On Robustness of Principal Component Regression

Anish Agarwal

Devavrat Shah, Dennis Shen, Dogyoon Song

MIT

1 What is PCR?

1 What is PCR?

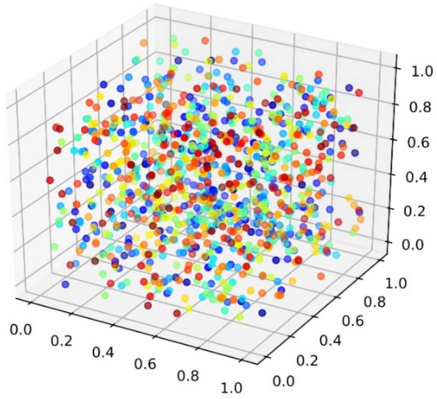
$$\boxed{Y} = \boxed{A} \boxed{\beta} + \boxed{\epsilon}$$

1 What is PCR?

$$Y = A \beta + \epsilon$$

Step 1: PCA

$$A$$



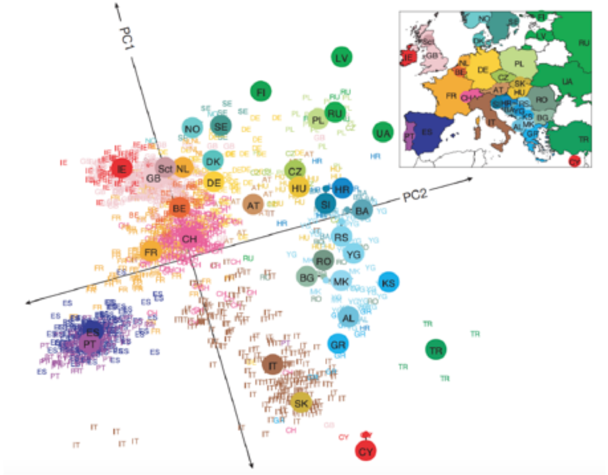
1 What is PCR?

$$Y = A \beta + \epsilon$$

Step 1: PCA

(k-components)

$$A_k$$



1 What is PCR?

$$Y = \boldsymbol{A}_k \beta + \epsilon$$

Step 2: Regression

$$\beta_{\boldsymbol{k}} = \underset{\boldsymbol{\theta}}{\text{minimize}} \left\| Y - \boldsymbol{A}_k \boldsymbol{\theta} \right\|_2^2$$

1 What is PCR?

$$Y = \boldsymbol{A}_k \beta_k + \epsilon$$

Step 3: Prediction

$$\widehat{Y} = \boldsymbol{A}_k \beta_k$$

2 When & Why Use PCR

2 Data Science Folklore

"IF DATA IS (APPROXIMATELY) LOW-DIMENSIONAL, USE PCR!"

-- Anonymous Data Scientists

When exactly should we be using PCR?

2 Key Questions We Answer

Theoretical properties of PCR?

Is dimension-reduction only benefit to PCR?

Our Theoretical Analysis of PCR helps answer following questions..

How low-rank do covariates need to be?

How many principal components to pick?

How well does PCR perform on a test data (i.e. generalization properties)?

Is Dimension-Reduction Only Benefit?

NO!

2 PCR (as is) works for a wide variety of settings!

Noisy

Missing

Mixed valued

Sensitive

We show PCR is surprisingly robust to problems that plague large-scale modern datasets

Main Contribution of this Work

Error-In Variable Regression

(Setting We Consider)

2 Classical (high-dimensional) Regression

$$\boldsymbol{Y} = \boldsymbol{A}\,\beta + \epsilon$$

2 Error-in-Variable (EIV) Regression

$$\boldsymbol{Y} = \boldsymbol{Z}\beta + \epsilon$$

Representative of modern datasets

Noisy

Missing

Mixed valued

Sensitive

2 EIV - Surprising Number of Applications

Time Series Analysis (measurement noise)

Causal Inference (Synthetic Control) (measurement noise)

