Information and Inference: A Journal of the IMA (2021) **00**, 1–23
https://doi.org/10.1093/imaiai/iaab016

A projector-based approach to quantifying total and excess uncertainties for sketched linear regression

Jocelyn T. Chi[†]
Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA
[†]Corresponding author. Email: jtchi@math.ucla.edu

AND

Ilse C. F. Ipsen
Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA

[Received on 20 January 2021; revised on 07 June 2021; accepted on 12 June 2021]

Linear regression is a classic method of data analysis. In recent years, sketching—a method of dimension reduction using random sampling, random projections or both—has gained popularity as an effective computational approximation when the number of observations greatly exceeds the number of variables. In this paper, we address the following question: how does sketching affect the statistical properties of the solution and key quantities derived from it? To answer this question, we present a projector-based approach to sketched linear regression that is exact and that requires minimal assumptions on the sketching matrix. Therefore, downstream analyses hold exactly and generally for all sketching schemes. Additionally, a projector-based approach enables derivation of key quantities from classic linear regression that account for the combined model- and algorithm-induced uncertainties. We demonstrate the usefulness of a projector-based approach in quantifying and enabling insight on excess uncertainties and bias-variance decompositions for sketched linear regression. Finally, we demonstrate how the insights from our projector-based analyses can be used to produce practical sketching diagnostics to aid the design of judicious sketching schemes.

Keywords: expectation; variance; bias; mean squared error; predictive risk.

1. Introduction

Linear regression is a classic method of data analysis that is ubiquitous across numerous domains. In recent years, sketching—a method of dimension reduction using random sampling, random projections or a combination of both—has gained popularity as an effective computational approximation when the number of observations greatly exceeds the number of variables. In this paper, we address the following question: how does sketching affect the statistical properties of the solution and key statistical quantities derived from it?

To answer this question, we begin with the simplest class of linear regression problems, those of full column rank because they are well-posed, and regularization is not required to ensure the existence of a unique solution in exact arithmetic. We present a projector-based approach to sketched linear regression that is exact and that requires no additional assumptions on the sketching matrix. Consequently, downstream analyses derived from this formulation of the sketched solution hold exactly and generally for all sketching schemes, while accounting for both model- and algorithmic-induced uncertainties.

Our paper extends previous work on the combined model- and algorithm-induced uncertainties of the sketched solution to exact expressions that hold generally for *all* sketching schemes. Specifically, we extend existing work on the total expectation and variance of the sketched solution from specific sampling schemes [22, 23] to all sketching schemes. Due to the assumptions and limitations of a Taylor expansion approach to the solution in [22, 23], the expressions for the total uncertainties there are restricted to specific sampling schemes. By contrast, our expressions hold for many commonly used sketching schemes not covered by [22, 23]. These include sketching with fast Fourier Johnston–Lindenstrauss transforms (FJLTs), Gaussian random matrices and random row-mixing transformations followed by uniform sampling.

We demonstrate the usefulness of a projector-based approach in quantifying and enabling insight on excess uncertainties arising from the randomness in the sketching algorithm. We highlight this through geometric insights and interpretation for the excess bias and variance and analyses of total and excess bias-variance decompositions for sketched linear regression. Finally, we demonstrate how the insights from our projector-based analyses can be used to produce practical sketching diagnostics to aid the design of judicious sketching schemes.

1.1 *Related work*

Randomized sketching is a form of preconditioning. It was introduced in [29] for data-oblivious random projections but first applied to least squares problems in [8] and explicitly presented as a preconditioner for least squares problems in [1, 28]. Its many variants can be classified [35, Section 1] according to whether they achieve row compression [2, 8, 9, 17, 22, 23, 27, 28, 39], column compression [1, 18, 25, 35, 41] or both [26]. We focus on *row-sketched linear regression*, where the number of observations greatly exceeds the number of variables. We refer to this simply as *sketched linear regression*.

Since sketched linear regression has roots in theoretical computer science and numerical analysis, much emphasis has been on analyzing the error due to algorithmic randomization. Recent works have made progress toward a combined statistical and algorithmic perspective. These include criteria for quantifying prediction and residual efficiency [27], bootstrap estimates for estimating the combined uncertainty [20], approximate expressions for the total expectation and variance of some randomized sampling estimators [22, 23] and asymptotic analysis of randomized sampling estimators [24].

1.2 *Overview*

We present results in terms of two regimes. The first regime requires no assumptions on the sketching matrix beyond its dimensions. Consequently, these results hold generally for all sketching matrices and provide a worst-case analysis since they hold even for poor choices of sketching schemes.

The second regime presents results conditioned on rank preservation so that the sketched matrix has the same rank as the original design matrix $\mathbf{X}$. Rank preservation implies that the sketching scheme successfully preserves the most relevant information in the original response $\mathbf{y}$ and design matrix $\mathbf{X}$. Although these results require an additional assumption, conditioning on rank preservation enables further insights on how the sketching process affects the solution and other key statistical quantities. Thus, results from this second regime provide insights from an ideal-case analysis.

2. Sketched Linear Regression

We begin by setting some notation for the rest of this paper. We then review the exact and sketched linear regression problems, their solutions and other relevant quantities.

2.1 *Preliminaries*

Let $\mathbf{X} \in \mathbb{R}^{n\times p}$ be observed with $\text{rank}(\mathbf{X}) = p$. Since $\mathbf{X}$ has full column rank, its Moore–Penrose inverse is a left inverse so that

$$\mathbf{X}^\dagger = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \qquad \text{and} \qquad \mathbf{X}^\dagger\mathbf{X} = \mathbf{I}_p.$$

Let $\|\mathbf{X}\|_2$ denote the Euclidean operator norm of $\mathbf{X}$. The two-norm condition number of $\mathbf{X}$ with regard to left inversion is

$$\kappa_2(\mathbf{X}) \equiv \|\mathbf{X}\|_2\|\mathbf{X}^\dagger\|_2.$$

We additionally use $\|\cdot\|_2$ to denote the Euclidean vector norm for vectors. The use of $\|\cdot\|_2$ to denote either the Euclidean operator or vector norm will be clear from the context. Let $\mathbf{I}_n$ denote the $n \times n$ identity matrix, and let $\mathbf{0}$ and $\mathbf{1}$ denote the vectors of all zeros and ones, respectively. Their lengths will be clear from the context.

2.2 *The exact problem and solution*

Given an observed pair $\mathbf{y} \in \mathbb{R}^n$ and $\mathbf{X} \in \mathbb{R}^{n\times p}$ with $\text{rank}(\mathbf{X}) = p$, we assume a Gaussian linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_n), \tag{2.1}$$

where $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is the true but unobserved coefficient vector and $\boldsymbol{\epsilon} \in \mathbb{R}^n$ is a noise vector with a zero mean multivariate normal distribution and $0 < \sigma^2 \in \mathbb{R}$. The unique maximum likelihood estimator of $\boldsymbol{\beta}_0$ is the solution $\hat{\boldsymbol{\beta}}$ of the exact linear regression problem

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2. \tag{2.2}$$

Since $\mathbf{X}$ has full column rank, this problem is well-posed and has the unique solution

$$\hat{\boldsymbol{\beta}} \equiv \mathbf{X}^\dagger\mathbf{y}.$$

The exact prediction and residual are

$$\hat{\mathbf{y}} \equiv \mathbf{X}\hat{\boldsymbol{\beta}} \qquad \text{and} \qquad \hat{\mathbf{e}} \equiv \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \hat{\mathbf{y}},$$

respectively. The orthogonal projector onto $\text{range}(\mathbf{X})$ along $\text{null}(\mathbf{X}^T)$ is

$$\mathbf{P}_\mathbf{x} \equiv \mathbf{X}\mathbf{X}^\dagger = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \in \mathbb{R}^{n\times n}$$

and is also known as the *hat matrix* [5, 14, 38]. We express the prediction and residual as

$$\hat{\mathbf{y}} = \mathbf{P}_\mathbf{x}\mathbf{y} \qquad \text{and} \qquad \hat{\mathbf{e}} = (\mathbf{I} - \mathbf{P}_\mathbf{x})\mathbf{y}.$$

2.3 *The sketched problem and solution*

Given an observed matrix-valued random variable $\mathbf{S} \in \mathbb{R}^{r\times n}$ with $p \leqslant r \leqslant n$, the sketched linear regression problem

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\mathbf{S}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\|_2^2 \tag{2.3}$$

has the minimum norm solution

$$\tilde{\boldsymbol{\beta}} \equiv (\mathbf{SX})^\dagger \mathbf{Sy},$$

where $\mathbf{S}$ is a *sketching matrix*. Since we make no assumptions on $\mathbf{S}$ beyond its dimensions, the sketched matrix $\mathbf{SX}$ may be rank deficient so that (2.3) may be ill-posed.

By design, $\mathbf{S}$ has fewer rows than $\mathbf{X}$. Therefore, the corresponding predictions $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ and $\mathbf{SX}\tilde{\boldsymbol{\beta}}$ have different dimension and cannot be directly compared. To remedy this, we follow previous works [8, 9, 27] and compare the predictions with regard to the *original* design matrix $\mathbf{X}$. Therefore, the sketched prediction and residual are

$$\tilde{\mathbf{y}} \equiv \mathbf{X}\tilde{\boldsymbol{\beta}} \qquad \text{and} \qquad \tilde{\mathbf{e}} \equiv \mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}} = \mathbf{y} - \tilde{\mathbf{y}}.$$

Sketching can be an effective approach in the highly over-constrained case [8, 9, 23, 27, 28, 39], where n greatly exceeds p. A standard method of computing the exact solution of (2.2) is based on a QR decomposition, which requires $O(n^2p)$ operations. Meanwhile, applying a general sketching matrix requires $O(rnp)$ operations (fewer when sketching with FJLTs or diagonal sampling matrices) and solving the reduced dimension problem (2.3) requires $O(r^2p)$ operations. Thus, computation of a general sketched solution requires $O(rnp)$ operations so that sketching can offer substantial computational savings for very large n with r significantly smaller than n.

We note that there are also other approaches to linear regression. For example, as pointed out by one of our reviewers, the Cholesky approach or sweep operator operate on the Gram matrix $\mathbf{X}^\mathsf{T}\mathbf{X}$. The computational cost for solutions involving these is $O(np^2)$ (to assemble the Gram matrix) plus $O(p^3)$ (to apply the Cholesky decomposition or sweep operator). Therefore, these algorithms scale readily to the big n, small p scenario, as they are essentially linear in n.

The solution of the normal equations $\mathbf{X}^\mathsf{T}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^\mathsf{T}\mathbf{y}$, however, can be numerically unstable since for nearly rank-deficient $\mathbf{X}$, 'the formation of $\mathbf{X}^T\mathbf{X}$ can result in a significant loss of information' [11, Section 5.3.2]. Moreover, the condition number is always $[\kappa_2(\mathbf{X})]^2$ even if the least squares residual is small. By contrast, the sensitivity from a QR-based solver is quantified by the condition number $\kappa_2(\mathbf{X})$ for small residuals [11, Theorem 5.3.1].

Finally, we note that the sweep operator [19, Sections 7.3–7.6] is designed for efficient computation of quadratic forms involving inverse covariances associated with multivariate normal distributions. When applied to a bordered 2×2 block matrix with $\mathbf{X}^\mathsf{T}\mathbf{X}$ in the (1, 1) block and expressions involving $\mathbf{y}$ in the remaining blocks, the sweep operator produces a block matrix with the variance in the (1, 1) block and the solution and least squares residual in the other blocks. We do not consider the sweep operator here as it appears to require—either explicitly or implicitly—the formation of $\mathbf{X}^\mathsf{T}\mathbf{X}$ with its attendant potential numerical instability.

3. A Projector-Based Approach

Given a sketching matrix $\mathbf{S}$, we view the sketched problem in (2.3) as a deterministic multiplicative perturbation of the exact problem in (2.2). Therefore, we derive structural bounds for the sketched quantities. We begin by presenting an oblique projector for the sketched problem in (2.3) that plays the role of $\mathbf{P}_\mathbf{x}$ in (2.2). This oblique projector enables comparisons between the sketched solution, prediction and residual and their higher-dimensional exact counterparts.

