Compressed Least Squares Regression revisited

Martin Slawski
Department of Statistics
George Mason University
Fairfax, VA 22030, USA
mslawsk3@gmu.edu

Abstract

We revisit compressed least squares (CLS) regression as originally analyzed in Maillard and Munos (2009) and later on in Kaban (2014) with some refinements. Given a set of high-dimensional inputs, CLS applies a random projection and then performs least squares regression based on the projected inputs of lower dimension. This approach can be beneficial with regard to both computation (yielding a smaller least squares problem) and statistical performance (reducing the estimation error). We will argue below that the outcome of previous analysis of the procedure is not meaningful in typical situations, yielding a bound on the prediction error that is inferior to ordinary least squares while requiring the dimension of the projected data to be of the same order as the original dimension. As a fix, we subsequently present a modified analysis with meaningful implications that much better reflects empirical results with simulated and real data.

1 Introduction

We consider a common setup of regression given data (y_i, x_i), with y_i taking values in $\mathbb{R}$ and x_i taking values in $\mathbb{R}^d$, $i \in [n]$, where for a positive integer m, we write $[m] := \{1, \ldots, m\}$. The inputs x_i are considered as fixed, and $y_i = f_i + \xi_i$, with $f_i = \mathbf{E}[y_i|x_i]$ and ξ_i following a distribution with mean zero and variance σ^2, $i \in [n]$. Moreover, the $\{\xi_i\}_{i=1}^n$ are assumed to be uncorrelated. More concisely, we write $y = f + \xi$, where $y = (y_i)_{i=1}^n$, etc.

The most basic approach to regression modelling is that of linear regression, in which one approximates $y \approx Xw$ for a suitable vector of coefficients $w \in \mathbb{R}^d$, and X is the real n-by-d matrix having the $\{x_i\}_{i=1}^n$ as its rows. The optimal linear predictor Xw^* of y given X with respect to squared loss is defined by the optimization problem

$$\min_{w \in \mathbb{R}^d} \mathbf{E}\left[\|y - Xw\|_2^2/n\right],$$

where the expectation is with respect to the noise ξ. Note that any minimizer w^* of the above problem satisfies $Xw^* = P_X f$, where for a matrix A whose column space is a subspace of $\mathbb{R}^n$, we write P_A for the projection operator onto that subspace; if there are multiple such w^* we choose the one with the smallest ℓ_2-norm. Accordingly, we define the excess risk of an estimator $\widehat{\theta} = \widehat{\theta}(X, y)$ of w^* by

$$\mathcal{E}(\widehat{\theta}) = \mathbf{E}[\|Xw^* - X\widehat{\theta}\|_2^2/n],$$

where the expectation is now taken with respect to $\widehat{\theta}$. If the linear model holds exactly (i.e., $P_X f = f$), $\mathcal{E}(\widehat{\theta})$ equals the in-sample mean squared prediction error that measures how well the $\{x_i^\top \widehat{\theta}\}_{i=1}^n$ predict the 'denoised' observations $\{x_i^\top w^*\}_{i=1}^n$ on average.

An ordinary least squares (OLS) estimator $\widehat{w}$ satisfies $X\widehat{w} = P_X y$. It is not hard to show that

$$\mathcal{E}(\widehat{w}) = \sigma^2 \mathrm{rank}(X)/n. \tag{1}$$

In this paper, we are interested in a high-dimensional setup in which $\mathrm{rank}(X)$ is of the same order of magnitude as n. To keep matters simple, we assume that X has full rank $d \wedge n$ unless otherwise stated. In a high-dimensional setup, OLS does not yield satisfactory statistical performance. In particular, $\mathcal{E}(\widehat{w})$ does not tend to zero as n grows which can be regarded as a basic consistency requirement. Moreover, if both n and d are large, obtaining $\widehat{w}$ or making predictions based on $\widehat{w}$ becomes computationally costly.

In light of these issues, it makes sense to consider alternatives that aim at leveraging some sort of low-dimensional structure. Scenarios in which w^* exhibits

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54.

one of various forms of sparsity are dominating in the literature, see the monograph of Hastie et al. (2015) for an overview. In the present paper, we follow another direction in which the inputs $\{x_i\}_{i=1}^n$ are linearly mapped into a lower-dimensional space, and linear least squares regression is then performed based on the subspace obtained in this way. Put differently, one considers a new design matrix $X_R = XR$ with R being a d-by-k matrix, $k \ll p$. On the statistical side, while yielding an increase of the approximation error (bias), one achieves a substantial reduction in estimation error. On the computational side, solving the OLS problem in the reduced space is less expensive. However, the success of this approach eventually depends on whether it is possible to find a suitable matrix R that keeps the bias in check for low k.

The traditional choice of X_R is based on the r-truncated SVD of X which yields $X_R = U_r\Sigma_r$, where U_r is the matrix of the top $k = r$ left singular vectors, Σ_r is a diagonal matrix containing the corresponding singular values, and $R = V_r$, where $V_r \in \mathbb{R}^{d\times r}$ contains the top r right singular vectors of X. When combined with subsequent least squares regression, the procedure is well-known under the name principal components regression (PCR) in the statistics literature (Kendall, 1957; Artemiou and Li, 2009).

A second approach that is in the focus of present paper is to choose R as a Johnson-Lindenstrauss transform (Johnson and Lindenstrauss, 1984; DasGupta, 2003; Vempala, 2005; Ailon and Chazelle, 2006; Matousek, 2008) which is realized by drawing the entries of R from a suitable distribution. Regression based on the thus obtained design matrix X_R is analyzed in Maillard and Munos (2009) who coined the term 'compressed least squares' (CLS) regression, and in a later paper by Kaban (2014). Another random mechanism for generating R based on b-bit minwise hashing (Li and König, 2011) that is suitable for sparse X is proposed in Shah and Meinshausen (2016). Selecting a subset of columns of X, either in a random or a systematic fashion, constitutes one more approach that falls into the same category.

