Technical Report

Random projections for Bayesian regression

Leo Geppert, Katja Ickstadt, Alexander Munteanu and Christian Sohler

04/2014

technische universität dortmund

SFB 876 Verfügbarkeit von Information durch Analyse unter Ressourcenbeschränkung

Part of the work on this technical report has been supported by Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center SFB 876 "Providing Information by Resource-Constrained Analysis", project C4.

Speaker: Prof. Dr. Katharina Morik
Address: TU Dortmund University
Joseph-von-Fraunhofer-Str. 23
D-44227 Dortmund
Web: http://sfb876.tu-dortmund.de

Abstract

This article introduces random projections applied as a data reduction technique for Bayesian regression analysis. We show sufficient conditions under which the entire d-dimensional distribution is preserved under random projections by reducing the number of data points from n to $k \in O(\text{poly}(d/\varepsilon))$ in the case $n \gg d$. Under mild assumptions, we prove that evaluating a Gaussian likelihood function based on the projected data instead of the original data yields a $(1+O(\varepsilon))$-approximation in the ℓ_2-Wasserstein distance. Our main result states that the posterior distribution of a Bayesian linear regression is approximated up to a small error depending on only an ε-fraction of its defining parameters when using either improper non-informative priors or arbitrary Gaussian priors. Our empirical evaluations involve different simulated settings of Bayesian linear regression. Our experiments underline that the proposed method is able to recover the regression model while considerably reducing the total run-time.

1 Introduction

Using a linear map $\Pi \in \mathbb{R}^{n\times k}$ whose choice is still to be defined, we transform the original data set $[X, Y] \in \mathbb{R}^{n\times(d+1)}$ into a sketch, i.e., a substitute data set, $[\Pi X, \Pi Y] \in \mathbb{R}^{k\times(d+1)}$ that is considerably smaller. Therefore, the likelihood function can be evaluated faster than on the original data. Moreover, we will show that the likelihood is very similar to the original one. In the context of Bayesian regression we have additional prior information $p_{\text{pre}}(\beta)$ in terms of a prior distribution over the parameters $\beta \in \mathbb{R}^d$ that we would like to estimate. Our main result will be to show that the resulting posterior distribution

$$p_{\text{post}}(\beta|X, Y) \propto \mathcal{L}(\beta|X, Y) \cdot p_{\text{pre}}(\beta)$$

will also be approximated within a small error.

The main idea of our approach is given in the following scheme:

$$\begin{array}{ccc} [X, Y] & \xrightarrow{\Pi} & [\Pi X, \Pi Y] \\ \downarrow & & \downarrow \\ p_{\text{post}}(\beta|X, Y) & \approx_\varepsilon & p_{\text{post}}(\beta|\Pi X, \Pi Y) \end{array}$$

More specifically we can choose $k \in O(\text{poly}(d/\varepsilon))$ which notably is independent of the number of data points n. Thus, the run-time of all subsequent calculations does not further depend on n. For instance, a Markov Chain Monte Carlo (MCMC) sampling algorithm may be used to obtain samples from an unknown distribution. Using the reduced data set will speed up the computations from several days to a few hours while the samples remain sufficiently accurate to resemble the original distribution and also to make statistical predictions that are nearly undistinguishable from the predictions that would have been made based on the full original sample.

2 Background and Related Work

Our proposed method projects the data set into a lower-dimensional subspace. Dimensionality reduction techniques, like e.g. principal component analysis [16], are commonly used in statistics. However, their focus is usually on reducing the number of variables. Our method aims to reduce the number of observations while keeping the algebraic structure of the data. This leads to a speed-up in the subsequent (frequentist or Bayesian) regression analysis, because the run-time of the common algorithms usually heavily depends on n.

Frequentist linear regression can be solved relatively straightforwardly using ordinary least squares. Bayesian regression, on the other hand, is typically computationally demanding. In some cases, calculating the posterior distribution analytically is possible, but in general, MCMC methods are standard in Bayesian analysis. They are reliable, but can take considerable time, before they converge and consequently sample from the desired posterior distribution. The run-time grows with the number of observations in the data-set.

There are multiple approaches trying to reduce the run-time of MCMC by employing more efficient algorithms. Approximate Bayesian Computing (ABC) and Integrated Nested Laplace Approximations (INLA) both fall into this category.

The main bottleneck of a lot of Bayesian analyses is the repeated evaluation of the likelihood. The main idea behind ABC is to avoid these evaluations by approximating the likelihood function using simulations [9]. INLA [24, 19] on the other hand is an approximation of the posterior distribution that is applicable to models that fall into the class of so-called latent Gaussian models. Both methods lead to a considerable speed-up compared to standard MCMC methods. Note however, that the speed-up is achieved by changing the algorithm, which is used to conduct the analysis. This is different in our approach, which reduces the number of observations in the data set while approximately retaining its statistical properties. The run-time of many algorithms including MCMC algorithms depends on the number of observations, which means that our proposed method also results in a speed-up of the analysis. In this article, we have only used MCMC methods for the analysis, but other algorithms that are based on the likelihood can also be used.

3 Preliminaries

A linear regression model is given in equation (1):

$$Y = X\beta + \xi. \tag{1}$$

$Y \in \mathbb{R}^n$ is a random variable containing the values of the response. n is the number of observations in the data set. $X \in \mathbb{R}^{n\times d}$ is a matrix containing the values of the d independent variables. $\xi \sim N(0, \sigma^2 I_n)$ is an n-dimensional random vector which models the unobservable error term. Y is also assumed to follow a normal distribution,

$Y \sim N(X\beta, \sigma^2 I_n)$. The corresponding probability density function is

$$f(z|X\beta, \Sigma) = (2\pi)^{-\frac{k}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(z - X\beta)'\Sigma(z - X\beta)\right), \tag{2}$$

where $\Sigma = \sigma^2 I_n$.

In a Bayesian setting, $\beta \in \mathbb{R}^d$ is the unknown parameter vector, which is assumed to follow an unknown distribution $p(\beta|X, Y)$ called the posterior distribution. Prior knowledge about β can be modeled using the prior distribution $p(\beta)$. The posterior distribution is a compromise between the prior distribution and the observed data.

In general, the posterior distribution cannot be calculated analytically. In this paper, we determine the posterior distribution employing Markov Chain Monte Carlo methods.

Before going into details about subspace embeddings, let us first define the Frobenius norm, which will be used as norm for matrices in this paper.

