Principal Component Regression with Semirandom Observations via Matrix Completion

Aditya Bhaskara
University of Utah

Kanchana Ruwanpathirana
University of Utah

Maheshakya Wijewardena
University of Utah

Abstract

Principal Component Regression (PCR) is a popular method for prediction from data, and is one way to address the so-called multi-collinearity problem in regression. It was shown recently that algorithms for PCR such as hard singular value thresholding (HSVT) are also quite robust, in that they can handle data that has missing or noisy covariates. However, such spectral approaches require strong distributional assumptions on which entries are observed. Specifically, every covariate is assumed to be observed with probability (exactly) p, for some value of p. Our goal in this work is to weaken this requirement, and as a step towards this, we study a "semi-random" model. In this model, every covariate is revealed with probability p, and then an adversary comes in and reveals additional covariates. While the model seems intuitively easier, it is well known that algorithms such as HSVT perform poorly. Our approach is based on studying the closely related problem of Noisy Matrix Completion in a semi-random setting. By considering a new semidefinite programming relaxation, we develop new guarantees for matrix completion, which is our core technical contribution.

1 Introduction

Regression is one of the fundamental problems in statistics and data analysis, with over two hundred years of history. We are given n observations, each consisting of d *prediction* or *regression* variables and one output that is a linear function of the prediction variables. Denoting the outputs by a vector $y \in \mathbb{R}^n$ and the prediction variables by a matrix $M \in \mathbb{R}^{n \times d}$ (each row corresponds to an observation), linear regression aims to model y as

$$y = M\beta + \eta, \tag{1}$$

where $\beta \in \mathbb{R}^d$ the vector of regression coefficients, and η is a vector of error terms, typically modeled as having independent Gaussian entries. The goal of the regression problem is to find the coefficients β. The standard procedure is to solve the least squares problem, $\min\|y - M\beta\|^2$.

One of the issues in high dimensional regression (large d) is the problem of *multi-collinearity*, where there are correlations (in values) between regression variables, leading to unstable values for the parameters β (see Gunst and Webster (1975); Jolliffe (1986)). One common approach to overcome this problem is to use subset selection methods (e.g., Hocking (1972); Draper and Smith (1966)). Another classic approach Hotelling (1957); Jolliffe (1986) is to perform regression after projecting to the space of the top principal components of the matrix M. This is called *Principal Component Regression (PCR)*. Formally, if $M^{(r)} = U^{(r)}\Sigma^{(r)}(V^{(r)})^T$ is the best rank r approximation of M, then the idea is to replace (1) with

$$y = MV^{(r)}\beta' + \eta. \tag{2}$$

The goal is to now find the best $\beta' \in \mathbb{R}^r$. This is known to yield more stable results in many settings (Mosteller and Tukey, 1977). It also has the advantage of being a "smaller" problem, thus it can be solved more efficiently. Recently, Agarwal et al. (2019) observed that in settings where the regression matrix is (close to) low rank, PCR also provides a way to deal with *missing and noisy* covariates — a well-known issue in applications (Little, 1992). Their main insight is that observing a random subset of the entries of M is good enough to obtain the top singular vectors $V^{(r)}$ to a high accuracy. Thus one can obtain a β' that yields a low-error regression model.

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

While the result of Agarwal et al. (2019) provides novel theoretical guarantees on regression with missing entries, it is restrictive: guarantees can only be obtained when every entry of M is observed *independently with (the same) probability* p. Given inherent dependencies in applications (e.g., some users may intrinsically reveal more information than others; certain covariates may be easier to measure than others, etc.), we ask: *Can we design algorithms for partially-observed PCR under milder assumptions on the observations?*
It is easy to see that unless sufficiently many entries are observed from each column, recovering the corresponding coordinate of β is impossible. This motivates having a *lower bound* on the probability of each entry being revealed. This leads us to considering a so-called *semi-random* model (Blum and Spencer, 1995; Feige and Krauthgamer, 2000; Makarychev et al., 2012; Cheng and Ge, 2018). In our context, such a model corresponds to the following: (a) initially, all the entries are revealed independently with probability p, (b) an adversary reveals *additional* elements (arbitrarily). Despite appearing to make the problem "easier" than the random case, since an algorithm has no idea if an entry was revealed in step (a) or (b), algorithms based on obtaining estimators (that rely on entries being observed with probability p) will fail.

Further motivation for the semirandom model. In both PCR with missing covariates and matrix completion, at a high level, we wish to understand what "observation patterns" allow for effective recovery. Without stochasticity assumptions, there are hardness results (even when one observes $\Omega(n^2)$ entries; Hardt et al. (2014)), and we know of positive results when each entry is observed independently with probability p. The semirandom model has been studied in the literature (for graph partitioning and also for matrix completion) because in spite of seeming easier as discussed above, it ends up causing spectral methods to fail. The semirandom model is equivalent to the setting where every entry has some "base" probability of being observed, but the probability could be higher for some (unknown to us) entries. E.g., in recommender systems, some users may provide more ratings for certain types of items than others, and this is typically unknown to the algorithm. The model is also closely related to the notion of *Massart noise* which is known to be challenging for a variety of learning problems (see, e.g., Diakonikolas et al. (2019)).

Our approach to partially-observed PCR will be based on the closely related problem of matrix completion (see Candes and Recht (2008); Keshavan et al. (2012); Bhojanapalli and Jain (2014) and references therein). The problems are related because intuitively, we can think of *filling in* the missing values in the covariate matrix and applying PCR. Further, matrix completion can also be formulated (and indeed has been studied in Cheng and Ge (2018)) in a semi-random setting. However, to the best of our knowledge, trade-offs between the error in the observations and the recovery error have not been studied in a semi-random model. Analyzing this via a semidefinite programming (SDP) relaxation is our main technical contribution. Interestingly, our analysis relies heavily on a recent non-convex approach to matrix completion (Chen et al., 2019).

