On Robustness of Principal Component Regression

Anish Agarwal
Department of EECS, MIT
Cambridge, MA 02139
anish90@mit.edu

Devavrat Shah
Department of EECS, MIT
Cambridge, MA 02139
devavrat@mit.edu

Dennis Shen
Department of EECS, MIT
Cambridge, MA 02139
deshen@mit.edu

Dogyoon Song
Department of EECS, MIT
Cambridge, MA 02139
dgsong@mit.edu

Abstract

Consider the setting of Linear Regression where the observed response variables, in expectation, are linear functions of the p-dimensional covariates. Then to achieve vanishing prediction error, the number of required samples, n, needs to scale faster than $p\sigma^2$ (i.e., $n \gg p\sigma^2$), where σ^2 is a bound on the noise variance. In a high-dimensional setting where p is large but the covariates admit a low-dimensional representation (say $r \ll p$), then Principal Component Regression (PCR), see [36], is an effective approach; here, the response variables are regressed with respect to the principal components of the covariates. The resulting number of required samples to achieve vanishing prediction error now only needs to scale faster than $r\sigma^2 (\ll p\sigma^2)$. Despite the tremendous utility of PCR, its ability to handle settings with noisy, missing, and mixed (discrete and continuous) valued covariates is not understood and remains an important open challenge, see [24]. As the main contribution of this work, we address this challenge by rigorously establishing that PCR is robust to noisy, sparse, and possibly mixed valued covariates. Specifically, under PCR, vanishing prediction error is achieved with the number of samples scaling as $n \gg r \max(\sigma^2, \rho^{-4} \log^5(p))$, where ρ denotes the fraction of observed (noisy) covariates. We establish generalization error bounds on the performance of PCR, which provides a systematic approach in selecting the correct number of components r in a data-driven manner. The key to our result is a simple, but powerful equivalence between (i) PCR and (ii) Linear Regression with covariate pre-processing via Hard Singular Value Thresholding (HSVT). From a technical standpoint, this work advances the state-of-the-art analysis for HSVT by establishing stronger guarantees with respect to the $\|\cdot\|_{2,\infty}$-error for the estimated matrix rather than the Frobenius norm/mean-squared error (MSE) as is commonly done in the matrix estimation / completion literature.

1 Introduction

In this paper, we are interested in developing a better understanding of a popular prediction method known as Principal Component Regression (PCR). In a typical prediction problem setup, we are given access to a labeled dataset $\{(Y_i, \boldsymbol{A}_{i,\cdot})\}$ over $i \geq 1$; here, $Y_i \in \mathbb{R}$ represents the response variable (also known as the label or target) we wish to predict and $\boldsymbol{A}_{i,\cdot} \in \mathbb{R}^{1\times p}$ represents the associated covariate (or feature) to be utilized in the prediction process.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Linear Regression. In Linear Regression, the data is believed to be generated as per a latent linear model and the goal is to learn the linear predictor. More precisely, for some $\beta^* \in \mathbb{R}^p$ and each $i \geq 1$, $Y_i = \boldsymbol{A}_{i,\cdot}\beta^* + \epsilon_i$, where ϵ_i denotes independent, zero-mean idiosyncratic noise with variance bounded by σ^2. Under generic noise distributions, the Ordinary Least Squares (OLS) estimator learnt using such observations yields an in-sample (or training) prediction error that vanishes to zero as long as the number of samples, n, scales faster than $p\sigma^2$ (i.e., $n \gg p\sigma^2$); e.g., see [51] and references therein. The same result holds true for the generalization prediction error under reasonable restrictions on the model class (e.g., see [51] and references therein).

Principal Component Regression. In the high-dimensional setting, the required number of samples may be too great since it scales with the number of features p, which is large. However, this problem can be circumvented when the covariates have a latent, low-dimensional representation. In particular, PCR, see [36], has been precisely designed to address such a setting. Using all observed covariates, PCR first finds an $r \ll p$ dimensional representation for each feature using the method of Principal Component Analysis (PCA); specifically, PCA projects every covariate $\boldsymbol{A}_{i,\cdot}$ onto the subspace spanned by the top r right singular vectors of the covariate matrix, the concatenation of all observed covariates. PCR then uses the r-dimensional features to perform linear regression. If the covariate matrix is indeed of rank r, then by the theory of Linear Regression, it follows that the number of samples required to achieve vanishing in- and out-of-sample prediction error need to scale faster than $r\sigma^2$ (i.e. $n \gg r\sigma^2$) , which is significantly smaller when $r \ll p$.

Noisy, missing, and mixed valued covariates. In many practical scenarios of interest (including high-dimensional settings where p is large), the covariates are not fully observed. Specifically, a common thread of many modern datasets is that only a small fraction of the covariates are observed, and the observations themselves are noisy versions of the true covariates. Moreover, as is standard in most real-world datasets, the observations may also be mixed (discrete and continuous) valued covariates. Despite the tremendous success of PCR in a variety of applications, its ability to handle such scenarios remains unknown, as noted in a recent survey [24].

In the context of Linear Regression, this scenario fits under the error-in-variable regression framework, where there has been exciting recent advancement, particularly in the high dimensional setting (see Section 1.2 for details). However, the current inventory of methods fall short in addressing the key challenge of handling noisy, sparse, and mixed valued covariates as the proposed estimators require detailed knowledge of the underlying noise model of the covariates.

1.1 Contributions

Summary of results. As the main contribution of this work, we argue that PCR, without any change, is robust to noise and missing values in the observed covariates. In particular, we demonstrate that PCR does not require *any* knowledge about the underlying noise model that corrupts the covariates in order to generate predicted responses. Formally, we argue that the (training) error decays to zero as long as the number of samples $n \gg r \max(\sigma^2, \rho^{-4}\log^5(p))$, where ρ denotes the fraction of observed (noisy) covariates. For a precise statement of the result, please see Theorem 4.1.

We also define an appropriate notion of generalization error for the particular setting of PCR. With respect to this notion, we establish that the testing prediction error of PCR scales similarly to that of the training error, i.e., the testing (or generalization) error is bounded above by the training prediction error plus a term that scales as $r^2(\log(np)/n)^{1/2}$; hence, the testing prediction error vanishes as long as the number of samples, $n \gg r \max(\sigma^2, \rho^{-4}\log^5(p))$. For a precise statement, see Theorem 4.2.