LEMMA 3.1. For the sketched problem in (2.3),

$$\mathbf{P} \equiv \mathbf{X}(\mathbf{SX})^\dagger \mathbf{S}$$

is an oblique projector where

$$\mathbf{P}_\mathbf{x}\mathbf{P} = \mathbf{P} \quad \text{and} \quad \mathbf{PX} = \mathbf{X} \text{ if } \operatorname{rank}(\mathbf{SX}) = p.$$

These properties follow from the definitions of $\mathbf{X}^\dagger$ and $(\mathbf{SX})^\dagger$. In general, we have

$$\operatorname{rank}(\mathbf{P}) = \operatorname{rank}(\mathbf{SX}) \leqslant \operatorname{rank}(\mathbf{X}) = \operatorname{rank}(\mathbf{P}_\mathbf{x}) = p,$$

so that $\operatorname{range}(\mathbf{P}) \subseteq \operatorname{range}(\mathbf{P}_\mathbf{x})$. If $\mathbf{S}$ preserves rank so that $\operatorname{rank}(\mathbf{SX}) = \operatorname{rank}(\mathbf{X})$, then $\operatorname{range}(\mathbf{P}) = \operatorname{range}(\mathbf{P}_\mathbf{x})$. However, $\operatorname{null}(\mathbf{P}) = \operatorname{null}(\mathbf{X}^\mathsf{T}\mathbf{S}^\mathsf{T}\mathbf{S})$ [37, Theorem 3.1], so that $\operatorname{null}(\mathbf{P}) \neq \operatorname{null}(\mathbf{P}_\mathbf{x})$ in general. Finally, if $\mathbf{S} = \mathbf{I}_n$, then $\mathbf{P} = \mathbf{P}_\mathbf{x}$.

Notice that $\mathbf{P}$ generalizes $\mathbf{P}_\mathbf{u} \equiv \mathbf{U}(\mathbf{SU})^\dagger\mathbf{S}$ in [27, (11)], where $\mathbf{U}$ is an orthonormal basis for $\operatorname{range}(\mathbf{X})$, for quantifying the *prediction efficiency* and *residual efficiency* of sketching algorithms. However, $\mathbf{P}_\mathbf{u}$ is only defined if $\operatorname{rank}(\mathbf{SX}) = \operatorname{rank}(\mathbf{X})$, and in that case, $\mathbf{P}_\mathbf{u} = \mathbf{P}$. Since our analyses extend to $\operatorname{rank}(\mathbf{SX}) < \operatorname{rank}(\mathbf{X})$, we employ the more general $\mathbf{P}$.

Oblique projectors also appear in other contexts. Examples include constrained least squares [33, 37], weighted least squares [3, 32], discrete inverse problems [12] and the discrete empirical interpolation method [10, Section 3.1] to name a few. We now present the sketched solution, prediction and residual for (2.3) in terms of $\mathbf{P}$.

THEOREM 3.1. For the sketched problem in (2.3), the minimum norm solution is

$$\tilde{\boldsymbol{\beta}} = \mathbf{X}^\dagger\mathbf{P}\mathbf{y} = \hat{\boldsymbol{\beta}} + \mathbf{X}^\dagger(\mathbf{P} - \mathbf{P}_\mathbf{x})\mathbf{y}.$$

Therefore, the sketched prediction $\tilde{\mathbf{y}} = \mathbf{X}\tilde{\boldsymbol{\beta}}$ and residual $\tilde{\mathbf{e}} = \mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}$ are

$$\tilde{\mathbf{y}} = \mathbf{P}\mathbf{y} = \hat{\mathbf{y}} + (\mathbf{P} - \mathbf{P}_\mathbf{x})\mathbf{y} \quad \text{and} \quad \tilde{\mathbf{e}} = (\mathbf{I} - \mathbf{P})\,\mathbf{y} = \hat{\mathbf{e}} + (\mathbf{P}_\mathbf{x} - \mathbf{P})\mathbf{y}.$$

The expressions for $\tilde{\boldsymbol{\beta}}$, $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{e}}$ follow from their definitions in Section 2 and the definitions of $\mathbf{P}$, $\mathbf{P}$, $\hat{\mathbf{y}}$, $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{e}}$. Although the expressions for $\tilde{\boldsymbol{\beta}}$, $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{e}}$ in Theorem 3.1 are straightforward, they are exact and hold generally for *all* sketching schemes.

The significance of Theorem 3.1 is that since it requires no assumptions on $\mathbf{S}$ (beyond its dimensions) or $\operatorname{rank}(\mathbf{SX})$, it enables expressions for the total uncertainty due to the combined model- and algorithm-induced randomness for *all* sketching schemes. These include many commonly

used sketching schemes not covered by previous work [22, 23]. We comparing Theorem 3.1 to a corresponding result in [23], reproduced below in Lemma 3.2.

LEMMA 3.2. ([23, Lemma 1]) For the sketched problem in (2.3), if the following additionally hold—(1) the sketching matrix $\mathbf{S}$ has a single non-zero entry per row, (2) the vector $\mathbf{w} \equiv \operatorname{diag}(\mathbf{S}^T\mathbf{S}) \in \mathbb{R}^n$ has a scaled multinomial distribution with expected value $\mathbb{E}[\mathbf{w}] = \mathbf{1}$, (3) $\mathbf{S}$ preserves rank so that $\operatorname{rank}(\mathbf{SX}) = \operatorname{rank}(\mathbf{X})$ and (4) the sketched solution admits a Taylor series expansion around $\mathbb{E}[\mathbf{w}]$—then

$$\tilde{\boldsymbol{\beta}}(\mathbf{w}) = \hat{\boldsymbol{\beta}} + \mathbf{X}^{\dagger} \operatorname{diag}(\hat{\mathbf{e}})(\mathbf{w} - \mathbf{1}) + R(\mathbf{w}),$$

where $R(\mathbf{w})$ is the remainder of the Taylor series expansion.

The assumptions in [23, Lemma 1] and its other versions in [23] limit their scope to sampling schemes where the expected value of the sampling weights vector is known. Consequently, downstream analysis of the total expectation and variance of the sketched solution using these in [23] are also limited to those same sampling schemes.

Therefore, Theorem 3.1 extends the pioneering work on quantifying the total uncertainties for sketched in linear regression in [22, 23] in the following ways.

1. First, Theorem 3.1 places no assumptions on $\mathbf{S}$ or $\operatorname{rank}(\mathbf{SX})$ so that it applies generally to *all* sketching schemes. In practice, a wide variety of sketching schemes are used. These include sketching with fast FJLTs, Gaussian transforms and combinations of FJLTs followed by uniform sampling, to name a few. Unfortunately, the analysis in [23] does not apply to these.
2. Secondly, Theorem 3.1 is exact so that downstream analysis with these expressions do not hinge on the assumptions required for approximations.
3. Thirdly, framing the sketched solution in terms of the difference between the orthogonal projector $\mathbf{P}_{\mathbf{x}}$ for the exact problem and oblique projector $\mathbf{P}$ for the sketched problem affords additional geometric insight that we detail later in Sections 4– 6.
4. Finally, a projector-based approach greatly simplifies the proofs so that Theorem 3.1 does not require the heavy-duty matrix algebra used to produce the approximate yet more restrictive existing results in [22, 23].

Applying Theorem 3.1 and [11, (5.3.16)], which implies that

$$\frac{\|\mathbf{y}\|_2}{\|\mathbf{X}\|_2\|\hat{\boldsymbol{\beta}}\|_2} \leqslant \frac{\|\mathbf{y}\|_2}{\|\mathbf{X}\hat{\boldsymbol{\beta}}\|_2} = \frac{1}{\cos\theta},$$

produces the following relative error bounds for the sketched solution and prediction.

COROLLARY 3.1. For the sketched problem in (2.3), let $0 < \theta < \frac{\pi}{2}$ be the angle between $\mathbf{y}$ and $\operatorname{range}(\mathbf{X})$. Then, the minimum norm sketched solution $\tilde{\boldsymbol{\beta}}$ satisfies

$$\frac{\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\|_2}{\|\hat{\boldsymbol{\beta}}\|_2} \leqslant \kappa_2(\mathbf{X}) \frac{\|\mathbf{y}\|_2}{\|\mathbf{X}\|_2\|\hat{\boldsymbol{\beta}}\|_2} \|\mathbf{P} - \mathbf{P}_{\mathbf{x}}\|_2 \leqslant \kappa_2(\mathbf{X}) \frac{\|\mathbf{P} - \mathbf{P}_{\mathbf{x}}\|_2}{\cos\theta}.$$

The sketched prediction $\tilde{\mathbf{y}} = \mathbf{X}\tilde{\boldsymbol{\beta}}$ satisfies

$$\frac{\|\tilde{\mathbf{y}} - \hat{\mathbf{y}}\|_2}{\|\hat{\mathbf{y}}\|_2} \leqslant \frac{\|\mathbf{P} - \mathbf{P}_\mathbf{x}\|_2}{\cos\theta}.$$

The bounds in Corollary 3.1 are tight for $\mathbf{S} = \mathbf{I}_n$. Corollary 3.1 implies that the sensitivity of $\tilde{\boldsymbol{\beta}}$ to multiplicative perturbations depends on the deviation of $\mathbf{P}$ from being an orthogonal projector onto range($\mathbf{X}$), quantified by $\|\mathbf{P} - \mathbf{P}_\mathbf{x}\|_2$. This distance is amplified, as expected, by the conditioning of $\mathbf{X}$ with regard to (left) inversion and by the closeness of $\mathbf{y}$ to range($\mathbf{X}$). Corollary 3.1 is an absolute and relative bound since $\|\mathbf{P}_\mathbf{x}\|_2 = 1$.

In contrast to multiplicative perturbation bounds for eigenvalue and singular value problems [15, 16], Corollary 3.1 does not require $\mathbf{S}$ to be non-singular or square. We do not view weighted least squares problems [11, Section 6.1] as multiplicative perturbations since they employ non-singular diagonal matrices $\mathbf{S}$ for regularization or scaling of discrepancies.

In contrast to additive perturbation bounds ([11, Section 5.3.6], [13, Section 20.1], [31, (3.4)]), Corollary 3.1 requires neither the square of the condition number nor rank($\mathbf{SX}$) = rank($\mathbf{X}$). Therefore, the minimum norm sketched solution $\tilde{\boldsymbol{\beta}}$ and its residual $\tilde{\mathbf{e}}$ are less sensitive to multiplicative perturbations than to additive perturbations. Extensive discussions of multiplicative perturbation bounds and comparisons to their additive counterparts are presented in [15, 16], where the purpose is the derivation of relative error bounds. Here, we employ multiplicative ones because they appear naturally since the perturbation arises from the multiplication of $\mathbf{X}$ by a sketching matrix.

Compared with [34], where the perturbation theory is targeted at additive perturbations, we also make extensive use of projectors. However, the bounds in [34, Chapter III] concern least squares problems in their most general form, where $\mathbf{X}$ can be rank deficient so that its Moore–Penrose inverse is ill-posed. Consequently, the derivations rely on expanding acute perturbations of the Moore–Penrose inverse as well as asymptotic forms and derivatives of the Moore–Penrose inverse.

Corollary 3.1 improves on existing structural bounds for sketched least squares algorithms, such as [9, Theorem 1] reproduced in Lemma 3.3 below.

LEMMA 3.3. ([9, Theorem 1]) For the sketched problem in (2.3), if $\|\mathbf{P}_\mathbf{x}\mathbf{y}\|_2 \geqslant \gamma \|\mathbf{y}\|_2$ for some $0 < \gamma \leqslant 1$ and $\|\tilde{\mathbf{e}}\|_2 \leqslant (1+\eta)\|\hat{\mathbf{e}}\|_2$, then

$$\frac{\|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\|_2}{\|\hat{\boldsymbol{\beta}}\|_2} \leqslant \kappa_2(\mathbf{X})\sqrt{\gamma^{-2} - 1}\,\sqrt{\eta}.$$

Corollary 3.1 improves on [9, Theorem 1] in the following ways. First, the bound for $\tilde{\boldsymbol{\beta}}$ in Corollary 3.1 is more general and tighter as it does not exhibit nonlinear dependencies on the perturbations. Secondly, Corollary 3.1 holds under weaker assumptions. The first inequality for the sketched solution in Corollary 3.1 requires only that $\hat{\boldsymbol{\beta}} \neq \mathbf{0}$. The second inequality for the sketched solution requires only that $\mathbf{y} \notin \text{range}(\mathbf{X})$ and $\mathbf{y} \notin \text{range}(\mathbf{X}^\perp)$.

