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 $\beta$ 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 $p$ for some known $p$.
- Can also be viewed as a two-step model: (a) every covariate is revealed with probability $p$, obtaining observations $\Omega$; (b) adversary reveals additional entries, obtaining $\tilde{\Omega}$.
- 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.

$$\text{Stop}(\delta) \cdot \min \|Z\|_1, \text{ subject to } |Z_{ij} - M_{ij}| \leq \delta \forall (i, j) \in \tilde{\Omega},$$

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

Assumptions
- Sufficiently large sampling complexity i.e. $np \geq Ck^2n^2 \log n$
- Sufficiently low noise i.e. $\sigma \leq \epsilon \frac{\log n}{\sqrt{n}}$

Our results

Algorithm

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

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.

$$\|M - Z\|_F \leq O_{\epsilon, \kappa, \rho, \sigma}(n^2 \sqrt{\log n} \cdot \sigma).$$

Principal Component Regression:
Under appropriate incoherence assumptions on $M^*$, there exists an efficient algorithm that, given a noisy and partially observed covariate matrix, outputs $\hat{\beta}$ whose mean-squared error (defined as $\|M\hat{\beta} - M^{*}\beta^*\|$) for the optimal coefficients $\beta^*$ is at most

$$O^2(\text{optimal MSE}) + O_{\epsilon, \kappa, \rho, \sigma}(\|\beta^*\|_F^2 n \log n \cdot \sigma).$$

Results

Recall

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