We extend our results for PCR's in- and out-of-sample prediction error even when the ground-truth covariates are not low-rank and the linear model itself may be misspecified (see Theorem 5.1 and Corollary 5.1). This result further suggests the robustness of PCR, reinforcing the utility of applying it in practice. Further, our result on generalization error provides a systematic way to select the correct number of principal components in a data-driven manner, i.e., to choose the value of r that minimizes the training error plus the generalization penalty term $r^2(\log(np)/n)^{1/2}$.

Finally, we describe various applications of our results, including synthetic control, privacy preserving regression, and regression with mixed valued covariates. Please refer to Section 6 for details.

Overview of techniques. To prove our results, we establish a simple, but powerful equivalence between (i) PCR and (ii) Linear Regression with covariate pre-processing via Hard Singular Value

Thresholding (HSVT) (see Proposition 3.1). The HSVT algorithm is commonly analyzed in literature, see [25], for matrix estimation / completion. In fact, there is significant literature establishing that HSVT is a noise-model-agnostic method that recovers the ground-truth matrix from its noisy observations. However, the current results concerning HSVT establish its estimation accuracy in terms of the mean-squared error or expected squared Frobenius norm of the error matrix. To establish our above mentioned results, we bound the expected squared $\|\cdot\|_{2,\infty}$ of the error matrix (see Lemma 5.1), which is a stronger guarantee than the Frobenius norm, as is standard in the literature (note $\frac{1}{np}\|\boldsymbol{E}\|_F^2 = \frac{1}{np}\sum_{i=1}^n\sum_{j=1}^p e_{ij}^2 \leq \frac{1}{n}\max_{j\in[p]}\sum_{i=1}^n e_{ij}^2 = \frac{1}{n}\|\boldsymbol{E}\|_{2,\infty}^2$). Thus, this result for HSVT may be of interest in its own right.

Our generalization error result utilizes the standard framework of Rademacher complexity, see [16] and references therein. However, there are two crucial differences that we need to overcome in order to obtain sharp, meaningful bounds. First, the notion of generalization we utilize to analyze PCR is slightly different from the traditional setup as the noisy test covariates (but not responses) are included in the training process, which complicates the analysis (see Section 2.3 for details); we relate this modified notion of generalization to that of the classical notion. Second, to obtain sharp bounds, we argue that the Rademacher complexity under PCR scales with the dimensionality of the number of principle components utilized rather than the ambient dimension p. To achieve this bound, we identify the Rademacher complexity of PCR with implicit ℓ_0-regularization.

1.2 Related works

We primarily focus on the literature pertaining to error-in-variable regression and PCR, but also include a brief discussion on the literature for matrix estimation / completion in Appendix A.1.

Error-in-variable regression. There exists a rich body of work regarding high-dimensional error-in-variable regression (see [41], [29], [45], [46], [17], [18], [26], [27], [37]). Two common threads of these works include: (1) a sparsity assumption on β^*; (2) error bounds with convergence rates for estimating β^* under different norms, i.e., $\|\widehat{\beta} - \beta^*\|_q$ where $\|\cdot\|_q$ denotes the ℓ_q-norm. In all of these works, the goal is to recover the underlying model, β^*. In contrast, as discussed, the goal of PCR is to primarily provide good prediction. Some notable works closest to our setup include [41], [29], [46], which are described in some more detail next.

In [41], a non-convex ℓ_1-penalization algorithm is proposed based on the plug-in principle to handle covariate measurement errors. This approach requires explicit knowledge of the unobserved noise covariance matrix $\Sigma_{\boldsymbol{H}} = \mathbb{E}\boldsymbol{H}^T\boldsymbol{H}$ and the estimator designed *changes* based on their assumption of $\Sigma_{\boldsymbol{H}}$. They also require explicit knowledge of a bound on $\|\beta^*\|_2$, the object they aim to estimate. In contrast, PCR does not require any such knowledge about the distribution of the noise matrix $\boldsymbol{H}$ (i.e., the algorithm does not explicitly use this information to make predictions).

The work of [29] builds upon [41], and propose a convex formulation of Lasso. Although the algorithm introduced does not require knowledge of $\|\beta^*\|_2$, similar assumptions on $\boldsymbol{Z}$ and $\boldsymbol{H}$ (e.g., sub-gaussianity and access to $\Sigma_{\boldsymbol{H}}$) are made. This renders their algorithm to be not model agnostic. In fact, many works (e.g., [45], [46], [17]) require either $\Sigma_{\boldsymbol{H}}$ to be known or the structure of $\boldsymbol{H}$ is such that it admits a data-driven estimator for its covariance matrix. This is so because these algorithms rely on correcting the bias for the matrix $\boldsymbol{Z}^T\boldsymbol{Z}$, which PCR does not need to compute.

It is worth noting that all these works in error-in-variable regression focus only on learning β^*, and not explicitly de-noising the noisy covariates. Thus even with the knowledge of β^*, it is not clear how to use it for producing predictions of response variable when given noisy covariates.

Principal Component Regression. The formal literature providing an analysis of PCR is surprisingly sparse, especially given its ubiquity in practice. A notable work is that of [15], which suggests a variation of PCR to infer the direction of the principal components. However, it stops short of providing meaningful finite sample analysis beyond what is naturally implied by that of standard Linear Regression. The regularization property of PCR is also well known due to its ability to reduce the variance. As a contribution, we provide rigorous finite sample guarantees of PCR: (i) under noisy, missing covariates; (ii) when the linear model is misspecified; (iii) when the covariate matrix is not exactly low-rank (see Theorem 5.1 and Corollary 5.1).

As a further contribution, we argue that the resulting regression model from PCR has sparse support (this is established using the equivalence between PCR and Linear Regression with covariate pre-

processing via HSVT); this sparsity allows for improved generalization error as the Rademacher complexity of the resulting model class scales with this sparsity parameter (i.e., the rank of the covariate matrix pre-processed with HSVT). Hence, PCR not only addresses the challenge of noisy and missing covariates, but also, in effect, performs multiple implicit regularization.

2 Problem Setup

The standard formulation for regression considers the setting where the covariates are noiseless and fully observed. In this work, our interest is in a more realistic setting where we observe a noisy and sparse version of the covariates. In particular, our interest is in the high-dimensional framework where the number of observations may be far fewer than the ambient dimension of the covariates.

2.1 Model

We describe the model in terms of the structural assumptions on the covariates and the generative process for the response variables. Let $N \geq 1$ denote the total number of observations of interest.