In the present paper, we revisit the analysis of CLS in Maillard and Munos (2009) and Kaban (2014). At this point, we point out that the notion of 'compressed regression' in those two as well as in the present paper needs to be distinguished from that in Zhou et al. (2009) in which the reduced design matrix is given by RX, i.e. R is multiplied from the left, and R is applied to the response y as well. This setup which is typically referred to as 'sketched regression' is studied in a substantial body of literature with major contributions from researchers in numerical linear algebra, theoretical computer science, and machine learning (Sarlos, 2006; Pilanci and Wainwright, 2015; Raskutti and Mahoney, 2015; Drineas and Mahoney, 2016). Sketched regression predominantly concerns the case of large n, but can also be motivated from privacy considerations (Zhou et al., 2009). By contrast, CLS is motivated by large d. Furthermore, sketched regression retains regression coefficients at the level of the original inputs, which can be beneficial for estimation and interpretation, and may be one of the reasons why CLS has received comparatively less attention.

The present paper is motivated by the observation that the existing analysis of CLS in Maillard and Munos (2009) and Kaban (2014) yields bounds on the excess risk that are too crude to be meaningful. Specifically, we show that according to these bounds, the statistical performance of CLS would be inferior to OLS except for specific low-signal situations as detailed below, and that the optimal choice of the reduced dimension k would have to be of the same order as $d \wedge n$. It turns out that the reason for this outcome lies in the analysis of the bias term. With a more careful estimation of the bias in dependence of the rate of decay of the singular values of X, we obtain a significantly improved (albeit not entirely sharp) bound according to which CLS can achieve a considerably lower excess risk than OLS even with small k. Our analysis comes with a side-by-side comparison of CLS to principal components regression (PCR). Subsequently, we outline a computationally favorable aproach for obtaining guidance regarding the choice of k in practice, before providing empirical results that illustrate central aspects of this work. We conclude with a brief discussion.

2 Existing excess risk bounds for CLS

For fixed $R \in \mathbb{R}^{p\times k}$, consider the excess risk of CLS

$$\mathcal{E}(R) := \mathcal{E}(R\widehat{w}_R) = \mathbf{E}\left[\|Xw^* - X_R\widehat{w}_R\|_2^2/n\right],$$

where $\widehat{w}_R$ is a least squares solution based on the reduced design matrix X_R, i.e., $\widehat{w}_R$ satisfies $X_R\widehat{w}_R = P_{X_R}y$, and the expectation is with respect to $\widehat{w}_R$. Straightforward calculations show that $\mathcal{E}(R)$ can be decomposed into a bias and a variance term:

$$\mathcal{E}(R) = \underbrace{\|(I - P_{X_R})Xw^*\|_2^2/n}_{\text{Bias}} + \underbrace{\sigma^2\text{rank}(X_R)/n}_{\text{Variance}}. \quad (2)$$

We commonly have $\text{rank}(X_R) = k$. The choice of k determines the trade-off between bias and variance. If k can be chosen significantly smaller than $d \wedge n$ while at the same time the magnitude of the bias can be controlled, a substantially better excess risk than that of OLS in (1) is obtained.

Kaban (2014) considers random R whose entries are drawn i.i.d. from a zero-mean symmetric distribution

with variance $1/k$ and finite fourth moment, and derives a bound on $\mathbf{E}[\mathcal{E}(R)]$, where the expectation is with respect to R. Letting $\Sigma = X^\top X/n$ denote the Gram matrix of the inputs, $Q = \mathrm{tr}(\Sigma)I + \Sigma$, and recalling that w^* contains the regression coefficients of the optimal linear predictor, her final result is of the form

$$\mathbf{E}[\mathcal{E}(R)] \leq \|w^*\|_Q^2/k \ + \sigma^2 k/n, \tag{3}$$

where for $v \in \mathbb{R}^d$, $\|v\|_Q = (v^\top Q v)^{1/2}$. Minimizing this bound with respect to k, the optimal value of k results as $k^* = \sqrt{n}\|w^*\|_Q/\sigma$. Back-substitution into (3) yields

$$\mathbf{E}[\mathcal{E}(R)] \leq 2\sigma\|w^*\|_Q/\sqrt{n}. \tag{4}$$

Maillard and Munos (2009) obtain a rather similar result for random X when R is a Johnson-Lindenstrauss transform, a smaller class of matrices as the one considered in Kaban (2014). The corresponding exces risk is qualitatively bounded as

$$O\left(\sigma\|w^*\|_2 \left(\mathbf{E}[\|x_1\|_2^2]\right)^{1/2} \sqrt{\log(n)/n}\right) \tag{5}$$

for $k = \Theta(\sqrt{n \log n}\|w^*\|_2 \, \mathbf{E}[\|x_1\|^2]^{1/2}/\sigma)$. Comparing the terms in (4) and (5), we note that for random design, $\mathbf{E}[\|x_1\|_2^2] = \mathbf{E}[\mathrm{tr}(\Sigma)]$. Moreover, for $1 \leq c \leq 2$,

$$\|w^*\|_Q^2 = \mathrm{tr}(\Sigma)\|w^*\|_2^2 + (w^*)^\top \Sigma w^* = c \cdot \mathrm{tr}(\Sigma)\|w^*\|_2^2,$$

We conclude that (4) and (5) only differ by an $O(\sqrt{\log n})$ term as does the underlying choice of k.

At first glance, (4) and (5) do no longer depend on d, which indicates that the approach has some merits in a high-dimensional setup even though the rate of decay $n^{-1/2}$ is slower than the usual n^{-1} rate, and the requirement $k = \Omega(n^{1/2})$ may be troubling as one may have hoped for stronger dimension reduction. However, closer inspection reveals that treating $\|w^*\|_Q$ respectively $\mathbf{E}[\|x_1\|_2^2]^{1/2}$ as $O(1)$ terms is not appropriate. In fact, for fixed design it is common to assume that the columns $\{X_j\}_{j=1}^d$ of X are scaled such that $\|X_j\|_2^2 = n$, whereas for standard random designs, e.g. X with i.i.d. rows from a zero-mean Gaussian distribution with unit variances, this scaling holds in expectation. In this case, $\mathrm{tr}(\Sigma)$ respectively $\mathbf{E}[\|x_1\|_2^2]$ evaluate as d which makes the bounds (4) and (5) of rather limited use. The bound (4) becomes

$$\mathbf{E}[\mathcal{E}(R)] \leq 2\sigma\|w^*\|_2\sqrt{d/n} \tag{6}$$

for $k = \sqrt{nd}\|w^*\|_2/\sigma$, which is of the same order (or may even exceed) $d \wedge n$, while the bound on the excess risk becomes inferior to that of OLS except for a low-signal situation as outlined below.