Differentially-private Regression (noise by design)

Mixed Valued Regression (structural noise)

2 EIV - Surprising Number of Applications

Time Series Analysis (measurement noise)

Causal Inference (Synthetic Control) (measurement noise)

Differentially-private Regression (noise by design)

Mixed Valued Regression (structural noise)

Formal Results

2 Theorem (Informal): Training Error

If principal components chosen correctly ($k = r$)

PCR implicitly denoises covariates!

number of covariates

$$\frac{1}{n}\mathbb{E}\left[\|\boldsymbol{Z}_{\boldsymbol{r}}\beta_{\boldsymbol{r}} - \boldsymbol{A}\beta\|_2^2\right] \sim \frac{\boldsymbol{\sigma}^2\boldsymbol{r}}{\boldsymbol{n}}\frac{\mathrm{poly}(\log p)}{\rho^4}$$

fraction of observations

OLS minmax error rate

(low-dimensional, noiseless, fully observed covariates)

2 Theorem (Informal): Testing Error

If principal components **not** chosen correctly $(k \neq r)$

$$\text{Test Error} \sim \underbrace{\text{Train Error with PCR}(k) + C\frac{k^{3/2}}{\sqrt{n}}}_{\text{Choose } k \text{ that minimizes above}}$$

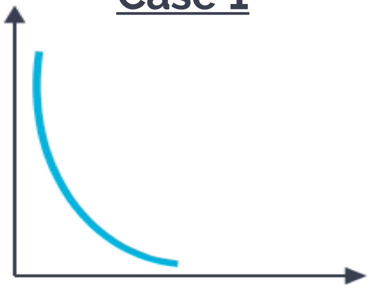
PCR implicitly performs l_0-regularization

PCR implicitly de-noises covariates

2 When To and Not to Use PCR? – Look at Spectrum

Use PCR!

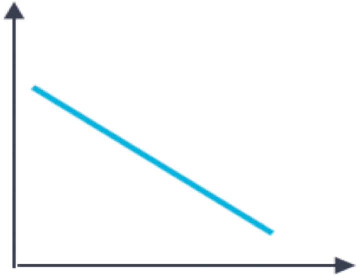
Case 1



Case 2

Don't Use PCR!