Covariates. Let $\boldsymbol{A} \in \mathbb{R}^{N \times p}$ denote the matrix of true covariates, where the number of predictors p is assumed to exceed N, i.e., $N \leq p$. We assume the entries of $\boldsymbol{A}$ are bounded:

Property 2.1. *There exists an absolute constant $\Gamma \geq 0$ such that $|A_{ij}| \leq \Gamma$ for all $(i, j) \in [N] \times [p]$.*

Additionally, we shall assume that the covariates have a lower-dimensional representation, which is formalized as follows:

Property 2.2. *The covariate matrix $\boldsymbol{A} \in \mathbb{R}^{N \times p}$ has rank $r < N \leq p$.*

Response Variables. For each $i \in [N]$, we let Y_i denote the random response variable that is linearly associated with the covariate $\boldsymbol{A}_{i,\cdot} \in \mathbb{R}^{1 \times p}$, i.e.,

$$Y_i = \boldsymbol{A}_{i,\cdot}\beta^* + \epsilon_i \tag{1}$$

where $\beta^* \in \mathbb{R}^p$ is the unknown model parameter and $\epsilon_i \in \mathbb{R}$ denotes the noise.

Property 2.3. *The response noise $\epsilon = [\epsilon_i] \in \mathbb{R}^N$ is a random vector with independent, mean zero entries such that each of its components has variance bounded above by σ^2.*

2.2 Observations

Rather than observing $\boldsymbol{A}$, we are given access to its corrupted version $\boldsymbol{Z}$, which contains noisy and missing values. Additionally, the observed response variables are restricted to a subset of the N observations.

Noisy covariates with missing values. We observe $\boldsymbol{Z} \in \mathbb{R}^{N \times p}$, which is assumed to satisfy the following property.

Property 2.4. *For all $(i, j) \in [N] \times [p]$, the (i, j)-th entry of $\boldsymbol{Z}$, denoted as Z_{ij}, is defined as $A_{ij} + \eta_{ij}$ with probability ρ and $\star$ with probability $1 - \rho$, for some $\rho \in (0, 1]$; here, $\star$ denotes a missing value and η_{ij} denotes the noise in the (i, j)-th entry.*

In words, Property 2.4 states that each entry Z_{ij} is observed with probability ρ, independently of other entries; however, even under observation, Z_{ij} is a noisy instantiation of the true covariate A_{ij}.

Let $\boldsymbol{H} = [\eta_{ij}] \in \mathbb{R}^{N \times p}$ denote the covariate noise matrix. For ease of notation, let us define $\boldsymbol{X} = \boldsymbol{A} + \boldsymbol{H}$ as the noisy perturbation of covariate matrix, without missing values. We assume the following property about the noise matrix $\boldsymbol{H}$ (see Appendix A.2 for the definition of the following ψ_α-random variables/vectors).

Property 2.5. *Let $\boldsymbol{H}$ be a matrix of independent, mean zero ψ_α -rows for some $\alpha \geq 1$, i.e., there exists an $\alpha \geq 1$ and $K_\alpha < \infty$ such that $\|\eta_{i,\cdot}\|_{\psi_\alpha} \leq K_\alpha$ for all $i \in [N]$. As a consequence, there exists a $\gamma^2 > 0$ such that $\|\mathbb{E}\eta_{i,\cdot}^T \eta_{i,\cdot}\| \leq \gamma^2$ for all $i \in [N]$ (note γ^2 can depend on both α and p).*

Response Variables. Let $\Omega \subset [N]$ with $|\Omega| = n < N$. We observe Y_i, where $i \in \Omega$.

2.3 Objective

Given noisy observations of all N covariates $\{\boldsymbol{Z}_{1,\cdot}, \ldots, \boldsymbol{Z}_{N,\cdot}\}$ and a subset of response variables $\{Y_i : i \in \Omega\}$, our aim is to produce an estimate $\widehat{Y} \in \mathbb{R}^N$ so that the prediction error is minimized. Specifically, we measure performance in terms of the *training error* and *testing error*:

$$\text{(training error)} \qquad \text{MSE}_\Omega(\widehat{Y}) = (1/n) \cdot \mathbb{E}\Big[\sum_{i \in \Omega} (\widehat{Y}_i - \boldsymbol{A}_{i,\cdot}\beta^*)^2\Big],$$

$$\text{(testing error)} \qquad \text{MSE}(\widehat{Y}) = (1/N) \cdot \mathbb{E}\Big[\sum_{i=1}^{N} (\widehat{Y}_i - \boldsymbol{A}_{i,\cdot}\beta^*)^2\Big].$$

Transductive semi-supervised learning setting. It is worth remarking that in our definition of test performance, the algorithm is given access to the observations associated with the covariates for both *training* and *testing* data during the training procedure; of course, however, the algorithm does not access the test response variables! Traditionally, the algorithm only has access to the training covariates and response variables during the training process. The reason for this difference is a simple consequence of the nature of the algorithm of interest, PCR. Specifically, PCR pre-processes the covariates using PCA, which changes the training procedure if only a subset of the covariates are utilized. Therefore, to allow for a meaningful evaluation, it is natural to allow the algorithm to have access to *all* available covariate information. This is commonly referred to in the literature as the transductive semi-supervised learning setting, where we want to infer the response variables for the specific unlabeled data. Indeed, as discussed in Section 6, having access to all covariates is entirely reasonable in many important real-world applications.

2.4 Notations, definitions, and a summary of assumptions.

For any matrix $\boldsymbol{B} \in \mathbb{R}^{N \times p}$ and index set $\Omega \subset [N]$, let $\boldsymbol{B}^\Omega$ denote the $|\Omega| \times p$ submatrix of $\boldsymbol{B}$ formed by stacking the rows of $\boldsymbol{B}$ according to Ω, i.e., $\boldsymbol{B}^\Omega$ is the concatenation of $\{\boldsymbol{B}_{i,\cdot} : i \in \Omega\}$. The superscript Ω is sometimes omitted if the matrix representation is clear from context. $\text{poly}(\alpha_1, \ldots, \alpha_k)$, denotes a function that scales at most polynomially in its arguments. Let $x \vee y = \max(x, y)$ and $x \wedge y = \min(x, y)$ for any $x, y \in \mathbb{R}$. Lastly, let $\mathbb{1}$ denote the indicator function.

3 Algorithm

We recall the description of PCR, as in [36]. In particular, we suggest a minor modification of PCR in the presence of missing data. Specifically, PCR is modified by simply rescaling the observed covariate matrix by the inverse of the fraction of observed data. We also describe Linear Regression with covariate pre-processing via Hard Singular Value Thresholding (HSVT). We observe that these two algorithms produce identical estimates of the response variable. This simple, but powerful equivalence will allow us to study the robustness property of PCR through the lens of HSVT.