4. Model- and Algorithm-Induced Uncertainties

The solution $\hat{\boldsymbol{\beta}}$ of the exact problem in (2.2) has desirable statistical properties since it is an unbiased estimator of the true coefficient vector $\boldsymbol{\beta}_0$, and it has minimal variance among all linear unbiased

estimators of $\boldsymbol{\beta}_0$, e.g. [30, Chapter 3, Section 3d]. A question one might ask is the following: how does sketching affect the statistical properties of the solution $\tilde{\boldsymbol{\beta}}$ of (2.3)?

To answer this question, we derive the total expectation and variance due to the combined model- and algorithm-induced uncertainties for the sketched solution $\tilde{\boldsymbol{\beta}}$ and compare them with those of the exact solution $\hat{\boldsymbol{\beta}}$. Since our expressions rely on Theorem 3.1, our results extend the work in [22, 23] to all sketching schemes.

We briefly review the model-induced uncertainty from a Gaussian linear model in Section 4.1. We then derive the expectation and variance of $\tilde{\boldsymbol{\beta}}$ conditioned on the algorithm-induced uncertainty in Section 4.2. Next, we employ the law of total expectation, e.g. [4, Theorem 4.4.3], to derive the total expectation and variance for the combined model- and algorithm-induced uncertainties in Section 4.3. Finally, we visit the total expectation and variance conditioned on sketching schemes that preserve rank in Section 4.4. While the latter require an additional assumption, they enable insights that we elaborate on later.

4.1 *Model-induced uncertainty*

We refer to the randomness implied by a Gaussian linear model as the *model-induced uncertainty*. Since the noise vector has mean and variance equal to

$$\mathbb{E}_{\mathbf{y}}[\boldsymbol{\epsilon}] = \mathbf{0} \quad \text{and} \quad \mathbb{V}\text{ar}_{\mathbf{y}}[\boldsymbol{\epsilon}] = \sigma^2 \mathbf{I}_n,$$

the exact solution $\hat{\boldsymbol{\beta}}$ has mean and variance equal to

$$\mathbb{E}_{\mathbf{y}}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}_0 \quad \text{and} \quad \mathbb{V}\text{ar}_{\mathbf{y}}[\hat{\boldsymbol{\beta}}] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \in \mathbb{R}^{p\times p}. \tag{4.1}$$

It is well known that the variance of $\hat{\boldsymbol{\beta}}$ depends on the conditioning of $\mathbf{X}$ [31, Section 5].

A difficulty in analyzing row-sketching (2.3), coupled with general concern regarding first-order expansions like the ones in [22, 23], is potential rank deficiency in the sketched matrix so that $\text{rank}(\mathbf{SX}) < \text{rank}(\mathbf{X})$. In this case, $(\mathbf{SX})^\dagger$ cannot be expressed in terms of $\mathbf{SX}$. Thus, we introduce a projector that quantifies the bias arising from rank deficiency in $\mathbf{SX}$.

For the sketched problem in (2.3),

$$\mathbf{P_0} \equiv (\mathbf{SX})^\dagger(\mathbf{SX}) \in \mathbb{R}^{p\times p}$$

is an orthogonal projector with the following consequences:

$$\mathbf{PX} = \mathbf{XP_0} \quad \text{and} \quad \mathbf{P_0} = \mathbf{I}_p \text{ if } \text{rank}(\mathbf{SX}) = p.$$

Orthogonality follows from $(\mathbf{P_0})^2 = \mathbf{P_0}$ and $(\mathbf{P_0})^T = \mathbf{P_0}$, which follow from the fact that $(\mathbf{SX})^\dagger$ is a Moore–Penrose generalized inverse. If $\text{rank}(\mathbf{SX}) < p$, then $\mathbf{P_0}$ characterizes the subspace of $\text{range}(\mathbf{X})$ onto which $\mathbf{P}$ projects. The name *bias projector* will become apparent in Theorem 4.1, where $\mathbf{P_0}$ quantifies the bias in $\tilde{\boldsymbol{\beta}}$.

4.2 *Conditional expectation and variance*

We condition on a given sketching matrix $\mathbf{S}$ and derive the conditional model-induced expectation and variance of the sketched solution $\tilde{\boldsymbol{\beta}}$. Theorem 4.1 below shows that the conditional expectation depends on the bias projector $\mathbf{P_0}$, while the conditional variance depends on the oblique projector $\mathbf{P}$.

For the sketched problem in (2.3), the solution $\tilde{\boldsymbol{\beta}}$ has conditional expectation

$$\mathbb{E}_{\mathbf{y}}[\tilde{\boldsymbol{\beta}} \mid \mathbf{S}] \;=\; \mathbf{P_0}\boldsymbol{\beta}_0 \;=\; \boldsymbol{\beta}_0 - (\mathbf{I} - \mathbf{P_0})\boldsymbol{\beta}_0,$$

where $\mathbf{I} - \mathbf{P_0}$ quantifies the rank deficiency of $\mathbf{SX}$ and conditional variance

$$\begin{aligned}\mathbb{V}\mathrm{ar}_{\mathbf{y}}[\tilde{\boldsymbol{\beta}} \mid \mathbf{S}] \;&=\; \sigma^2 \left(\mathbf{X}^{\dagger}\mathbf{P}\right)\left(\mathbf{X}^{\dagger}\mathbf{P}\right)^T \\ &=\; \mathbb{V}\mathrm{ar}_{\mathbf{y}}[\hat{\boldsymbol{\beta}}] + \sigma^2\, \mathbf{X}^{\dagger}\left(\mathbf{P}\mathbf{P}^T - \mathbf{P}_{\mathbf{x}}\right)(\mathbf{X}^{\dagger})^T,\end{aligned}$$

where $\mathbf{P}\mathbf{P}^T - \mathbf{P}_{\mathbf{x}}$ represents the deviation of $\mathbf{P}$ from being an orthogonal projector onto range($\mathbf{X}$).

Proof. For the conditional expectation, we employ the second expression for $\tilde{\boldsymbol{\beta}}$ in Theorem 3.1. The result follows from the fact that $\mathbf{X}^{\dagger}$ is a left inverse for $\mathbf{X}$ and the definition of $\mathbf{P_0}$.

For the first expression for the conditional variance, we apply the definition of the variance conditioned on $\mathbf{S}$ to the first expression for $\tilde{\boldsymbol{\beta}}$ in Theorem 3.1. We combine this with the expression for the conditional expectation for $\tilde{\boldsymbol{\beta}}$ to obtain

$$\begin{aligned}\mathbb{V}\mathrm{ar}_{\mathbf{y}}[\tilde{\boldsymbol{\beta}} \mid \mathbf{S}] \;&=\; \mathbb{E}_{\mathbf{y}}[\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}^T \mid \mathbf{S}] - \mathbb{E}_{\mathbf{y}}[\tilde{\boldsymbol{\beta}} \mid \mathbf{S}]\, \mathbb{E}_{\mathbf{y}}[\tilde{\boldsymbol{\beta}} \mid \mathbf{S}]^T \\ &=\; \left(\mathbf{X}^{\dagger}\mathbf{P}\right)\mathbb{E}_{\mathbf{y}}[\mathbf{y}\mathbf{y}^T]\left(\mathbf{X}^{\dagger}\mathbf{P}\right)^T - (\mathbf{P_0}\boldsymbol{\beta}_0)(\mathbf{P_0}\boldsymbol{\beta}_0)^T. \qquad (4.2)\end{aligned}$$

Expanding the middle term in the first summand gives

$$\begin{aligned}\mathbb{E}_{\mathbf{y}}[\mathbf{y}\mathbf{y}^T] \;&=\; (\mathbf{X}\boldsymbol{\beta}_0)(\mathbf{X}\boldsymbol{\beta}_0)^T + \mathbb{E}_{\mathbf{y}}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] \\ &=\; (\mathbf{X}\boldsymbol{\beta}_0)(\mathbf{X}\boldsymbol{\beta}_0)^T + \sigma^2\mathbf{I}_n. \qquad (4.3)\end{aligned}$$

We then substitute (4.3) into (4.2). Using the fact that $\mathbf{X}^{\dagger}\mathbf{P}\mathbf{X} = \mathbf{P_0}$ and canceling terms produces the first expression. For the second expression for the conditional variance, we use the facts that

$$\mathbf{X}^{\dagger}\mathbf{P}_{\mathbf{x}} = \mathbf{X}^{\dagger} \quad \text{and} \quad \mathbf{X}^{\dagger}(\mathbf{X}^{\dagger})^T = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}$$

to rewrite $\mathbb{V}\mathrm{ar}_{\mathbf{y}}[\hat{\boldsymbol{\beta}}]$ in (4.1) as

$$\mathbb{V}\mathrm{ar}_{\mathbf{y}}[\hat{\boldsymbol{\beta}}] = \sigma^2\, \mathbf{X}^{\dagger}\mathbf{P}_{\mathbf{x}}(\mathbf{X}^{\dagger})^T. \qquad (4.4)$$

The result follows from adding and subtracting (4.4) in the first expression for the conditional variance.

For the interpretation of $\mathbf{I} - \mathbf{P_0}$, notice that if $\mathbf{SX}$ has full column rank, then $\mathbf{P_0} = \mathbf{I}$. Therefore, $\mathbf{I} - \mathbf{P_0}$ represents the deviation of $\mathbf{SX}$ from having full column rank.

For the interpretation of $\mathbf{P}\mathbf{P}^T - \mathbf{P}_\mathbf{X}$, notice that since range($\mathbf{P}$) $\subseteq$ range($\mathbf{P}_\mathbf{X}$), $\mathbf{P}$ projects onto a subspace of range($\mathbf{X}$). If additionally, $\mathbf{P}$ is an orthogonal projector, symmetry requires $\mathbf{S} = \mathbf{I}_n$ so that $\mathbf{P} = \mathbf{P}\mathbf{P}^T = \mathbf{P}_\mathbf{X}$. Therefore, $\mathbf{P}\mathbf{P}^T - \mathbf{P}_\mathbf{X}$ represents the deviation of $\mathbf{P}$ from being an orthogonal projector onto range($\mathbf{X}$). □

Theorem 4.1 shows that the conditional expectation of $\tilde{\boldsymbol{\beta}}$ depends on the rank deficiency of $\mathbf{SX}$. In particular, the conditional bias of $\tilde{\boldsymbol{\beta}}$ is proportional to the deviation $\mathbf{I} - \mathbf{P}_\mathbf{0}$ of $\mathbf{SX}$ from having full column rank. To see this, notice that conditioned on $\mathbf{SX}$ having full column rank, $\mathbf{P}_\mathbf{0} = \mathbf{I}$. In this case, $\mathbf{I} - \mathbf{P}_\mathbf{0}$ vanishes and $\tilde{\boldsymbol{\beta}}$ is a conditionally unbiased estimator of $\boldsymbol{\beta}_0$ with

$$\mathbb{E}_\mathbf{y}[\tilde{\boldsymbol{\beta}} \mid \operatorname{rank}(\mathbf{SX}) = \operatorname{rank}(\mathbf{X})] = \boldsymbol{\beta}_0.$$

Since this holds for any $\mathbf{S}$, the conditional bias of $\tilde{\boldsymbol{\beta}}$ depends only on rank($\mathbf{SX}$).