Case 3

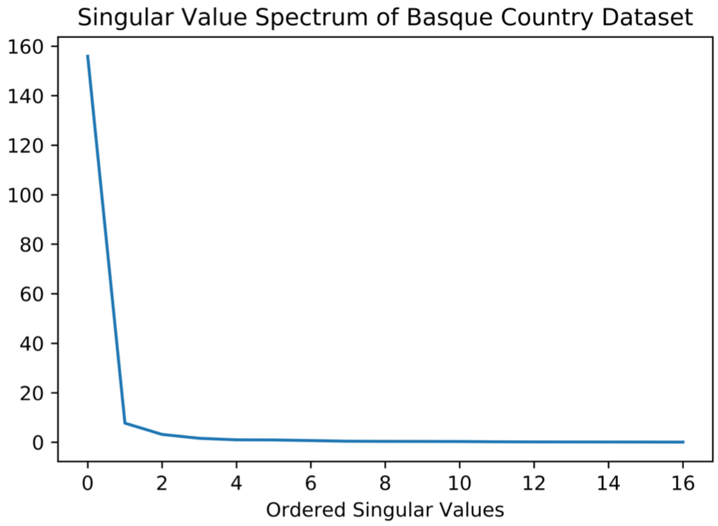


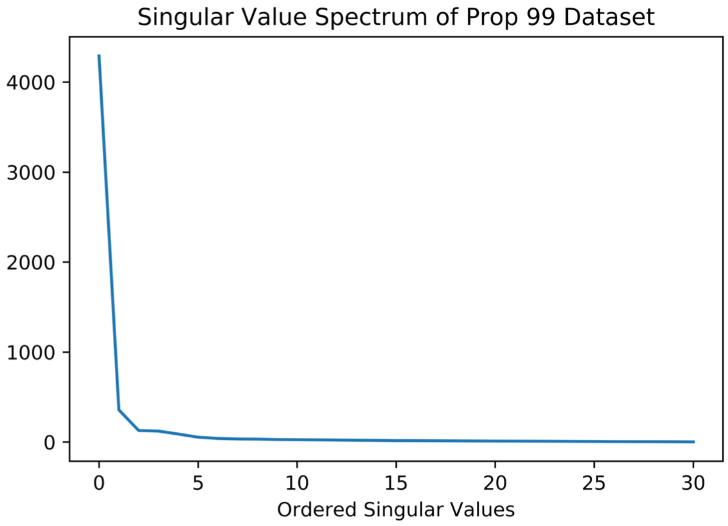
Case 4

2

Exponential-decaying spectrum is ubiquitous in real-world data

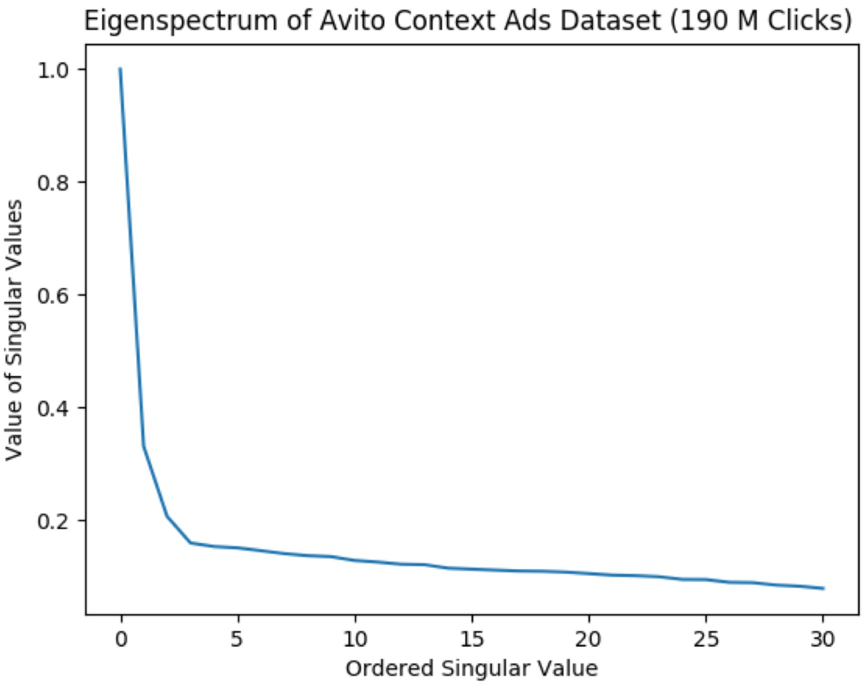
GDP Trajectories (Macroeconomics)