1.1 Our results

We discuss now our results about matrix completion and the implications to PCR with missing and noisy entries. Formal definitions and the full setup is deferred to Section 2.

Result about matrix completion. Recall that in matrix completion, we have an unknown rank-r matrix M^* (dimensions $n \times n$), which we wish to infer given noisy and partial observations. $\widetilde{\Omega}$ denotes the set of observed indices, and σ denotes the standard deviation of the noise added to each observation. $\widetilde{\Omega}$ is chosen using a semi-random process (first forming Ω in which every index is present with probability p and then adversarially adding indices).

Theorem (informal). Under appropriate incoherence assumptions on M^*, there exists a polynomial time algorithm that finds an estimate Z that satisfies

$$\|M - Z\|_F \leq O_{\kappa,p,\mu}\left(nr \log n \cdot \sigma\right).$$

The result holds with high probability, and the formal statement along with conditions for p, σ is presented in Theorem 1. The parameters κ, μ capture the condition number and incoherence properties of M^*. To the best of our knowledge, such a bound relating the noise and recovery error is not previously known in the semi-random model for matrix completion. The prior work of (Cheng and Ge, 2018) does not consider the noisy case.

Our recovery error bound is a factor $\sqrt{n}$ worse than the best (and optimal) bounds for matrix completion with *random* observations (Chen et al., 2019; Keshavan et al., 2012). However, it is better than some of the earlier results of (Candes and Plan, 2009). Specifically, the work of (Candes and Plan, 2009) imposes the constraint $\|\mathcal{P}_\Omega(Z - M)\|_F \leq \delta$ and shows that $\|Z - M^*\|_F \leq \sqrt{\frac{n}{p}}\delta$ (up to constants). If Ω corresponds to i.i.d. observations with probability p, we must set $\delta = n\sqrt{p} \cdot \sigma$ for the SDP to be feasible. This leads to a recovery error bound of $n^{3/2}\sigma$. In the semi-random model, the SDP feasibility constraint must now depend on $\widetilde{\Omega}$ and if we have $\gg n^2 p$ observations,

the $n^{3/2}\sigma$ bound becomes even worse. Chen et al. (2019) improve the error bound above by a factor of $n\sqrt{p}$ in the random case. Our result can be viewed as being in between the two results while holding true even in the semirandom model. An interesting open problem is thus to close the $\sqrt{n}$ gap between the random and semirandom settings.

Our main proof ingredient is a semidefinite programming (SDP) relaxation that we can write down given $\widetilde{\Omega}$, but whose analysis can be conducted using tools from the random case. In particular, we use a technique of Chen et al. (2019) in which the solution to a non-convex program is used to derive a dual certificate.

Result about PCR. Using our algorithm for matrix completion as a subroutine lets us obtain new guarantees for PCR under noisy and partial observations. Under appropriate assumptions on the true covariate matrix M^*, we can relax the i.i.d. assumption on the observations to semi-random observations, while still obtaining bounds similar to those of Agarwal et al. (2019).

Theorem (informal). Under appropriate incoherence assumptions on M^*, there exists an efficient algorithm that, given a noisy and partially observed covariate matrix, outputs $\widehat{\beta}$ whose mean-squared-error (defined as $\frac{1}{n}\|M^*\widehat{\beta} - M^*\beta^*\|_2^2$ for the optimal coefficients β^*; see Section 2.2) is at most

$$O(\text{“optimal MSE”}) + O_{\kappa,\mu,p}\left(\|\beta^*\|_2^2 r^2 n \log n \cdot \sigma^2\right).$$

The optimal MSE is a term that turns out to be unavoidable (with high probability) even if we knew M^*. We also note that our bound is in general incomparable with that of Agarwal et al. (2019), but it avoids additive-error terms. Specifically, when $\sigma = 0$, our bounds truly converge to the optimal MSE, even if the matrix is only partially observed (which is not true for the HSVT-based algorithm of Agarwal et al. (2019)).

Experimental results. In our experiments, we discuss the efficacy of our techniques along multiple axes. First, we focus on matrix completion and (a) evaluate the robustness of our SDP formulation to adding more observations (i.e., compare the random setting to the semi-random one) and (b) compare the recovery error in matrix completion with that of HSVT. Next, for matrix completion, we compare our SDP with the *best* SDP when the observations are purely random. Here, we observe that our SDP does slightly worse. This indicates a “price to pay” for being robust; it also suggests that the loss of $\sqrt{n}$ compared to the best-known bounds in the random case may not be an artefact of our analysis. Finally, we compare our SDP-based algorithm with HSVT in the PCR context and show significant improvement both with and without revealing additional entries.

1.2 Related work

There has been extensive literature on PCR and matrix completion that we do not review for lack of space. For PCR, we refer to the early work of Jolliffe (1986), the recent work Agarwal et al. (2019) and references therein. We highlight that unlike classic regression, where the goal is to find β and error is measured in terms of the distance $\|\widehat{\beta} - \beta^*\|$, in PCR we implicitly consider situations where there are multiple good β^*s, and the focus is on the mean-squared error (see Section 2.2).

The HSVT algorithm is a classic method for matrix completion, and Chatterjee (2015) presents near-tight results for it. The early works on matrix completion such as Candes and Plan (2009); Keshavan et al. (2012) obtain weaker dependencies between the noise (per observation) and the recovery error, even with random observations. To the best of our knowledge, the recent work of Cheng and Ge (2018) is the only one to develop guarantees for matrix completion in the semi-random model (albeit without error). There is a large body of work on such models for other problems such as graph partitioning and community detection (see Feige and Krauthgamer (2000); Makarychev et al. (2012) and references therein).