Theorem 4.1 also shows that the conditional variance of $\tilde{\boldsymbol{\beta}}$ depends on the deviation of $\mathbf{P}$ from being an orthogonal projector onto range($\mathbf{X}$). In particular, the conditional variance $\mathbb{V}\mathrm{ar}_\mathbf{y}[\tilde{\boldsymbol{\beta}} \mid \mathbf{S}]$ is close to the model variance $\mathbb{V}\mathrm{ar}_\mathbf{y}[\hat{\boldsymbol{\beta}}]$ if $\mathbf{P}$ is close to $\mathbf{P}_\mathbf{X}$. In the extreme case that $\mathbf{S} = \mathbf{I}_n$, the conditional variance is identical to the model variance. Corollary 4.1 follows directly from Theorem 4.1 and further highlights the relevance of $\mathbf{I} - \mathbf{P}_\mathbf{0}$ and $\mathbf{P}\mathbf{P}^T - \mathbf{P}_\mathbf{X}$.

Given the assumptions in Theorem 4.1, we have

$$\| \mathbb{E}_\mathbf{y}[\tilde{\boldsymbol{\beta}} \mid \mathbf{S}] - \boldsymbol{\beta}_0 \|_2 \;\leqslant\; \|\mathbf{I} - \mathbf{P}_\mathbf{0}\|_2 \, \|\boldsymbol{\beta}_0\|_2$$

and

$$\frac{\| \mathbb{V}\mathrm{ar}_\mathbf{y}[\tilde{\boldsymbol{\beta}} \mid \mathbf{S}] - \mathbb{V}\mathrm{ar}_\mathbf{y}[\hat{\boldsymbol{\beta}}] \|_2}{\| \mathbb{V}\mathrm{ar}_\mathbf{y}[\hat{\boldsymbol{\beta}}] \|_2} \;\leqslant\; \|\mathbf{P}\mathbf{P}^T - \mathbf{P}_\mathbf{X}\|_2.$$

The relative conditional variance follows from Theorem 4.1 and the facts that $\|\mathbf{X}^\dagger\|_2 \, \|(\mathbf{X}^\dagger)^T\|_2 = \|\mathbf{X}^\dagger(\mathbf{X}^\dagger)^T\|_2$, $\mathbf{X}^\dagger(\mathbf{X}^\dagger)^T = (\mathbf{X}^T\mathbf{X})^{-1}$ and $\sigma^2 > 0$ so that $\| \mathbb{V}\mathrm{ar}_\mathbf{y}[\hat{\boldsymbol{\beta}}] \|_2 \neq 0$.

Corollary 4.1 shows that the relative differences in the conditional bias and variance can be expressed solely in terms of $\mathbf{I} - \mathbf{P}_\mathbf{0}$ and $\mathbf{P}\mathbf{P}^T - \mathbf{P}_\mathbf{X}$. In particular, the conditional bias of $\tilde{\boldsymbol{\beta}}$ increases with rank deficiency in $\mathbf{SX}$. Additionally, the relative difference between conditional and model variances increases with the deviation of $\mathbf{P}$ from $\mathbf{P}_\mathbf{X}$.

Therefore, Corollary 4.1 shows that unbiasedness is more readily achievable since it requires only that $\mathbf{SX}$ have full column rank. Meanwhile, the conditional variance of $\tilde{\boldsymbol{\beta}}$ is guaranteed to be at least as large as $\mathbb{V}\mathrm{ar}_\mathbf{y}[\hat{\boldsymbol{\beta}}]$, with equality only when $\mathbf{S} = \mathbf{I}_n$ so that $\mathbf{P} = \mathbf{P}_\mathbf{X}$. In this case, the sketched problem in (2.3) becomes the exact problem in (2.2).

4.3 *Total expectation and variance*

We now view the sketching matrix $\mathbf{S}$ as a matrix-valued random variable and derive the total expectation and variance of the sketched solution $\tilde{\boldsymbol{\beta}}$. We employ the expressions for the conditional expectation and variance in Section 4.2 and the law of total expectation.

For the sketched problem in (2.3), the solution $\tilde{\boldsymbol{\beta}}$ has total expectation

$$\mathbb{E}[\tilde{\boldsymbol{\beta}}] \;=\; \boldsymbol{\beta}_0 - \left(\mathbf{I} - \mathbb{E}_{\mathbf{s}}[\mathbf{P_0}]\right)\boldsymbol{\beta}_0$$

and total variance

$$\mathbb{V}\mathrm{ar}[\tilde{\boldsymbol{\beta}}] \;=\; \mathbb{V}\mathrm{ar}[\hat{\boldsymbol{\beta}}] + \sigma^2\,\mathbf{X}^{\dagger}\left(\mathbb{E}_{\mathbf{s}}[\mathbf{P}\mathbf{P}^T] - \mathbf{P}_{\mathbf{x}}\right)(\mathbf{X}^{\dagger})^T + \mathbb{V}\mathrm{ar}_{\mathbf{s}}[\mathbf{P_0}\boldsymbol{\beta}_0].$$

Proof. For the total expectation, we combine our expression for $\mathbb{E}_{\mathbf{y}}[\tilde{\boldsymbol{\beta}} \mid \mathbf{S}]$ from Theorem 4.1 with the law of total expectation. For the total variance, we apply the expression for the total expectation in the definition of the variance to obtain

$$\begin{aligned}\mathbb{V}\mathrm{ar}[\tilde{\boldsymbol{\beta}}] \;&=\; \mathbb{E}[\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}^T] - \mathbb{E}[\tilde{\boldsymbol{\beta}}]\,\mathbb{E}[\tilde{\boldsymbol{\beta}}]^T \\ &=\; \mathbb{E}_{\mathbf{s}}\left[\mathbb{E}_{\mathbf{y}}\left[\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}^T \,\middle|\, \mathbf{S}\right]\right] - \left(\mathbb{E}_{\mathbf{s}}[\mathbf{P_0}]\boldsymbol{\beta}_0\right)\left(\mathbb{E}_{\mathbf{s}}[\mathbf{P_0}]\boldsymbol{\beta}_0\right)^T. \end{aligned} \tag{4.5}$$

From (4.2) and (4.3), we have

$$\mathbb{E}_{\mathbf{y}}\left[\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}^T \,\middle|\, \mathbf{S}\right] = \sigma^2\mathbf{X}^{\dagger}\mathbf{P}\mathbf{P}^T(\mathbf{X}^{\dagger})^T + (\mathbf{P_0}\boldsymbol{\beta}_0)(\mathbf{P_0}\boldsymbol{\beta}_0)^T. \tag{4.6}$$

Inserting (4.6) into (4.5) then gives us

$$\begin{aligned}\mathbb{V}\mathrm{ar}[\tilde{\boldsymbol{\beta}}] \;=\; &\sigma^2\mathbf{X}^{\dagger}\,\mathbb{E}_{\mathbf{s}}\left[\mathbf{P}\mathbf{P}^T\right](\mathbf{X}^{\dagger})^T \\ &+ \underbrace{\mathbb{E}_{\mathbf{s}}\left[\left(\mathbf{P_0}\boldsymbol{\beta}_0\right)\left(\mathbf{P_0}\boldsymbol{\beta}_0\right)^T\right] - \left(\mathbb{E}_{\mathbf{s}}[\mathbf{P_0}]\boldsymbol{\beta}_0\right)\left(\mathbb{E}_{\mathbf{s}}[\mathbf{P_0}]\boldsymbol{\beta}_0\right)^T}_{\mathbb{V}\mathrm{ar}_{\mathbf{s}}[\mathbf{P_0}\boldsymbol{\beta}_0]},\end{aligned}$$

where the latter two terms in the above expression are equal to $\mathbb{V}\mathrm{ar}_{\mathbf{s}}[\mathbf{P_0}\boldsymbol{\beta}_0]$. Finally, using the fact that $\mathbf{X}^{\dagger}\mathbf{P}_{\mathbf{x}}(\mathbf{X}^{\dagger})^{\mathsf{T}} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}$, we add and subtract $\mathbb{V}\mathrm{ar}[\hat{\boldsymbol{\beta}}]$ from the above expression to obtain the result. □

Theorem 4.2 shows that the total bias of $\tilde{\boldsymbol{\beta}}$ is proportional to the expected deviation of the matrix-valued random variable $\mathbf{SX}$ from having full column rank. Therefore, after accounting for both the model- and algorithm-induced uncertainties, the bias of $\tilde{\boldsymbol{\beta}}$ depends on the expected value of $\mathbf{P_0}$. Notice, however, that the expectation $\mathbb{E}_{\mathbf{s}}[\mathbf{P_0}]$ of a projector $\mathbf{P_0}$ is not a projector in general.

Theorem 4.2 also shows that the total variance of $\tilde{\boldsymbol{\beta}}$ can be decomposed into the following three components:

1. the inherent model variance in $\hat{\boldsymbol{\beta}}$,
2. the expected deviation of the matrix-valued random variable $\mathbf{P}$ from being an orthogonal projector onto range($\mathbf{X}$) and
3. the variance in the rank deficiency of the matrix-valued random variable $\mathbf{SX}$ as captured through the bias projector $\mathbf{P_0}$.

Corollary 4.2 follows from Theorem 4.2. It shows how rank deficiency, as quantified by $\mathbf{I} - \mathbf{P_0}$, and the deviation of $\mathbf{P}$ from being an orthogonal projector, as quantified by $\mathbf{P}\mathbf{P}^T - \mathbf{P_x}$, affect the relative differences between the total and model uncertainties.

Given the assumptions in Theorem 4.2, we have

$$\| \mathbb{E}[\tilde{\boldsymbol{\beta}}] - \boldsymbol{\beta}_0 \|_2 \;\leqslant\; \| \mathbf{I} - \mathbb{E}_{\mathbf{s}}[\mathbf{P_0}] \|_2 \, \| \boldsymbol{\beta}_0 \|_2$$

and

$$\frac{\| \mathbb{V}\mathrm{ar}[\tilde{\boldsymbol{\beta}}] - \mathbb{V}\mathrm{ar}_{\mathbf{y}}[\hat{\boldsymbol{\beta}}] \|_2}{\| \mathbb{V}\mathrm{ar}_{\mathbf{y}}[\hat{\boldsymbol{\beta}}] \|_2} \;\leqslant\; \| \mathbb{E}_{\mathbf{s}}[\mathbf{P}\mathbf{P}^T] - \mathbf{P_x} \|_2 + \frac{\| \mathbb{V}\mathrm{ar}_{\mathbf{s}}[(\mathbf{I} - \mathbf{P_0})\boldsymbol{\beta}_0] \|_2}{\| \mathbb{V}\mathrm{ar}_{\mathbf{y}}[\hat{\boldsymbol{\beta}}] \|_2}.$$

Compared with Corollary 4.1, where the difference between the conditional and model variance depends only on $\mathbf{P}\mathbf{P}^T - \mathbf{P_x}$, Corollary 4.2 shows that the difference between the total and model variance depends on two sources. The first is the expected deviation of $\mathbf{P}$ from being an orthogonal projector as quantified in $\mathbb{E}_{\mathbf{s}}[\mathbf{P}\mathbf{P}^T] - \mathbf{P_x}$. The second is the ratio of the variance of the estimation distortion due to rank deficiency to the model variance. If the variance in the distortion due to rank deficiency is small relative to the model variance, then this latter term is likewise small.

4.4 *Total uncertainties conditioned on rank preservation*

In the previous sections, we worked toward deriving unconditional expressions quantifying the combined model- and algorithm-induced uncertainties in sketched linear regression. Since those expressions require no assumptions on the sketching matrix $\mathbf{S}$ beyond its dimensions, they hold exactly and in general for all sketching schemes.

We now present results that condition on sketching matrices that preserve rank so that $\mathrm{rank}(\mathbf{SX}) = \mathrm{rank}(\mathbf{X})$. Although these results require an additional assumption, conditioning on rank preservation enables further insight, which we detail below and in other following sections.

