{\mathrm{S}} - \beta_{\mathrm{F}}). \tag{7}$$

The one-step estimator can be interpreted as a single step of the iterative Hessian sketch proposed by Pilanci & Wainwright (2016), initialized at β_{S}. Setting $\tilde{H} = (\tilde{X}^{\mathrm{T}}\tilde{X})^{-1}X^{\mathrm{T}}X$, it follows from (7) and Theorem 1(i) that

$$E_{\mathrm{S}}(\|\beta_{\mathrm{H}} - \beta_{\mathrm{F}}\|_2^2 \mid y, X) = E_{\tilde{X}}\Big[\mathrm{tr}\big\{k^{-1}\mathrm{RSS}_{\mathrm{F}}(\tilde{X}^{\mathrm{T}}\tilde{X})^{-1}(I - \tilde{H})^{\mathrm{T}}(I - \tilde{H})\big\}\Big]. \tag{8}$$

The key terms in (8) are the random matrices $(\tilde{X}^{\mathrm{T}}\tilde{X})^{-1}$ and $\tilde{H} = (\tilde{X}^{\mathrm{T}}\tilde{X})^{-1}X^{\mathrm{T}}X$. As $(\tilde{X}^{\mathrm{T}}\tilde{X})^{-1} \mid y, X \sim \mathrm{IW}\{k, k(X^{\mathrm{T}}X)^{-1}\}$, it is possible to evaluate the expectation in (8) using the first, second and third moments of the inverse Wishart distribution. The exact expression for (8) is lengthy and is given in the Supplementary Material. The main conclusions are that the one-step estimator β_{H} can have a larger mean squared error than β_{S} when the ratio k/p of sketch size to number of variables is close to 1. As k/p increases, the one-step estimator becomes more efficient than both β_{S} and β_{C} with the optimal weight ϕ_{opt}. The relative efficiency of β_{C} to β_{S} is at most 2. The relative efficiency of β_{H} to β_{S} can be much higher, provided that k/p is sufficiently large.

4. Asymptotics

4.1. *Preliminaries*

Finite-sample distributions of random projection estimators can be mathematically intractable, and thus asymptotic analysis can be a powerful tool (Diaconis & Freedman, 1984; Li et al., 2006). It is very difficult to establish meaningful finite-sample results for the Hadamard and Clarkson–Woodruff sketches, as they are discrete distributions over an enormous combinatorial space. Instead, it is useful to study the large-n distribution of the estimators β_{S} and β_{P} to obtain an interpretable expression.

As β_{F} is the estimand in sketching algorithms, conditioning on the source data is required in the asymptotic analysis. To elaborate, let $A_{(n)} = (y_{(n)}, X_{(n)})$ represent the $n \times d$ source data matrix of full column rank. Any source data matrix $A_{(n)}$ has a set of associated least squares coefficients,

which will be denoted by $\beta_{\mathrm{F}}^{(n)}$ here. The overall goal is to determine the asymptotic form of the distributions $p(\beta_{\mathrm{S}} \mid A_{(n)})$ and $p(\beta_{\mathrm{P}}^{*} \mid A_{(n)})$ for some arbitrary large dataset $A_{(n)}$. To take limits, we employ a fixed sequence of $n \times d$ datasets, all of rank d.

Some related work has been done by Ma et al. (2015), who developed Taylor series approximations for the bias and variance of data-aware sketched regression estimators, where the asymptotic expansion is taken in the sketch size k. In independent work, Dobriban & Liu (2019) examined the behaviour of data-oblivious sketching algorithms in the asymptotic regime where $k, d \to \infty$, using elements of random matrix theory. Our work is novel, as we study data-oblivious random projections in the regime where k and d are fixed, while taking limits in the number n of source observations.

4.2. *Sketching central limit theorem*

A central limit theorem for sparse sketching matrices with independent entries is given in Li et al. (2006). The Clarkson–Woodruff sketch and the Hadamard sketch have dependent entries, so we use a different method of proof. Under some regularity conditions, the Hadamard and Clarkson–Woodruff sketches produce sketched data that asymptotically have the same matrix normal distribution as under the Gaussian sketch.