2 Notation and preliminaries

In what follows, all the matrices that we consider will be square ($n \times n$). Our results immediately apply to rectangular matrices, as long as we use the maximum of the two dimensions. The rank parameter r is assumed to be $\ll n$.

The matrix M^* will denote the unknown (or ground truth) of rank r. We assume that the low-rank decomposition is $M^* = U^*\Sigma^*(V^*)^T$, where $U^*, V^* \in \mathbb{R}^{n\times r}$. The decomposition is said to satisfy the μ-incoherence property if for all $i \in [n]$, we have $\|U_i^*\| \le \sqrt{\frac{\mu r}{n}}$. The incohrence property ensures that the mass of the matrix (as well as “information” about the decomposition) is well spread.

High probability. All our theorems hold with high probability, and by this we mean probability over the randomness in the support Ω (which gets appended to form $\widetilde{\Omega}$ in the semirandom model), as well as the randomness in the noise terms, which are distributed as independent Gaussians.

2.1 Model for semi-random matrix completion

Let M^* be the ground truth matrix with rank r. We assume that the low-rank decomposition of $M^* =$

$U^*\Sigma^*(V^*)^T$ satisfies the μ-incoherence property defined above. Further, we let σ_i denote the ith largest singular value of M^*, and for convenience, denote $\sigma_{\min} = \sigma_r$. The condition number $\kappa := \frac{\sigma_1}{\sigma_{\min}}$.

Finally, what we are given are partial observations from a matrix $M := M^* + E$ where $E_{ij} \sim \mathcal{N}(0, \sigma^2)$. The set of observed entries is denoted by $\widetilde{\Omega}$. These are generated via a *semi-random model* where (a) $\Omega \subset [n] \times [n]$ is chosen selecting each index uniformly at random with probability p, and (b) an adversary adds an arbitrary number of indices to Ω to yield $\widetilde{\Omega}$. Crucially, the adversary *sees the matrices* M and M^* before deciding the entries to reveal.

2.2 The PCR Model

Let M^* be the ground-truth matrix of covariates. Let y be the response variables where each y_i is linearly associated with $M^*_{i,\cdot}$, i.e.,

$$y_i = M^*_{i,\cdot}\beta^* + \epsilon_i \tag{3}$$

where β^* is the unknown model parameter and the ϵ_i is an independent noise distributed as $\mathcal{N}(0, \gamma^2)$. We are given a noisy, partially observed version of M^*. Specifically, let $M = M^* + E$, where E is an error matrix whose entries are again independent and distributed as $\mathcal{N}(0, \sigma^2)$. To this matrix, a *mask* is applied according to the semi-random model discussed earlier. $\mathcal{P}_{\widetilde{\Omega}}(M)$ is the matrix that is given as input.

The objective in PCR is to produce a $\widehat{\beta}$ so as to minimize the recovery error:

$$MSE(\widehat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\left(M^*_{i,\cdot}\beta^*_i - M^*_{i,\cdot}\widehat{\beta}_i\right)^2. \tag{4}$$

In other words, the combination $\widehat{\beta}$ that we produce must lead to a good approximation of the "ideal" (noise-free) observations. Even if we were *given* the rank r matrix M^*, the recovery error via least-squares is $\Theta\left(\frac{\gamma^2 r}{n}\right)$ (see, e.g., Agarwal et al. (2019)), and thus this is the "best possible" error we can achieve.

3 Robust semi-random matrix completion

We will use the model and notation as discussed in Section 2.1. While many of the known SDP relaxations (e.g., Candes and Plan (2009); Chen et al. (2019)) involve terms such as $\|\mathcal{P}_\Omega(M - Z)\|_F$ in the constraints, using an analogous constraint with $\widetilde{\Omega}$ makes it impossible to carry over the results, as they rely crucially on the randomness of Ω. Our starting point is thus a different convex optimization problem:

$$\begin{aligned} \textsc{Sdp}(\delta): \quad & \min \|Z\|_* \qquad \text{subject to} \\ & |Z_{ij} - M_{ij}| \le \delta \quad \forall\ (i,j) \in \widetilde{\Omega}, \end{aligned} \tag{5}$$

where $\delta = 4\sigma\sqrt{\log n}$. The choice of δ ensures that M^* is feasible. Our key observation (as also shown in our experiments) is that this SDP is quite stable to the addition of observations (which translate to additional constraints in the matrix).

Parameter choices and assumptions. In what follows, we will set $\lambda = \delta n\sqrt{p}$, and the lemmas will assume that

$$\sigma \le c\,\frac{\sigma_{\min}\log n}{n^{3/2}} \tag{6}$$

$$n^2 p \ge C\kappa^4\mu^2 r^2 n \log^3 n, \tag{7}$$

where c is a sufficiently small and C a sufficiently large constant.

Our main technical result bounds the error between the solution to $\textsc{Sdp}(\delta)$ and the unknown low-rank matrix M^*.