Definition 1 (matrix norms)**.** *For a matrix $A \in \mathbb{R}^{n\times d}$ the Frobenius norm is defined as*

$$\|A\|_F = \left(\sum_{i=1}^{n}\sum_{j=1}^{d} A_{i,j}^2\right)^{1/2}$$

and the spectral norm is defined by

$$\|A\|_2 = \sup_{x\in\mathbb{R}^d\setminus\{0\}} \frac{\|Ax\|_F}{\|x\|_F}\ .$$

In the special case of a vector $y \in \mathbb{R}^d$, both matrix norms coincide with the Euclidean length of y, i.e.,

$$\|y\|_F = \|y\|_2 = \left(\sum_{i=1}^{d} y_i^2\right)^{1/2}\ .$$

In several places throughout the paper it may be helpful to think only in terms of matrix norms and to treat the Euclidean norm of a d-dimensional vector as spectral norm of a $d \times 1$ matrix.

The following definition of so called ε-subspace embeddings will be central to our work. Such an embedding can be used to reduce the size of a given data matrix while preserving the algebraic structure of its spanned subspace up to $(1 \pm \varepsilon)$ distortion. Before we summarize several methods to construct a subspace embedding for a given input matrix we give a formal definition. Here and in the rest of the paper we assume $0 < \varepsilon \leq 1/2$.

Definition 2 (ε-subspace embedding)**.** *Given a matrix $U \in \mathbb{R}^{n\times d}$ with orthonormal columns, an integer $k \leq n$ and an approximation parameter $0 < \varepsilon \leq 1/2$, an ε-subspace embedding for U is a map $\Pi : \mathbb{R}^n \to \mathbb{R}^k$ such that*

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

holds for all $x \in \mathbb{R}^d$, or, equivalently

$$\|U^T\Pi^T\Pi U - I\|_2 \leq \varepsilon. \tag{4}$$

There are several ways to construct an ε-subspace embedding. One of the more recent methods is using a so called *graph-sparsifier* that was initially introduced for the efficient construction of sparse sub-graphs with good expansion properties [3]. The work of [4] adapted the technique to work for ordinary least-squares regression. While the initial construction was deterministic they also gave alternative constructions combining the deterministic decision rules with non-uniform random sampling techniques. Subspace preserving sampling of rows from the data matrix for ℓ_2-regression was introduced in [11] and generalized to more general subspace sampling for the p-norm. All the aforementioned methods have in common, that their (possibly random) construction depends on the input itself. For the streaming setting introduced in [20] this is not desirable because for instance one needs two passes over the data to perform the subspace sampling procedures, one for pre-computing the probabilities and another for the actual sampling. In order to make ε-subspace embeddings suitable for the design of single-pass streaming algorithms, we consider a different approach of so called *oblivious* subspace embeddings in this paper. These can be viewed as distributions over appropriately structured $k \times n$ matrices from which we can draw a realization Π independent of the input matrix. It is then guaranteed that for any fixed matrix U as in Definition 2, Π is an ϵ-subspace embedding with probability at least $1 - \delta$.

In this paper we consider three different approaches for obtaining oblivious ϵ-subspace embeddings:

1. The Rademacher Matrix (BCH): Π is obtained by choosing each entry independently from $\{-1, 1\}$ with equal probability. The matrix is then rescaled by $\frac{1}{\sqrt{k}}$. This method has been shown in [26] to form an ε-subspace embedding for essentially $k = O(\frac{d \log(d/\delta)}{\varepsilon^2})$. This was later improved in [7] to $k = O(\frac{d+\log(1/\delta)}{\varepsilon^2})$. While this is the best reduction among the methods that we used in the present work, the BCH embedding has the disadvantage that we need $\Theta(ndk)$ time to apply it to an $n \times d$ matrix.

2. The Subsampled Randomized Hadamard Transform (SRHT) (originally from [1]) is an embedding that is chosen to be $\Pi = RH_nD$ where D is an $m \times m$ diagonal matrix where each entry is independently sampled from $\{-1, 1\}$ with equal probability. The value of m is assumed to be a power of two. It is convenient to choose the smallest such number that is not smaller than n. H_m is the *Hadamard-matrix* of order m and R is a $k \times n$ row sampling matrix. That is, each row of R contains exactly one 1-entry and is 0 everywhere else. The index of the 1-entry is chosen uniformly from $[m]$ i.i.d. for every row. The matrix is then rescaled by $\frac{1}{\sqrt{k}}$. The target dimension of this family of matrices was shown to be $k = O(\frac{(\sqrt{d}+\sqrt{\log n})^2 \log(d/\delta)}{\varepsilon^2})$ [5] which improved upon previous results from [12]. Compared to the BCH method this is worse by essentially a factor of $O(\log d)$. It can be shown that $k = \Omega(d \log d)$ is necessary due to the sampling based approach by reduction from the coupon collectors theorem, see [15] for details. The benefit that we get is that due to the inductive structure of the Hadamard matrix, the embedding can be applied in $O(nd \log k)$ time which is considerably faster.

3. The most recent construction that we considered in this article is called the Clarkson Woodruff (CW) sketch [8]. In this case the embedding is obtained as $\Pi = \Phi D$. D is constructed in the same way as the diagonal matrix in the SRHT case. Given a random

map $h : [n] \to [k]$ such that for every $i \in [n]$ its image is chosen to be $h(i) = t \in [k]$ with probability $\frac{1}{k}$. Φ is again a binary matrix whose 1-entries can be defined by $\Phi_{h(i),i} = 1$. All other entries are 0. This is obviously the fastest embedding. It can be applied to any matrix $X \in \mathbb{R}^{n \times d}$ in $O(\text{nnz}(X)) = O(nd)$ time, where $\text{nnz}(X)$ denotes the number of non-zero entries in X. This is referred to as *input sparsity time* and is clearly optimal since this is the time needed to even read the input. However, its disadvantage is that the target dimension is $k = \Omega(d^2)$ [21]. Basically this is necessary due to the need to obliviously hash the standard basis vectors for $\mathbb{R}^d$ perfectly. Improved bounds over the original ones from [8] show that $k = O(\frac{d^2}{\varepsilon^2 \delta})$ is enough to draw an ϵ-subspace embedding from this distribution of matrices [22].

The results of our work are always conditioned on the event that the map Π is an ε-subspace embedding omitting to further mention the error probability of δ. The reader should keep in mind that there is the aforementioned possibility of failure during the phase of sketching the data.

Note that while the size of the resulting sketches does not depend on n, this is not true for the embedding matrices $\Pi \in \mathbb{R}^{k \times n}$. However, due to the structured constructions that we have surveyed above, we stress that the sketching matrices can be stored implicitly by the use of hash functions of bounded independence this has been proved for the different constructions for example in [2, 7, 22]. Common choices are the BCH scheme also used in [2] and the fast universal hashing scheme introduced in [10]. A survey on different methods can be found in [25]. These hash functions can be evaluated very efficiently using bit-wise operations and can be stored using a seed whose size is only $O(\log n)$. Note that even this small dependency on n is needed only for the sketching phase. After the sketch has been computed, the space requirements will be independent of n.