3.1 Principal Component Regression

Let $\widehat{\rho}$ denote the fraction of observed entries of $\boldsymbol{Z}$, i.e., $\widehat{\rho} = \frac{1}{Np}\sum_{i=1}^{N}\sum_{j=1}^{p} \mathbb{1}(Z_{ij} \neq \star) \vee \frac{1}{Np}$. Let $\widetilde{\boldsymbol{Z}} \in \mathbb{R}^{N \times p}$ represent the rescaled version of $\boldsymbol{Z}$, where every unobserved value $\star$ is replaced by 0, i.e., for all $i \in [N]$ and $j \in [p]$, $\widetilde{Z}_{ij} = Z_{ij}/\widehat{\rho}$ if $Z_{ij} \neq \star$ and 0 otherwise.

The Singular Value Decomposition (SVD) of $\widetilde{\boldsymbol{Z}}$ is denoted as $\widetilde{\boldsymbol{Z}} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T = \sum_{i=1}^{N} s_i u_i v_i^T$ where $\boldsymbol{U} \in \mathbb{R}^{N \times N}$, $\boldsymbol{S} \in \mathbb{R}^{N \times p}$, and $\boldsymbol{V} \in \mathbb{R}^{p \times p}$. Without loss of generality, assume that the singular values s_i's are arranged in decreasing order, i.e., $s_1 \geq \ldots \geq s_N \geq 0$. Note that $\boldsymbol{U} = [u_1, \ldots, u_N]$ and $\boldsymbol{V} = [v_1, \ldots, v_p]$ are orthogonal matrices, i.e., the u_i's and v_j's are orthonormal vectors.

For any $k \in [N]$, let $\boldsymbol{U}_k = [u_1, \ldots, u_k]$, $\boldsymbol{V}_k = [v_1, \ldots, v_k]$, and $\boldsymbol{S}_k = \text{diag}(s_1, \ldots, s_k)$. Then, the k-dimensional representation of $\widetilde{\boldsymbol{Z}}$, as per PCA, is given by $\boldsymbol{Z}^{\text{PCR},k} = \widetilde{\boldsymbol{Z}}\boldsymbol{V}_k$. Let $\beta^{\text{PCR},k} \in \mathbb{R}^k$ be the solution to the Linear Regression problem under $\boldsymbol{Z}^{\text{PCR},k}$, i.e., $\beta^{\text{PCR},k}$ solves minimize $\sum_{i \in \Omega}(Y_i - \boldsymbol{Z}^{\text{PCR},k}_{i\cdot} w)^2$ over $w \in \mathbb{R}^k$. Then, the estimated response vector $\widehat{Y}^{\text{PCR},k} = \boldsymbol{Z}^{\text{PCR},k}\beta^{\text{PCR},k}$.

3.2 Ordinary Least Square Linear Regression after Hard Singular Value Thresholding

Here, we describe Linear Regression with covariate pre-processing via Hard Singular Value Thresholding (HSVT). To that end, given any $\lambda > 0$, we define the map $\text{HSVT}_\lambda : \mathbb{R}^{N \times p} \to \mathbb{R}^{N \times p}$, which simply shaves off the input matrix's singular values that are below the threshold λ. Precisely, given $\boldsymbol{B} = \sum_{i=1}^{N} \sigma_i x_i y_i^T$, let $\text{HSVT}_\lambda(\boldsymbol{B}) = \sum_{i=1}^{N} \sigma_i \mathbb{1}(\sigma_i \geq \lambda) x_i y_i^T$.

For any $k \in [N]$, given $\widetilde{\boldsymbol{Z}}$ as before, define $\boldsymbol{Z}^{\text{HSVT},k} = \text{HSVT}_{s_k}(\widetilde{\boldsymbol{Z}})$. Let $\beta^{\text{HSVT},k} \in \mathbb{R}^p$ be a solution of Linear Regression under $\boldsymbol{Z}^{\text{HSVT},k}$, i.e., $\beta^{\text{HSVT},k}$ solves minimize $\sum_{i \in \Omega}(Y_i - \boldsymbol{Z}_{i\cdot}^{\text{HSVT},k} w)^2$ over $w \in \mathbb{R}^p$. Then, the estimated response vector $\widehat{Y}^{\text{HSVT},k} = \boldsymbol{Z}^{\text{HSVT},k} \beta^{\text{HSVT},k}$.

3.3 Equivalence between the Response Estimates by PCR and HSVT-OLS

We now state a key relation between the above two algorithms. Precisely, the two algorithms produce identical estimated response vectors. Refer to Appendix C for a proof of Proposition 3.1.

Proposition 3.1. *For any $k \leq N$, $\widehat{Y}^{\text{PCR},k} = \widehat{Y}^{\text{HSVT},k}$.*

4 PCR Prediction Error: Low-Rank Covariates

We now state our main results in terms of the training and testing error for PCR. For a review on vector and matrix norms, see Appendix A.2.

4.1 Theorem Statements

Training prediction error. We state the following result for PCR when the covariate matrix is low-rank, i.e., $\boldsymbol{A}$ admits a low-dimensional representation, and PCR chooses the correct number of principal components.

Theorem 4.1 (Training Error of PCR)**.** *Let Properties 2.1, 2.2, 2.3, 2.4, and 2.5 hold and let r denote the rank of $\boldsymbol{A}$. Suppose PCR chooses the correct number of principal components $k = r$. Let $\rho \geq \frac{64 \log(Np)}{Np}$ and $n = \Theta(N)$. Then for any given $\Omega \subset [N]$,*

$$\text{MSE}_\Omega(\widehat{Y}) \leq \frac{4\sigma^2 r}{n} + C(\alpha) \frac{C' \log^2(np)}{n\rho^2} \|\beta^*\|_1^2 \cdot \left(r + \frac{(n^2\rho + np) \log^3(np)}{\rho^2 \tau_r^2} \right), \tag{2}$$

where $C' = (1 + \gamma + \Gamma + K_\alpha)^4$, τ_r is r-th singular value of true covariate matrix $\boldsymbol{A}$, and $C(\alpha) > 0$ a constant that may depend on $\alpha \geq 1$.

The proof of Theorem 4.1 follows from general results presented in Section 5, e.g., Corollary 5.1 by letting $\phi = 0$, $k = r$, $\boldsymbol{A}^k = \boldsymbol{A}$ and $\tau_{k+1} = \tau_{r+1} = 0$.