Theorem 1. *(Error bound for* Sdp*) Suppose* $\delta = 4\sigma\sqrt{\log n}$. *Let* μ *be the incoherence parameter and let* σ *be the variance of the noise. Suppose each* M_{ij} *is observed with probability at least* p. *Then for a sufficiently large constant* c, *the solution* Z *to optimization problem (5) satisfies*

$$\|Z - M^*\|_F \le c \cdot \frac{\kappa^3 r^3 \mu^2 n\sqrt{\log n}}{p} \cdot \sigma \tag{8}$$

We note that the bound does not depend on the number of additional element revealed by the adversary. It is also a factor of $\sqrt{n}$ weaker than the optimal recovery bounds (Keshavan et al., 2012; Chen et al., 2019). Our proof of the theorem goes via an elegant argument used in Chen et al. (2019). In the random support setting (where we are given Ω) they show that the solution to the unconstrained SDP:

$$\min \quad \frac{1}{2}\|\mathcal{P}_\Omega(Z - M)\|_F^2 + \lambda\|Z\|_* \tag{9}$$

is closely related to the solution of an appropriate non-convex optimization problem:

$$\min_{X,Y \in \mathbb{R}^{n\times r}} \quad \frac{1}{2}\|\mathcal{P}_\Omega(XY^T - M)\|_F^2 + \lambda\mathcal{R}(X,Y), \tag{10}$$

where $\mathcal{R}(X,Y)$ is an appropriate regularization term. In both the problems above, λ is a parameter (which may be viewed as an appropriate Lagrange multiplier), that is set to roughly $\sigma\sqrt{np}$ for their result.

Unfortunately, the strong connection between the two optimization problems seems to fail to hold in the semi-random setting, i.e., when Ω in the formulations is replaced by $\widetilde{\Omega}$. The analysis of Chen et al. (2019) that the non-convex optimum is close to M^* also strongly relies on the randomness of Ω. However, the *existence* of $X, Y \in \mathbb{R}^{n\times r}$ with appropriate structural properties turns out to be very useful for us. We now state the lemmas from Chen et al. (2019) that we will use.

Lemma 1 (Properties from Chen et al. (2019))**.** *There exist matrices $X, Y \in \mathbb{R}^{n\times r}$ (specifically the output of an appropriate gradient descent algorithm for the non-convex problem above) with the following properties. In what follows, T refers to the tangent space defined as*

$$T = \{XW^T + W'Y^T \ : \ W, W' \in \mathbb{R}^{n\times r}\}.$$

$\mathcal{P}_T$ and $\mathcal{P}_{T^\perp}$ refer to the projections to T and $T^\perp$ respectively.

1. *(Claim 2 of Chen et al. (2019)). Let $XY^T = U\Sigma V^T$ be the SVD. Then*

$$\mathcal{P}_\Omega(XY^T - M) = -\lambda UV^T + R \tag{11}$$

where R is a residual matrix that satisfies

$$\|\mathcal{P}_T(R)\|_F \le \frac{72\kappa\lambda}{n^5} \text{ and } \|\mathcal{P}_{T^\perp}(R)\| \le \lambda/2,$$

2. *(Strong injectivity) The projection $\mathcal{P}_\Omega$ satisfies:*

$$\frac{1}{p}\|\mathcal{P}_\Omega(H)\|_F^2 \ge \frac{1}{32\kappa}\|H\|_F^2 \quad \forall H \in T \tag{12}$$

3. *(Lemma 5 of Chen et al. (2019)) XY^T is close to the ground truth M^* in the following sense:*

$$\begin{aligned}
&\left\|XY^T - M^*\right\|_F \le C_F\frac{\lambda\kappa^2\mu\sqrt{r}}{p},\\
&\left\|XY^T - M^*\right\|_* \le C_{op}\frac{2r\lambda\kappa}{p},\\
&\left\|\mathcal{P}_\Omega(XY^T - M^*)\right\|_F \le C_\infty\frac{\lambda\sqrt{\mu^3r^3\kappa^5}}{\sqrt{p}},
\end{aligned} \tag{13}$$

where C_F, C_{op}, C_∞ are appropriate constants.

Remarks. Roughly, Lemma 1 shows the existence of a low-rank matrix XY^T that has *nicer* properties than even the ground-truth $X^*(Y^*)^T$, while being close to it in different norms. We also note that parts (1-3) of Lemma 1 are shown in Chen et al. (2019) for a slightly different value of λ. In the supplementary material, Section B, we verify that all the claims also hold for our choice of λ. This lemma is where the incoherence property plays a crucial role, especially in part 2 (Equation 12).

We now use this lemma to show our main result. The proof is inspired by duality-based arguments in many prior works (including Candes and Plan (2009); Chen et al. (2019)).

3.1 Proof of Theorem 1

Let Z denote the optimum solution to the optimization problem (5). Our goal will be to prove that Z is close to XY^T. By part (3) of Lemma 1, it will follow that Z is also close to XY^T, by the triangle inequality.

To this end, define $Z = XY^T + \Delta$. In order to bound $\|\Delta\|_F$, we first write

$$\|\Delta\|_F \le \|\mathcal{P}_T\Delta\|_F + \|\mathcal{P}_{T^\perp}\Delta\|_F, \tag{14}$$

and bound the two terms. The first step is to relate the first term on the RHS with the second.

$$\begin{aligned}
\|\mathcal{P}_\Omega(\Delta)\|_F &= \|\mathcal{P}_\Omega\mathcal{P}_T(\Delta) + \mathcal{P}_\Omega\mathcal{P}_{T^\perp}(\Delta)\|_F\\
&\ge \|\mathcal{P}_\Omega\mathcal{P}_T(\Delta)\|_F - \|\mathcal{P}_\Omega\mathcal{P}_{T^\perp}(\Delta)\|_F\\
&\ge \sqrt{\frac{p}{32\kappa}}\|\mathcal{P}_T(\Delta)\|_F - \|\mathcal{P}_{T^\perp}(\Delta)\|_F
\end{aligned} \tag{15}$$

In the last step, we used the strong injectivity property from Lemma 1. This implies that

$$\|\mathcal{P}_T(\Delta)\|_F \le \sqrt{\frac{32\kappa}{p}}\left(\|\mathcal{P}_{T^\perp}(\Delta)\|_F + \|\mathcal{P}_\Omega(\Delta)\|_F\right) \tag{16}$$