The linearity of the embeddings allows for efficient application in sequential streaming and in distributed environments, see e.g. [7, 18, 29]. The sketches can be updated in the most flexible dynamic setting, which is commonly referred to as the *turnstile* model [20] and allows for additive updates of rows, columns or even single entries. Note that by using negative updates, even deletions are possible in this setting. For distributed computations note that the embedding matrices can be communicated efficiently to every machine in the computing cluster. This is due to the small implicit representation by hash functions. Thus, every machine can compute a sketch on its own share of the data and communicate it to one dedicated central server. A sketch of the entire dataset can be obtained by simply summing up the single sketches since

$$X = \sum_{i=1}^{l} X_i$$

and consequently

$$\Pi X = \sum_{i=1}^{l} \Pi X_i.$$

For two probability measures γ, ν over $\mathbb{R}^d$, let $\Lambda(\gamma, \nu)$ denote the set of all joint probability measures over $\mathbb{R}^d \times \mathbb{R}^d$ with marginals γ and ν respectively.

Definition 3 (Wasserstein distance, cf. [28])**.** *For $p \in [1, \infty)$ let $(\mathbb{R}^d, \|\cdot\|_p)$ denote d-dimensional ℓ_p-space. Given two probability measures γ, ν on $\mathbb{R}^d$ the Wasserstein distance of order p between γ and ν is defined as*

$$\begin{aligned}&\mathcal{W}_p(\gamma, \nu)\\ &= \left(\inf_{\lambda \in \Lambda(\gamma,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|_p^p \ \mathrm{d}\lambda(x, y)\right)^{\frac{1}{p}}\end{aligned}$$

or equivalently

$$\begin{aligned}&\mathcal{W}_p(\gamma, \nu)\\ &= \inf\left\{\mathbb{E}\left[\|x - y\|_p^p\right]^{\frac{1}{p}} \,\middle|\, x \sim \gamma, y \sim \nu\right\}\end{aligned}$$

From the definition of the Wasserstein distance we can derive a measure of how much points drawn from a given distribution will spread from 0. The *Wasserstein weight* can be thought of as a norm of a probability measure.

Definition 4 (Wasserstein weight)**.** *For $p \in [1, \infty)$ let $(\mathbb{R}^d, \|\cdot\|_p)$ denote d-dimensional ℓ_p-space. We define the Wasserstein weight of order p of a probability measure γ as*

$$\begin{aligned}\mathcal{W}_p(\gamma) &= \mathcal{W}_p(\gamma, \delta)\\ &= \left(\int_{\mathbb{R}^d} \|x\|_p^p \ \mathrm{d}\gamma\right)^{\frac{1}{p}} = \mathbb{E}\left[\|x\|_p^p\right]^{\frac{1}{p}}\end{aligned}$$

where δ denotes the Dirac delta function.

4 Theory

4.1 Embedding the likelihood

In this section we introduce and develop the theoretical foundations of our approach and will combine existing results on ordinary least squares regression to bound the Wasserstein distance between the real likelihood function and its counterpart evaluated only on the considerably smaller sketch. Empirical evaluations supporting our theoretical results will be conducted in the subsequent section.

The following Observation is standard (cf. [14, 17]) and will be helpful in bounding the ℓ_2-Wasserstein distance of two Gaussian measures. It allows us to derive such a bound by inspecting their means and their covariances separately.

Observation 1. *Let $Z_1, Z_2 \in \mathbb{R}^d$ be random variables with finite first moments $\mu_1, \mu_2 < \infty$ and let $Z_1^m = Z_1 - \mu_1$ respectively $Z_2^m = Z_2 - \mu_2$ be their mean-centered counterparts. Then it holds that*

$$\mathbb{E}\left[\|Z_1 - Z_2\|_2^2\right] = \|\mu_1 - \mu_2\|_2^2 + \mathbb{E}\left[\|Z_1^m - Z_2^m\|_2^2\right]$$

Proof.

$$\begin{aligned}
& \mathbb{E}\left[\|Z_1 - Z_2\|_2^2\right] \\
= \; & \mathbb{E}\left[\|Z_1 - \mu_1 + \mu_1 - Z_2 + \mu_2 - \mu_2\|_2^2\right] \\
= \; & \mathbb{E}\left[\|Z_1^m - Z_2^m + \mu_1 - \mu_2\|_2^2\right] \\
= \; & \mathbb{E}\left[\|Z_1^m - Z_2^m\|_2^2 + \|\mu_1 - \mu_2\|_2^2\right] \\
& + 2\,(\mu_1 - \mu_2)^T \underbrace{\mathbb{E}\left[Z_1^m - Z_2^m\right]}_{=0} \\
= \; & \mathbb{E}\left[\|Z_1^m - Z_2^m\|_2^2\right] + \|\mu_1 - \mu_2\|_2^2
\end{aligned}$$

□

In our first lemma we show that using an ε-subspace embedding Π for the columnspace of $[X, Y]$, we can approximate the least squares regression problem up to a factor of $1+\varepsilon$. That is, we can find a solution ν by projecting ΠY into the columnspace of ΠX such that $\|X\nu - Y\|_2 \leq (1+\varepsilon)\min_{\beta\in\mathbb{R}^d}\|X\beta - Y\|_2$. Similar proofs can be found in [6, 4]. We repeat the result here for completeness.

Lemma 5. *Given $X \in \mathbb{R}^{n\times d}, Y \in \mathbb{R}^n$, let Π be an $(\varepsilon/3)$-subspace embedding for the columnspace of $[X, Y]$. Let $\gamma = \operatorname{argmin}_{\beta\in\mathbb{R}^d}\|X\beta - Y\|_2^2$ and let $\nu = \operatorname{argmin}_{\beta\in\mathbb{R}^d}\|\Pi(X\beta - Y)\|_2^2$. Then*

$$\|X\nu - Y\|_2^2 \leq (1+\varepsilon)\|X\gamma - Y\|_2^2$$

Proof. Let $[X, Y] = U\Sigma V^T$ denote the SVD of $[X, Y]$. Now define $\eta_1 = \Sigma V^T[\gamma^T, -1]^T$ and $\eta_2 = \Sigma V^T[\nu^T, -1]^T$. Using this notation we can rewrite $U\eta_1 = X\gamma - Y$ and similarly $U\eta_2 = X\nu - Y$. We have that

$$\begin{aligned}
(1-\varepsilon/3)\|U\eta_2\|_2^2 &\leq \|\Pi U\eta_2\|_2^2 \\
&\leq \|\Pi U\eta_1\|_2^2 \\
&\leq (1+\varepsilon/3)\|U\eta_1\|_2^2.
\end{aligned}$$

The first and the last inequalities are direct applications of the subspace embedding property 3, where the middle inequality follows from the optimality of ν in the embedded subspace.