Test prediction error. We now evaluate the generalization performance of PCR. As previously mentioned, the emphasis of this work is to provide a rigorous analysis on the prediction properties of the PCR algorithm through the lens of HSVT. Recall from Proposition 3.1, PCR with parameter r is equivalent to Linear Regression with pre-processing of the noisy covariates using HSVT (more specifically, retaining the top r singular values). To that end, we study candidate vectors $\beta^{\text{HSVT},r} = \boldsymbol{V}_r \cdot \beta^{\text{PCR},r} \in \mathbb{R}^p$. In light of this observation, we establish the following result that suggests restricting our model class to sparse linear models only (the proof of which can be found in Appendix D).

Proposition 4.1. *Let $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ and $M = \boldsymbol{X} v$ for some $v \in \mathbb{R}^p$. If $\text{rank}(\boldsymbol{X}) = r$, then there exists $v^* \in \mathbb{R}^p$ such that $M = \boldsymbol{X} v^*$ and $\|v^*\|_0 = r$.*

By Proposition 4.1, for any $\boldsymbol{Z}^{\text{HSVT},r}$ and $\beta^{\text{HSVT},r} = \boldsymbol{V}_r \cdot \beta^{\text{PCR},r}$, there exists a $\beta' \in \mathbb{R}^p$ such that $\boldsymbol{Z}^{\text{HSVT},r} \cdot \beta^{\text{HSVT},r} = \boldsymbol{Z}^{\text{HSVT},r} \cdot \beta'$ where $\|\beta'\|_0 \leq r$. Thus, for analyzing the test error of PCR with parameter r, or equivalently Linear regression with covariate pre-processing using HSVT with rank r thresholding, it suffices to restrict our model class to linear predictors with sparsity r. Specifically,

$$\mathcal{F} = \{\beta \in \mathbb{R}^p : \|\beta\|_2 \leq B,\ \|\beta\|_0 \leq r\},$$

where $B > 0$ is a positive constant (we consider candidate vectors with bounded ℓ_2-norm as is commonly assumed in generalization error analysis).

Refer to Appendix E for a proof of Theorem 4.2 and a more rigorous theoretical justification of our model class of interest.

Theorem 4.2 (Test Error of PCR)**.** *Let the conditions of Theorem 4.1 hold. Further, let $\widehat{\beta} \in \mathcal{F}$. Then,*

$$\mathbb{E}_\Omega\left[\text{MSE}(\widehat{Y})\right] \leq \mathbb{E}_\Omega\left[\text{MSE}_\Omega(\widehat{Y})\right] + \frac{C' \cdot r^{3/2} \cdot \widehat{\alpha}^2}{\sqrt{n}}\|\beta^*\|_1$$

where $C' = CB^2\Gamma$ with $C > 0$ a universal constant; $\widehat{\alpha}^2 = \mathbb{E}[\|\widehat{\boldsymbol{A}}\|_{\max}^2]$; and $\mathbb{E}_\Omega$ denotes the expectation taken with respect to $\Omega \subset [N]$ (of size n), which is chosen uniformly at random without replacement.

Since Theorem 4.1 holds for any Ω, we note that $\mathbb{E}_\Omega[\text{MSE}_\Omega(\widehat{Y})]$ is also bounded above by the right-hand side of (2).

Implications. The statement of Theorem 4.1 requires that the *correct* number of principal components are chosen in PCR. In settings where all r singular values of $\boldsymbol{A}$ are roughly equal (see the discussion below for such an example), i.e., $\tau_1 \approx \tau_2 \approx \ldots \approx \tau_r = \Theta(\sqrt{Np/r})$, the training prediction error vanishes as long as $n \gg \max(\sigma^2 r, \rho^{-4} r \log^5 p)$. Further, as long as $r = O(\log^{\frac{1}{4}} p)$, the testing error also vanishes with the same scaling of n, with $n = \Theta(N)$.

4.2 Example

Embedded random Gaussian features. We now present a classical example that justifies algorithms such as PCR (or PCA). Consider the setting where the matrix of interest $\boldsymbol{A} \in \mathbb{R}^{N \times p}$ is generated by sampling its rows from a distribution on $\mathbb{R}^p$, which in turn, is an embedding of some underlying latent distribution on $\mathbb{R}^r$. Specifically, consider the example in Proposition 4.2, which describes how the rows of $\boldsymbol{A}$ are generated; this is similar in spirit to the probabilistic model for PCA, cf. [19, 48] (refer to Appendix J.1 for its proof).

Proposition 4.2. *Let $\boldsymbol{A} = \tilde{\boldsymbol{A}}\tilde{\boldsymbol{R}}$ where $\tilde{\boldsymbol{A}} \in \mathbb{R}^{N \times r}$ is a random matrix whose entries are independent standard normal random variables, i.e., $\tilde{A}_{ij} \sim \mathcal{N}(0,1)$ and $\tilde{\boldsymbol{R}} \in \mathbb{R}^{r \times p}$ is another random matrix with independent entries such that $\tilde{R}_{ij} = 1/\sqrt{r}$ with probability $1/2$ and $\tilde{R}_{ij} = -1/\sqrt{r}$ with probability $1/2$. Suppose, $r \leq \frac{\sqrt{p}}{4\sqrt{2\log p}} + 1$. Then,*

$$\text{MSE}_\Omega(\widehat{Y}) \leq \frac{4\sigma^2 r}{n} + C''\|\beta^*\|_1^2 \frac{r \log^7(np)}{n\rho^4}$$

where $C'' > 0$ is a constant that may depend on model parameters γ, $\alpha \geq 1$, and K_α.

5 PCR Prediction Error: Beyond Low-Rank Covariates, Mismatched Model

We state a bound on the prediction error for PCR in the general setting where the covariates are not necessarily low-rank. We also consider the scenario where the response variables may satisfy the linear model but with error, i.e., the linear model is mismatched. Precisely, rather than satisfying (1), we assume the response variables are generated in the following manner: for each $i \in [N]$, the random response Y_i is associated with the covariate $\boldsymbol{A}_{i,\cdot} \in \mathbb{R}^{1 \times p}$ such that

$$Y_i = \boldsymbol{A}_{i,\cdot}\beta^* + \phi_i + \epsilon_i, \tag{3}$$

where $\beta^* \in \mathbb{R}^p$ remains the unknown model parameter, $\epsilon_i \in \mathbb{R}$ again denotes the zero mean response noise satisfying Property (2.3), and $\phi_i \in \mathbb{R}$ is the arbitrary mismatch error; for simplicity, we assume the mismatch error is deterministic. In contrast to Property 2.2, we do not assume the covariate matrix $\boldsymbol{A}$ is necessarily low-rank. However, as in Section 2, we assume the other properties hold, i.e., the conditions on the observed (noisy) covariate matrix $\boldsymbol{Z}$ and training subset $\Omega \in [N]$ of size n. As before, our interest is in bounding the prediction error of PCR, but we now do so in the general setting where $\boldsymbol{A}$ is not necessarily low-rank and there exists a mismatch error in the linear model.