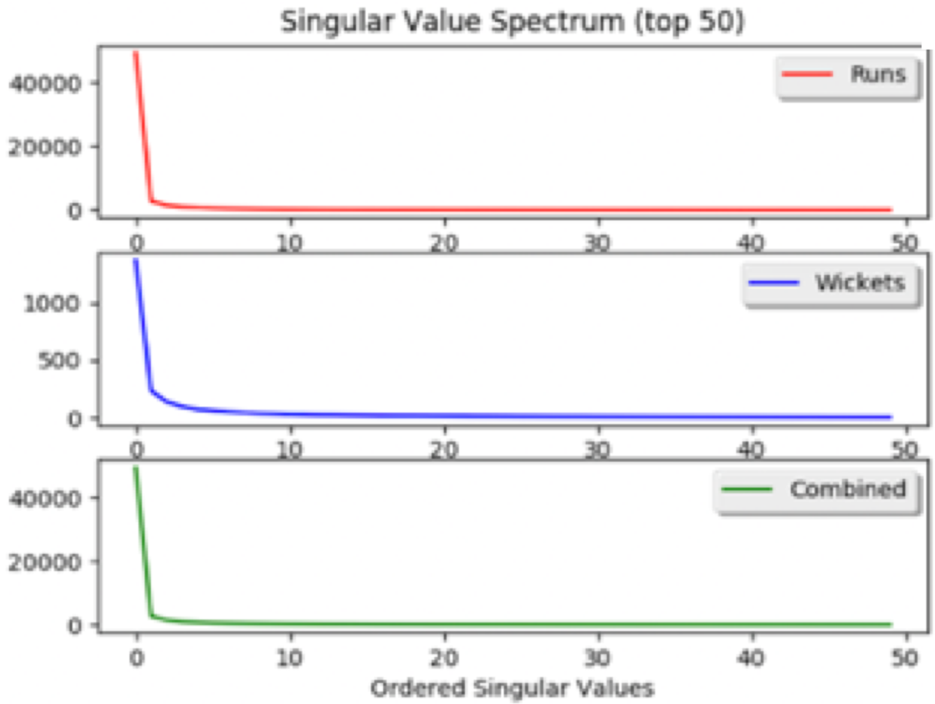
2 Exponential-decaying spectrum is ubiquitous in real-world data

Avito Ad-Click Dataset (E-Commerce)



2 Exponential-decaying spectrum is ubiquitous in real-world data

Cricket Trajectories (Sports)



3 Surprising Applications of PCR

3 Applications of Error-In-Variable Regression

Time Series Analysis (measurement noise)

Causal Inference (Synthetic Control) (measurement noise)

Differentially-private Regression (noise by design)

Mixed Valued Regression (structural noise)

Data privacy is top-of-mind as we increasingly apply ML on sensitive user data (genetic data, purchase history etc.)

Standard Notion of Privacy in ML

ε-Differential Privacy

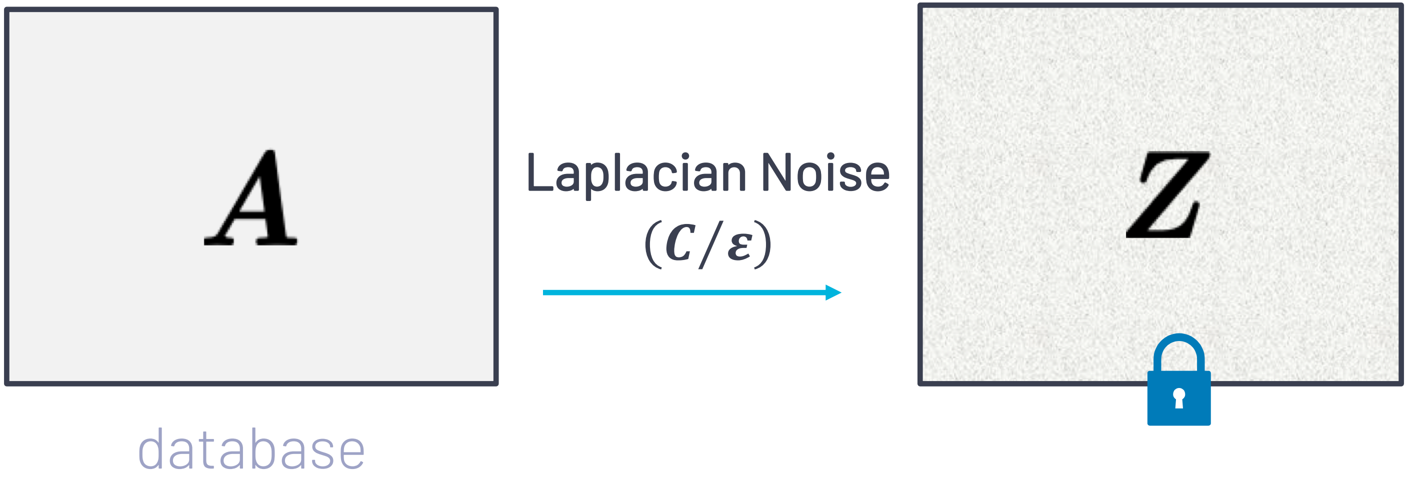
Intuitively, an algorithm is **ε**-differentially private if **outcome of a statistical query** on a database **cannot change by more than ε** due to **presence/absence of any user** data record

Example of Statistical Query:

"Average Income of all users between ages 25 and 30"

How to achieve ε-differentially privacy?