The $k \times d$ random matrix $\tilde{A}$ is the output of a stochastic process governed by the fixed $n \times d$ source dataset $A_{(n)}$ and the distribution of the random $k \times n$ sketching matrix S. Each column of the sketched dataset is a linear combination of random vectors, the number of which increases with n. Under an assumption on the limiting leverage scores of the source data matrix, we can establish a central limit theorem for the sketched dataset. The leverage scores of the observations in the source data matrix have been identified as an important structural property of sketching algorithms (Mahoney & Drineas, 2016). Assumption 1 highlights their role in establishing asymptotic normality of the sketched data matrix.

Assumption 1. Let $A_{(n)} = U_{(n)} D_{(n)} V_{(n)}^{\mathrm{T}}$ be the singular value decomposition of the $n \times d$ source dataset, and let $u_{(n)i}^{\mathrm{T}}$ be the ith row in $U_{(n)}$. The maximum leverage score tends to zero, that is,

$$\lim_{n \to \infty} \max_{i=1,\ldots,n} \|u_{(n)i}\|_2^2 = 0.$$

Theorem 2 is the sketching central limit theorem. Its proof is given in the Supplementary Material.

THEOREM 2. *Consider a sequence of arbitrary $n \times d$ data matrices $A_{(n)}$, where d is fixed. Let $A_{(n)} = U_{(n)} D_{(n)} V_{(n)}^{\mathrm{T}}$ be the singular value decomposition of $A_{(n)}$, and let S be a $k \times n$ Hadamard or Clarkson–Woodruff sketching matrix where k is also fixed. Suppose that Assumption 1 is satisfied. Then, as n tends to infinity, the following convergence in distribution holds:*

$$\{\tilde{A} V_{(n)} D_{(n)}^{-1} \mid A_{(n)}\} \to \mathrm{MN}(0, I_k, I_d/k),$$

where MN *denotes the matrix normal distribution.*

4.3. *Sketching estimators*

The central limit theorem for the sketched data suggests that the results on β_{S} and β_{P} for the Gaussian sketch will also hold approximately for the Hadamard and Clarkson–Woodruff sketches

for large n. To establish convergence of the estimators it helps to make an extra assumption on the sequence of source datasets.

Assumption 2. We have that

$$\lim_{n\to\infty} n^{-1} \begin{pmatrix} y_{(n)}^{\mathrm{T}} y_{(n)} & y_{(n)}^{\mathrm{T}} X_{(n)} \\ X_{(n)}^{\mathrm{T}} y_{(n)} & X_{(n)}^{\mathrm{T}} X_{(n)} \end{pmatrix} = Q$$

for some positive-definite matrix Q.

The limiting matrix Q allows one to avoid specifying a probability model for the source dataset, without overcomplicating the mathematical analysis. Under Assumptions 1 and 2, it is possible to establish an asymptotic result for β_{S} and β_{P}.

THEOREM 3. *Suppose that Assumptions 1 and 2 hold, $k \geqslant p$, and β_{S} is computed using a Hadamard or Clarkson–Woodruff sketch. Let $(\tilde{X}^{\mathrm{T}}\tilde{X})^{+}$ denote the Moore–Penrose pseudo-inverse of $(\tilde{X}^{\mathrm{T}}\tilde{X})$. Let*

$$\tilde{C}_{(n)} = \frac{\mathrm{RSS}_{\mathrm{F}}^{(n)}}{k} (\tilde{X}^{\mathrm{T}}\tilde{X})^{+}, \quad C_{(n)} = \frac{\mathrm{RSS}_{\mathrm{F}}^{(n)}}{k-p+1} (X_{(n)}^{\mathrm{T}} X_{(n)})^{-1}.$$

Then, as $n \to \infty$, the following convergence results hold in distribution:

(i) $\{C_{(n)}^{-1/2}(\beta_{\mathrm{S}} - \beta_{\mathrm{F}}^{(n)}) \mid A_{(n)}\} \to \mathrm{Student}(0, I_p, k-p+1)$;
(ii) $\{\tilde{C}_{(n)}^{-1/2}(\beta_{\mathrm{S}} - \beta_{\mathrm{F}}^{(n)}) \mid A_{(n)}\} \to N(0, I_p)$.

