Sparse PCA in High Dimensions

Jing Lei, Department of Statistics, Carnegie Mellon

Workshop on Big Data and Differential Privacy
Simons Institute, Dec, 2013

(Based on joint work with V. Q. Vu, J. Cho, and K. Rohe)
Overview

• Sparse PCA and subspace estimation.
• A convex relaxation.
• Consistency and sparsistency.
• Sparse PCA with differential privacy.
Principal Components Analysis

• I have iid data points $X_1, \ldots, X_n$ on $p$ variables.

• $p$ may be large, so I want to use principal components analysis (PCA) for dimension reduction.
Principal Components Analysis

- $\Sigma = \mathbb{E}(XX^T)$ is the population covariance matrix (say $\mathbb{E}X = 0$).
- Eigen-decomposition

$$\Sigma = \lambda_1 v_1 v_1^T + \lambda_2 v_2 v_2^T + \ldots + \lambda_p v_p v_p^T$$

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0,$$ (eigenvalues)

$$v_i^T v_j = \delta_{ij},$$ (eigenvectors)

- “Optimal” $d$-dimensional projection: $X \rightarrow \Pi_d X$

$$\Pi_d = V_d V_d^T, \quad V_d = (v_1, v_2, \ldots, v_d).$$
Classical Estimator

- Sample covariance matrix: \( \hat{\Sigma} = n^{-1}(X_1X_1^T + \ldots + X_nX_n^T) \).
- Estimate \((\hat{\lambda}_j, \hat{\nu}_j)\) by eigen-decomposition of \(\hat{\Sigma}\).
  \(\hat{V}_d = (\hat{\nu}_1, \ldots, \hat{\nu}_d)\), \(\hat{\Pi}_d = \hat{V}_d \hat{V}_d^T\).
- These are consistent and asymptotically normal when \(p\) is fixed and \(n \to \infty\).
High-Dimensional PCA: Challenges

- When $\frac{p}{n} \rightarrow c \in (0, \infty]$, PCA can be inconsistent (Johnstone & Lu 09), and/or hard to interpret.
- **Sparse PCA** offers dimension reduction with better statistical properties and interpretability.
Subspace Sparsity [Vu & L 2013]

- **Identifiability.** If $\lambda_1 = \lambda_2 = \ldots = \lambda_d$, then one cannot distinguish $V_d$ and $V_dQ$ from observed data for any orthogonal $Q$.
- **Intuition:** a good notion of sparsity must be rotation invariant.
- **Row sparsity:**

  At most $s$ rows of $\Pi_d$ (and hence $V_d$) are non-zero. $s \ll p$.

- **Interpretation:** the projection involves at most $s$ variables.
The Sparse PCA Model

\[
\Sigma = \begin{pmatrix} \overbrace{UDU^T}^{s} & \overbrace{0}^{p-s} \\
0 & 0 \end{pmatrix} + \begin{pmatrix} \Gamma_1 & \Gamma_{12} \\
\Gamma_{21} & \Gamma_2 \end{pmatrix}, \quad \Pi_d = \begin{pmatrix} UU^T & 0 \\
0 & 0 \end{pmatrix}
\]

- "signal" = \(\lambda_1 v_1 v_1^T + \ldots + \lambda_d v_d v_d^T\).
- "noise" = \(\lambda_{d+1} v_{d+1} v_{d+1}^T + \ldots + \lambda_p v_p v_p^T\).
- \(U \in \mathbb{R}^{s \times d}\) is the non-zero block of \(V_d\).
- \(D = \text{diag}(\lambda_1, \ldots, \lambda_d)\).
- This decomposition is unique when \(\lambda_d > \lambda_{d+1}\).
**Sparsity Reduces the Error Rate**