Now by rearranging and resubstituting terms this yields

$$\begin{aligned}
\|X\nu - Y\|_2^2 &\leq \left(\frac{1+\varepsilon/3}{1-\varepsilon/3}\right)\|X\gamma - Y\|_2^2 \\
&\leq (1+\varepsilon)\|X\gamma - Y\|_2^2
\end{aligned}$$

□

In the following we investigate the distributions proportional to the likelihood functions $p \propto \mathcal{L}(\beta|X, Y)$ and $p' \propto \mathcal{L}(\beta|\Pi X, \Pi Y)$ and bound their Wasserstein distance.

We begin with a bound on the distance of their means γ and ν respectively.

Lemma 6. *Let X, Y, γ, ν be defined as in Lemma 5. Then*

$$\|\gamma - \nu\|_2^2 \leq \frac{\varepsilon}{\sigma_{\min}^2(X)} \|X\gamma - Y\|_2^2.$$

Proof. Let $X = U\Sigma V^T$ denote the SVD of X. Let $\eta = V^T(\gamma - \nu)$. First note that γ and ν are both contained in the columnspace of V (cf. [26]) which means that V^T is a proper rotation with respect to $\gamma - \nu$. Thus,

$$\begin{aligned}
\|X(\gamma - \nu)\|_2^2 &= \|U\Sigma V^T(\gamma - \nu)\|_2^2 \\
&= \|\Sigma V^T(\gamma - \nu)\|_2^2 \\
&= \sum \sigma_i^2(X)\eta_i^2 \\
&\geq \sum \sigma_{\min}^2(X)\eta_i^2 \\
&= \sigma_{\min}^2(X) \|V^T(\gamma - \nu)\|_2^2 \\
&= \sigma_{\min}^2(X) \|\gamma - \nu\|_2^2 .
\end{aligned}$$

Consequently, it remains to bound $\|X(\gamma - \nu)\|_2^2$. This can be done by using the fact that the minimizer γ is obtained by projecting Y orthogonally onto the columnspace of X. Therefore we have $X^T(X\gamma - Y) = 0$ (cf. [7]). Furthermore by Lemma 5 it holds that $\|X\nu - Y\|_2^2 \leq (1 + \varepsilon)\|X\gamma - Y\|_2^2$. Now by plugging this into the Pythagorean theorem and rearranging we get that

$$\begin{aligned}
\|X(\gamma - \nu)\|_2^2 &= \|X\nu - Y\|_2^2 - \|X\gamma - Y\|_2^2 \\
&\leq \varepsilon \|X\gamma - Y\|_2^2
\end{aligned}$$

Putting all together this yields the proposition:

$$\begin{aligned}
\|\gamma - \nu\|_2^2 &\leq \frac{1}{\sigma_{\min}^2(X)} \|X(\gamma - \nu)\|_2^2 \\
&\leq \frac{\varepsilon}{\sigma_{\min}^2(X)} \|X\gamma - Y\|_2
\end{aligned}$$

□

Now using Observation 1 we may assume w.l.o.g. that $\gamma = \nu = 0$ holds. We still have to bound $\inf \mathbb{E}\left[\|Z_1^m - Z_2^m\|_2^2\right]$, the least expected squared Euclidean distance of two points drawn from a joint distribution whose marginals are the original distribution and its embedded counterpart. Of course we can bound this quantity by explicitly defining a properly chosen joint distribution and bounding the expected squared distance for its particular choice.

Lemma 7. *Let $p \propto \mathcal{L}(\beta|X, Y)$ and $p' \propto \mathcal{L}(\beta|\Pi X, \Pi Y)$. Let Z_1^m, Z_2^m be the mean-centered versions of the random variables $Z_1 \sim p$ and $Z_2 \sim p'$ that are distributed according to p and p' respectively. Then we have*

$$\inf \mathbb{E}\left[\|Z_1^m - Z_2^m\|_2^2\right] \leq \varepsilon^2 \operatorname{tr}\left((X^T X)^{-1}\right).$$

Proof. Our plan is to design a joint distribution that deterministically maps points from one distribution to another in such a way that we can bound the distance of every pair of points. This can be done by utilizing the Dirac delta function $\delta(\cdot)$, which is a degenerate probability density function that concentrates all probability mass at zero and has zero density otherwise. Given a bijection $f : \mathbb{R}^d \to \mathbb{R}^d$ we can define such a joint distribution $\lambda \in \Lambda(p, p')$ through its conditional distributions $\lambda(x \,|\, y) = \delta(x - f(y))$ for every $y \in \mathbb{R}^d$. It therefore remains to define f.

Using the embedding Π and applying (4), the columnspace of a matrix is expanded or contracted respectively by a factor of at most $1 \pm \varepsilon$. Let $A = U\Sigma V^T$ and $\Pi A = \tilde{U}\tilde{\Sigma}\tilde{V}^T$ denote the SVDs of A and ΠA respectively. Now consider vectors $x, x', y, y' \in \mathbb{R}^d$ where x' and y' are contained in the columspaces of V and $\tilde{V}$ respectively. Additionally assume the following properties:

1. $\exists c \geq 0 : \|x'\|_2 = \|y'\|_2 = c$
2. $x = \Sigma V^T x'$
3. $y = \tilde{\Sigma}\tilde{V}^T y'$
4. $\exists \tau > 0 : x = \tau y$

Observe that by the first property x' and y' lie on a d-dimensional sphere with radius c centered at 0. Therefore, there exists a rotation matrix $R \in \mathbb{R}^{d \times d}$ such that $y' = Rx'$.

The second item defines a map of such spheres to ellipsoids (also centered at 0) given by the bijection ΣV^T. The third property is defined analogously.

The fourth property urges that x and y both lie on a ray starting from 0. Note that any such ray intersects each ellipsoid exactly once.

Our bijection can be defined accordingly as

$$\begin{aligned} f : \mathbb{R}^d &\to \mathbb{R}^d \\ x &\mapsto \tilde{\Sigma}\tilde{V}^T R V \Sigma^{-1} x \end{aligned}$$

by composing the map ΣV^T, defined in the second item, with the rotation R and finally with $\tilde{\Sigma}\tilde{V}^T$ from the third property. The map is bijective since it is obtained as the composition of bijections.

Now in order to bound the distance $\|Z_1^m - Z_2^m\|_2^2$ for any realization of (Z_1^m, Z_2^m) according to their joint distribution defined above, we can derive a bound on the parameter τ.