The proof is given in the Supplementary Material. For large n we expect β_{S} to be approximately distributed as per Theorem 1 for both the Hadamard and the Clarkson–Woodruff sketches.

It is harder to establish a comparable limit theorem for β_{P}^{*}, because of the nonstandard distribution of β_{P}^{*} when a Gaussian sketch is used. Instead, we wish to show that the partial sketching estimators under the Hadamard and Clarkson–Woodruff sketches have similar mean and variance properties to the Gaussian partial sketching estimator. Convergence in moments can be established given a stability condition on the singular values of the sketched data matrix.

Assumption 3. The sequence of source datasets is such that $E_{\mathrm{S}}\{1/\sigma_{\min}^{4}(n^{-1}\tilde{X}^{\mathrm{T}}\tilde{X}) \mid y, X\}$ is finite for large enough n, where $\sigma_{\min}(\cdot)$ denotes the minimum singular value of a matrix.

This additional regularity condition enables a formal limit theorem regarding the moments of β_{P}^{*} to be established.

THEOREM 4. *Suppose that Assumptions 1–3 hold, $k > p+3$, and β_{P}^{*} is computed using a Hadamard or Clarkson–Woodruff sketch. Let*

$$C_{(n)} = \frac{(k-p-1)}{(k-p)(k-p-3)} \left\{ \mathrm{MSS}_{\mathrm{F}}^{(n)} (X_{(n)}^{\mathrm{T}} X_{(n)})^{-1} + \frac{k-p+1}{k-p-1} \beta_{\mathrm{F}}^{(n)} \beta_{\mathrm{F}}^{(n)\mathrm{T}} \right\}.$$

Then, as $n \to \infty$:

(i) $E_{\mathrm{S}}\{\beta_{\mathrm{P}}^{*} - \beta_{\mathrm{F}}^{(n)} \mid A_{(n)}\} \to 0$;
(ii) $\mathrm{var}_{\mathrm{S}}\{C_{(n)}^{-1/2}(\beta_{\mathrm{P}}^{*} - \beta_{\mathrm{F}}^{(n)}) \mid A_{(n)}\} \to I_p$.

The proof is given in the Supplementary Material. This theorem suggests that the conditional bias and variance of β_{P}^* under the Clarkson–Woodruff and Hadamard sketches should be approximately equal to those under the Gaussian sketch. The results here are meant to provide useful heuristics for assessing the uncertainty associated with the output of the randomized approximation algorithm. There is a need to quantify the approximation error of sketching algorithms and communicate it to end users (Lopes et al., 2018), for which the asymptotic results developed in this section may be helpful.

5. Unconditional results

The previous analysis treated the source dataset as fixed to isolate the approximation error introduced by the random projection. When sketching is used for statistical inference, the hierarchical model of § 3.1 can be extended to include a source of variation at the population level. We take the design matrix X to be fixed and treat the response y as random. The assumed data-generating process is $y = X\beta_0 + \varepsilon$, where ε is a vector of n independent and identically distributed random variables with mean zero and variance σ^2. Let γ^2 represent the average mean function sum of squares, so $\gamma^2 = \|X\beta_0\|_2^2/n$. As shown in Searle (1997), at the population level the ordinary least squares estimator satisfies $E_y(\beta_{\mathrm{F}} \mid X) = \beta_0$, $\mathrm{var}_y(\beta_{\mathrm{F}} \mid X) = \sigma^2(X^{\mathrm{T}}X)^{-1}$, $E_y(\mathrm{RSS}_{\mathrm{F}} \mid X) = (n-p)\sigma^2$ and $E_y(\mathrm{MSS}_{\mathrm{F}} \mid X) = p\sigma^2 + n\gamma^2$. Taking iterated expectations, it can be seen that the Gaussian sketch gives an unbiased estimator of the population parameter β_0: $E_y(\beta_{\mathrm{S}} \mid X) = E_y\{E_{\mathrm{S}}(\beta_{\mathrm{S}} \mid y, X)\} = E_y(\beta_{\mathrm{F}} \mid X) = \beta_0$. The same argument shows that $E_y(\beta_{\mathrm{P}}^* \mid X) = \beta_0$. In the Supplementary Material, we use the law of total variance to determine the unconditional variances