5.1 Theorem Statements

Training Prediction Error. We first state a somewhat abstract result, Theorem 5.1 (proof in Appendix F). Next, we state a technical property of HSVT, Lemma 5.1 (proof in Appendix H). Together, they yield a concrete result, Corollary 5.1. For definitions on vector/matrix norms, see Appendix A.2.

Theorem 5.1 (Training Error of PCR: Generic Result)**.** *Consider PCR with parameter $k \geq 1$. Suppose Property 2.3 holds. Then, under the model described by (3),*

$$\text{MSE}_\Omega(\widehat{Y}) \leq \frac{4\sigma^2 k}{n} + \frac{3\|\beta^*\|_1^2 \mathbb{E}\|\boldsymbol{A}^\Omega - \widehat{\boldsymbol{A}}^\Omega\|_{2,\infty}^2}{n} + 5\|\phi\|_\infty^2. \tag{4}$$

Interpretation. The bound in (4) has three terms on the right hand side: (a) $\sigma^2 k/n$ represents the standard "regression" prediction error, which scales with the model complexity k and inversely with number of samples n; (b) $\|\beta^*\|_1^2 \cdot \mathbb{E}\|\boldsymbol{A}^\Omega - \widehat{\boldsymbol{A}}^\Omega\|_{2,\infty}^2/n$, which is a consequence of the corruption of $\boldsymbol{A}$ (if $\boldsymbol{A}$ was fully observed and rank k, then this error term would vanish); (c) $\|\phi\|_\infty^2$ represents the (inevitable) impact of the model mismatch.

Quantification. To quantify (4), we need to evaluate $\mathbb{E}[\|\boldsymbol{Z}^{\text{HSVT},k,\Omega} - \boldsymbol{A}^\Omega\|_{2,\infty}^2]$ under HSVT. In effect, HSVT produces the estimate $\boldsymbol{Z}^{\text{HSVT},k}$ of $\boldsymbol{A}$ from its noisy and sparse instantiation $\boldsymbol{Z}$. Our interest is in evaluating the estimation error with respect to the $\ell_{2,\infty}$-error. It is worth remarking that the estimation error for HSVT is typically evaluated with respect to the Frobenius norm; hence, this quantity is well understood, see [25]. On the other hand, the error bound with respect to $\ell_{2,\infty}$-norm is unknown. To that end, we provide a novel characterization of this error in Lemma 5.1 below.

Let the SVD of the covariate matrix be $\boldsymbol{A} = \sum_{i=1}^N \tau_i u_i v_i^T$ with the singular values τ_i arranged in descending order. Let $\boldsymbol{A}^k = \sum_{i=1}^k \tau_i u_i v_i^T$ denote the truncation of $\boldsymbol{A}$ obtained by retaining the top k components. Then for $C(\alpha) > 0$, an absolute constant that depends only on α, we define the quantity,

$$\Delta = \sqrt{N\rho}\sqrt{\rho\gamma^2 + (1-\rho)\Gamma^2} + 2C(\alpha)\sqrt{p}(K_\alpha + \Gamma)\Big(1 + 9\log(Np)\Big)^{\frac{1}{\alpha}}\sqrt{\log(Np)}. \tag{5}$$

Lemma 5.1 ($\|\cdot\|_{2,\infty}$ error bound for HSVT)**.** *Let Properties 2.1, 2.3, 2.4, and 2.5 hold. Let τ_k and τ_{k+1} denote the k-th and $(k+1)$-st singular values of $\boldsymbol{A}$, respectively. Suppose $\rho \geq \frac{64\log(Np)}{Np}$. Then, for $C > 0$, a universal constant.*

$$\mathbb{E}[\|\boldsymbol{Z}^{\text{HSVT},k} - \boldsymbol{A}\|_{2,\infty}^2] \leq \frac{C(K_\alpha^2 + \Gamma^2)}{\rho^2}\Big(k + \frac{N\Delta^2}{\rho^2(\tau_k - \tau_{k+1})^2}\Big)\log^{\frac{2}{\alpha}} Np + 2\|\boldsymbol{A}^k - \boldsymbol{A}\|_{2,\infty}^2.$$

Corollary 5.1 (Training Error of PCR: Generic Result)**.** *Let the conditions of Theorem 5.1 and Lemma 5.1 hold. Let $n = \Theta(N)$. Then, for $C' = (1 + \gamma + \Gamma + K_\alpha)^4$ and $C(\alpha) > 0$ is a constant that may depend on $\alpha \geq 1$, we have*

$$\text{MSE}_\Omega(\widehat{Y}) \leq \frac{4\sigma^2 k}{n} + \frac{C(\alpha)C'\|\beta^*\|_1^2 \log^2 np}{n\rho^2}\Big(\frac{(n^2\rho + np)\log^3 np}{\rho^2(\tau_k - \tau_{k+1})^2} + k\Big) + \frac{6\|\beta^*\|_1^2}{n}\|\boldsymbol{A}^k - \boldsymbol{A}\|_{2,\infty}^2 + 5\|\phi\|_\infty^2. \tag{6}$$

Test Prediction Error. Theorem 4.2 holds with r replaced by a general k.

How do we pick a good k in practice? The purpose of test prediction error, such as that implied by Theorem 4.2, is to precisely resolve such a question. Specifically, Theorem 4.2 suggests that the overall error is at most the training error plus a term that scales as $k^2(\log(np)/n)^{1/2}$. Therefore, one should choose the k that minimizes this bound. Naturally, as k increases, the training error is likely to decrease, but the additional term $k^2(\log(Np)/n)^{1/2}$ will increase; therefore, a unique minima in terms of the value of k exists and can be found in a data-driven manner.