**Theorem: (Vu & L 2013)**

Under the sparse PCA model, the optimal error rate of estimating $\Pi_d$ is

$$\|\hat{\Pi}_d - \Pi_d\|_F^2 \approx s \frac{\lambda_1 \lambda_{d+1}}{(\lambda_d - \lambda_{d+1})^2} \frac{d + \log p}{n},$$

and can be achieved by

$$\hat{\Pi}_d = \arg \max_{\Pi} \text{Tr}(\hat{\Sigma} \Pi),$$

where the maximization is over all $s$-sparse $d$-dimensional projection matrices.
Proof of Upper Bound

- Curvature Lemma

\[(\lambda_d - \lambda_{d+1}) \| \hat{\Pi}_d - \Pi_d \|_F^2 \leq 2 \text{Tr}(\Sigma (\Pi_d - \hat{\Pi}_d))\]

- \(\hat{\Pi}_d\) optimizes the objective function.

\[0 \leq \text{Tr}(\hat{\Sigma}(\hat{\Pi}_d - \Pi_d))\]

- Combine the above two.

\[\| \hat{\Pi}_d - \Pi_d \|_F^2 \leq \frac{2}{\lambda_d - \lambda_{d+1}} \text{Tr} \left[ (\hat{\Sigma} - \Sigma)(\hat{\Pi}_d - \Pi_d) \right]\]

- Empirical process ...
This theorem gives optimal dependence on
\((n, p, s, d, \lambda_1, \lambda_d, \lambda_{d+1})\).

No additional structural assumptions on \(\Gamma\) (a popular assumption \(\Gamma = \sigma^2 I\) is known as the spiked covariance model).

But the proposed minimax optimal estimator is NP-hard to compute.

Convex relaxation?
Fantope Projection and Selection (FPS) [VCLR13]

\[
\max_{Z} \quad \text{Tr}(\hat{\Sigma}Z) - \rho \|Z\|_1, \\
\quad \text{s.t.} \quad 0 \preceq Z \preceq I, \quad \text{Tr}(Z) = d.
\]

The constraint set \( \mathcal{F}_{p,d} = \{Z : 0 \preceq Z \preceq I, \text{Tr}(Z) = d\} \) is called the Fantope (Fillmore & Williams 71, Dattorro 05), named after Ky Fan.

FPS can be solved efficiently using alternating direction method of multipliers (ADMM).
\[ \ell_2 \text{ Error Bound for FPS} \]

**Theorem: FPS Error Bound [VCLLR 2013]**

Under the PCA model with \( s \)-sparsity on \( \Pi_d \), if (for \( C \) large enough)

\[
\rho = C \sqrt{\frac{p}{n}},
\]

the global optimizer \( \hat{Z} \) of FPS satisfies (w.h.p)

\[
\| \hat{Z} - \Pi_d \|^2_F \lesssim s^2 \frac{\lambda_1 \lambda_{d+1}}{(\lambda_d - \lambda_{d+1})^2} \frac{\log p}{n}.
\]

Roughly, this has an extra factor of \( s \) (compare to minimax rate), which may be unavoidable for polynomial time algorithms [BR13].
Proof

Curvature Lemma extends to the Fantope!
Same trick as before (use $\rho \geq \|\hat{\Sigma} - \Sigma\|_\infty$)

$$\frac{\lambda_d - \lambda_{d+1}}{2} \|\hat{Z} - \Pi_d\|^2_F \lesssim \text{Tr} [(\hat{\Sigma} - \Sigma)(\hat{Z} - \Pi_d)] - \rho (\|\hat{Z}\|_1 - \|\Pi_d\|_1)$$

$$\leq \rho \|\hat{Z} - \Pi_d\|_1 - \rho (\|\hat{Z}\|_1 - \|\Pi_d\|_1)$$

Then apply triangle inequality and Cauchy-Schwartz.
Do no need empirical process.
Variable Selection

- Can we estimate the set of relevant variables in $\Pi_d$?
- The case of $d = 1$ is analyzed by Amini & Wainwright (2009).
- We are able to
  1. remove a common assumption $\Gamma_{21} = 0$ (zero correlation between relevant and irrelevant variables);
  2. extend to $d > 1$. 
Variable Selection Consistency of FPS