For the sketched problem in (2.3) conditioned on $\mathrm{rank}(\mathbf{SX}) = \mathrm{rank}(\mathbf{X})$, the solution $\tilde{\boldsymbol{\beta}}$ has total expectation

$$\mathbb{E}[\tilde{\boldsymbol{\beta}}] \;=\; \boldsymbol{\beta}_0$$

and total variance

$$\mathbb{V}\mathrm{ar}[\tilde{\boldsymbol{\beta}}] \;=\; \mathbb{V}\mathrm{ar}[\hat{\boldsymbol{\beta}}] + \sigma^2 \, \mathbf{X}^{\dagger} \left(\mathbb{E}_{\mathbf{s}}[\mathbf{P}\mathbf{P}^T] - \mathbf{P_x} \right) (\mathbf{X}^{\dagger})^T.$$

The expressions for the total expectation and variance follow from Theorem 4.2 and the fact that $\mathbb{E}_{\mathbf{s}}[\mathbf{P_0} \mid \mathrm{rank}(\mathbf{SX}) = \mathrm{rank}(\mathbf{X})] = \mathbf{I}$. Corollary 4.3 shows that conditioning on rank preservation, the sketched solution $\tilde{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}_0$. Later, in Corollary 5.3, we will find that even in these cases, however, the total variance of $\tilde{\boldsymbol{\beta}}$ is at least as great as the model variance $\mathbb{V}\mathrm{ar}[\hat{\boldsymbol{\beta}}]$.

Compared with [23, Lemma 2] which also assumes rank preservation, Corollary 4.3 is more general in that it holds for all sketching matrices, without restriction to specific kinds of sampling matrices. Additionally, [23, Lemma 2] has an additional term due to the variance of the Taylor expansion

remainder. Corollary 4.3 lacks this term since the projector-based formulation of the $\tilde{\boldsymbol{\beta}}$ in Theorem 3.1 holds exactly without any additional assumptions.

5. Total Excess Bias and Variance

We summarize and interpret the *excess bias* and *excess variance* attributable to algorithm-induced uncertainties. These represent the additional bias and variance in the sketched solution $\tilde{\boldsymbol{\beta}}$ beyond the model bias $\text{Bias}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0)$ and model variance $\mathbb{V}\text{ar}(\hat{\boldsymbol{\beta}})$ arising from the assumptions of a Gaussian linear model. We show that the projector-based approach in Theorem 3.1 enables insight and understanding into the sources of excess bias and variance.

For the problem in (2.3), the solution $\tilde{\boldsymbol{\beta}}$ has total excess bias equal to

$$\mathscr{B} \equiv (\mathbb{E}_{\mathbf{S}}[\mathbf{P_0}] - \mathbf{I})\boldsymbol{\beta}_0$$

and total excess variance equal to

$$\mathscr{V} \equiv \underbrace{\sigma^2 \mathbf{X}^\dagger \left(\mathbb{E}_{\mathbf{S}}[\mathbf{P}\mathbf{P}^T] - \mathbf{P}_{\mathbf{X}}\right) (\mathbf{X}^\dagger)^T}_{\mathscr{V}_{\mathbf{P}}} + \underbrace{\mathbb{V}\text{ar}_{\mathbf{S}}[\mathbf{P_0}\boldsymbol{\beta}_0]}_{\mathscr{V}_{\mathbf{P_0}}}.$$

Corollary 5.1 follows from Theorem 4.2 and the fact that the exact solution $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}_0$. Recall that $\mathbb{E}_{\mathbf{S}}[\mathbf{P_0}] - \mathbf{I}$ represents the expected deviation of the sketched matrix $\mathbf{SX}$ from having full column rank. Therefore, the *excess bias* $\mathscr{B}$ represents the expected estimation distortion under rank deficiency from sketching.

Corollary 5.1 shows that we can decompose the *excess variance* $\mathscr{V}$ due to randomness in the sketching algorithm into two sources. The first source $\mathscr{V}_{\mathbf{P}}$ is due to the expected deviation of the oblique projector $\mathbf{P}$ from being an orthogonal projector onto range($\mathbf{X}$). The second source $\mathscr{V}_{\mathbf{P_0}}$ arises from the variance of the estimation distortion under rank deficiency from sketching. Conditioning on rank preservation so that $\text{rank}(\mathbf{SX}) = \text{rank}(\mathbf{X})$ presents simplifications that enable additional insights on the total excess bias and variance.

For the problem in (2.3) conditioned on $\text{rank}(\mathbf{SX}) = \text{rank}(\mathbf{X})$, the solution $\tilde{\boldsymbol{\beta}}$ has zero total excess bias and total excess variance equal to

$$\mathscr{V}' \equiv \underbrace{\sigma^2 \mathbf{X}^\dagger \left(\mathbb{E}_{\mathbf{S}}[\mathbf{P}\mathbf{P}^T] - \mathbf{P}_{\mathbf{X}}\right) (\mathbf{X}^\dagger)^T}_{\mathscr{V}_{\mathbf{P}}}.$$

Corollary 5.2 follows from Corollary 4.3. Conditioning on rank preservation, both the excess bias $\mathscr{B}$ and the excess variance due to rank deficiency $\mathscr{V}_{\mathbf{P_0}}$ vanish. Therefore, the excess variance conditioned on rank preservation $\mathscr{V}'$ is equal to $\mathscr{V}_{\mathbf{P}}$, which quantifies the excess variance arising from the expected deviation of $\mathbf{P}$ from $\mathbf{P}_{\mathbf{X}}$.

For further interpretation of $\mathscr{V}_{\mathbf{P}_{\mathbf{X}}}$, we revisit the range and null spaces of $\mathbf{P}$ and $\mathbf{P}_{\mathbf{X}}$. Recall that if $\text{rank}(\mathbf{SX}) = \text{rank}(\mathbf{X})$, we have

$$\text{range}(\mathbf{P}) = \text{range}(\mathbf{P}_{\mathbf{X}}).$$

The fact that range($\mathbf{P}$) $\subseteq$ range($\mathbf{P_X}$) follows from the identity $\mathbf{P_X P} = \mathbf{P}$. Additionally, the fact that range($\mathbf{P_X}$) $\subseteq$ range($\mathbf{P}$) follows from the identity $\mathbf{P P_X} = \mathbf{P_X}$. Equality therefore follows from double containment. Meanwhile, from [37, Theorem 3.1], we have

$$\text{null}(\mathbf{P}) = \text{null}(\mathbf{X}^\mathsf{T}\mathbf{S}^\mathsf{T}\mathbf{S}) \neq \text{null}(\mathbf{X}^\mathsf{T}) = \text{null}(\mathbf{P_X})$$

in general. Thus, we observe how sketching perturbs the subspaces from the exact problem. If rank($\mathbf{SX}$) = rank($\mathbf{X}$), the sketching and orthogonal projectors, $\mathbf{P}$ and $\mathbf{P_X}$, have the same range. However, the dimension reduction achieved through sketching comes at the cost of a perturbation of null($\mathbf{P_X}$).

Therefore, the excess variance arising from the deviation of $\mathbf{P}$ from $\mathbf{P_X}$ reflects the perturbation of the original subspaces due to algorithm-induced randomness. Specifically, the deviation of $\mathbf{P}$ from $\mathbf{P_X}$ in $\mathscr{V}_{\mathbf{P_X}}$ conditioned on rank preservation reflects the deviation of null($\mathbf{P}$) from null($\mathbf{P_X}$).

For the problem in (2.3) conditioned on rank($\mathbf{SX}$) = rank($\mathbf{X}$), we have

$$\mathbb{V}\text{ar}[\tilde{\boldsymbol{\beta}}] \succcurlyeq \mathbb{V}\text{ar}[\hat{\boldsymbol{\beta}}],$$

where the $\succcurlyeq$ operator denotes the Loewner ordering for symmetric matrices of the same dimension. Additionally, we have

$$\text{trace}(\mathscr{V}_{\mathbf{P}}) \geqslant 0 \quad \text{so that} \quad \text{trace}(\mathbb{V}\text{ar}[\tilde{\boldsymbol{\beta}}]) \geqslant \text{trace}(\mathbb{V}\text{ar}[\hat{\boldsymbol{\beta}}]).$$

Proof. Corollary 5.3 follows from the fact that conditioning on rank preservation gives the identity $\mathbf{P P_X P}^\mathsf{T} = \mathbf{P_X}$. Therefore, $\mathscr{V}_{\mathbf{P}}$ is positive semi-definite since $\mathbf{I} - \mathbf{P_X}$ is idempotent. The variance inequalities follow from the fact that positive semi-definite matrices have non-negative trace. □

The facts that $\mathbb{V}\text{ar}[\tilde{\boldsymbol{\beta}}] \succcurlyeq \mathbb{V}\text{ar}[\hat{\boldsymbol{\beta}}]$ and $\text{trace}(\mathbb{V}\text{ar}[\tilde{\boldsymbol{\beta}}]) \geqslant \text{trace}(\mathbb{V}\text{ar}[\hat{\boldsymbol{\beta}}])$ are unsurprising in themselves since $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator of $\boldsymbol{\beta}_0$, e.g. [30, Chapter 3, Section 3d]. What is surprising, however, is that the projector-based approach shows directly that the additional variance is due to the expected deviation of null($\mathbf{P}$) from null($\mathbf{P_X}$).

6. Bias-Variance Decompositions

We show that the projector-based approach combined with the total uncertainty quantities from Section 4.3 further enable bias-variance decompositions that hold generally for all sketching schemes. We begin by analyzing the mean squared error for the true parameter $\boldsymbol{\beta}_0$. We then examine the predictive risk, which in this case is the mean squared error for the true prediction $\mathbf{X}\boldsymbol{\beta}_0$. We employ the $\text{MSE}(\cdot,\cdot)$ and $\text{R}(\cdot,\cdot)$ operators to denote the mean squared error and predictive risk between two vectors of the same dimension, respectively.

For the problem in (2.3), the solution $\tilde{\boldsymbol{\beta}}$ has total mean squared error equal to

$$\begin{aligned}\text{MSE}(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}_0) \;&=\; \text{trace}\{\mathbb{V}\text{ar}[\hat{\boldsymbol{\beta}}]\} + \sigma^2\, \text{trace}\{\mathbf{X}^\dagger \left(\mathbb{E}_{\mathbf{S}}[\mathbf{P}\mathbf{P}^T] - \mathbf{P_X}\right) (\mathbf{X}^\dagger)^T\} \\ &\quad + \text{trace}\{\mathbb{V}\text{ar}_{\mathbf{S}}[\mathbf{P_0}\boldsymbol{\beta}_0]\} + \| \left(\mathbf{I} - \mathbb{E}_{\mathbf{S}}[\mathbf{P_0}]\right) \boldsymbol{\beta}_0 \|_2^2.\end{aligned}$$

Proof. We employ the properties of the trace operator and linearity of the trace and expectation to obtain the well-known bias-variance trade-off in terms of the trace operator

$$\begin{aligned} MSE(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}_0) &= \mathbb{E}[\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2] \\ &= \mathbb{E}[\|\tilde{\boldsymbol{\beta}} - \mathbb{E}[\tilde{\boldsymbol{\beta}}]\|_2^2] + \|\,\mathbb{E}[\tilde{\boldsymbol{\beta}}] - \boldsymbol{\beta}_0\|_2^2 \\ &= \operatorname{trace}\{\mathbb{V}\mathrm{ar}[\tilde{\boldsymbol{\beta}}]\} + \|\mathrm{Bias}(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}_0)\|_2^2. \end{aligned}$$

The result follows directly from applying the expressions for the total variance and bias of $\tilde{\boldsymbol{\beta}}$ from Theorem 4.2. □

Corollary 6.1 directly states how the bias and variance of $\tilde{\boldsymbol{\beta}}$ contribute to the total mean squared error. Specifically, the portion of the total mean squared error due to variance includes the following: (1) $\operatorname{trace}\{\mathbb{V}\mathrm{ar}[\hat{\boldsymbol{\beta}}]\}$—the variance due to randomness from the model assumptions; (2) $\sigma^2 \operatorname{trace}\{\mathbf{X}^\dagger \left(\mathbb{E}_{\mathbf{s}}[\mathbf{P}\mathbf{P}^T] - \mathbf{P}_{\mathbf{x}}\right) (\mathbf{X}^\dagger)^T\}$—the excess variance due to the deviation of the oblique projector $\mathbf{P}$ from being an orthogonal projector onto range($\mathbf{X}$); and (3) $\operatorname{trace}\{\mathbb{V}\mathrm{ar}_{\mathbf{s}}[\mathbf{P}_{\mathbf{0}}\boldsymbol{\beta}_0]\}$—the excess variance due to rank deficiency arising from randomness in the sketching algorithm. Additionally, the bias portion of the total mean squared error represents the excess bias due to rank deficiency from the sketching process.