The term $\|\mathcal{P}_\Omega(\Delta)\|_F$ turns out to be easy to bound, using the properties of the SDP:

$$\begin{aligned}
\|\mathcal{P}_\Omega(\Delta)\|_F &= \left\|\mathcal{P}_\Omega(Z - XY^T)\right\|_F\\
&\le \|\mathcal{P}_\Omega(Z - M^*)\|_F + \left\|\mathcal{P}_\Omega(M^* - XY^T)\right\|_F\\
&\le 2\lambda + C_\infty\frac{\lambda\sqrt{\mu^3r^3\kappa^5}}{\sqrt{p}}.
\end{aligned} \tag{17}$$

The last inequality is due to the fact that M^* is a feasible solution to the SDP (5) (which implies that both Z and M^* are within distance λ from M), combined with the last inequality in (13).

Combining (14), (15) and (17), it follows that we only need to bound $\|\mathcal{P}_{T^\perp}(\Delta)\|_F$. We will indeed show a stronger bound, on the quantity $\|\mathcal{P}_{T^\perp}(\Delta)\|_*$, which is always $\ge \|\mathcal{P}_{T^\perp}(\Delta)\|_F$.

The bulk of the argument is thus in bounding $\|\mathcal{P}_{T^\perp}(\Delta)\|_*$. The key claim is the following, relating this term to the matrices U, V (defined using the SVD $XY^T = U\Sigma V^T$).

Claim 1. We have

$$\|\mathcal{P}_{T^\perp}(\Delta)\|_* \leq \|XY^T + \Delta\|_* - \|XY^T\|_* - \langle UV^T, \Delta\rangle.$$

Proof of Claim 1. By definition, for any $W \in T^\perp$ with $\|W\| \leq 1$, we have that $UV^T + W$ is in the subgradient of $\|\cdot\|_*$ at XY^T. Thus for any such W, we have

$$\left\|XY^T + \Delta\right\|_* \geq \left\|XY^T\right\|_* + \langle UV^T + W, \Delta\rangle.$$

The observation is that we can pick W such that $\langle W, \Delta\rangle = \|\mathcal{P}_{T^\perp}(\Delta)\|_*$. This is well-known, and follows by choosing W using the SVD directions of $\mathcal{P}_{T^\perp}(\Delta)$ (and this will lie entirely in T). Rearranging now implies the claim. □

The next claim shows the following bound.

Claim 2. $\|XY^T + \Delta\|_* - \|XY^T\|_* \leq C_{op}\frac{2r\lambda\kappa}{p}$.

Proof of Claim 2. Since $Z = XY^T + \Delta$ is the optimum solution to the SDP (5) and M^* is another feasible solution, we have (using also (13) from Lemma 1),

$$\begin{aligned}\|Z\|_* &\leq \|M^*\|_* \\ &\leq \left\|M^* - XY^T\right\|_* + \left\|XY^T\right\|_* \\ &\leq C_{op}\frac{2r\lambda\kappa}{p} + \left\|XY^T\right\|_*\end{aligned} \tag{18}$$

Rearranging now implies the claim. □

Using the two claims, it follows that our goal should be to prove that $\langle UV^T, \Delta\rangle$ is not too negative. This is done somewhat indirectly.

Claim 3. From our choice of λ, we have:

$$\|\mathcal{P}_\Omega(XY^T + \Delta - M)\|_F^2 \leq \|\mathcal{P}_\Omega(XY^T - M)\|_F^2. \tag{19}$$

Proof of Claim 3. Because $Z = XY^T + \Delta$ is feasible for our SDP, we have that the LHS of (19) is $\leq \lambda^2$, by our choice of λ.

Next, using the property of the non-convex solution (part 1 of Lemma 1), we have $\mathcal{P}_\Omega(XY^T - M) = -\lambda UV^T + R$, for some $R \in T^\perp$. This implies that $\|\mathcal{P}_\Omega(XY^T - M)\|_F^2 \geq r\lambda^2 \geq \lambda^2$. Thus the claim follows. □

We can now expand Eq. (19) to obtain

$$\frac{1}{2}\|\mathcal{P}_\Omega(\Delta)\|_F^2 + \langle \mathcal{P}_\Omega(XY^T - M), \Delta\rangle \leq 0.$$

Plugging in part 1 of Lemma 1 again, we get

$$\langle \lambda UV^T - R, \Delta\rangle \geq \frac{1}{2}\|\mathcal{P}_\Omega(\Delta)\|_F^2 \geq 0.$$

This implies that $\langle UV^T, \Delta\rangle \geq \frac{1}{\lambda}\langle R, \Delta\rangle$.

Now the key observation is that

$$\begin{aligned}|\langle R, \Delta\rangle| &\leq |\langle \mathcal{P}_{T^\perp}R, \mathcal{P}_{T^\perp}\Delta\rangle| + |\langle \mathcal{P}_T R, \mathcal{P}_T\Delta\rangle| \\ &\leq \frac{\lambda}{2}\|\mathcal{P}_{T^\perp}\Delta\|_* + \frac{72\kappa\lambda}{n^5}\|\mathcal{P}_T\Delta\|_*.\end{aligned} \tag{20}$$

This immediately implies that

$$-\langle UV^T, \Delta\rangle \leq \frac{1}{2}\|\mathcal{P}_{T^\perp}\Delta\|_* + \frac{72\kappa}{n^5}\|\mathcal{P}_T\Delta\|_*.$$

Plugging this into Claim 1 and using Claim 2, we have that

$$\begin{aligned}&\|\mathcal{P}_{T^\perp}(\Delta)\|_* \leq C_{op}\frac{2r\lambda\kappa}{p} + \frac{1}{2}\|\mathcal{P}_{T^\perp}\Delta\|_* + \frac{72\kappa}{n^5}\|\mathcal{P}_T\Delta\|_* \\ &\implies \|\mathcal{P}_{T^\perp}(\Delta)\|_* \leq C_{op}\frac{4r\lambda\kappa}{p} + \frac{144\kappa}{n^5}\|\mathcal{P}_T\Delta\|_*.\end{aligned} \tag{21}$$

This is where we crucially used the factor of $1/2$ from part 1 of Lemma 1.