$$\mathrm{var}_y(\beta_{\mathrm{S}} \mid X) = \sigma^2(X^{\mathrm{T}}X)^{-1} + \frac{(n-p)\sigma^2}{(k-p-1)}(X^{\mathrm{T}}X)^{-1},$$

$$\mathrm{var}_y(\beta_{\mathrm{P}}^* \mid X) = \sigma^2(X^{\mathrm{T}}X)^{-1} + \frac{(k-p-1)}{(k-p)(k-p-3)}\Bigg[(p\sigma^2 + n\gamma^2)(X^{\mathrm{T}}X)^{-1} + \frac{k-p+1}{k-p-1}\{\sigma^2(X^{\mathrm{T}}X)^{-1} + \beta_0\beta_0^{\mathrm{T}}\}\Bigg].$$

For large n, the most significant term in the unconditional variance of β_{S} is $n\sigma^2(X^{\mathrm{T}}X)^{-1}$. The dominating term in the unconditional variance of β_{P}^* is $n\gamma^2(X^{\mathrm{T}}X)^{-1}$, a function of the average model sum of squares γ^2. We reach conclusions similar to those of the conditional analysis in § 3.3, in that β_{S} is expected to be more efficient when the signal-to-noise ratio is high, while β_{P}^* is expected to be more efficient when the signal-to-noise ratio is low. Under Assumptions 1–3, the variance expressions give asymptotic approximations for the Hadamard and Clarkson–Woodruff projections. These results can be extended to account for more complicated error models on ϵ if it is still possible to determine $E_y(\beta_{\mathrm{F}} \mid X)$, $\mathrm{var}_y(\beta_{\mathrm{F}} \mid X)$, $E_y(\mathrm{RSS}_{\mathrm{F}} \mid X)$ and $E_y(\mathrm{MSS}_{\mathrm{F}} \mid X)$. Raskutti & Mahoney (2016) provides further results on the performance of sketching estimators from an inferential perspective.

6. Data application

6.1. *Human leukocyte antigen locus dataset*

We compare the performance of the sketching estimators on a genetic dataset from the UK Biobank database. We use a small extract of the data in Astle et al. (2016). The selected response