(1) $n \geq d$: the excess risk of OLS becomes $\sigma^2 d/n$, which is superior to (6) whenever $\|w^*\|_2 \geq \sigma\sqrt{d/n}$.

(2) $n < d$: the excess risk of OLS becomes σ^2 which is superior to (6) whenever $\|w^*\|_2 \geq \sigma\sqrt{n/d}$.

In summary, we conclude that the existing analysis of CLS fails to establish that the approach achieves a reasonable statistical performance. This raises the question whether this is intrinsic to the approach, or an artifact of the analysis. From (3), we find that the correct variance term $\sigma^2 k/n$ is present. On the other hand, the estimate of the bias term as $\|w^*\|_Q^2/k$ deserves further investigation, as it can be seen from (2) that for $k = d \wedge n$ the bias is actually zero whenever the entries of R are drawn i.i.d. from an absolutely continuous distribution on the real line as in this case the column space of X_R coincides with the column space of X with probability one.

3 Excess risk of Principal Components Regression (PCR)

Our improved analysis of CLS presented below is motivated from an excess risk bound for PCR that is derived in the sequel. Let $X = U\Sigma V^\top$ be the SVD of X, where $U \in \mathbb{R}^{n \times d\wedge n}$, $U^\top U = I$, is the matrix of left singular vectors, $\Sigma \in \mathbb{R}^{d\wedge n \times d \wedge n}$ is the diagonal matrix whose diagonal contains the decreasingly ordered sequence of singular values $\sigma_1 \geq \ldots \geq \sigma_{d\wedge n}$, and $V \in \mathbb{R}^{d \times d\wedge n}$, $V^\top V = I$, is the matrix of right singular vectors. For $r \in \{1, \ldots, d \wedge n\}$, consider

$$U = [U_r \ U_{r+}], \quad \Sigma = \begin{bmatrix} \Sigma_r & 0 \\ 0 & \Sigma_{r+} \end{bmatrix}, \quad V = [V_r \ V_{r+}], \tag{7}$$

where U_r and $V_r \in \mathbb{R}^{d \times r}$ contain the top r left respectively right singular vectors, and Σ_r contains the corresponding singular values. The remaining singular vectors respectively singular values are contained in U_{r+}, V_{r+} and Σ_{r+}. Letting $R = V_r$, we have

$$X_R = XV_r = (U_r\Sigma_r V_r^\top + U_{r+}\Sigma_{r+}V_{r+}^\top)V_r = U_r\Sigma_r.$$

The corresponding projection operator is given by $P_{X_R} = U_r U_r^\top$ and the bias term in (2) results from

$$(I - P_{X_R})Xw^* = (I - U_r U_r^\top)Xw^* = U_{r+}\Sigma_{r+}V_{r+}^\top w^*.$$

Denote $\alpha^* = V^\top w^* \in \mathbb{R}^{d\wedge n}$, and partition $\alpha^* = [(\alpha_r^*)^\top \ (\alpha_{r+}^*)^\top]$ as in (7). We then have

$$\begin{aligned}
\mathcal{E}(V_r) &= \|U_{r+}\Sigma_{r+}V_{r+}^\top w^*\|_2^2/n + \sigma^2 r/n \\
&= \|\Sigma_{r+}\alpha_{r+}^*\|_2^2/n + \sigma^2 r/n \\
&= \frac{1}{n}\sum_{j=r+1}^{d\wedge n} \sigma_j^2(\alpha_j^*)^2 + \sigma^2\frac{r}{n} \qquad (8) \\
&\leq \|\alpha_{r+}^*\|_\infty^2 \frac{1}{n}\sum_{j=r+1}^{d\wedge n}\sigma_j^2 + \sigma^2\frac{r}{n} \\
&= \|\alpha_{r+}^*\|_\infty^2 \|X - \mathcal{T}_r(X)\|_F^2/n + \sigma^2 r/n, \qquad (9)
\end{aligned}$$

where $\mathcal{T}_r(X) = U_r \Sigma_r V_r^\top$ denotes the best rank r-approximation of X w.r.t. the Frobenius norm $\|\cdot\|_F$.

From (8) respectively (9), we see that the excess risk of PCR behaves favorably if (i) the tail of the squared singular values at truncation level r is small (i.e., X can be well approximated by a matrix of rank r) and (ii) if there are no large coefficients in α^* outside the its top r entries corresponding to the leading singular vectors. Condition (ii) constitutes the main source of criticism of PCR as there is often no good reason to assume that α^* is well-aligned with the leading singular values of X; see Artemiou and Li (2009) and the references therein. In the sequel, we disregard this issue and focus on condition (i), assuming that α^* is dense in the sense that $\max_j |\alpha_j^*| / \min_j |\alpha_j^*| \leq \beta$ for some positive universal constant β, in which case we do not lose much by working with the bound

$$\mathcal{E}(V_r) \leq \|\alpha^*\|_\infty^2 \|\Delta_r\|_F^2/n + \sigma^2 r/n, \quad \Delta_r := X - \mathcal{T}_r(X), \tag{10}$$

in place of (8)/(9). Depending on the rate of decay of the singular values, we can determine the optimal choice of r and the resulting excess risk. For what follows, we assume that X is scaled such that $\|X\|_F^2 = \sum_{j=1}^{d \wedge n} \sigma_j^2 = n \cdot d$ (or equivalently, $\mathrm{tr}(\Sigma) = d$).