Plugging this back into (16) and using (17) along with the fact that $\|\cdot\|_F \leq \|\cdot\|_* < n\|\cdot\|_F$, we obtain:

$$\begin{aligned}\|\mathcal{P}_T(\Delta)\|_F \leq \sqrt{\frac{32\kappa}{p}}\Bigg(&C_{op}\frac{4r\lambda\kappa}{p} + \frac{144\kappa}{n^5}\|\mathcal{P}_T\Delta\|_* \\ &+ 2\lambda + C_\infty\frac{\lambda\sqrt{\mu^3 r^3\kappa^5}}{\sqrt{p}}\Bigg).\end{aligned}$$

Simplifying, we get

$$\begin{aligned}&\|\mathcal{P}_T(\Delta)\|_F\left(1 - \frac{900\kappa^{3/2}}{\sqrt{p}n^4}\right) \\ &\leq \frac{16r\lambda\kappa^{3/2}}{p}\left(\frac{2C_{op}}{\sqrt{p}} + C_\infty\sqrt{\kappa^3\mu^3 r^3}\right).\end{aligned}$$

Using the fact that n is large enough, the coefficient on the LHS is $> 1/2$, we get

$$\|\mathcal{P}_T(\Delta)\|_F \leq \frac{32r\lambda\kappa^{3/2}}{p}\left(\frac{2C_{op}}{\sqrt{p}} + C_\infty\sqrt{\kappa^3\mu^3 r^3}\right). \tag{22}$$

Combining equations (22) and (21) with (14), the theorem follows.

4 Principal Component Regression

We now present the application of our results on matrix completion to the PCR problem. Recall that we have an unknown rank r covariate matrix M^*, and we are given a perturbed, partially observed version of M^*, which we will denote by $\mathcal{P}_{\widetilde{\Omega}}(M)$ (see Section 2.2). Also, as before, Ω denotes the *random part* of the observed indices, and $\widetilde{\Omega}$ is the given set of indices, where

Ω is appended with some (adversarially chosen) subset of indices.

Recall also that our goal is to find a $\widehat{\beta}$ whose MSE $\frac{1}{n}\|M^*\beta^* - M^*\widehat{\beta}\|$ is minimized.

Approach of Agarwal et al. (2019) The work of Agarwal et al. (2019) studies a natural approach for PCR, when we are given $\mathcal{P}_\Omega(M)$ (and not the projection to $\widetilde{\Omega}$). The idea is to first "complete" and denoise this matrix, thus obtaining a low-rank estimate $\hat{M}$ for M^*, and then using ordinary least squares with $\hat{M}, y$ to obtain $\widehat{\beta}$.

They then prove that the MSE of this $\widehat{\beta}$ can be bounded in terms of an appropriate function of the error $(\hat{M} - M^*)$ in estimating M^*. To obtain $\hat{M}$, they prove that simply using the best rank-r approximation of the matrix $\frac{1}{p}\mathcal{P}_\Omega(M)$ suffices. This procedure is referred to as hard singular-value thresholding, or HSVT.

The semi-random setting. The issue with HSVT is that it crucially relies on $\frac{1}{p}\mathcal{P}_\Omega(M)$ being an "unbiased estimator" of M, and since the error $M - M^*$ is i.i.d. Gaussian, also of M^*. The guarantee for HSVT would completely fail to hold if $\widetilde{\Omega}$, for instance, is imbalanced so that entries in some columns have a probability $> p$ of being revealed. We will demonstrate this also in our experiments (Section 5). Our main idea is thus to replace the HSVT with the noisy matrix completion subroutine developed in Section 3.

We start by presenting our overall algorithm for PCR. The algorithm assumes knowledge of n, and more crucially, of r and σ. (As discussed in Agarwal et al. (2019), when $r \ll n$, dividing the input randomly and performing cross validation is a general way to overcome this issue.)

Algorithm 1 PCR via Matrix Completion

Input: A matrix $\mathcal{P}_{\widetilde{\Omega}}(M)$ with missing entries and noise with variance σ^2.
Output: A feature weight vector $\widehat{\beta}$

1: Solve the SDP 5 (using σ) and get the optimal solution Z
2: Define $Z^{(r)} \leftarrow$ rank-r approximation of Z (obtained via SVD)
3: Carry out ordinary least squares using $Z^{(r)}$ and the given y, return the obtained $\widehat{\beta}$

We show the following result.

Theorem 2. *Suppose the ground truth matrix M^* has rank r and has a decomposition that satisfies the properties described in Section 2. Suppose also that the noise parameter σ satisfies the conditions of Theorem 1. Finally, suppose the values y satisfy the following, for some unknown β^*:*

$$Y_i = M^*_{i,\cdot}\beta^* + \phi_i + \epsilon_i,$$

where ϕ_i is a model mismatch parameter. Then with high probability, the output $\widehat{\beta}$ of Algorithm 1 satisfies

$$MSE(\widehat{\beta}) \le \frac{4\gamma^2 r}{n} + \frac{20\,\|\phi\|_2^2}{n} + \frac{3\,\|\beta^*\|_2^2}{n}\left(\frac{4C^2\kappa^6 r^6 \mu^4 \sigma^2 n^2 \log n}{p^2}\right).$$

where γ is the variance of the noise in the regression model and the rest of the parameters are defined as in previous sections.