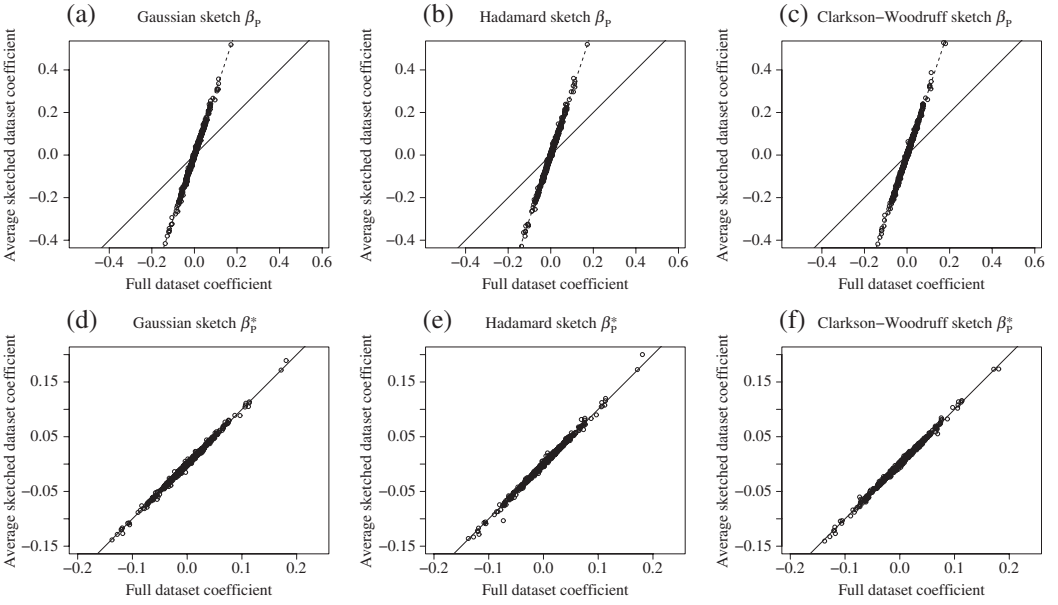

Fig. 1. Bias of partial sketching estimators on the HLA dataset: panels (a)–(c) show results for β_P and panels (d)–(f) results for the bias-corrected estimator β_P^*; mean estimates are plotted against the true values. In this scenario $n = 132\,353$, $p = 1000$ and $k = 1500$. The solid line in each panel is the identity line, and the dashed line in panels (a)–(c) represents the theoretical bias factor.

variable is mean red cell volume, taken from the full blood count assay and with adjustments for various technical and environmental covariates. Genome-wide imputed genotype data in expected allele dose format were available on $n = 132\,353$ study subjects (Howie et al., 2009; Bycroft et al., 2018). We consider 1000 genetic variants in the human leukocyte antigen, HLA, region of chromosome 6, selected so that no pair of variants had squared Pearson correlation of posterior expected allele doses greater than 0.8. We chose to focus on this region because many associations have been discovered in a genome-wide scan using univariable models; these associations were with variants having different allele frequencies, which suggests multiple distinct causal variants in the region. The aim is to perform a multivariable regression analysis to obtain variant effect size estimates that are conditional on the other variants in the region.

An early theoretical finding was that the partial sketching estimator β_P is biased. One thousand sketches were taken to estimate the bias $E_S(\beta_P - \beta_F \mid y, X)$ with $k = 1500$. We also computed the bias-corrected estimator β_P^* in each replication. Figure 1 plots the average value of the estimators against the true value of the least squares coefficient using the full dataset. The top row shows results for β_P, and the bottom row shows results for β_P^*. The left, middle and right columns display results for the Gaussian, Hadamard and Clarkson–Woodruff sketches, respectively. The solid line in each panel is the identity line. The dashed line in the top row represents the theoretical bias, with slope $k/(k - p - 1)$.

The results in panels (a)–(c) show that β_P is biased for each of the random projections. The bias closely matches the theoretical factor. Panels (d)–(f) show that the adjusted estimator β_P^* appears to be unbiased, with the mean values falling close to the identity line.

We also compared the complete and partial sketching estimators in terms of mean squared error and coverage of confidence intervals at $k = 1500$ and $k = 10\,000$. Moreover, we compared the data-oblivious sketches to simple uniform subsampling with replacement. Table 1 reports the mean squared error for each of the estimators. The signal-to-noise ratio is quite low for

Table 1. *Mean squared errors of sketching estimators on the HLA dataset*

	$k = 1500$			$k = 10\,000$		
	β_S	β_P	β_P^*	β_S	β_P	β_P^*
Gaussian	238 (3)	39 (0.7)	3.8 (0.08)	13.3 (0.17)	0.28 (0.004)	0.21 (0.002)
Hadamard	238 (4)	39 (0.7)	3.8 (0.07)	12.5 (0.16)	0.26 (0.003)	0.20 (0.002)
Clarkson–Woodruff	241 (3)	38 (0.8)	4.0 (0.05)	13.2 (0.16)	0.28 (0.004)	0.21 (0.002)
Uniform	375 (15)	105 (7.6)	10.7 (0.55)	13.8 (0.20)	0.38 (0.007)	0.29 (0.005)

