CMSE 820: Mathematical Foundations of Data Science

Lecture 04
PCA denoising

- Data drawn from normal distribution $\mathcal{N}(\Sigma + \epsilon, \mu)$

- $\Sigma \in \mathbb{R}^{p \times p}$ with rank($\Sigma$) = 10

- $\epsilon \in \mathbb{R}^{p \times p}$ with rank($\epsilon$) = $p$, but $||\epsilon|| \ll ||\Sigma||$

- Example with $n = 10000$ and $p = 100$
PCA denoising

Without noise
PCA denoising

With noise
PCA in high dimensions

- \( \mathcal{X}_n = \{x_i\}_{i \leq n} \subset \mathbb{R}^p \)
- \( X_n = [x_1 \cdots x_n] \in \mathbb{R}^{p \times n} \)
- \( x_i \sim \mathcal{N}(0, I) \)
- \( S_n = (1/n)X_nX_n^T \)
- Examine spectral properties of \( S_n \) when \( p, n \to \infty \) and \( p/n = \gamma \leq 1 \)
- In other words, how does PCA perform in high dimensions?
PCA in high dimensions

\[ p = 500, \ n = 1000 \]

Eigenvalues of \( S_n = (1/n)X_nX_n^T \)
PCA in high dimensions

$p = 500, \ n = 1000$

Histogram of eigenvalues of $S_n = (1/n)X_nX_n^T$
Spike model

- $X_n = \{x_i\}_{i \leq n} \subset \mathbb{R}^p$
- $X_n = [x_1 \cdots x_n] \in \mathbb{R}^{p \times n}$
- $x_i \sim \mathcal{N}(0, I + \beta vv^T)$, where $\beta \geq 0$ and $v \in \mathbb{R}^p$
- $S_n = (1/n)X_nX_n^T$
- Examine spectral properties of $S_n$ when $p, n \to \infty$ and $p/n = \gamma \leq 1$
- Can PCA find the 1-dimensional structure of the data with variance $1 + \beta$ and along the direction $v$?
Spike model

$p = 500, \ n = 1000, \ \beta = 2$

Histogram of eigenvalues of $S_n = (1/n)X_nX_n^T$
$p = 500$, $n = 1000$, $\beta = 1/2$

Histogram of eigenvalues of $S_n = (1/n)X_nX_n^T$