Remark. In our proof below as well as in previous works on robust PCR, we recover the incomplete matrix before training the regression model rather than directly approximating the target $M\beta$. The reason for this is partly that the top r components of M must be estimated either directly or indirectly, and bounding the error seems to require bounding the error in this estimation.

4.1 Proof of Theorem 2

We show that a slight modification of the proof of Agarwal et al. (2019), combined with our argument from Section 3 yields the theorem.

At a high level, the argument in Agarwal et al. (2019) proceeds by proving an upper bound on the MSE in terms of $\|\beta^*\|$ and the error $\|\hat{M} - M^*\|$, where $\hat{M}$ is the *completed* matrix. However, the proof uses the fact that $\hat{M}$ is also of rank r. This is the reason we use $Z^{(r)}$ instead of the SDP solution Z itself in Algorithm 1.

A simple lemma shows that $\|Z^{(r)} - M^*\|$ is also small.

Lemma 2. *Let $Z^{(r)}$ be the best rank-r approximation of Z (where $r = rank(M^*)$), as computed in the algorithm. Then with high probability, we have*

$$\left\|Z^{(r)} - M^*\right\|_F \le 2C \cdot \frac{\kappa^3 r^3 \mu^2 n \sigma \sqrt{\log n}}{p}. \tag{23}$$

The proof is based on using the matrix XY^T from Section 3 to show that Z has a good low-rank approximation. The proof is deferred to Section A of the supplement.

We can thus prove our main result of the section.

Proof of Theorem 2. We invoke the proof of Theorem 6 in Agarwal et al. (2019). This relies on the fact that

the estimate $\hat{M}$ is rank r (denoted $\widehat{A}$ in the paper; it is the result of $HSVT$ on the normalized matrix). In our case, we can carry out the same arguments with $Z^{(r)}$ instead of $\widehat{A}$. This is used to obtain: (see Eq. (29) of Agarwal et al. (2019))

$$nMSE(\widehat{\beta}) \leq 4\gamma^2 r + 3\|(Z^{(r)} - M^*)\beta^*\|_2^2 + 20\|\phi\|_2^2.$$

To obtain the desired bound from this, we can use Hölder's inequality, which yields:

$$\left\|\left(Z^{(r)} - M^*\right)\beta^*\right\|_2^2 \leq \|\beta^*\|_2^2 \left\|Z^{(r)} - M^*\right\|_F^2$$

This along with the bound above gives us

$$\frac{1}{n}\left\|Z^{(r)}\widehat{\beta} - M^*\beta^*\right\|_2^2 \leq \frac{4\gamma^2 r}{n} + \frac{20\,\|\phi\|_2^2}{n} + \frac{3\,\|\beta^*\|_2^2}{n}\left\|Z^{(r)} - M^*\right\|_F^2$$

Combining this with Lemma 2 completes the proof of the theorem. □

5 Experiments

In this section we empirically evaluate our results on matrix completion and the application to PCR using synthetic datasets. Additional experiments are deferred to section A of the supplement.

5.1 Matrix completion

We first compare the error rates of the solutions to SDPs (5) (elementwise constrained) and (9) (unconstrained) with uniformly random observations with varying noise levels (measured in variance σ). The error is measured as $\|Z - M^*\|_F/\|M^*\|_F$ where Z is the solution to each SDP and M^* is the ground truth matrix. Here we consider a ground truth matrix with $n = 100$ and $r = 5$ where each singular value is 1(figure 1). Covariates are sampled with probability $p = 0.4$ and results are averaged over 5 iterations. λ in SDP (9) is set to $5\sigma\sqrt{np}$.

When the covariates are observed uniformly at random, as seen in the figure 1, solving the SDP (5) gives slightly worse error compared to the SDP from (9), which is known to give the tight bounds.

In the second experiment, we compare the error rates of the SDP (5) and HSVT with random and semirandom observations. Similar to the previous experiment, error is measured in Frobenius norm scaled by $1/\|M^*\|_F$ with varying noise levels. We consider two ground truth matrices, with ranks 1 and 5. Each covariate is first observed with probability 0.4 then first 90 rows of first 30 columns are opened later as additional observations. The results are averaged over 5 iterations (figure 2).

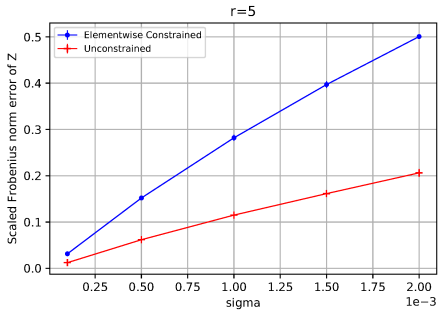

Figure 1: Scaled Frobenius norm error of recovered matrices when observations are uniformly random.

Error rates of HSVT are higher in general than those obtainable via the SDP. Further, the gap between the random and semi-random cases is much worse in the case of HSVT. This suggests that the solutions to elementwise constrained SDP (5) remain stable under semirandom perturbation.

5.2 PCR using matrix completion vs. PCR with HSVT

We compare the performance our algorithm with HSVT in principle component regression with covariates observed with random and semi-random manner. Similar to the experiments in the previous section we generate ground truth covariate matrices with $n = 100$ and $r = 1, 5$ (M^*). We generate a regression coefficient vector β^* of size n. In rank 1 case each covariate is observed with probability $p = 0.4$ at first then first 90 rows of first 30 columns are opened later as additional observations. In rank 5 case each covariate is observed with probability $p = 0.5$ at first then every row of first 10 columns is opened later as additional observations. We evaluate the regression error as in equation 4 for the recovered matrices using our algorithm and HSVT with varying noise levels (figure 3). The results are average over 5 iterations.