Substituting the first two properties into the third one, we get that

$$\Sigma V^T x' = \tau \tilde{\Sigma}\tilde{V}^T y'$$

which can be rearranged to

$$\begin{aligned} y'^T y'\tau &= (y'^T \tilde{V})\tilde{\Sigma}^{-1}\Sigma(V^T x') \\ &= \sum (y'^T \tilde{V})_i (V^T x')_i \frac{\sigma_i}{\tilde{\sigma}_i} \\ &\leq \sum (y'^T \tilde{V})_i (V^T x')_i \frac{\sigma_i}{\sigma_i\sqrt{1-\varepsilon}} \\ &\leq (1+\varepsilon)\sum (y'^T \tilde{V})_i (V^T x')_i \\ &= (1+\varepsilon)\, c^2. \end{aligned}$$

The first inequality follows from $\tilde{\sigma}_i \geq \sqrt{1-\varepsilon}\ \sigma_i$ and the second from the assumption $\varepsilon \leq 1/2$. This eventually means that $\tau \leq (1+\varepsilon)$ since $y'^T y' = c^2$ by the first property.

A lower bound of $\tau \geq (1-\varepsilon)$ can be derived analogously by using $\tilde{\sigma}_i \leq \sqrt{1+\varepsilon}\ \sigma_i$.

Now we can conclude our proof. It follows that

$$\begin{aligned} \inf \mathbb{E}\left[\|Z_1^m - Z_2^m\|_2^2\right] &\leq \mathbb{E}_\lambda\left[\|Z_1^m - Z_2^m\|_2^2\right] \\ &= \mathbb{E}_\lambda\left[\|\varepsilon Z_1^m\|_2^2\right] \\ &= \varepsilon^2\, \mathbb{E}_\lambda\left[\|Z_1^m\|_2^2\right] \\ &= \varepsilon^2\ \mathrm{tr}\left((X^T X)^{-1}\right) \end{aligned}$$

The last equality holds since the expected squared norm of the centralized random variable is just the trace of its covariance matrix. □

Combining the above results we get the following lemma.

Lemma 8. *Let Π be an $(\varepsilon/3)$-subspace embedding for the columnspace of X. Let $p \propto \mathcal{L}(\beta|X,Y)$ and $p' \propto \mathcal{L}(\beta|\Pi X, \Pi Y)$. Then*

$$\mathcal{W}_2^2(p,p') \leq \frac{\varepsilon}{\sigma_{\min}^{-2}}\|X\mu - Y\|_2^2 + \varepsilon^2\,\mathrm{tr}\left((X^T X)^{-1}\right)$$

Proof. The lemma follows from Observation 1, Lemma 6 and Lemma 7. □

Under mild assumptions we can argue that this leads to a $(1+O(\varepsilon))$-approximation of the likelihood with respect to the Wasserstein weight (see Definition 4).

Corollary 9. *Let Π be an $(\varepsilon/3)$-subspace embedding for the columnspace of X. Let $p \propto \mathcal{L}(\beta|X,Y)$ and $p' \propto \mathcal{L}(\beta|\Pi X, \Pi Y)$. Let $\kappa(X) = \sigma_{\max}(X)/\sigma_{\min}(X)$ be the condition number of X. Assume that for some $\rho \in (0,1]$ we have $\|X\mu\|_2 \geq \rho\|Y\|_2$. Then*

$$\mathcal{W}_2(p') \leq \left(1 + \frac{\kappa(X)}{\rho}\sqrt{\varepsilon}\right)\ \mathcal{W}_2(p).$$

Proof. By definition, the squared Wasserstein weight of order 2 of p equals its second moment. Since p is a Gaussian measure with mean μ and covariance matrix $(X^T X)^{-1}$, we thus have

$$\mathcal{W}_2^2(p) = \|\mu\|_2^2 + \mathrm{tr}\left((X^T X)^{-1}\right).$$

Similarly we have that

$$\mathcal{W}_2^2(p') = \|\nu\|_2^2 + \operatorname{tr}\left((X^T\Pi^T\Pi X)^{-1}\right).$$

Since Π is an ε-subspace embedding for the columnspace of X we know from its definition (3), that all the squared singular values of X are approximated up to $(1 \pm \varepsilon)$ error and so are their inverses. Therefore we have that

$$\operatorname{tr}\left((X^T\Pi^T\Pi X)^{-1}\right) \leq (1+\varepsilon)\operatorname{tr}\left((X^TX)^{-1}\right).$$

It remains to bound $\|\nu\|_2^2$. To this end we use the assumption that for some $\rho \in (0,1]$ we have $\|X\mu\|_2 \geq \rho\|Y\|_2$. By the Pythagorean Theorem this means that

$$\begin{aligned}
\|X\mu - Y\|_2^2 &= \|Y\|_2^2 - \|X\mu\|_2^2 \\
&\leq \|X\mu\|_2^2\left(\frac{1}{\rho^2} - 1\right) \\
&\leq \frac{\|X\mu\|_2^2}{\rho^2}
\end{aligned}$$

Now we can apply the triangle inequality and Lemma 6 to get

$$\begin{aligned}
\|\nu\|_2 &\leq \|\mu\|_2 + \|\nu - \mu\|_2 \\
&\leq \|\mu\|_2 + \frac{\sqrt{\varepsilon}}{\sigma_{\min}(X)}\|X\mu - Y\|_2 \\
&\leq \|\mu\|_2 + \frac{\sqrt{\varepsilon}}{\rho\sigma_{\min}(X)}\|X\mu\|_2 \\
&\leq \|\mu\|_2 + \frac{\sqrt{\varepsilon}}{\rho\sigma_{\min}(X)}\|X\|_2\|\mu\|_2 \\
&= \|\mu\|_2 + \frac{\sqrt{\varepsilon}}{\rho}\kappa(X)\|\mu\|_2 \\
&= \left(1 + \frac{\kappa(X)}{\rho}\sqrt{\varepsilon}\right)\|\mu\|_2
\end{aligned}$$

Note that $\frac{\kappa(X)}{\rho} \geq 1$ and $\varepsilon \leq \sqrt{\varepsilon}$. The claim follows since this implies $(1+\varepsilon) \leq (1 + \frac{\kappa(X)}{\rho}\sqrt{\varepsilon})^2$ □

We stress that the assumption that there exists some constant $\rho \in (0,1]$ such that $\|X\mu\|_2 \geq \rho\|Y\|_2$ is very natural and mild in the setting of linear regression since it means that at least a constant fraction of the dependent variable Y can be explained within the columnspace of the data X (cf. [11]).

4.2 Bayesian Regression

So far we have shown that using subspace embeddings to compress a given dataset for regression yields a good approximation to the likelihood. Note that in a Bayesian regression setting Lemma 8 already implies a similar approximation error for the posterior

distribution if the prior is chosen to be an improper, non-informative uniform distribution over $\mathbb{R}^d$. For regression models and especially for regression models on data sets with large n, this covers a considerable amount of the cases of interest, confer [13]. We will extend this to arbitrary Gaussian priors leading to our main result: an approximation guarantee for Gaussian Bayesian regression in its most general form.