The *total excess mean squared error* denotes the portion of the mean squared error attributable to randomness in the sketching algorithm. This represents the portion of $MSE(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}_0)$ exceeding $MSE(\hat{\boldsymbol{\beta}}, \mathrm{MSE}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0)$, the mean squared error due to model-induced randomness. Using the notation in Section 5, we can rewrite the total mean squared error for the sketched solution $\tilde{\boldsymbol{\beta}}$ as

$$MSE(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}_0) = MSE(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0) + \underbrace{\operatorname{trace}\{\mathscr{V}_{\mathbf{P}}\} + \operatorname{trace}\{\mathscr{V}_{\mathbf{P_0}}\} + \|\mathscr{B}\|_2^2}_{\mathscr{M}},$$

where $\mathscr{M}$ denotes the *total excess mean squared error*. Thus, the excess total mean squared error can be decomposed into three sources with interpretation as stated above. Conditioning on sketching schemes that preserve rank provides simplifications and additional insights on the total mean squared error.

For the problem in (2.3) conditioned on $\operatorname{rank}(\mathbf{SX}) = \operatorname{rank}(\mathbf{X})$, the solution $\tilde{\boldsymbol{\beta}}$ has total mean squared error

$$MSE(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}_0) = \operatorname{trace}\{\mathbb{V}\mathrm{ar}[\hat{\boldsymbol{\beta}}]\} + \sigma^2 \operatorname{trace}\{\mathbf{X}^\dagger \left(\mathbb{E}_{\mathbf{s}}[\mathbf{P}\mathbf{P}^T] - \mathbf{P}_{\mathbf{x}}\right) (\mathbf{X}^\dagger)^T\}.$$

Therefore, we additionally have

$$MSE(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}_0) \geqslant MSE(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0).$$

Proof. The expression for the mean squared error follows from the fact that both $\tilde{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}$ are unbiased estimators of $\boldsymbol{\beta}_0$ in this case. Therefore, the mean squared error is the trace of the variance. For the inequality, we again employ the properties of the trace operator and linearity of the trace and expectation

to obtain

$$\begin{aligned} \mathrm{MSE}(\tilde{\boldsymbol{\beta}}, \boldsymbol{\beta}_0) &= \mathbb{E}[\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2] \\ &= \operatorname{trace}\{\mathbb{E}[(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T]\} = \operatorname{trace}\{\mathbb{V}\mathrm{ar}(\tilde{\boldsymbol{\beta}})\} \\ &= \sigma^2 \operatorname{trace}\{(\mathbf{X}^T\mathbf{X})^{-1}\} + \sigma^2 \operatorname{trace}\{\mathbf{X}^\dagger \left(\mathbb{E}_{\mathbf{S}}[\mathbf{P}\mathbf{P}^T] - \mathbf{P}_{\mathbf{X}}\right) (\mathbf{X}^\dagger)^T\} \\ &\geqslant \mathrm{MSE}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0). \end{aligned}$$

Once again, conditioning on rank preservation gives us $\mathbf{P}\mathbf{P}_{\mathbf{X}}\mathbf{P}^{\mathsf{T}} = \mathbf{P}_{\mathbf{X}}$ so that $\mathscr{V}_{\mathbf{P}}$ is positive semi-definite since $\mathbf{I} - \mathbf{P}_{\mathbf{X}}$ is idempotent. Since the trace of a positive semi-definite matrix is non-negative, the result follows from the fact that $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}_0$. □

Corollary 6.2 shows that when conditioning on rank preservation, the excess bias and variance due to rank deficiency, $\mathscr{B}$ and $\mathscr{V}_{\mathbf{P}_0}$, vanish. Therefore, the *excess total mean squared error* in this case is simply

$$\mathscr{M}' \equiv \sigma^2 \operatorname{trace}\{\mathbf{X}^\dagger \left(\mathbb{E}_{\mathbf{S}}[\mathbf{P}\mathbf{P}^T] - \mathbf{P}_{\mathbf{X}}\right) (\mathbf{X}^\dagger)^T\} = \operatorname{trace}\{\mathscr{V}_{\mathbf{P}_{\mathbf{X}}}\}.$$

As we saw in the explanation of $\mathscr{V}_{\mathbf{P}_{\mathbf{X}}}$ following Corollary 5.2, $\mathscr{V}_{\mathbf{P}_{\mathbf{X}}}$ in this case quantifies the excess variance due to the deviation of $\operatorname{null}(\mathbf{P})$ from $\operatorname{null}(\mathbf{P}_{\mathbf{X}})$.

Corollary 6.2 also shows that even conditioning on rank preservation so that $\tilde{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}_0$, the total mean squared error of $\tilde{\boldsymbol{\beta}}$ is at least as great as that of $\hat{\boldsymbol{\beta}}$. The decomposition of the total mean squared error in Corollary 6.2 shows that there are two reasons for this. First, $\tilde{\boldsymbol{\beta}}$ inherits the model variance $\mathbb{V}\mathrm{ar}(\hat{\boldsymbol{\beta}})$. Second, $\tilde{\boldsymbol{\beta}}$ additionally acquires excess variance $\mathscr{V}_{\mathbf{P}_{\mathbf{X}}}$ from the perturbation of $\operatorname{null}(\mathbf{P}_{\mathbf{X}})$ through sketching.

For the problem in (2.3), the solution $\tilde{\boldsymbol{\beta}}$ has total predictive risk equal to

$$\begin{aligned} \mathrm{R}(\tilde{\mathbf{y}}, \mathbf{X}\boldsymbol{\beta}_0) &= \mathrm{R}(\hat{\mathbf{y}}, \mathbf{X}\boldsymbol{\beta}_0) + \sigma^2 \operatorname{trace}\{\mathbb{E}_{\mathbf{S}}[\mathbf{P}\mathbf{P}^{\mathsf{T}}] - \mathbf{P}_{\mathbf{X}}\} \\ &\quad + \|(\mathbb{E}_{\mathbf{S}}[\mathbf{P}\mathbf{P}^{\mathsf{T}}] - \mathbf{P}_{\mathbf{X}})\mathbf{X}\boldsymbol{\beta}_0\|_2^2. \end{aligned}$$

Proof. Using the properties of the trace operator and the linearity of the trace and expectation, we obtain the following bias-variance decomposition for the predictive risk:

$$\mathrm{R}(\tilde{\mathbf{y}}, \mathbf{X}\boldsymbol{\beta}_0) = \mathbb{E}[\|\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}_0\|_2^2] = \operatorname{trace}\{\mathbb{V}\mathrm{ar}[\tilde{\mathbf{y}}]\} + \|\mathrm{Bias}(\tilde{\mathbf{y}}, \mathbf{X}\boldsymbol{\beta}_0)\|_2^2.$$

The total variance of $\tilde{\mathbf{y}}$ follows from applying the law of total expectation to the sketched prediction $\mathbf{P}\mathbf{y}$. The result follows from the facts that $\hat{\mathbf{y}}$ is an unbiased estimator for $\mathbf{X}\boldsymbol{\beta}_0$ so that $\mathrm{R}(\hat{\mathbf{y}}, \mathbf{X}\boldsymbol{\beta}_0) = \mathbb{V}\mathrm{ar}[\hat{\mathbf{y}}]$ and $\mathbf{P}_{\mathbf{X}}\mathbf{X} = \mathbf{X}$. □

Corollary 6.3 shows that the predictive risk can be decomposed into the following three sources: (1) $\mathrm{R}(\hat{\mathbf{y}}, \mathbf{X}\boldsymbol{\beta}_0)$—the prediction variance inherent in the model; (2) $\sigma^2 \operatorname{trace}\{\mathbb{E}_{\mathbf{S}}[\mathbf{P}\mathbf{P}^{\mathsf{T}}] - \mathbf{P}_{\mathbf{X}}\}$—the excess prediction variance due to the expected deviation of $\mathbf{P}$ from $\mathbf{P}_{\mathbf{X}}$; and (3) $\|(\mathbb{E}_{\mathbf{S}}[\mathbf{P}\mathbf{P}^{\mathsf{T}}] - \mathbf{P}_{\mathbf{X}})\mathbf{X}\boldsymbol{\beta}_0\|_2^2$—the excess prediction bias arising from the expected deviation of $\mathbf{P}$ from $\mathbf{P}_{\mathbf{X}}$.

The *excess predictive risk* represents the portion of the predictive risk attributable to randomness in the sketching algorithm. Corollary 6.3 shows that it is equal to

$$\mathscr{R} \equiv \underbrace{\sigma^2 \operatorname{trace}\{\mathbb{E}_{\mathbf{S}}[\mathbf{P}\mathbf{P}^{\mathsf{T}}] - \mathbf{P}_{\mathbf{X}}\}}_{\mathscr{R}_{\mathbf{V}}} + \underbrace{\|(\mathbb{E}_{\mathbf{S}}[\mathbf{P}\mathbf{P}^{\mathsf{T}}] - \mathbf{P}_{\mathbf{X}})\mathbf{X}\boldsymbol{\beta}_0\|_2^2}_{\mathscr{R}_{\mathbf{B}}},$$

where the excess predictive variance $\mathscr{R}_{\mathbf{V}}$ and excess predictive bias $\mathscr{R}_{\mathbf{B}}$ have interpretation as stated above.

Notice that the bias projector $\mathbf{P}_{\mathbf{0}}$ does not appear in expressions for the total predictive risk. Therefore, the predictive risk remains unaffected by expected rank preservation and the effects of algorithmic-induced randomness on it are restricted to the deviation of $\mathbf{P}$ from $\mathbf{P}_{\mathbf{X}}$. Thus, compared with the total variance and mean squared error for the true parameter, the total predictive risk is less affected by algorithmic-induced randomness.

For the problem in (2.3) conditioned on $\operatorname{rank}(\mathbf{SX}) = \operatorname{rank}(\mathbf{X})$, the solution $\tilde{\boldsymbol{\beta}}$ has total predictive risk equal to

$$\mathrm{R}(\tilde{\mathbf{y}}, \mathbf{X}\boldsymbol{\beta}_0) = \mathrm{R}(\hat{\mathbf{y}}, \mathbf{X}\boldsymbol{\beta}_0) + \sigma^2 \operatorname{trace}\{\mathbb{E}_{\mathbf{S}}[\mathbf{P}\mathbf{P}^{\mathsf{T}}] - \mathbf{P}_{\mathbf{X}}\}.$$

Therefore, we additionally have

$$\mathrm{R}(\tilde{\mathbf{y}}, \mathbf{X}\boldsymbol{\beta}_0) \geqslant \mathrm{R}(\hat{\mathbf{y}}, \mathbf{X}\boldsymbol{\beta}_0).$$

Corollary 6.4 follows from the following facts when conditioning on $\operatorname{rank}(\mathbf{SX}) = \operatorname{rank}(\mathbf{X})$. First, $\mathbf{Py}$ is an unbiased estimator for $\mathbf{X}\boldsymbol{\beta}_0$ so that the excess predictive bias $\mathscr{R}_{\mathbf{B}}$ vanishes. Secondly, $\mathbf{P}\mathbf{P}_{\mathbf{X}}\mathbf{P}^{\mathsf{T}} = \mathbf{P}_{\mathbf{X}}$ so that the excess predictive variance $\mathscr{R}_{\mathbf{V}}$ is positive semi-definite.

The *excess predictive risk* in this case is given by

$$\mathscr{R}' \equiv \sigma^2 \operatorname{trace}\{\mathbb{E}_{\mathbf{S}}[\mathbf{P}\mathbf{P}^{\mathsf{T}}] - \mathbf{P}_{\mathbf{X}}\} = \mathscr{R}_{\mathbf{V}},$$

representing the excess predictive variance due to the deviation of $\operatorname{null}(\mathbf{P})$ from $\operatorname{null}(\mathbf{P}_{\mathbf{X}})$. Notice that although the bias projector $\mathbf{P}_{\mathbf{0}}$ does not appear in the unconditional total predictive risk in Corollary 6.3, the predictive risk still decreases when conditioning on rank preservation. This is because the predictive bias $\operatorname{Bias}(\tilde{\mathbf{y}}, \mathbf{X}\boldsymbol{\beta}_0)$ depends only on the deviation of $\operatorname{range}(\mathbf{P})$ from $\operatorname{range}(\mathbf{P}_{\mathbf{X}})$. Since these are equal when conditioning on $\operatorname{rank}(\mathbf{SX}) = \operatorname{rank}(\mathbf{X})$, the predictive bias vanishes in this case.