5.2 Example

To explain the utility of Theorem 5.1 and Corollary 5.1, we consider a setting where $\boldsymbol{A}$ is an approximately low-rank matrix with geometrically decaying singular values. To that end, let $e_{\cdot,j} \in \mathbb{R}^p$ denote the j-th canonical basis vector. Let u_i, v_i, and τ_i denote the left singular vectors, right singular vectors, and singular values of $\boldsymbol{A}$, respectively. Let $\tau_1 = C_1\sqrt{Np}$ for some constant $C_1 > 0$. Further, suppose $\tau_k = \tau_1\theta^{k-1}$ for all $k \in [N]$ with $\theta \in (0,1)$, and let $v_i^T e_j = O(1/\sqrt{p})$ for all $i, j \in [p]$.

The conditions stated above are self-explanatory with potentially one exception: $v_i^T e_j = O(1/\sqrt{p})$. In effect, this assumption states that the right singular vectors of $\boldsymbol{A}$ satisfy an "incoherence" condition,

cf. [22], with the natural basis; or, equivalently, all elements of the right singular vectors are roughly of the same magnitude, $O(1/\sqrt{p})$. Under this setting, we state the following (proof in Appendix J.2):

Proposition 5.1. *Let $\boldsymbol{A}$ be generated as above and let conditions of Corollary 5.1 hold. Suppose PCR chooses parameter $k = C_2 \frac{\log\log(np)}{\log(1/\theta)}$ for absolute constant $C_2 > 0$. Then, for $C' = (1+\gamma+\Gamma+K_\alpha)^4$, $C'(\alpha,\theta) > 0$ a constant dependent only on α and θ; and $C'' > 0$, a universal constant, we have*

$$\text{MSE}_\Omega(\widehat{Y}) \leq \frac{2C_2\sigma^2 \log\log np}{n\log(1/\theta)} + \frac{C'(\alpha,\theta)C'\|\beta^*\|_1^2 \log^{5+4C_2} np}{n\rho^4} + \frac{C''\|\beta^*\|_1^2}{\log^{2C_2} np} + 5\|\phi\|_\infty^2, \tag{7}$$

From Proposition 5.1, it follows that if the number of principal components is chosen as $\Theta\left(\frac{\log\log(np)}{\log(1/\theta)}\right)$ and $n = \Omega(\rho^{-4}\text{poly}(\log p))$, then the training prediction error is effectively $\|\phi\|_\infty^2$ for sufficiently large n, p. This is precisely the unavoidable model mismatch error.

Existence of such a matrix. Here, we show that there exists a matrix with exponentially decaying singular values that also satisfies the required properties of our theorem.

We will construct an example based on the incoherence between the canonical basis and the Discrete Fourier Transform (DFT) basis. Suppose that $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$, where (i) $\boldsymbol{\Sigma}$ is a diagonal matrix such that $\Sigma_{11} = C\sqrt{Np}$ for some $C > 0$ and the diagonal entries of $\boldsymbol{\Sigma}$ satisfy $0 \leq (\Sigma_{i+1,i+1})(\Sigma_{i,i}) \leq \theta$ for all $i \in [N \wedge p - 1]$ and for some $\theta \in (0,1)$; (ii) $\boldsymbol{U} \in \mathbb{R}^{N\times N}$ is a DFT matrix such that $U_{ij} = (1/\sqrt{N}) \cdot e^{\boldsymbol{i}\frac{2\pi}{N}(i-1)(j-1)}$ for all $i, j \in [N]$, where $\boldsymbol{i}$ denotes the imaginary unit; (iii) $\boldsymbol{V} \in \mathbb{R}^{p\times p}$ is a DFT matrix such that $V_{ij} = (1/\sqrt{p}) \cdot e^{\boldsymbol{i}\frac{2\pi}{p}(i-1)(j-1)}$ for all $i, j \in [p]$.

The entries of the resulting matrix $\boldsymbol{A}$ are complex numbers, but one could also construct $\boldsymbol{A}$ by taking $\boldsymbol{U}$ and $\boldsymbol{V}$ as discrete cosine (or sine) transform matrices. Further, observe that $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices; hence, $\sigma_i(\boldsymbol{A}) = \sigma_i(\boldsymbol{\Sigma})$ for all $i \in [N \wedge p]$. Next, we argue that $\|\boldsymbol{A}\|_{\max} \leq C'$ for some constant $C' > 0$ (the proof of which can be found in Appendix J.3).

Proposition 5.2. *Let $\boldsymbol{A}$ be generated as above. Then, $\|\boldsymbol{A}\|_{\max} \leq \frac{C}{1-\theta}$. Here, $C > 0$ and $\theta \in (0,1)$ are the constants that appear in the description of $\boldsymbol{\Sigma}$.*

6 Applications

Given the ubiquity of PCR in practice, we describe four concrete, important applications that are enabled (and theoretically justified) by our formulation and the associated finite sample analyses results: (i) causal inference (synthetic control); (ii) privacy preserving learning; (iii) regression with mixed valued covariates. We choose these examples as they showcase the broad meaning of "error" with respect to the covariates. (i) is related to measurement error (as is commonly assumed with temporal data); (ii) is when noise is added to the covariates by design (in this example, to ensure differential privacy); (iii) is when the structure of the covariates restricts our observations to only its noisy instantiations (in this example, the latent covariate of interest is a "continuous" Bernoulli parameter, but we only observe its discrete $0/1$ categorical instantiation).

Due to space constraints, we defer the detailed description of these applications to Appendix B.

7 Conclusion

In conclusion (i) our work addresses a long-standing problem of demonstrating PCR is effective when we only have access to high-dimensional noisy, sparse, and mixed valued covariates - in particular, we provide non-asymptotic bounds for both training and testing error (transductive semi-supervised learning) for these settings as well as when the covariate matrix is not low-rank and and the linear model is misspecified; (ii) we establish a simple, but powerful equivalence between PCR and linear regression with covariate pre-processing via HSVT, and provide a novel error analysis of matrix estimation via HSVT with respect to the $\|\cdot\|_{2,\infty}$-norm; (iii) we formally connect our results with important applications to demonstrate the broad meaning of "noisy covariates": (a) synthetic control (measurement noise); (b) differentially-private regression (noise added by design); (c) mixed covariates ("structural" noise).

References

[1] A. Abadie, A. Diamond, and J. Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of californiaâs tobacco control program. *Journal of the American Statistical Association*, 2010.

[2] A. Abadie and J. Gardeazabal. The economic costs of conflict: A case study of the basque country. *American Economic Review*, 2003.

[3] E. Abbe and C. Sandon. Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 670–688. IEEE, 2015.