To this end, note that since the posterior distribution is given by

$$p_{\text{post}}(\beta|X,Y) \propto \mathcal{L}(X,Y|\beta) \cdot p_{\text{pre}}(\beta)$$

we know that up to some constants, the logarithm of the posterior can be described by

$$\|X\beta - Y\|_2^2 + \|S(\beta - m)\|_2^2 \tag{5}$$

where m is the mean of the prior distribution and S is derived from its covariance matrix by $\Sigma = (S^T S)^{-1}$. Now let

$$Z = \begin{bmatrix} X \\ S \end{bmatrix} \quad \text{and} \quad z = \begin{bmatrix} Y \\ Sm \end{bmatrix}.$$

With these definitions we can rewrite equation (5) above to get $\|Z\beta - z\|_2^2$. This, in turn, can be treated as a frequentist regression problem in the same way as we did in the proof of Lemma 8. For this we only have to use a subspace embedding for the columnspace of $[Z, z]$ instead of only embedding $[X, Y]$. We will see that this is not necessary. More precisely, embedding only the data matrix is sufficient to have a subspace embedding for the entire columnspace defined by the data and the prior information and therefore to have a proper approximation of the posterior distribution. This can be formalized in the following lemma.

Lemma 10. *Let $M = [M_1, M_2]^T \in \mathbb{R}^{(n_1+n_2)\times d}$ be an arbitrary matrix. Suppose Π is an ε-subspace embedding for the columnspace of M_1. Let $I_{n_2} \in \mathbb{R}^{(n_2 \times n_2)}$ be the identity matrix. Then*

$$P = \begin{bmatrix} \Pi & 0 \\ 0 & I_{n_2} \end{bmatrix} \in \mathbb{R}^{(k+n_2)\times(n_1+n_2)}$$

is an ε-subspace embedding for the columnspace of M.

Proof. Fix an arbitrary $x \in \mathbb{R}^d$. We have

$$\begin{aligned}
& |\|PMx\|_2^2 - \|Mx\|_2^2| \\
= \; & |\|\Pi M_1 x\|_2^2 + \|M_2 x\|_2^2 - \|M_1 x\|_2^2 - \|M_2 x\|_2^2| \\
= \; & |\|\Pi M_1 x\|_2^2 - \|M_1 x\|_2^2| \\
\leq \; & \varepsilon \|M_1 x\|_2^2 \\
\leq \; & \varepsilon \|M_1 x\|_2^2 + \|M_2 x\|_2^2 \\
= \; & \varepsilon \|Mx\|_2^2
\end{aligned}$$

which concludes the proof by linearity. □

This lemma finally enables us to prove our main theoretical result.

Theorem 11. *Let Π be an $(\varepsilon/3)$-subspace embedding for the columnspace of X. Let $p_{\text{pre}}(\beta)$ be an arbitrary normal distribution with expected value m and covariance matrix $\Sigma = (S^T S)^{-1}$. Let*

$$Z = \begin{bmatrix} X \\ S \end{bmatrix} \quad \text{and} \quad z = \begin{bmatrix} Y \\ Sm \end{bmatrix}.$$

Let $\mu = \operatorname{argmin}_{\beta \in \mathbb{R}^d} \|Z\beta - z\|_2$ be the posterior expected value. Let $p \propto \mathcal{L}(\beta|X, Y) \cdot p_{\text{pre}}(\beta)$ and $p' \propto \mathcal{L}(\beta|\Pi X, \Pi Y) \cdot p_{\text{pre}}(\beta)$. Then

$$\begin{aligned} & \mathcal{W}_2^2(p, p') \\ \leq \; & \frac{\varepsilon}{\sigma_{\min}^{-2}(Z)} \|Z\mu - z\|_2^2 + \varepsilon^2 \operatorname{tr}\left((Z^T Z)^{-1}\right). \end{aligned}$$

Proof. From our previous reasoning we know that approximating the posterior distribution can be reduced to approximating a likelihood function that is defined in terms of the data as well as the parameters of the prior distribution. This has been shown by rewriting Equation (5) above as $\|Z\beta - z\|_2^2$. Therefore we can apply Lemma 8 to get the desired result given an $(\epsilon/3)$-subspace embedding for the columnspace of Z. Using Lemma 10 we know that it is sufficient to use an $(\epsilon/3)$-subspace embedding for the columnspace of $[X, Y]$ independent of the covariance and mean that define the prior distribution. □

Similar to Corollary 12 we have the following result concerning the posterior distribution.

Corollary 12. *Let Π be an $(\varepsilon/3)$-subspace embedding for the columnspace of X. Let $p_{\text{pre}}(\beta)$ be an arbitrary normal distribution with expected value m and covariance matrix $\Sigma = (S^T S)^{-1}$. Let*

$$Z = \begin{bmatrix} X \\ S \end{bmatrix} \quad \text{and} \quad z = \begin{bmatrix} Y \\ Sm \end{bmatrix}.$$

Let $\mu = \operatorname{argmin}_{\beta \in \mathbb{R}^d} \|Z\beta - z\|_2$ be the posterior expected value. Let $p \propto \mathcal{L}(\beta|X, Y) \cdot p_{\text{pre}}(\beta)$ and $p' \propto \mathcal{L}(\beta|\Pi X, \Pi Y) \cdot p_{\text{pre}}(\beta)$. Let $\kappa(Z)$ be the condition number of Z. Assume that for some $\rho \in (0, 1]$ we have $\|Z\mu\|_2 \geq \rho \|z\|_2$. Then we have

$$\mathcal{W}_2(p') \leq \left(1 + \frac{\kappa(Z)}{\rho} \sqrt{\varepsilon}\right) \mathcal{W}_2(p).$$

5 Simulation Study

We use simulated data to validate the proposed method empirically. The data sets consist of $n = 50\,000$ observations and $d = 50$ variables. This is by no means a very large data set, however, we need to be able to analyze both the sketched versions of the data set and the original data set. Four different data sets have been analyzed. The main difference between these is the simulated error variance σ_ε. σ_ε takes values of 1, 2, 5, and 10 respectively. This is done to check whether the differences in the ℓ_2-norm grow with growing variance, as is expected according to the theory.

In the simulated data sets, some or all of the variables may have an influence on the dependent variable Y. Additionally, an intercept is modeled, giving a total of 51 variables.

