Principal Component Regression with Semirandom Observations

Aditya Bhaskara Aravinda Kanchana Ruwanpathirana Maheshakya Wijewardena

School of Computing, University of Utah

Problem Statement

Principal Component Regression (PCR)

- **Regression:** given covariate matrix M and observations y, find β s.t. $M\beta = y$
- Multi-collinearity problem: linearly dependent $M \implies$ ill-conditioned system
- **Idea behind PCR:** project M to top k principal components (obtaining M_k) and solve system with M_k

Intuition: PCR handles multi-collinearity by reducing the number of predictors to a smaller number of uncorrelated ones.

Semirandom Observation Model

- **Setting:** partially observed covariates. Each covariate is observed with probability $\geq p$ for some known p.
- Can also be viewed as a two-step model: (a) every covariate is revealed with prob (exactly) p, obtaining observations Ω (b) adversary reveals additional entries, obtaining $\widetilde{\Omega}$.
- **Note:** Similar in spirit to Massart noise in classification problems.

Why is the semirandom model more challenging?

- Semirandom observation model is natural model in applications like recommender systems, where (e.g.) some certain users may review more products than others.
- In spite of seeming easier, semirandom model causes spectral methods to fail.
- **Problem:** without step (b), can obtain unbiased estimator of M by re-weighting observed entries. Adversary revealing new entries makes it impossible to obtain unbiased estimator of M.

Our Contributions

- Introduce a new semidefinite programming relaxation for **noisy** matrix completion – every observed entry has $\mathcal{N}(0, \sigma^2)$ added.

$$\textsc{Sdp}(\delta): \quad \min \|Z\|_* \quad \text{subject to} \quad |Z_{ij} - M_{ij}| \leq \delta \quad \forall (i,j) \in \widetilde{\Omega}, \tag{1}$$

- "Per entry" constraint with $\delta = \widetilde{O}(\sigma)$. Main technical result: providing theoretical guarantees for the recovery error for matrix completion.
- Implies new theoretical guarantees for Principal Component Regression using the low rank completion.

Assumptions

- Sufficiently large sampling complexity,i.e. $np \geq C\kappa^4\mu^2 r^2 \log^3 n$
- Sufficiently low noise,i.e. $\sigma \leq c\, \frac{\sigma_{\min} \log n}{n^{3/2}}$

Our results

Algorithm

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

Theoretical Guarantees

Matrix Completion:
Under appropriate incoherence assumptions on M^* (the covariate matrix without noise), there exists a polynomial time algorithm that finds an estimate Z s.t.

$$\|M - Z\|_F \leq O_{\kappa,p,\mu}\left(nr^3\sqrt{\log n} \cdot \sigma\right).$$

Principal Component Regression:

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^*) is at most

$$O(\text{“optimal MSE”}) + O_{\kappa,\mu,p}\left(\|\beta^*\|_2^2 r^6 n \log n \cdot \sigma^2\right).$$

Idea of the proof

- Our analysis goes through drawing a connection between the solution of the SDP 1 and the factorization derived for the non-convex optimization problem discussed in [1].
- We can show that the claims required for our analysis, made in [1] holds true even with the relaxed regularization parameter given necessary conditions on noise and the sample complexity.
- We can use the properties of the factorization derived following the aforementioned conditions and the proof ideas used in [2] to derive stronger bounds that are roughly $O(\sqrt{np})$ better than the bounds derived in [2].

