Robust Sparse Principal Component Regression

Fang Han
Johns Hopkins University

Han Liu
Princeton University

Neural Information Processing Systems (NIPS)
Lake Tahoe, Nevada, 2013

Poster number: Fri36

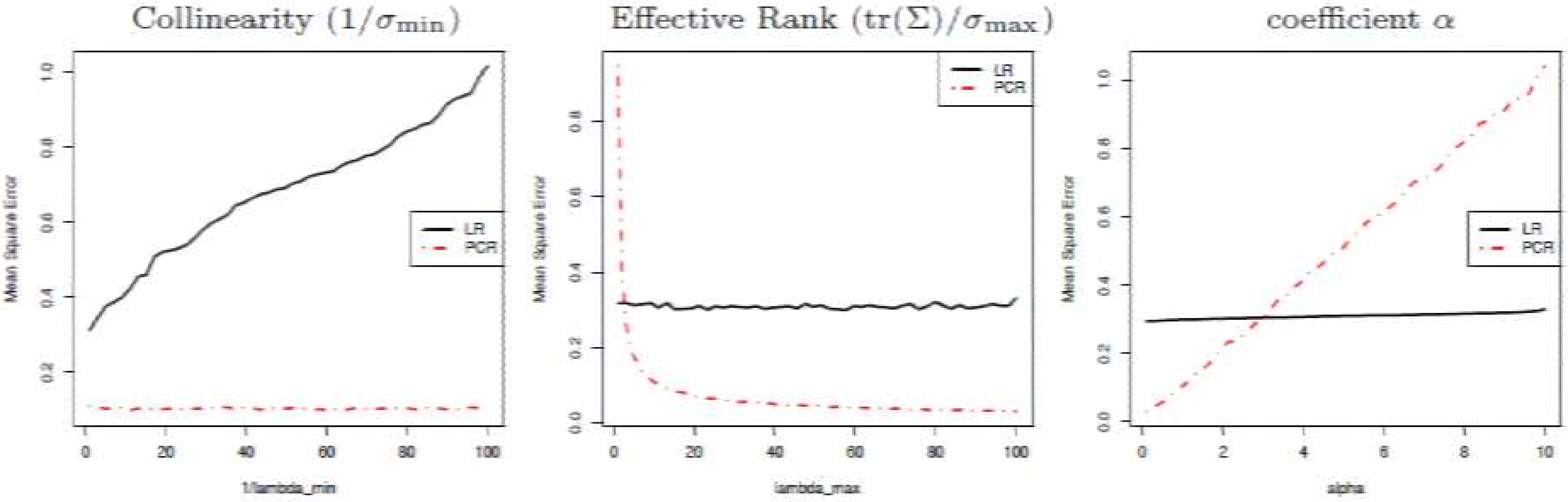
Advantages of Simple PCR over LSE

- Principal Component Regression (PCR) model is a subset of the linear regression model:

$$\text{(Classical linear regression model)} \qquad \boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon};$$

$$\text{(Simple PCR Model)} \qquad \boldsymbol{Y} = \alpha \mathbf{X} \boldsymbol{u}_1 + \boldsymbol{\epsilon}.$$

- Three advantages of Simple PCR over LSE:



Robust Sparse PCR

$$Y_{n\times 1} = \underbrace{\mathbf{X}_{n\times d}}_{\sim(\mathbf{0},\boldsymbol{\Sigma})} \quad \underbrace{\mathbf{U}_{d\times k}}_{\text{eigenspace of } \boldsymbol{\Sigma}} \quad \boldsymbol{\alpha}_{k\times 1} + \boldsymbol{\epsilon}_{n\times 1}, \quad d \gg n.$$

Heavy tailed models on $\mathbf{X}$ and $\boldsymbol{\epsilon}$: $\mathbf{X}$ is elliptically distributed (stretched Gaussian: $\boldsymbol{\mu} + \xi \cdot N(\mathbf{0}, \mathbf{S})$), $\boldsymbol{\epsilon}$ has finite fourth moments.

Sparsity on $\mathbf{U}$ and $\boldsymbol{\alpha}$: model constrained for identifiability: $\|\mathbf{U}\|_0 + \|\boldsymbol{\alpha}\|_0 = s + k \leq n$. Procedure should induce sparsity on $\mathbf{U}$.

Robust Sparse PCR: Estimating $\boldsymbol{\beta} = \mathbf{U}\boldsymbol{\alpha}$ in optimal (parametric) rate, robust to potential outliers.

Robust Sparse PCR

1. Model: Random design matrix, drawn from elliptical distribution. (Light ⇒ heavy tails)
2. Procedure 1: Conduct sparse PCA on the pairwise self-normalized data $(\boldsymbol{x}_i - \boldsymbol{x}_j)/\|\boldsymbol{x}_i - \boldsymbol{x}_j\|$. (Elliptical: self-normalization preserves the eigenspace)
3. Procedure 2: Project the data to the eigenspace learnt in the second step, then conduct LSE to estimate $\boldsymbol{\alpha}$.
4. Theory: $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| = O_P(\sqrt{s \log d / n})$. (optimal rate)