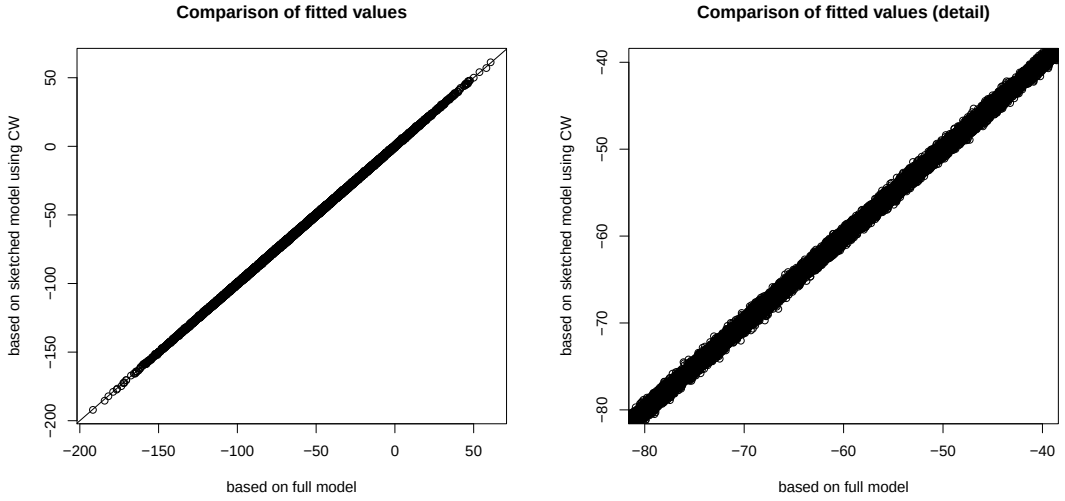

Figure 1: Comparison of fitted values of the regression model based on the original data set (full model) and on a sketched data set obtained using the CW approach

The MCMC sampling was done using software programs R [23] and R-package rstan [27]. The sketches are calculated using our R package. The number of observations of the three sketched data sets is $k_{BCH} = 20\,546$, $k_{SRHT} = 20\,547$, and $k_{CW} = 16\,384$, respectively. Please note that the sizes of the BCH sketch and the SRHT sketch are almost the same, while the CW sketch is smaller. This might seem contrary to chapter 3, where we state that the target dimension of the CW sketch is higher compared to the others. However, the reason for this is the relatively small number of variables d. For higher values of d, the CW sketch will result in a higher k compared to the other sketching approaches.

Following the theory, the posterior distributions, represented by the MCMC samples, should be very close to one another. To evaluate this empirically, we compare the MCMC samples resulting from the full data set (referred to as "full model") with each of the MCMC samples resulting from a sketched model. Figure 1 shows the fitted values of the full model on the x-axis, using the mean of the full MCMC sampler for each of the variables. On the y-axis are the predicted values for the full data set we get when using the mean of the sketched MCMC sampler based on sketching approach CW for each variable. All pairs of fitted values are on or close to the bisecting line, indicating that the sketched model gives a very similar result to the full model.

Figure 2 shows the difference between the fitted values based on the sketched models and the fitted values based on the original model for one data set. For all three sketching approaches, the median of the differences is 0, the largest values show an absolute deviation of around 2. The boxes only cover a small area around 0, indicating that all three sketching approaches predict the dependent variable very accurately (compared to the original model) for a majority of the observations. The differences show a similar structure for all three methods. The box and spread of the whiskers is slightly larger for SRHT, but generally speaking, the location and variation of all boxes are almost the same.

We compare the resulting posterior distributions of all parameters using boxplots. Two

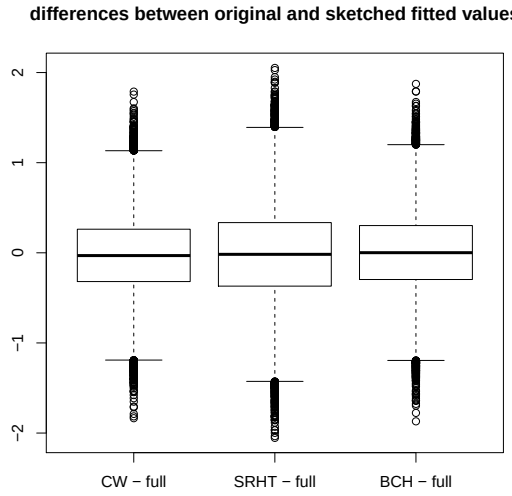

Figure 2: Differences between fitted values based on the full model and fitted values based on three sketched models obtained using the approaches CW, SRHT, and BCH

examples are shown in Figure 3. In this boxplot, we can see that our method adds some variation to the MCMC sample, the boxes and the spread of the samples are larger. The extra amount of variation for the sketches based on SRHT and BCH is very similar, while the sketched based on CW shows slightly more extra variation. Given that this sketching approach results in the smallest sketched data set for our simulated data sets (not in the general case), this indicates a trade-off between the amount of reduction and the additional variation in the parameters.

The medians may differ from the full model, however, we did not find evidence of systematic bias for any of the sketching approaches. The position of the boxes may sometimes show no overlap (as for the full model versus the model based on an SRHT sketch in Figure 3), but the general location remains the same. When looking at the 95% credible intervals, 0 either lies in the interval for all four models or it does not lie in the interval for all of them, which means that one would not consider a variable "important" for the explanation of the dependent variable in one model, but consider it "unimportant" in another model.

Table 1 shows the run time for the Bayesian analysis of the data set. For the original data set, this only includes the run-time of the MCMC algorithm. For the sketched data sets, the value gives the total run-time, including both our sketching algorithm and the MCMC algorithm. The speed-up gained by applying our sketching algorithm first is evident. Note that the run-times for the respective sketching algorithms are given in seconds, while the run-time for the complete analysis is given in hours. This underlines that the sketching requires only a very small part of the total run-time.

If we add more observations to our data set, we expect the run-time of the MCMC analysis to take longer. However, the size of the sketched data set is independent of the number of observations in the original data set. For that reason, the MCMC analysis on the sketched data sets would take around the same time. It would take longer to

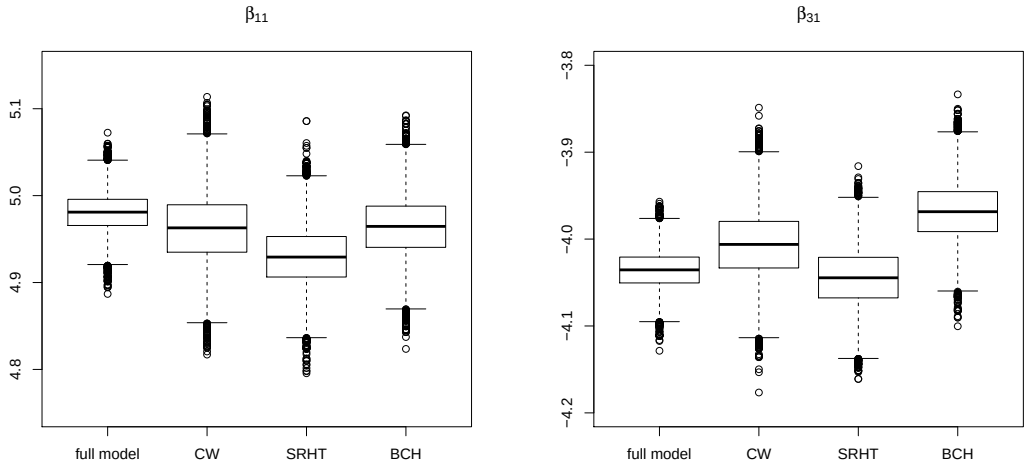

Figure 3: Boxplots of the distribution of MCMC samples for two selected parameters based on the model on the original data set (full model) and sketched data sets using the sketching approaches CW, SRHT, and BCH

sketching approach	run-time for analysis (hours)	sketching time (seconds)
full	338.872	-
CW	93.353	0.573
SRHT	111.010	1.338
BCH	114.576	97.475

Table 1: Comparison of run-times

calculate the sketches, but the effect would be barely noticeable. The comparative run-time advantage of our method grows with the number of observations, while the difference between the posterior distributions stays the same.