Table 2. *Coverage of confidence intervals; the largest standard error is* 0.004

	HLA $k = 1500$		HLA $k = 10\,000$		Flights $k = 1500$	
	β_S	β_P^*	β_S	β_P^*	β_S	β_P^*
Gaussian	0.950	0.953	0.950	0.951	0.948	0.951
Hadamard	0.949	0.949	0.954	0.954	0.950	0.948
Clarkson–Woodruff	0.947	0.952	0.951	0.950	0.948	0.947

Table 3. *Timings for sketching: average times to compute the sketched dataset $\tilde{A} = SA$, in seconds*

	HLA		Flights
	$k = 1500$	$k = 10\,000$	$k = 5000$
Gaussian	522	3479	404
Hadamard	57	65	5.8
Clarkson–Woodruff	5.3	5.4	0.2

this dataset, with $R_F^2 = 0.02$. We expect that partial sketching will be much more efficient than complete sketching on this dataset given the low signal-to-noise ratio. The simulation results support this prediction, with β_P^* having a mean squared error roughly 60 times smaller than β_S at both values of k. The results are very similar for each of the random projections, suggesting that the asymptotic approximations are reasonable for this dataset. For $k = 1500$, the mean squared error of β_P is approximately 10 times that of β_P^*. For $k = 10\,000$ there is less of a difference, as the ratio $k/(k - p - 1)$ is closer to 1.

Table 2 summarizes the coverage of 95% confidence intervals for the sketching estimators. We report the overall proportion of intervals containing the true value of the least squares estimate β_F over the 250 sketches and $p = 1000$ coefficients. The observed coverage is close to the nominal level of 0.95 at both levels of k. The different random projections give very similar results, suggesting that the use of asymptotic approximations is again reasonable for this dataset. The intervals for the Hadamard sketch appear to be slightly conservative at $k = 10\,000$.

Table 3 reports the average sketching times for the data-oblivious sketches. We computed 10 sketches using each projection. The Gaussian sketch is an order of magnitude slower than the Hadamard projection and two orders of magnitude slower than the Clarkson–Woodruff sketch.

6.2. *New York flights dataset*

We also evaluated the sketching algorithms on the New York flights dataset available in the `R` (R Development Core Team, 2021) package `nycflights13` (Wickham, 2014). Arrival delay was taken as the response, and departure delay, distance, departure time, origin, and month and

Table 4. *Mean squared errors of sketching estimators (with standard errors in parentheses) on the flights dataset with* $k = 5000$

	β_{S}	β_{P}	β_{P}^{*}
Gaussian	60 (2)	14900 (400)	14900 (400)
Hadamard	63 (2)	14800 (500)	13900 (400)
Clarkson–Woodruff	66 (2)	15000 (500)	13800 (400)
Uniform	64 (2)	14600 (500)	14600 (400)

day were chosen to be the covariates. Rows of the dataset with missing data were omitted, so that we were left with $n = 327\,346$ and $d = 47$. The goal is to compare the accuracy of the various sketches on real data rather than to build a statistical model for the flights dataset. We compare the mean squared error of the estimators and the coverage of confidence intervals for $k = 5000$. In contrast to the HLA dataset, the flights dataset has a very high R_{F}^2 value of 0.99. We took 500 sketches to compare complete and partial sketching. See Table 4 for details.

7. Discussion

In recent years work has been done to adapt sketching methods for statistical inference in large datasets, building upon the worst-case bounds developed in the computer science literature. Geppert et al. (2017) and Bardenet & Maillard (2015) investigated sketching algorithms for Bayesian regression, and derived bounds on the difference between the sketched posterior distribution and the full-data posterior distribution. Only complete sketching was considered in those works. The results on the advantages of partial sketching in this paper could motivate adaptations that make use of the exact marginal associations $X^{\mathrm{T}}y$. Sketching ideas have been used to develop methods for approximate nonlinear regression (Banerjee et al., 2013; Avron et al., 2014). The goodness of fit of the model may also influence the relative efficiency of different sketching algorithms in more complex regression tasks. A related branch of work uses random projections to reduce the number of predictors in regression and classification problems (Shah & Meinshausen, 2018; Guhaniyogi & Dunson, 2015; Cannings & Samworth, 2017).