Regression error rates using HSVT imitate the error rates displayed in matrix completion with the same technique while replacing the HSVT matrix by the solution to elementwise constrained SDP (5) maintains stable regression error rates.

6 Conclusion

We investigate the principal component regression problem with noisy and missing entries, where the set of observed entries is not i.i.d., but comes from a semirandom model. In this case, existing algorithms based on spectral methods fail, and our contribution is a semidefinite programming (SDP) based algorithm

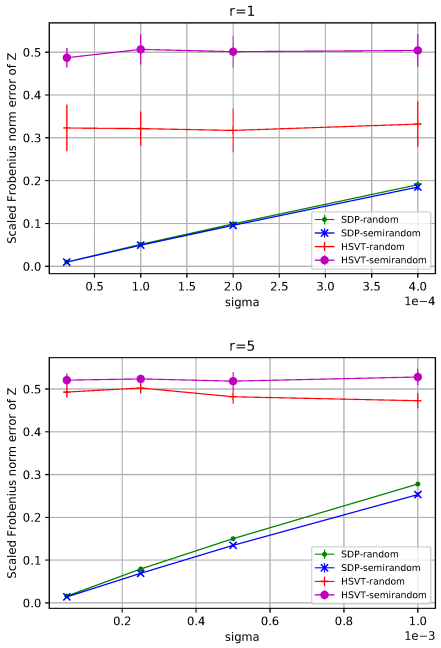

Figure 2: Comparison of scaled Frobenius norm error of recovered matrices when observations are random and semirandom.

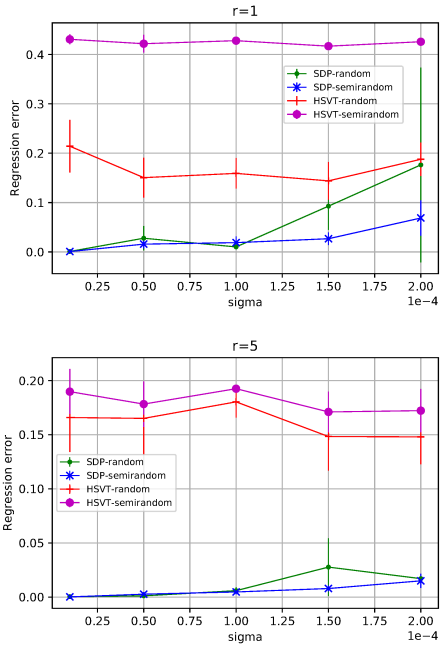

Figure 3: Comparison of regression error when covariates are observed in random and semirandom manner.

that is robust to such cases. The key technical contribution in this paper is an analysis of matrix completion using an elementwise constrained SDP and we develop the first recovery bounds under noisy and semirandom observations. We complement our results with experiments demonstrating the pros and cons of our approach.

References

Agarwal, A., Shah, D., Shen, D., and Song, D. (2019). On robustness of principal component regression.

Bhojanapalli, S. and Jain, P. (2014). Universal matrix completion. In Xing, E. P. and Jebara, T., editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1881–1889, Bejing, China. PMLR.

Blum, A. and Spencer, J. (1995). Coloring random and semi-random k-colorable graphs. *J. Algorithms*, 19(2):204–234.

Candes, E. J. and Plan, Y. (2009). Matrix completion with noise.

Candes, E. J. and Recht, B. (2008). Exact matrix completion via convex optimization.

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

Chen, Y., Chi, Y., Fan, J., Ma, C., and Yan, Y. (2019). Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization.

Cheng, Y. and Ge, R. (2018). Non-convex matrix completion against a semi-random adversary. In Bubeck, S., Perchet, V., and Rigollet, P., editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1362–1394. PMLR.

Diakonikolas, I., Gouleakis, T., and Tzamos, C. (2019). Distribution-independent pac learning of halfspaces with massart noise. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Draper, N. and Smith, H. (1966). *Applied regression analysis*. Wiley series in probability and mathematical statistics. Wiley, New York [u.a.].

Feige, U. and Krauthgamer, R. (2000). Finding and certifying a large hidden clique in a semirandom graph. *Random Structures & Algorithms*, 16(2):195–208.

Gunst, R. and Webster, J. (1975). Regression analysis and problems of multicollinearity. *Communications in Statistics*, 4(3):277–292.

Hardt, M., Meka, R., Raghavendra, P., and Weitz,

B. (2014). Computational limits for matrix completion. In Balcan, M. F., Feldman, V., and Szepesvári, C., editors, *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 703–725, Barcelona, Spain. PMLR.

Hocking, R. R. (1972). Criteria for selection of a subset regression: Which one should be used? *Technometrics*, 14(4):967–976.

Hotelling, H. (1957). The relations of the newer multivariate statistical methods to factor analysis. *British Journal of Statistical Psychology*, 10(2):69–79.

Jolliffe, I. (1986). *Principal Component Analysis*. Springer Verlag.

Keshavan, R. H., Montanari, A., and Oh, S. (2012). Matrix completion from noisy entries.

Little, R. J. A. (1992). Regression with missing x's: A review. *Journal of the American Statistical Association*, 87(420):1227–1237.

Makarychev, K., Makarychev, Y., and Vijayaraghavan, A. (2012). Approximation algorithms for semirandom graph partitioning problems.

Mosteller, F. and Tukey, J. W. (1977). *Data Analysis and Regression: a Second Course in Statistics*. Addison-Wesley Publishing Company.