Note that our method assumes that the model is correctly specified, i.e. that all independent variables have a linear influence on the dependent variable or no influence at all. This is the case for our simulated data sets, but it will not be true in general. For this reason, the regression model needs to be built carefully, doing some pre-checks before calculating the sketch. If a wrong model is used, e.g. by treating a variable with logarithmic influence as a linear variable, the results will still be close to the non-sketched linear model. However, diagnostic plots such as residual plots may show different features, which may make identifying problems in the model harder.

6 Conclusion

Our paper deals with random projections as a data reduction technique for Bayesian regression. In a series of theoretical results, we have shown that we can apply random projections to achieve a so-called oblivious subspace embedding for the columnspace of

the given data matrix if the likelihood is modeled using standard linear regression with a Gaussian error term. The size of the sketched and reduced dataset is of the number n of observations in the original data set. Therefore, subsequent computations can operate within time and space bounds that are also independent of n regardless of the algorithm that is actually used. We show that the likelihood function is approximated within small error. Furthermore, if an improper, non-informative uniform distribution over $\mathbb{R}$ or an arbitrary Gaussian distribution is used as prior distribution, the desired posterior distribution is also well approximated within small error. We also show our results to be $(1 + O(\varepsilon))$ approximations to the distributions of interest in the context of Bayesian linear regression.

In our simulation experiments, we found that the approximation works much better than the theoretical bounds guarantee. Important values like the mean, median, and quantiles are recovered

For future research, we would like to generalize our results to other classes of distributions for the likelihood and to more general priors. The recent results on frequentist ℓ_p regression might give rise to efficient streaming algorithms also in the Bayesian regression setting.

References

[1] N. Ailon and E. Liberty. Fast dimension reduction using Rademacher series on dual bch codes. *Discrete & Computational Geometry*, 42(4):615–630, 2009.

[2] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *J. Comput. Syst. Sci.*, 58(1):137–147, 1999.

[3] J. D. Batson, D. A. Spielman, and N. Srivastava. Twice-Ramanujan sparsifiers. *SIAM J. Comput.*, 41(6):1704–1721, 2012.

[4] C. Boutsidis, P. Drineas, and M. Magdon-Ismail. Near-optimal coresets for least-squares regression. *IEEE Transactions on Information Theory*, 59(10):6880–6892, 2013.

[5] C. Boutsidis and A. Gittens. Improved matrix algorithms via the subsampled randomized Hadamard transform. *SIAM J. Matrix Analysis Applications*, 34(3):1301–1340, 2013.

[6] K. L. Clarkson, P. Drineas, M. Magdon-Ismail, M. W. Mahoney, X. Meng, and D. P. Woodruff. The fast Cauchy transform: with applications to basis construction, regression, and subspace approximation in l1. *CoRR*, abs/1207.4684, 2012.

[7] K. L. Clarkson and D. P. Woodruff. Numerical linear algebra in the streaming model. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing, STOC*, pages 205–214, 2009.

[8] K. L. Clarkson and D. P. Woodruff. Low rank approximation and regression in input sparsity time. In *Symposium on Theory of Computing Conference, STOC'13*, pages 81–90, 2013.

[9] K. Csillery, M. Blum, O. Gaggiotti, and O. Francois. Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology and Evolution*, 25(7):410–418, 2010.

[10] M. Dietzfelbinger, T. Hagerup, J. Katajainen, and M. Penttonen. A reliable randomized algorithm for the closest-pair problem. *J. Algorithms*, 25(1):19–51, 1997.

[11] P. Drineas, M. W. Mahoney, and S. Muthukrishnan. Sampling algorithms for ℓ_2 regression and applications. In *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2006*, pages 1127–1136, 2006.

[12] P. Drineas, M. W. Mahoney, S. Muthukrishnan, and T. Sarlós. Faster least squares approximation. *Numer. Math.*, 117(2):219–249, Feb. 2011.

[13] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Texts in Statistical Science. Chapman & Hall/CRC, 2 edition, 2004.

[14] C. R. Givens and R. M. Shortt. A class of wasserstein metrics for probability distributions. *The Michigan Mathematical Journal*, 31(2):231–240, 1984.

[15] N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.

[16] I. Jolliffe. *Principal component analysis*. Springer, 2 edition, 2002.

[17] R. Kannan and S. Vempala. Spectral algorithms. *Foundations and Trends in Theoretical Computer Science*, 4(3-4):157–288, 2009.

[18] R. Kannan, S. Vempala, and D. P. Woodruff. Principal component analysis and higher correlations for distributed data. In *COLT*, pages 1040–1057, 2014.

[19] T. Martins, D. Simpson, F. Lindgren, and H. Rue. Bayesian computing with INLA: New features. *Computational Statistics and Data Analysis*, 67:68–83, 2013.

[20] S. Muthukrishnan. Data streams: Algorithms and applications. *Foundations and Trends in Theoretical Computer Science*, 1(2), 2005.

[21] J. Nelson and H. L. Nguyen. Lower bounds for oblivious subspace embeddings. *CoRR*, abs/1308.3280, 2013.

[22] J. Nelson and H. L. Nguyen. Osnap: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *54th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2013*, pages 117–126, 2013.

[23] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[24] H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 71:319–392, 2009.

[25] F. Rusu and A. Dobra. Pseudo-random number generation for sketch-based estimations. *ACM Trans. Database Syst.*, 32(2):11, 2007.

[26] T. Sarlós. Improved approximation algorithms for large matrices via random projections. In *FOCS*, pages 143–152. IEEE Computer Society, 2006.

[27] Stan Development Team. *Stan: A C++ Library for Probability and Sampling, Version 1.3.*, 2013.

[28] C. Villani. *Optimal transport: Old and new.* Grundlehren der mathematischen Wissenschaften. Springer, Berlin, 2009.

[29] D. P. Woodruff and Q. Zhang. Subspace embeddings and ℓ_p-regression using exponential random variables. In *COLT*, pages 546–567, 2013.