Scenario (F): perfectly flat spectrum

For $n \geq d$, this means that the columns of X are orthogonal with $\sigma_j = \sqrt{n}$, $j \in [d]$, and according to (10), $\mathcal{E}(V_r) \leq \|\alpha^*\|_\infty^2 (d - r) + \sigma^2 r/n$. The optimal value of r is given by $r^* = d$ if $\|\alpha^*\|_\infty > \sigma/\sqrt{n}$ and $r^* = 0$ else. In the first case, we recover the OLS solution. For $n < d$, $\sigma_j = \sqrt{d}$, $j \in [n]$, so that $\mathcal{E}(V_r) \leq (d/n) \cdot (n - r)\|\alpha^*\|_\infty^2 + \sigma^2 r/n$ and $r^* = d$ if $\|\alpha^*\|_\infty > \sigma/\sqrt{d}$ and $r^* = 0$ else.

In summary, except for low-signal situations, we do not gain anything compared to OLS, which is expected, as for a flat spectrum dimension reduction by means of a truncated SVD is not effective.

Next, we consider the situation in which the sequence of singular values decays at certain rates. To this end, for $1 \leq s \leq d \wedge n$, let us define $\gamma(s) = \sum_{j=1}^s \sigma_j^2$ and $\tau(s) = \{\gamma(d \wedge n) - \gamma(s)\}/\gamma(d \wedge n)$. We then have

$$\begin{aligned} \|\Delta_r\|_F^2/n &= \{\gamma(d \wedge n) - \gamma(r)\}/n \\ &= \tau(r)\gamma(d \wedge n)/n \\ &= \tau(r) \cdot d, \end{aligned}$$

recalling that $\gamma(d \wedge n) = \sum_{j=1}^{d \wedge n} \sigma_j^2 = \|X\|_F^2 = n \cdot d$.

Scenario (P): polynomial decay

Suppose that $\sigma_j^2 = C \cdot j^{-q}$, $j \in [d]$, for $q \geq 2$ arbitrary and a constant $C > 0$. By comparing series and integrals, one shows that $\gamma(d \wedge n) - \gamma(r) \leq C(r^{-(q-1)} - (d \wedge n)^{-(q-1)})$ and $\gamma(d \wedge n) \geq C$, so that $\tau(r) \leq r^{-(q-1)}$. It follows that

$$\mathcal{E}(V_r) \leq r^{-(q-1)} \cdot d \cdot \|\alpha^*\|_\infty^2 + \sigma^2 r/n.$$

Minimizing the right hand side w.r.t. r, we obtain

$$\begin{aligned} r^* &= \big\{(q-1)\|\alpha^*\|_\infty^2 (n \cdot d)/\sigma^2\big\}^{1/q} \\ \mathcal{E}(V_{r^*}) &\leq 2(q-1)^{1/q} \left(d\|\alpha^*\|_\infty^2\right)^{1/q} \left(\sigma^2/n\right)^{(q-1)/q}. \end{aligned} \tag{11}$$

Let $q = 2$. For a dense unit vector α^*, we have $\|\alpha^*\|_\infty^2 = O(1/d)$ in which case (11) yields $r^* = O(n^{1/2})$ and $\mathcal{E}(V_{r^*}) \leq O(1/\sqrt{n})$. Recall that this is the result claimed by Maillard and Munos (2009) and Kaban (2014) for CLS, *without* making any assumption on the singular values of X, which is one more indication that their claims cannot be valid in general. From (11), we also find that as q grows, r^* essentially becomes $O(1)$ and $\mathcal{E}(V_r)$ becomes proportional to σ^2/n.

Scenario (E): exponential decay

Suppose that $\sigma_j^2 = C_0 \theta^j$ for $\theta \in (0, 1)$. Then, $\tau(r) \leq C_0 \frac{\theta}{1-\theta} \theta^r = C_1 \exp(-cr)$, say. The optimal choice of r^* and the corresponding bound on $\mathcal{E}(V_{r^*})$ result as

$$\begin{aligned} r^* &= \frac{1}{c} \log\left(C_2 \|\alpha^*\|_\infty^2 \, n \, d/\sigma^2\right) \\ \mathcal{E}(V_{r^*}) &\leq \frac{2}{c} \log\left(C_2 \|\alpha^*\|_\infty^2 \, n \, d/\sigma^2\right) \sigma^2/n. \end{aligned} \tag{12}$$

Scenarios (P) and (E) tell us that PCR may improve significantly over OLS in terms of excess risk. The key condition for this to happen is a decaying spectrum of the design matrix X.

4 Improved analysis of CLS

We have seen in Section 2 that the reason for the existing analysis of CLS not to yield a particularly useful result lies in the bias term. In Kaban (2014), the expected bias (w.r.t. R) is bounded as

$$\begin{aligned} \mathbf{E}[\|(I - P_{X_R}) X w^*\|_2^2/n] &= \mathbf{E}\left[\min_{v \in \mathbb{R}^k} \|X w^* - XRv\|_2^2/n\right] \\ &\leq \mathbf{E}[\|X w^* - XRR^\top w^*\|_2^2/n] \end{aligned}$$

It turns out that replacing the minimizing v by $R^\top w^*$ is too crude, and makes it impossible to exploit potential decay in the sequence of singular values, which is the starting point of our analysis.

The main observation is the following: it is known from the literature on randomized numerical linear algebra (see e.g. Sarlos (2006); Mahoney (2011); Halko et al.

(2011)) that if R is a Johnson-Lindenstrauss transform, the subspace spanned by the top r singular vectors of X_R is close to that of the leading r singular vectors of X, provided k is large enough (in fact, k can be chosen proportional to r, modulo logarithmic factor). This suggests that for CLS, one should be able to obtain excess risk bounds not far from those of PCR in the previous section. This is the route taken in the sequel.