Notice additionally that although $\operatorname{range}(\mathbf{P}) = \operatorname{range}(\mathbf{P}_{\mathbf{X}})$ in this case, we still have $\operatorname{null}(\mathbf{P}) \neq \operatorname{null}(\mathbf{P}_{\mathbf{X}})$ in general. Therefore, the predictive risk contains excess predictive variance $\mathscr{R}_{\mathbf{V}}$ arising from the expected deviation of $\operatorname{null}(\mathbf{P})$ from $\operatorname{null}(\mathbf{P}_{\mathbf{X}})$.

Corollary 6.4 shows that even when conditioning on sketching schemes that preserve rank so that $\tilde{\mathbf{y}}$ is an unbiased estimator of $\mathbf{X}\boldsymbol{\beta}_0$, the total predictive risk of $\tilde{\mathbf{y}}$ is at least as great as that of $\hat{\mathbf{y}}$. This is because $\tilde{\mathbf{y}}$ inherits the predictive variance due to model-induced randomness. Additionally, it acquires excess predictive variance arising from the perturbation of $\operatorname{null}(\mathbf{P}_{\mathbf{X}})$ under sketching.

7. Sketching Diagnostics

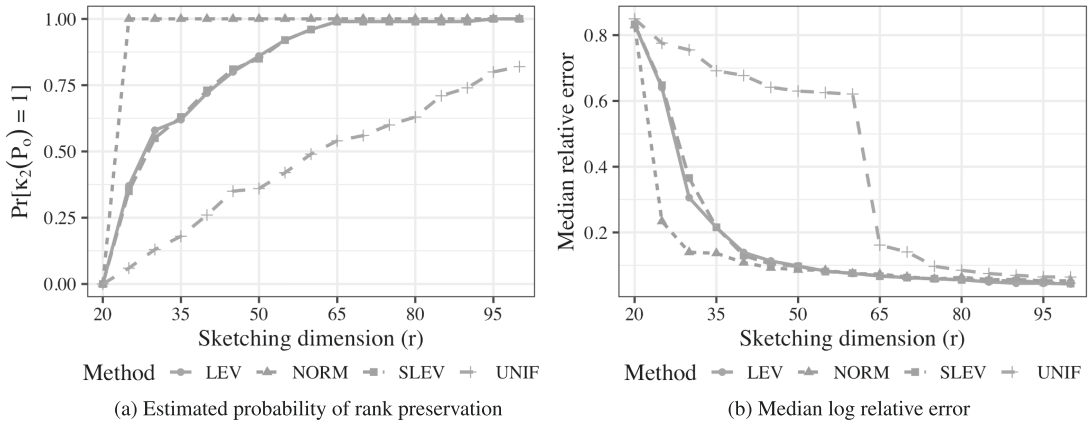
In the previous sections, we observed that the bias, and hence expected accuracy, of the sketched solution and prediction hinge on rank preservation. A natural consequence is that the bias projector $\mathbf{P_0}$ proves ideal for use in a sketching diagnostic. Compared with $\mathbf{P} \in \mathbb{R}^{n\times n}$, which may be computationally expensive for large n, $\mathbf{P_0} \in \mathbb{R}^{p\times p}$ can be computed quickly and inexpensively. Moreover, if rank is preserved, $\mathbf{P_0} = \mathbf{I}_p$ so that its two-norm condition number $\kappa_2(\mathbf{P_0})$ becomes a simple diagnostic for rank preservation: if $\kappa_2(\mathbf{P_0}) = 1$, then the sketching process preserves rank. Otherwise, it does not.

We illustrate how one can employ $\mathbf{P_0}$ as a sketching diagnostic to aid in the practical design of judicious sketching schemes when $\mathbf{y}$ is well represented by the column space of $\mathbf{X}$ as indicated by the angle between $\mathbf{y}$ and range($\mathbf{X}$) in Corollary 3.1. We also show that in this case, $\mathbf{P_0}$ can be utilized in selecting a suitable sketching dimension r. Examples illustrating the scenario in which $\mathbf{y}$ is not well represented by the column space of $\mathbf{X}$ can be found in [6, Section 2.7]. In those cases, however, the least squares residual is large so that even numerical accuracy for the full data problem is not guaranteed. Consequently, we do not highlight experiments involving sketching for those scenarios here.

To simulate realistic regression data satisfying a Gaussian linear model, we build a linear model based on data from the 2018 American Community Survey (ACS) 1-year Public Use Microdata Sample (PUMS) from the US Census Bureau. The ACS collects population and housing information on individuals and households across the USA to help guide policy-making. Technical details regarding the ACS PUMS files can be found at [36]. We employ the ACS PUMS from California as a foundation for realistic survey data from a large and diverse population.

For our initial response $\mathbf{y}'$, we utilize the gross rent as a percentage of annual household income, and subset for respondents with responses for this variable. For our initial design $\mathbf{X}'$, we employ the following economic, language and household status variables: food stamp program participation, primary household language, limited English proficiency status as a household, multigenerational household status and citizenship status. We also employ the following control variables: age, sex, marital status and education level of the respondent. We obtain our final design $\mathbf{X}$ with $n = 105{,}142$ respondents and $p = 21$ variables after standard recoding for categorical variables and appending a column of ones for the intercept. To obtain a Gaussian linear model, we simulate $\mathbf{y}$ as follows. We obtain $\boldsymbol{\beta}_0$ by regressing $\mathbf{y}'$ onto $\mathbf{X}$ and then setting entries in the resulting estimator corresponding to non-significant variables to zero. We then obtain $\mathbf{y} \equiv \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ follows a zero mean multivariate Gaussian distribution with $\sigma^2 = 10^{-12}$. While problems of size $n = 105{,}142$ can be readily solved with modern statistical software such as R or SAS, this value of n is a function of the example dataset. We employ it to illustrate the usefulness of $\mathbf{P_0}$ as a diagnostic when $\mathbf{y}$ is well represented by the column space of $\mathbf{X}$. We also highlight that our analysis applies even in the non-asymptotic regime.

We conduct numerical simulations with $\mathbf{y}$ and $\mathbf{X}$ and compare each $\tilde{\boldsymbol{\beta}}$ to $\hat{\boldsymbol{\beta}}$ obtained on the same data. We compare performance on four sketching schemes: (1) uniform sampling with replacement (UNIF) with sampling probabilities $\pi_{\text{unif}_i} = \frac{1}{n}$ for $1 \leqslant i \leqslant n$; (2) leverage score sampling with replacement (LEV) [22, 23] with sampling probabilities $\pi_{\text{lev}_i} = \frac{l_i}{p}$, where l_i denotes the ith leverage score for $1 \leqslant i \leqslant n$; (3) shrinkage leverage score sampling (SLEV), a convex combination of LEV and UNIF with sampling probabilities $\pi_{\text{slev}_i} = \alpha\,\pi_{\text{lev}_i} + (1-\alpha)\,\pi_{\text{unif}_i}$ where $\alpha = 0.9$ as recommended in [23] for $1 \leqslant i \leqslant n$; and 4) random projections with a matrix whose entries are standard Gaussian random variables (NORM). These sketching schemes enter our analysis through the random sketching matrix $\mathbf{S}$. For UNIF, LEV and SLEV, the rows of $\mathbf{S}$ are the rows sampled from $\mathbf{X}$. For NORM, the entries of $\mathbf{S}$ are standard Gaussian random variables.



(a) Estimated probability of rank preservation

(b) Median log relative error

FIG. 1. *Simulation results illustrate the pattern between (a) rank preservation and (b) median log relative error of* $\tilde{\boldsymbol{\beta}}$ *with respect to* $\hat{\boldsymbol{\beta}}$ *as a function of sketching method and dimension.*

To illustrate how rank preservation varies with r, we perform simulations over a range of sketching dimensions. These range from $r = 20 < 21 = p$, so that all simulations perform poorly, to $r = 100$, where most simulations perform well. We run 100 replicates of each scenario.

Figure 1(a) depicts $\Pr[\kappa_2(\mathbf{P_0}) = 1]$, the estimated probability of rank preservation, over the 100 replicates for each scenario. We observe that the r at $\Pr[\kappa_2(\mathbf{P_0}) = 1] > 0.50$ corresponds to the r where the relative error transitions from high to low in Fig. 1(b). NORM, LEV and SLEV achieve $\Pr[\kappa_2(\mathbf{P_0}) = 1] > 0.5$ at $r = 25$, $r = 30$ and $r = 30$, respectively, and their relative errors likewise drop then. UNIF achieves $\Pr[\kappa_2(\mathbf{P_0}) = 1] > 0.5$ at $r = 65$ so it transitions to low relative error at $r = 65$. Since we employ $\alpha = 0.9$ as recommended in [23] for the SLEV sampling probabilities, SLEV is very similar to LEV and it is unsurprising that they perform similarly with respect to rank preservation and median relative error.

Figure 1 illustrates that since $\kappa_2(\mathbf{P_0}) = 1$ correlates with low relative error, it can provide an inexpensive diagnostic for candidate sketching matrices. Figure 1 also shows that given a class of sketching matrices, one can employ $\Pr[\kappa_2(\mathbf{P_0}) = 1]$ in selecting an appropriate r. For example, in this illustrative problem, the numerical results shown in Fig. 1 would suggest selecting $r = 25$ if employing Gaussian sketching. This may be useful in solving large iterative linear systems where it may be impractical to hand-select a sketching matrix at each iteration.

8. Extensions to High-Dimensional Scenarios

In this work, we focus on the simplest class of least squares problems, those of full column rank since they are well-posed and regularization is not required to ensure the existence of a unique solution in exact arithmetic. Meanwhile, there are many machine learning algorithms for high-dimensional problems, where the number of columns p greatly exceeds the number of rows n. There already exists some analysis for sketching in such scenarios. For example, [7, Section 6.2] analyzes the accuracy of the sketched solution with respect to algorithmic uncertainty only.

During the review of this paper, a referee asked a natural question: can we extend our analysis to the high-dimensional case by adding a ridge penalty? Indeed, an important open problem is the

quantification of the combined model- and algorithm-induced uncertainties in sketched linear regression for high-dimensional problems. Therefore, for completeness, we describe potential extensions for analyzing under-determined regression scenarios with our projector-based framework from Section 3 here. As an example, we focus on ridge regression for the $n < p$ case, which involves regularization with the squared ℓ_2-norm. Previous work on sketched ridge regression include both the $n > p$ case [40] and the $n < p$ case [21], where the authors provide a bound for the inflation of the ridge regression risk function due to sketching with the subsampled randomized Hadamard transform.