Acknowledgement

This research was conducted using the UK Biobank resource. Richardson was supported by the UKRI Medical Research Council and the Alan Turing Institute. Astle was supported by NHS Blood and Transplant and the National Institute for Health Research Blood and Transplant Research Unit. Many thanks to Rajen Shah for helpful discussions, and the reviewers and associate editor for insightful comments that have improved the quality of the manuscript.

Supplementary material

Supplementary material available at *Biometrika* online includes proofs of all the theorems.

References

Ailon, N. & Chazelle, B. (2009). The fast Johnson Lindenstrauss transform and approximate nearest neighbors. *SIAM J. Comp.* **39**, 302–22.

ASTLE, W. J., ELDING, H., JIANG, T., ALLEN, D., RUKLISA, D., MANN, A. L., MEAD, D., BOUMAN, H., RIVEROS-MCKAY, F., KOSTADIMA, M. A. et al. (2016). The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* **167**, 1415–29.

AVRON, H., NGUYEN, H. & WOODRUFF, D. (2014). Subspace embeddings for the polynomial kernel. In *Proc. 27th Int. Conf. Neural Information Processing Systems (NIPS'14)*, vol. 2. Cambridge, Massachusetts: MIT Press, pp. 2258–66.

BANERJEE, A., DUNSON, D. B. & TOKDAR, S. T. (2013). Efficient Gaussian process regression for large datasets. *Biometrika* **100**, 75–89.

BARDENET, R. & MAILLARD, O.-A. (2015). A note on replacing uniform subsampling by random projections in MCMC for linear regression of tall datasets. *HAL* preprint 01248841, https://hal.archives-ouvertes.fr/hal-01248841.

BECKER, S., KAWAS, B., PETRIK, M. & RAMAMURTHY, K. (2015). Robust partially-compressed least-squares. *arXiv:* 1510.04905v1.

BYCROFT, C., FREEMAN, C., PETKOVA, D., BAND, G., ELLIOTT, L. T., SHARP, K., MOTYER, A., VUKCEVIC, D., DELANEAU, O., O'CONNELL, J. et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–9.

CANNINGS, T. I. & SAMWORTH, R. J. (2017). Random-projection ensemble classification. *J. R. Statist. Soc.* B **79**, 959–1035.

CLARKSON, K. L. & WOODRUFF, D. P. (2013). Low rank approximation and regression in input sparsity time. In *Proc. 45th Annual ACM Sympos. Theory of Computing (STOC'13)*. New York: Association for Computing Machinery, pp. 81–90.

CORMODE, G. (2011). Sketch techniques for approximate query processing. In *Foundations and Trends in Databases*. Hanover, Massachusetts: NOW Publishers.

DHILLON, P., LU, Y., FOSTER, D. P. & UNGAR, L. (2013). New subsampling algorithms for fast least squares regression. In *Proc. 26th Int. Conf. Neural Information Processing Systems (NIPS'13)*. Red Hook, New York: Curran Associates, pp. 360–8.

DIACONIS, P. & FREEDMAN, D. (1984). Asymptotics of graphical projection pursuit. *Ann. Statist.* **12**, 793–815.

DOBRIBAN, E. & LIU, S. (2019). Asymptotics for sketching in least squares regression. In *Advances in Neural Information Processing Systems 32 (Proc. NeurIPS 2019)*. La Jolla, California: Neural Information Processing Systems Foundation, pp. 3675–85.

DRINEAS, P., MAHONEY, M. W. & MUTHUKRISHNAN, S. (2006). Sampling algorithms for ℓ_2 regression and applications. In *Proc. 17th Annual ACM-SIAM Sympos. Discrete Algorithm (SODA '06)*. Philadelphia: Society for Industrial and Applied Mathematics, pp. 1127–36.