Let us reconsider the bias term in (2):

$$\begin{aligned}\|(I - P_{X_R})Xw^*\|_2^2/n &\leq \|(I - P_{X_R})X\|_2^2/n \cdot \|w^*\|_2^2, \\ &\leq \|(I - P_{X_R})X\|_F^2/n \cdot \|w^*\|_2^2,\end{aligned}$$

where $\|M\|_2$ denotes the operator norm of a matrix M. For R with i.i.d. entries from a standard Gaussian distribution, Halko et al. (2011) provide bounds on both $\mathbf{E}[\|(I - P_{X_R})X\|_2^2]$ and $\mathbf{E}[\|(I - P_{X_R})X\|_F^2]$. Regarding the latter, for $r \in [d \wedge n]$ and $k \geq r + 2$, the following result is obtained:

$$\mathbf{E}[\|(I - P_{X_R})X\|_F^2] \leq \left(1 + \frac{r}{k - r - 1}\right)\|\Delta_r\|_F^2. \quad (13)$$

In particular, for $k = 2r + 1$, the approximation error in Frobenius norm is within a factor two of what is attained by the r-truncated SVD. The bound on $\mathbf{E}[\|(I - P_{X_R})X\|_2^2]$ in Halko et al. (2011) does not improve much over the Frobenius norm bound (13) as it still depends on $\|\Delta_r\|_F^2$ (albeit with a better prefactor). Below, we state a result of a similar flavor that holds for a broader class of matrices that satisfy the following two conditions:

(**C1**) Let $r \in [d \wedge n]$ and let $\mathcal{V}_r \subset \mathbb{R}^d$ denote the column space of V_r. For some $\delta \in (0, 1)$, it then holds that $(1-\delta)\|v\|_2^2 \leq \|R^\top v\|_2^2 \leq (1+\delta)\|v\|_2^2$ for all $v \in \mathcal{V}_r$.

This is a restricted isometry-type condition as it appears on the literature on sparse estimation, with the difference that the condition is milder in the sense that approximate norm preservation is required only for a single subspace (as opposed to the union over subspaces of r-sparse vectors).

(**C2**) For $\varepsilon \in (0, 1)$, $R^\top$ is an $\varepsilon/\sqrt{r}$-Johnson-Lindenstrauss transform w.r.t. a fixed set of vectors $\mathcal{S} \subset \mathbb{R}^d$ of cardinality $2n \cdot r$, i.e. it holds that $(1 - \varepsilon/\sqrt{r})\|v\|_2^2 \leq \|R^\top v\|_2^2 \leq (1 + \varepsilon/\sqrt{r})\|v\|_2^2$ for all $v \in \mathcal{S}$.

Conditions (**C1**) and (**C2**) are naturally satisfied with high probability for sub-Gaussian matrices.

Proposition 1. *Let R have entries drawn i.i.d. from a zero-mean sub-Gaussian distribution and variance $1/k$. If $k = \Omega(\varepsilon^{-2}r\{\log(r) + \log(n)\} + \delta^{-2}\log(\delta^{-1})r)$, then $R^\top$ satisfies conditions (**C1**), (**C2**) with probability at least $1 - \exp(-c\log(\delta^{-1})r) - \exp(-c'\log(n\,r))$ for absolute constants $c, c' > 0$.*

The merits of conditions (**C1**) and (**C2**) can be seen from the following result.

Theorem 1. *Let $R^\top$ satisfy condition (**C1**) and (**C2**). We then have*

$$\|(I - P_{X_R})X\|_F^2 \leq \left(1 + \frac{\varepsilon^2}{(1-\delta)^2}\right)\|\Delta_r\|_F^2.$$

An implication for CLS is then as follows.

Corollary 1. *Under the condition of Theorem 1, the excess risk of CLS can be bounded as*

$$\mathcal{E}(R) \leq \left(1 + \frac{\varepsilon^2}{(1-\delta)^2}\right)\|w^*\|_2^2\frac{\|\Delta_r\|_F^2}{n} + \sigma^2\frac{k}{n}.$$

In particular, if k can be chosen proportional to r, the conclusions for PCR with polynomial (11) *and exponential* (12) *decay continue to hold, up to a constant factor and with $\|\alpha^*\|_\infty^2$ replaced by $\|w^*\|_2^2$.*

Regarding the second part of the corollary, Proposition 1 requires k to be chosen slightly larger than proportional to r. The extra log factor is likely to be an artifact of our analysis as is indicated by the result (13) in Halko et al. (2011) for Gaussian matrices. Putting this issue aside, the conclusion is that CLS can roughly match the excess risk of PCR even though there is a slack between $\|\alpha^*\|_\infty^2$ on one side and $\|w^*\|_2^2$ on the other which may be of the order of d. In light of (11) and (12) this may not have much of an effect as long as the spectrum of X exhibits strong decay.

In the worst case, however, the ratio of the excess risk of PCR and CLS can be arbitrarily large as can be seen from the finer bound (9): if α^* happens to be perfectly aligned with the top r singular values so that $\alpha^*_{r+} = 0$, we have $\mathcal{E}(V_r) = 0$. On the other hand, the column space of X_R does not contain that of U_r unless $k = d \wedge n$, hence in this rather specific case CLS falls short of PCR. It remains an open question whether there are scenarios in which CLS can substantially improve over PCR.

From a computational perspective, CLS avoids computation of the truncated SVD of X. Obtaining X_R only requires a single a matrix-matrix multiplication, an operation that amounts to $O(ndr)$ flops and that is trivially parallelizable. On the other hand, as discussed in Halko et al. (2011), state-of-the-art algorithms for computing the r-truncated SVD also require $O(ndr)$ flops on average, with a worse complexity for some problem instances.

5 Estimation of the bias term

The choice of the dimension k is a crucial issue in practice. In order to get an idea of the bias of CLS, it makes

sense to evaluate the quantity $\delta_R^2 = \|(I - P_{X_R})X\|_F^2$. We have $P_{X_R} = \mathcal{U}\mathcal{U}^\top$ for $\mathcal{U} \in \mathbb{R}^{n\times k}$ unitary which can be computed from an SVD of X_R using $O(nk^2)$ flops. However, forming $P_{X_R}X = \mathcal{U}\mathcal{U}^\top X$ requires $O(ndk)$ flops which is as expensive as computing X_R. Hence, we would like to circumvent this operation. We here propose an approach based on randomization that delivers a reasonably accurate estimate of δ_R^2 while achieving a reduction to $O(nd)$ flops.