**Theorem: (L & Vu 2013)**

FPS correctly selects the relevant variables with high probability, if

\[
n \gtrsim s^2 \log p, \quad \text{(sample complexity)}
\]

\[
\| \Gamma_{21}(j,:) \| \lesssim s^{-1}, \quad \forall j, \quad \text{(incoherence)}
\]

\[
\min_{1 \leq j \leq s} \Pi_{jj} \gtrsim s \sqrt{\frac{\log p}{n}}, \quad \text{(signal strength)}
\]

\[
\rho = C \sqrt{\frac{\log p}{n}}. \quad \text{(tuning parameter)}
\]

**Remarks**

- The information-theoretic lower bound is \( n \gtrsim s \log p \) [AW09].
- The omitted constants depend on the eigenvalues of \( \Sigma \).
Key Ingredients of Proof

Also only needs $||\hat{\Sigma} - \Sigma||_\infty$ to be small.

- Strong duality and KKT.
- Curvature lemma.
- Linear algebra, perturbation theory.
The analysis of FPS only needs $\hat{\Sigma}$ to satisfy entry-wise accuracy):

\[
\max_{jk} |\hat{\Sigma}_{jk} - \Sigma_{jk}| = O_P \left( \sqrt{\frac{\log p}{n}} \right).
\]

Proof: Bernstein + union bound.

The results for FPS still hold if we add entry-wise perturbations to $\hat{\Sigma}$, on the order of $\sqrt{\log p/n}$. \\

\textit{FPS with Differential Privacy}
Method 0: Laplace Noise

- **Goal**: d.p. release of $\hat{\Sigma}$, with entry-wise accuracy $\sqrt{\log p/n}$.
- Assume $\mathbb{E}X = 0$, $|X_{ij}| \leq 1$.
- Naive idea: adding entry-wise independent double exponential noise.
- The entry-wise noise is of order $p^2/n$. 
Method 1: Counting Queries

- **Goal**: d.p. release of $\hat{\Sigma}$, with entry-wise accuracy $\sqrt{\log p/n}$.
- Assume $\mathbb{E}X = 0$, $|X_{ij}| \leq 1$.
- Observation: each entry of $\hat{\Sigma}$ is a sample average (counting query).
- The method of [Hardt, Ligett, & McSherry 12] reduces the entry-wise error to $O\left(\sqrt{\frac{\log p}{n\epsilon}} (\log p \log \frac{1}{\delta})^{1/4}\right)$ for $(\epsilon, \delta)$-d.p.
Method 2: Stability Test

- **Perturbation stability**: for a given $\rho$ (a good one), how many data points need to be modified in order to obtain a different variable selection result?

- Applied to the LASSO in [Smith & Thakurta 13]. See also [Dwork & L 09].

- Idea: Estimate $\Pi_d$ with d.p. after variable selection.

- **Challenge**: the query is insensitive but may be hard to compute in general.
Method 3: Random Projection

- Let \( X \in \mathbb{R}^{n \times p} \) be the data matrix, then \( \hat{\Sigma} = n^{-1}X^TX \).
- \( \hat{\Sigma}_{ij} \) measures the covariance/correlation between variables \( j \) and \( k \).
- Johnson-Lindenstrauss Transform has been proved to preserve pairwise similarity and d.p. [Kenthapadi, Korolova, Mironov, & Mishra, 12], [Blocki, Blum, Datta, & Sheffet, 12].
- Idea: Use sample covariance of \( Y = RX(\Delta) \), where \( R \) and \( \Delta \) are random matrices (iid normal).
Sparse PCA is an important topic with interesting structure and lots of recent developments.

The statistical analysis of sparse PCA fits well into some existing differential privacy methods.

1. D.p. release of $p^2$ related counting queries in continuous space.
2. Stability test for sparse PCA (and more general settings).
4. D.p. ADMM (?).
Thank You!