[4] E. Abbe and C. Sandon. Recovering communities in the general stochastic block model without knowing the parameters. In *Advances in neural information processing systems*, 2015.

[5] E. Abbe and C. Sandon. Detection in the stochastic block model with multiple clusters: proof of the achievability conjectures, acyclic bp, and the information-computation gap. *Advances in neural information processing systems*, 2016.

[6] R. Adamczak, A. E. Litvak, A. Pajor, and N. Tomczak-Jaegermann. Sharp bounds on the rate of convergence of the empirical covariance matrix. *Comptes Rendus Mathematique*, 349(3-4):195–200, 2011.

[7] A. Agarwal, M. J. Amjad, D. Shah, and D. Shen. Model agnostic time series analysis via matrix estimation. *POMACS*, 2(3):40–79, 2018.

[8] E. M. Airoldi, T. B. Costa, and S. H. Chan. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems*, pages 692–700, 2013.

[9] D. J. Aldous. Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598, 1981.

[10] M. Amjad, V. Mishra, D. Shah, and D. Shen. mrsc: Multi-dimensional robust synthetic control. *arXiv preprint arXiv:1905.06400*, 2019.

[11] M. J. Amjad, D. Shah, and D. Shen. Robust synthetic control. *Journal of Machine Learning Research*, 19:1–51, 2018.

[12] A. Anandkumar, R. Ge, D. Hsu, and S. Kakade. A tensor spectral approach to learning mixed membership community models. In *Conference on Learning Theory*, pages 867–881, 2013.

[13] S. Athey, M. Bayati, N. Doudchenko, and G. Imbens. Matrix completion methods for causal panel data models. 2017.

[14] S. Athey and G. Imbens. The state of applied econometrics - causality and policy evaluation. *The Journal of Economic Perspectives*, 31(2):3–32, 2016.

[15] E. Bair, T. Hastie, D. Paul, and R. Tibshirani. Prediction by supervised principal components. *Journal of the American Statistical Association*, 101(473):119–137, 2006.

[16] P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.*, 3:463–482, Mar. 2003.

[17] A. Belloni, V. Chernozhukov, A. Kaul, M. Rosenbaum, and A. B. Tsybakov. Pivotal estimation via self-normalization for high-dimensional linear models with errors in variables. *arXiv:1708.08353*, 2017.

[18] A. Belloni, M. Rosenbaum, and A. B. Tsybakov. Linear and conic programming approaches to high-dimensional errors-in-variables models. *Journal of the Royal Statistical Society*, 79:939–956, 2017.

[19] C. M. Bishop. Bayesian pca. In *Advances in neural information processing systems*, pages 382–388, 1999.

[20] C. Borgs, J. Chayes, C. E. Lee, and D. Shah. Thy friend is my friend: Iterative collaborative filtering for sparse matrix estimation. In *Advances in Neural Information Processing Systems*, pages 4718–4729, 2017.

[21] C. Borgs, J. T. Chayes, H. Cohn, and S. Ganguly. Consistent nonparametric estimation for heavy-tailed sparse graphs. *arXiv preprint arXiv:1508.06675*, 2015.

[22] E. Candes and J. Romberg. Sparsity and incoherence in compressive sampling. *Inverse problems*, 23(3):969, 2007.

[23] E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[24] G. Chao, Y. Luo, and W. Ding. Recent advances in supervised dimension reduction: A survey. *Machine Learning and Knowledge Extraction*, 1(1):341–358, 2019.

[25] S. Chatterjee. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2015.

[26] Y. Chen and C. Caramanis. Orthogonal matching pursuit with noisy and missing data: Low and high dimensional results. *arXiv preprint arXiv:1206.0823*, 2012.

[27] Y. Chen and C. Caramanis. Noisy and missing data regression: Distribution-oblivious support recovery. In *International Conference on Machine Learning*, pages 383–391, 2013.

[28] Y. Chen and M. J. Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

[29] A. Datta and H. Zou. Cocolasso for high-dimensional error-in-variables regression. *The Annals of Statistics*, 45(6):2400–2426, 2017.

[30] M. A. Davenport, Y. Plan, E. van den Berg, and M. Wootters. 1-bit matrix completion. *Information and Inference*, 3(3):189–223, 2014.

[31] C. Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[32] N. Doudchenko and G. Imbens. Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. *NBER Working Paper No. 22791*, 2016.

[33] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[34] S. B. Hopkins and D. Steurer. Efficient bayesian estimation from few samples: community detection and related problems. In *Foundations of Computer Science (FOCS), 2017 IEEE 58th Annual Symposium on*, pages 379–390. IEEE, 2017.

[35] C. Hsiao, S.-K. Wan, and Y. Xie. Panel data approach vs synthetic control method. *Economics Letters*, 164:121–123, 2018.

[36] I. T. Jolliffe. A note on the use of principal components in regression. *Journal of the Royal Statistical Society*, 31(3):300–303, 1982.

[37] A. Kaul and H. L. Koul. Weighted ℓ_1-penalized corrected quantile regression for high dimensional measurement error models. *Journal of Multivariate Analysis*, 140:72–91, 2015.

[38] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

[39] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11(Jul):2057–2078, 2010.

[40] C. E. Lee, Y. Li, D. Shah, and D. Song. Blind regression: Nonparametric regression for latent variable models via collaborative filtering. In *Advances in Neural Information Processing Systems 29*, pages 2155–2163, 2016.

[41] P.-l. Loh and M. J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664, 2012.

[42] R. Meir and T. Zhang. Generalization error bounds for bayesian mixture algorithms. *Journal of Machine Learning Research*, 4:839–860, 2003.

[43] S. Negahban and M. J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, pages 1069–1097, 2011.

[44] B. Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(Dec):3413–3430, 2011.

[45] M. Rosenbaum and A. B. Tsybakov. Sparse recovery under matrix estimation. *The Annals of Statistics*, 38(5):2620–2651, 2010.

[46] M. Rosenbaum and A. B. Tsybakov. Improved matrix uncertainty selector. *From Probability to Statistics and Back: High-Dimensional Models and Processes*, 9:276–290, 2013.

[47] D. Shah and D. Song. Learning mixture model with missing values and its application to rankings. *arXiv preprint arXiv:1812.11917*, 2018.

[48] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[49] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[50] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

[51] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[52] P.-Å. Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

[53] Y. Zhang, E. Levina, and J. Zhu. Estimating network edge probabilities by neighborhood smoothing. *arXiv preprint arXiv:1509.08588*, 2015.