Proposition 2. *Consider a collection of L i.i.d. d-dimensional standard Gaussian random vectors $\{\omega_l\}_{l=1}^L$ independent of R and the estimator*

$$\widehat{\delta}_R^2 = \frac{1}{L}\sum_{l=1}^{L}\|X\omega_l - P_{X_R}X\omega_l\|_2^2.$$

Then, for any $c \in (0,1)$ and any $C > 1$, as long as

$$L \geq \max\left\{\frac{16}{(1-c)^2}, \frac{144}{(C-1)^2}\right\}$$

it holds that $\mathbf{P}\left(c\delta_R^2 \leq \widehat{\delta}_R^2 \leq C\delta_R^2\right) \geq 0.96$.

For example, setting $C = 3$, $c = 1/3$, we would need $L = 36$ to estimate δ_R^2 within a multiplicative factor of 3 with probability near 1. The constants here may not necessarily be optimal. Note that computing $P_{X_R}X\omega_l = \mathcal{U}(\mathcal{U}^\top(X\omega_l))$ for a single l only amounts to $O(nd)$ flops. Proposition 2 is of independent interest as it provides a general scheme for estimating the Frobenius norm of a product of matrices.

6 Experiments

In this section, we present the results of experiments with synthetic and real data in order to support the main observations made in the previous sections.

6.1 Synthetic data

We start by generating a random n-by-d matrix X_0 with $n = 1000$, $d = 500$, where the entries of X_0 are drawn i.i.d. from the standard Gaussian distribution. The SVD of X_0 is given by

$$X_0 = U_0\Sigma_0 V_0^\top, \quad U_0 \in \mathbb{R}^{n\times d},\ \Sigma_0 \in \mathbb{R}^{d\times d},\ V_0 \in \mathbb{R}^{d\times d}.$$

We then replace Σ_0 by a diagonal matrix Σ whose diagonal elements $\{\sigma_j\}_{j=1}^d$ are chosen in a deterministic fashion according to one of the following regimes:

$$\text{constant}: \quad \sigma_j \propto 1,\ j \in [d],$$

$$\text{polynomial}: \ \sigma_j \propto j^{-q},\ q \in \left\{\frac{1}{2}, \frac{3}{4}, 1, \frac{3}{2}, 2, 4\right\},\ j \in [d],$$

$$\text{exponential}: \sigma_j \propto 0.9^j,\ j \in [d],$$

where the constant of proportionality is determined by the scaling $\sum_{j=1}^d \sigma_j^2 = n \cdot d$. We subsequently work with $X = U_0\Sigma V_0^\top$, generating data from the model

$$y = Xw^* + \sigma\xi, \tag{14}$$

where w^* is drawn uniformly from the unit sphere in $\mathbb{R}^d$, $\sigma \in 2^p$, $p \in \{-1, -0.5, \ldots, 1\}$, and ξ has i.i.d. standard Gaussian entries.

Given data (X, y), we then perform PCR with ten different choices of r, using an equi-spaced grid of values depending on the regime according to which X has been generated. For CLS, R is chosen as a standard d-by-k Gaussian matrix with $k = \alpha r$, where the oversampling factor $\alpha \in \{1, 1.2, 1.5, 2, 2.5, 3\}$. We conduct 100 independent replications for each regime. Our main interest is in the bias and the prediction error of PCR and CLS:

$$\|(I - P_{U_r})Xw^*\|_2^2/n \quad \text{vs.} \quad \|(I - P_{X_R})Xw^*\|_2^2/n,$$
$$\|Xw^* - XV_r\widehat{w}_{V_r}\|_2^2/n \quad \text{vs.} \quad \|Xw^* - X_R\widehat{w}_R\|_2^2/n,$$

where $\widehat{w}_{V_r}$ and $\widehat{w}_R$ denote the least squares estimator for data (XV_r, y) and (X_R, y), respectively.

A subset of the results involving the three different regimes of decay is shown in Figure 1.

In the constant regime (top row), the performance of CLS and PCR is not distinguishable. In fact, it can be shown that under model (14) and Gaussian R, the expected bias is the same for both approaches: for constant spectrum, Xw^* is uniformly distributed on the d-dimensional subspace of $\mathbb{R}^n$ spanned by U_0 intersected with the sphere of radius $\sqrt{n}$, and the column spaces of both XV_r (PCR) and X_R with $k = r$ are uniformly distributed on the Grassmanian $\mathsf{Gr}(r, \mathbb{R}^n)$, hence in both cases the expected bias is proportional to the expected distance of a random point from the sphere in $\mathbb{R}^d$ and a random element from $\mathsf{Gr}(r, \mathbb{R}^d)$.

In the regime of polynomial decay ($q = 1$), we observe that the bias of CLS is roughly proportional to that of PCR (or alternatively, we need to choose k as a suitable multiple of r to achieve the same bias). Accordingly, the dip in the prediction error curve occurs for $k = 2r^*$ with $r^* = 40$ yielding the smallest prediction error for PCR. In both low and high noise settings, PCR and CLS improve significantly over OLS in terms of prediction error ($\approx$0.02 and $\approx$0.04 vs. 0.125 and $\approx$ 0.1 and $\approx$ 0.15 vs. 2).

In the regime of exponential decay, the bias of CLS is not quite proportional to that of PCR for small values of r, but this improves once r reaches 20. Overall, the results agree well with what is suggested by the theory.