EATON, M. L. (2007). *Multivariate Statistics: A Vector Space Approach*. Beachwood, Ohio: Institute of Mathematical Statistics.

GELMAN, A., CARLIN, J. B., STERN, H. S., DUNSON, D. B., VEHTARI, A. & RUBIN, D. B. (2014). *Bayesian Data Analysis*. Boca Raton, Florida: Chapman & Hall, 3rd ed.

GEPPERT, L. N., ICKSTADT, K., MUNTEANU, A., QUEDENFELD, J. & SOHLER, C. (2017). Random projections for Bayesian regression. *Statist. Comp.* **27**, 79–101.

GUHANIYOGI, R. & DUNSON, D. B. (2015). Bayesian compressed regression. *J. Am. Statist. Assoc.* **110**, 1500–14.

HALKO, N., MARTINSSON, P. G. & TROPP, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.* **53**, 217–88.

HOWIE, B. N., DONNELLY, P. & MARCHINI, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529.

LI, P., HASTIE, T. J. & CHURCH, K. W. (2006). Very sparse random projections. In *Proc. 12th ACM-SIGKDD Int. Conf. Knowledge Discovery and Data Mining*. New York: Association for Computing Machinery, pp. 287–96.

LOPES, M., WANG, S. & MAHONEY, M. (2018). Error estimation for randomized least-squares algorithms via the bootstrap. In *Proc. 35th Int. Conf. Machine Learning*, J. Dy & A. Krause, eds., vol. 80 of *Proceedings of Machine Learning Research*. PMLR, pp. 3217–26.

MA, P., MAHONEY, M. W. & YU, B. (2015). A statistical perspective on algorithmic leveraging. *J. Mach. Learn. Res.* **16**, 861–911.

MA, P. & SUN, X. (2015). Leveraging for big data regression. *WIREs Comp. Statist.* **7**, 70–6.

MAHONEY, M. (2011). Randomized algorithms for matrices and data. In *Foundations and Trends in Machine Learning*, vol. 3. Hanover, Massachusetts: NOW Publishers, pp. 123–224.

MAHONEY, M. & DRINEAS, P. (2016). Structural properties underlying high-quality randomized numerical linear algebra algorithms. In *Handbook of Big Data*, P. Buhlmann, P. Drineas, M. Kane & M. van de Laan, eds. Boca Raton, Florida: Chapman & Hall, pp. 137–54.

PILANCI, M. & WAINWRIGHT, M. J. (2016). Iterative Hessian sketch: Fast and accurate solution approximation for constrained least-squares. *J. Mach. Learn. Res.* **17**, 1842–79.

R DEVELOPMENT CORE TEAM (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.

RASKUTTI, G. & MAHONEY, M. W. (2016). A statistical perspective on randomized sketching for ordinary least-squares. *J. Mach. Learn. Res.* **17**, 7508–38.

Sarlos, T. (2006). Improved approximation algorithms for large matrices via random projections. In *47th Annual IEEE Sympos. Foundations of Computer Science (FOCS'06)*. New York: Institute of Electrical and Electronics Engineers, pp. 143–52.

Searle, S. R. (1997). *Linear Models*. New York: Wiley.

Shah, R. D. & Meinshausen, N. (2018). Min-wise hashing for large-scale regression and classification with sparse data. *arXiv:* 1308.1269v4.

Wickham, H. (2014). *nycflights13: Flights that Departed NYC in 2013*. R package version 0.1, available at https://cran.r-project.org/web/packages/nycflights13/.

Woodruff, D. P. (2014). Sketching as a tool for numerical linear algebra. In *Foundations and Trends in Theoretical Computer Science*, vol. 10. Hanover, Massachusetts: NOW Publishers, pp. 1–157.

[*Received on* 10 *March* 2017. *Editorial decision on* 27 *May* 2020]