Experiments

Matrix Recovery Error

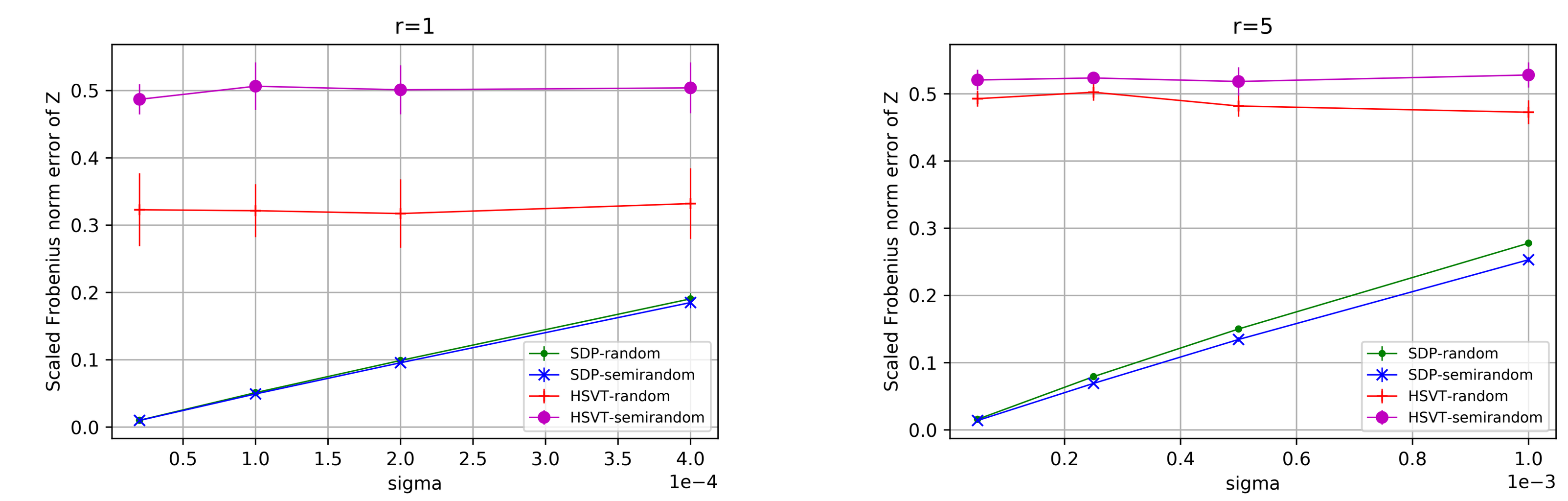

Figure: Comparison of scaled Frobenius norm error of recovered matrices when observations are random and semirandom.

Regression Error

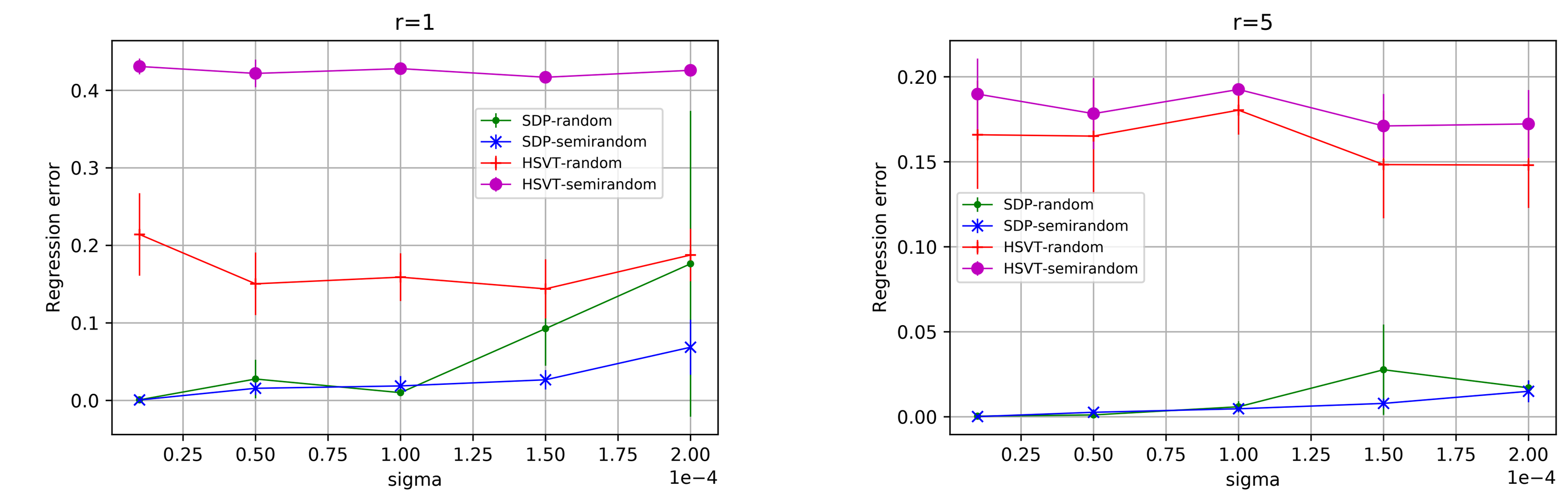

Figure: Comparison of regression error when covariates are observed in random and semirandom manner.

Recap

We were able to:

- Provide the first recovery guarantee for matrix completion under semi-random observations.
- Obtain a more robust algorithm for PCR via solving the SDP 1.
- Avoid additive error terms in the regression error (in contrast to [3]) and provide a bound that converges to the optimal error in the absence of noise.

References

[1] Yuxin Chen et al. *Noisy Matrix Completion: Understanding Statistical Guarantees for Convex Relaxation via Nonconvex Optimization.* 2019. arXiv: 1902.07698 [stat.ML].

[2] Emmanuel J. Candes and Yaniv Plan. *Matrix Completion With Noise.* 2009. arXiv: 0903.3131 [cs.IT].

[3] Anish Agarwal et al. *On Robustness of Principal Component Regression.* 2019. arXiv: 1902.10920 [cs.LG].