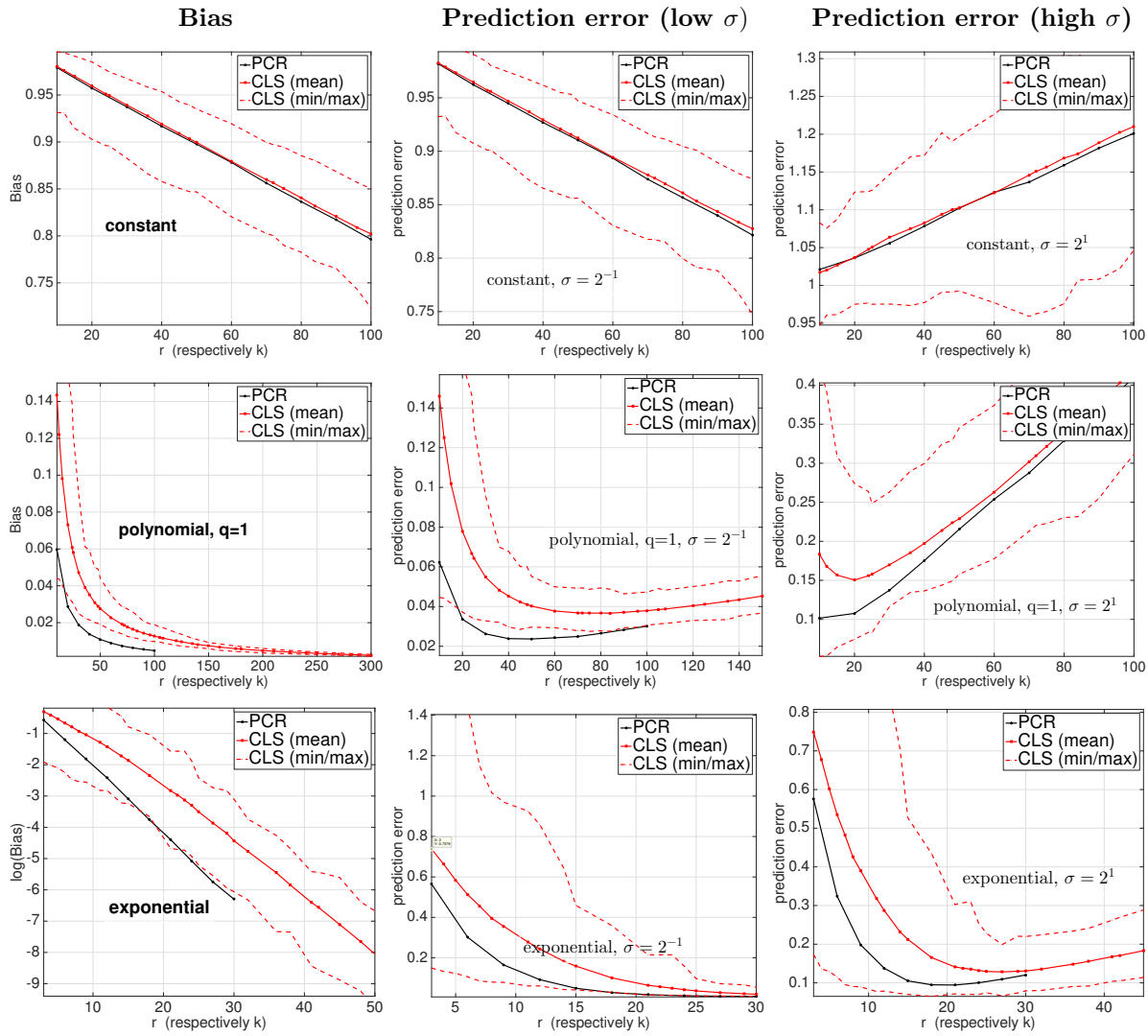

Figure 1: Results of the synthetic data experiment. From left to right: the bias $\|(I - P_{X_R})Xw^*\|_2^2/n$, and the mean squared prediction error $\|Xw^* - X_R\widehat{w}_R\|_2^2/n$ for $\sigma = 1/2$ (middle) and $\sigma = 2$ (right) in dependence of k (horizontal axis) for CLS in relation to PCR. Solid curves are averages, dashed curves minima and maxima (only CLS) over 100 replications. From top to bottom: constant spectrum, spectrum with polynomial decay ($q = 1$), exponential decay. Note that for the latter, the bias is plotted on a log scale. For comparison, the mean squared prediction error of OLS $\sigma^2 d/n$ equals $1/8 = .125$ for the middle column and 2 for the right column.

6.2 Real data

We analyze a subset of the Twitter social media buzz dataset available from the UCI machine learning repository, in a way that is similar to the analysis in Lu and Foster (2014). This is a regression problem in which the goal is to predict the popularity of topics as quantified by its mean number of active discussions given 77 predictor variables such as number of authors contributing to the topic over time, average discussion lengths, number of interactions between authors etc. We here only work with the first 8000 observations. Several of the original predictor variables as well as the response variable are log-transformed prior to analysis. Following Lu and Foster (2014), we add quadratic interactions which yields $d = 3080$ predictors in total. We consider 50 random partitions into a training set of size 6000 and a test set of size 2000 which is used to evaluate the prediction error. Training and test set are centered so that the response and predictors have zero mean, and the predictors are scaled to unit norm. We compare the mean squared prediction error on the test set of PCR with $r \in \{5, 10, \ldots, 50, 60, \ldots, 100, 120, \ldots, 200\}$ and CLS with $k = r\alpha$, where the grid for the factor α is as for the synthetic data. For CLS, we take R as a Gaussian

matrix. For each training set, we obtain 10 i.i.d. copies of R and perform regression with each of the resulting matrices X_R. The main results are summarized in Figure 2. Looking at the top panel, we find that the decay of the singular values is noticeable in the sense that the input matrices X can be well-approximated by a truncated SVD of small to moderate rank, but the rate of decay is still rather polynomial than exponential when compared to the bottom left panel of Figure 1. As already seen for the synthetic data experiments, CLS requires only a moderate amount of of oversampling to achieve the approximation error of PCR. Turning to the bottom part, we see that PCR with $r = 20$ achieves the lowest test error of about 2.3. The performance of CLS is not far off, with an optimal test error achieved for k about 40. From the point of view of computation, CLS is on average faster by a factor of about 50 when using the vanilla svd function in MATLAB for PCR.

While the example confirms that PCR and CLS achieve comparable performance, it turns out that both are not competitive for the given problem, being outperformed by the lasso by a margin (with an average test error that is smaller by a factor of 10 for an optimal choice of the regularization parameter).