Laplace Mechanism



Predictive Accuracy vs. Privacy Tradeoff

Can we achieve good prediction error and still maintain privacy?

$$Y = \boldsymbol{Z}\beta + \epsilon$$

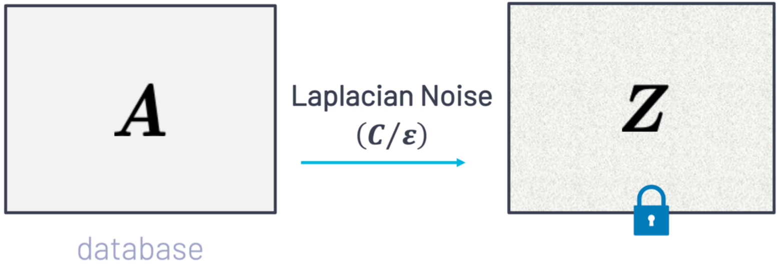
Yes!

Predictive Accuracy vs. Privacy Tradeoff

Can we achieve good prediction error and still maintain privacy?

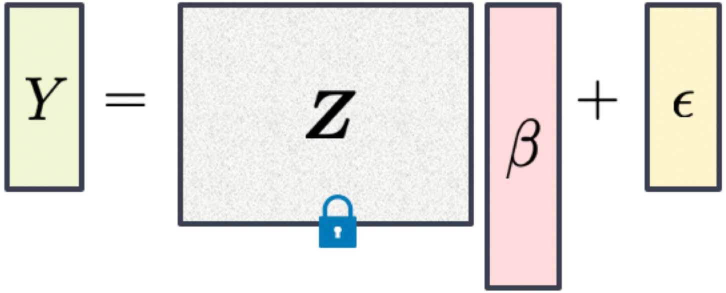
Step 1:

Data Owner adds Laplacian Noise



Step 2:

Analyst Performs PCR



Done!

What is sample complexity cost for ε-differential privacy?

$$\text{Prediction Error} \sim \frac{\sigma^2 r}{n} \frac{\text{poly}(\log p)}{\rho^4} \boxed{\left(\frac{1}{\varepsilon^2}\right)}$$

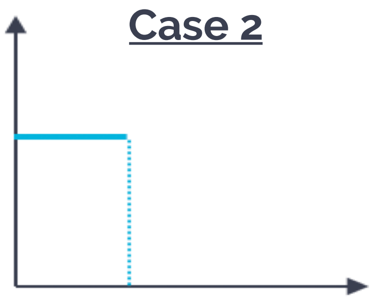
Does de-noising step (PCA) break privacy?

No, PCA only de-noises covariates *on average*

with respect to the $\|\cdot\|_{2,\infty}$ - norm

4 Conclusion

Inspect spectrum of your covariate matrix



Use PCR!

de-noises

regularizes

Possible Implications for Modern ML

Linear Case

Step 1: Dimension Reduction

$$\boldsymbol{A} \xrightarrow{\text{PCA}} \boldsymbol{A}_k$$

Linear low-dimensional covariate pre-processing has many implicit benefits (e.g. de-noising, regularizing)

Non-Linear Case

$$\boldsymbol{A} \xrightarrow{\text{GANs?}}$$

Does non-linear covariate pre-processing (e.g. GANs) have similar benefits for unstructured data?

Come Meet Us At Our Poster

Poster #3 – East Exhibition Hall B + C, 5-7pm, Thursday

Shameless Plug :)

PCR for Time Series Analysis: **tspdb.mit.edu**

PCR for Causal Inference: **github.com/Romcos/SC_demo**