Let $f(\boldsymbol{\beta}_r)$ denote the ridge regression loss function, where

$$f(\boldsymbol{\beta}_r) \;=\; \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_r\|_2^2 + \frac{\lambda}{2}\|\boldsymbol{\beta}_r\|_2^2. \tag{8.1}$$

It is well known that the ridge regression solution for the full data problem is

$$\hat{\boldsymbol{\beta}}_r \;=\; \left[(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} + \lambda\mathbf{I}\right]^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}. \tag{8.2}$$

Noting that (8.1) can be equivalently written as

$$f(\boldsymbol{\beta}_r) \;=\; \frac{1}{2}\left\|\begin{pmatrix}\mathbf{y}\\ \mathbf{0}\end{pmatrix} - \begin{pmatrix}\mathbf{X}\\ \sqrt{\lambda}\,\mathbf{I}\end{pmatrix}\boldsymbol{\beta}_r\right\|_2^2$$

and applying a random sketching matrix $\mathbf{S}$ on the left in the same manner as in (2.3) gives us the following sketched ridge regression loss function:

$$g(\boldsymbol{\beta}_r) \;=\; \frac{1}{2}\left\|\mathbf{S}\begin{pmatrix}\mathbf{y}\\ \mathbf{0}\end{pmatrix} - \mathbf{S}\begin{pmatrix}\mathbf{X}\\ \sqrt{\lambda}\,\mathbf{I}\end{pmatrix}\boldsymbol{\beta}_r\right\|_2^2 \;=\; \frac{1}{2}\|\mathbf{S}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_r)\|_2^2 + \frac{\lambda}{2}\|\mathbf{S}\boldsymbol{\beta}_r\|_2^2.$$

Differentiating g with respect to $\boldsymbol{\beta}_r$, setting it equal to zero and solving for $\boldsymbol{\beta}_r$ gives the following sketched ridge regression solution:

$$\tilde{\boldsymbol{\beta}}_r \;=\; \left[(\mathbf{SX})^\mathsf{T}\mathbf{SX} + \lambda\,\mathbf{S}\right]^{-1}(\mathbf{SX})^\dagger\mathbf{Sy}. \tag{8.3}$$

Rewriting (8.2) and (8.3) in terms of the projectors in Section 3 gives us

$$\hat{\boldsymbol{\beta}}_r \;=\; \left[(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} + \lambda\mathbf{I}\right]^{-1}\mathbf{X}^\mathsf{T}\mathbf{P}_\mathbf{x}\mathbf{y} \quad\text{and}\quad \tilde{\boldsymbol{\beta}}_r \;=\; \left[(\mathbf{SX})^\mathsf{T}\mathbf{SX} + \lambda\,\mathbf{S}\right]^{-1}\mathbf{X}^\dagger\mathbf{Py} \tag{8.4}$$

for the full data and sketched solutions, respectively. Therefore, we can express the sketched ridge regression solution in terms of the projectors from Section 3 as

$$\tilde{\boldsymbol{\beta}}_r \;=\; \hat{\boldsymbol{\beta}}_r + \left\{\left[(\mathbf{SX})^\mathsf{T}\mathbf{SX} + \lambda\,\mathbf{S}\right]^{-1}\mathbf{X}^\dagger\mathbf{P} - \left[(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} + \lambda\mathbf{I}\right]^{-1}\mathbf{X}^\mathsf{T}\mathbf{P}_\mathbf{x}\right\}\mathbf{y}.$$

This formulation makes clear why a corresponding analysis of the combined uncertainties for sketched ridge regression is nontrivial. First, we observe that $\mathbf{S}$ enters both multiplicatively and additively within an inverse. Secondly, $\mathbf{S}$ additionally appears in $\mathbf{P}$ and the inverse term is multiplied on the left of a term containing $\mathbf{P}$. Nonetheless, sketching for high-dimensional linear regression problem is relevant in

many machine learning algorithms. An analysis of its combined uncertainties is an important direction for future work.

9. Discussion

We presented a projector-based approach for sketched linear regression and analyzed the combined uncertainties on the sketched solution $\tilde{\boldsymbol{\beta}}$ from both statistical noise in the model and randomness from the sketching algorithm. Our results show that the total expectation and variance of $\tilde{\boldsymbol{\beta}}$ are governed by the spatial geometry of the sketching process, rather than by structural properties of specific sketching matrices. Surprisingly, the condition number $\kappa_2(\mathbf{X})$ with respect to (left) inversion has far less impact on the statistical measures than it has on the numerical errors.

Our results demonstrate the usefulness of a projector-based approach in enabling expressions for quantifying the total and excess uncertainties that hold generally for *all* sketching schemes. A projector-based approach also enables insights and interpretations on how the sketching process affects the solution and other key statistical quantities. Our numerical experiments illustrate the practicality of the bias projector $\mathbf{P_0}$ as a computationally inexpensive and effective sketching diagnostic under a Gaussian linear model when $\mathbf{y}$ is well represented by the column space of $\mathbf{X}$.

Finally, we began with the simplest class of least squares problems in this work: those of full column rank because they are well-posed and regularization is not required to ensure the existence of a unique solution in exact arithmetic. Consequently, there are many avenues for future work in extending our work to more complex least squares problems. These include the highly under-determined scenario as well as their corresponding regularized versions for high-dimensional problems as discussed in Section 8.

Data Availability Statement

The data underlying this article were derived from sources in the public domain: US Census Bureau, American Community Survey, Public Use Microdata Sample files available at https://www.census.gov/programs-surveys/acs/microdata/access.2018.html.

Acknowledgements

We thank the anonymous reviewers for their helpful comments and suggestions.

Funding

National Science Foundation (DMS-1760374 and DMS-1745654).

References

1. Avron, H., Maymounkov, P. & Toledo, S. (2010) Blendenpik: supercharging LAPACK's least squares solver. *SIAM J. Sci. Comput.*, **32**, 1217–1236.
2. Boutsidis, C. & Drineas, P. (2009) Random projections for the nonnegative least squares problem. *Linear Algebra Appl.*, **431**, 760–771.
3. Brust, J. J., Marcia, R. F. & Petra, C. G. (2020) Computationally efficient decompositions of oblique projection matrices. *SIAM J. Matrix Anal. Appl.*, **41**, 852–870.
4. Casella, G. & Berger, R. L. (2002) *Statistical Inference, vol. 2*. Pacific Grove, CA, USA: Duxbury Press.

5. Chatterjee, S. & Hadi, A. S. (1986) Influential observations, high leverage points, and outliers in linear regression. *Statist. Sci.*, **1**, 379–416 With discussion.
6. Chi, J. T. (2021) *Algorithms and analysis for the efficient solution of large-scale linear least squares problems. Ph.D. Thesis*. Raleigh, NC: North Carolina State University.
7. Drineas, P., Magdon-Ismail, M., Mahoney, M. W. & Woodruff, D. P. (2012) Fast approximation of matrix coherence and statistical leverage. *J. Mach. Learn. Res.*, **13**, 3475–3506.
8. Drineas, P., Mahoney, M. W. & Muthukrishnan, S. (2006) *Sampling algorithms for l_2 regression and applications. Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. New York: ACM, pp. 1127–1136.
9. Drineas, P., Mahoney, M. W., Muthukrishnan, S. & Sarlós, T. (2011) Faster least squares approximation. *Numer. Math.*, **117**, 219–249.
10. Drmač, Z. & Saibaba, A. K. (2018) The discrete empirical interpolation method: canonical structure and formulation in weighted inner product spaces. *SIAM J. Matrix Anal. Appl.*, **39**, 1152–1180.
11. Golub, G. H. & Van Loan, C. F. (2013) *Matrix Computations*, 4th edn. Baltimore: The Johns Hopkins University Press.
12. Hansen, P. C. (2013) Oblique projections and standard-form transformations for discrete inverse problems. *Numer. Linear Algebra Appl.*, **20**, 250–258.
13. Higham, N. J. (2002) *Accuracy and Stability of Numerical Algorithms*, 2nd edn. Philadelphia: SIAM.
14. Hoaglin, D. C. & Welsch, R. E. (1978) The hat matrix in regression and ANOVA. *Amer. Statist.*, **32**, 17–22.
15. Ipsen, I. C. F. (1998) Relative perturbation results for matrix eigenvalues and singular values. *Acta Numer.*, **7**, 151–201.
16. Ipsen, I. C. F. (2000) An overview of relative $\sin\Theta$ theorems for invariant subspaces of complex matrices. *J. Comput. Appl. Math.*, **123**, 131–153. Invited paper for the special issue Numerical Analysis 2000: Vol. III—Linear Algebra.
17. Ipsen, I. C. F. & Wentworth, T. (2014) The effect of coherence on sampling from matrices with orthonormal columns and preconditioned least squares problems. *SIAM J. Matrix Anal. Appl.*, **35**, 1490–1520.
18. Kabán, A. (2014) New bounds on compressive linear least squares regression. *Artificial Intelligence and Statistics*. Reykjavik, Iceland: Journal of Machine Learning Research-Proceedings Track, pp. 448–456.
19. Lange, K. (2010) *Numerical Analysis for Statisticians, Statistics and Computing*, 2nd edn. New York: Springer.
20. Lopes, M., Wang, S. & Mahoney, M. (2018) Error estimation for randomized least-squares algorithms via the bootstrap. *International Conference on Machine Learning*. Stockholm, Sweden: PMLR, pp. 3217–3226.
21. Lu, Y., Dhillon, P. S., Foster, D. P. & Ungar, L. H. (2013) Faster ridge regression via the subsampled randomized Hadamard transform. *Advances in Neural Information Processing Systems*, vol. 32. Lake Tahoe, NV: Curran Associates, Inc.
22. Ma, P., Mahoney, M. W. & Yu, B. (2014) A statistical perspective on algorithmic leveraging. *Proceedings of the 31st International Conference on International Conference on Machine Learning (ICML)*, vol. 32. Beijing, China: JMLR, pp. 91–99.
23. Ma, P., Mahoney, M. W. & Yu, B. (2015) A statistical perspective on algorithmic leveraging. *J. Mach. Learn. Res.*, **16**, 861–911.
24. Ma, P., Zhang, X., Xing, X., Ma, J. & Mahoney, M. W. (2020) Asymptotic analysis of sampling estimators for randomized numerical linear algebra algorithms. *International Conference on Artificial Intelligence and Statistics*. Palermo, Sicily, Italy: PMLR, pp. 1026–1035.
25. Maillard, O. & Munos, R. (2009) Compressed least squares regression. *Advances in Neural Information Processing Systems*. Vancouver, B.C., Canada: Curran Associates, Inc., pp. 1213–1221.
26. Meng, X., Saunders, M. A. & Mahoney, M. W. (2014) LSRN: a parallel iterative solver for strongly over- or underdetermined systems. *SIAM J. Sci. Comput.*, **36**, C95–C118.
27. Raskutti, G. & Mahoney, M. W. (2016) A statistical perspective on randomized sketching for ordinary least squares. *J. Mach. Learn. Res.*, **17**, 7508–7538.

28. Rokhlin, V. & Tygert, M. (2008) A fast randomized algorithm for overdetermined linear least squares regression. *Proc. Natl. Acad. Sci. USA*, **105**, 13212–13217.

29. Sarlós, T. (2006) Improved approximation algorithms for large matrices via random projections. *47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*. Berkeley, CA: IEEE, pp. 143–152.

30. Searle, S. R. & Gruber, M. H. (2016) *Linear Models*. Hoboken, New Jersey: Wiley.

31. Stewart, G. W. (1987) Collinearity and least squares regression. *Statist. Sci.*, **2**, 68–100. With discussion.

32. Stewart, G. W. (1989) On scaled projections and pseudoinverses. *Linear Algebra Appl.*, **112**, 189–193.

33. Stewart, G. W. (2011) On the numerical analysis of oblique projectors. *SIAM J. Matrix Anal. Appl.*, **32**, 309–348.

34. Stewart, G. W. & Sun, J. (1990) *Matrix Perturbation Theory*. San Diego: Academic Press.

35. Thanei, G.-A., Heinze, C. & Meinshausen, N. (2017) Random projections for large-scale regression. *Big and Complex Data Analysis*. Cham, Switzerland: Springer, pp. 51–68.

36. US Census Bureau (2018) *American Community Survey 1-Year Public Use Microdata Sample*. Washington D.C.: U.S. Census Bureau.

37. Černý, A. (2009) Characterization of the oblique projector $U(VU)^{\dagger}V$ with application to constrained least squares. *Linear Algebra Appl.*, **431**, 1564–1570.

38. Velleman, P. F. & Welsch, R. E. (1981) Efficient computing of regression diagnostics. *Amer. Statist.*, **35**, 234–242.

39. Wang, H., Zhu, R. & Ma, P. (2018) Optimal subsampling for large scale logistic regression. *J. Amer. Statist. Assoc.*, **113**, 829–844.

40. Wang, S., Gittens, A. & Mahoney, M. W. (2017) Sketched ridge regression: optimization perspective, statistical perspective, and model averaging. *International Conference on Machine Learning*. Sydney, Australia: PMLR, pp. 3608–3616.

41. Zhou, S., Wasserman, L. & Lafferty, J. D. (2008) Compressed regression. *Advances in Neural Information Processing Systems*. Vancouver, B.C., Canada: Curran Associates, Inc., pp. 1713–1720.