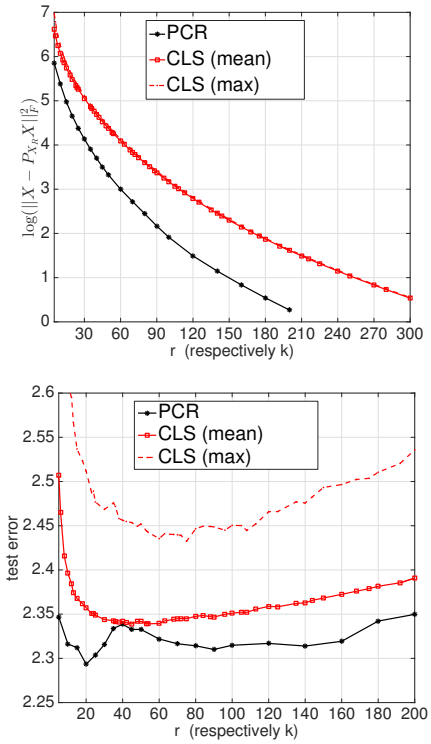

Figure 2: Top: Average approximation errors $\log(\|X - P_{U_r}X\|_F^2)$ (PCR) respectively $\log(\|X - P_{X_R}X\|_F^2)$ (CLS) over the 50 training sets from the Twitter data in dependence of r (or k). Bottom: Average test errors vs. r or k. For CLS, we plot both the mean and the maximum over then 10 realizations of R per iteration. For the top panel, the max curve essentially overlaps with the mean curve.

7 Conclusion

Linear dimension reduction by means of a Johnson-Lindenstrauss transform is commonly used in many standard machine learning problems (DasGupta, 2000; Bingham and Mannila, 2001; Fradkin and Madigan, 2003; Vempala, 2005). In linear regression, this approach is typically used by applying the transform to both the inputs and the responses, in which case one speaks of sketched regression. There a several recent and thorough analyses of sketched regression as mentioned in the introduction. By contrast, in this paper we have considered the situation in which the transform is only applied to the inputs while maintaining the original responses, for which Maillard and Munos (2009) coined the term compressed regression. Prior analysis in the latter work and in Kaban (2014) seems to suggest that the approach achieves a $O(1/\sqrt{n})$ bound on the excess risk without any assumptions on the design matrix, which is also referenced in recent work by Shah and Meinshausen (2016).

The analysis in the present paper is not affirmative. Instead, we show herein that the statistical performance of CLS can roughly match that of traditional PCR which can be reasonable even in a high-dimensional setup if the design matrix is approximately of low rank and if the regression coefficients are dense or at least not misaligned with the leading subspace. It is an open question whether there are interesting scenarios in which CLS substantially improves over PCR. CLS seems to have merits from the computational rather than the statistical side by achieving a reduction to a linear model of potentially small dimension without requiring an SVD or non-linear optimization as do approaches based on sparsity.

Regarding future work, exploring the statistical performance limit of compressed regression with choices of R different from Johnson-Lindenstrauss transforms (Shi et al., 2009; Shah and Meinshausen, 2016) under different assumptions on X like sparsity appears to be a worthwhile endeavor.

Supplement.

The supplement contains the proofs of Theorem 1 and Propositions 1 and 2.

Acknowledgment.

The experiments were run on ARGO, a research computing cluster provided by the Office of Research Computing at George Mason University, VA. (URL: http://orc.gmu.edu).

References

N. Ailon and B. Chazelle. Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. In *Proceedings of the Symposium on Theory of Computing (STOC)*, pages 557–563, 2006.

A. Artemiou and B. Li. On principal components and regression: a statistical explanation of a natural phenomenon. *Statistica Sinica*, 19:1557–1565, 2009.

E. Bingham and H. Mannila. Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the International Conference on Knowledge discovery and Data mining (KDD)*, pages 245–250, 2001.

S. DasGupta. Experiments with Random Projection. In *Uncertainty in Artificial Intelligence (UAI)*, 2000.

S. DasGupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures and Algorithms*, 22:60–65, 2003.

P. Drineas and M. Mahoney. RandNLA: Randomized Numerical Linear Algebra. *Communications of the ACM*, 59:80–90, 2016.

D. Fradkin and D. Madigan. Experiments with random projections for machine learning. In *Proceedings of International Conference on Knowledge discovery and Data mining (KDD)*, pages 517–522, 2003.

N. Halko, P. Martinsson, and J. Tropp. Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. *SIAM Review*, pages 217–288, 2011.

T. Hastie, R. Tisbhirani, and M. Wainwright. *Statistical Learning with Sparsity*. CRC Press, 2015.

W. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, pages 189–206, 1984.

A. Kaban. New Bounds on Compressive Linear Least Squares Regression. In *Artificial Intelligence and Statistics (AISTATS)*, pages 448–456, 2014.

M. Kendall. *A course in Multivariate Analysis*. Griffith, London, 1957.

P. Li and A.C. König. Theory and Applications of b-bit minwise hashing. *Communications of the ACM*, 54:101–109, 2011.

Y. Lu and D. Foster. Fast Ridge Regression with Randomized Principal Component Analysis and Gradient Descent. In *Uncertainty in Artificial Intelligence (UAI)*, 2014.

M. Mahoney. Randomized Algorithms for Matrices and Data. *Foundations and Trends in Machine Learning*, 3:123–224, 2011.

O. Maillard and R. Munos. Compressed least-squares regression. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1213–1221. 2009.

J. Matousek. On variants of the Johnson-Lindenstrauss lemma. *Random Structures and Algorithms*, 33:142–156, 2008.

M. Pilanci and M. Wainwright. Randomized Sketches of Convex Programs With Sharp Guarantees. *IEEE Transactions on Information Theory*, 61:5096–5115, 2015.

G. Raskutti and M. Mahoney. A Statistical Perspective on Randomized Sketching for Ordinary Least-Squares. arXiv:1406.5986, 2015.

T. Sarlos. Improved approximation algorithms for large matrices via random projections. In *Foundations of Computer Science (FOCS)*, pages 143–152, 2006.

R. Shah and N. Meinshausen. On b-bit min-wise hashing for large-scale regression and classification with sparse data. arXiv:1308.1269, 2016.

Q. Shi, J. Petterson, G. Dror, J. Langford, A. Smola, and SVN Vishwanathan. Hash kernels for structured data. *Journal of Machine Learning Research*, 10: 2615–2637, 2009.

S. Vempala. *The Random Projection Method*. American Mathematical Society, 2005.

S. Zhou, J. Lafferty, and L. Wasserman. Compressed and privacy-sensitive regression. *IEEE Transactions on Information Theory*, 55:846–866, 2009.