# *Estimating Sparse Principal Components and Subspaces*

Jing Lei

*Department of Statistics, CMU*

Joint work with V. Q. Vu (OSU), J. Cho, and K. Rohe (U. of Wisc.)

July 1, 2013

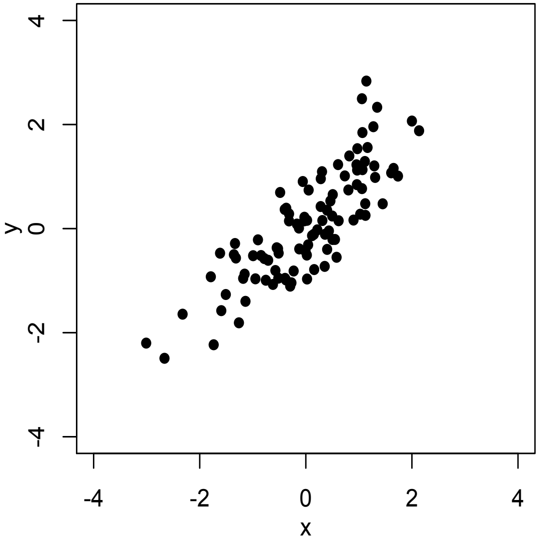
## *Outline*

- PCA in high dimensions.
- Sparsity of principal components.
- Consistent estimation and minimax theory.
- Feasible algorithms using convex relaxation.

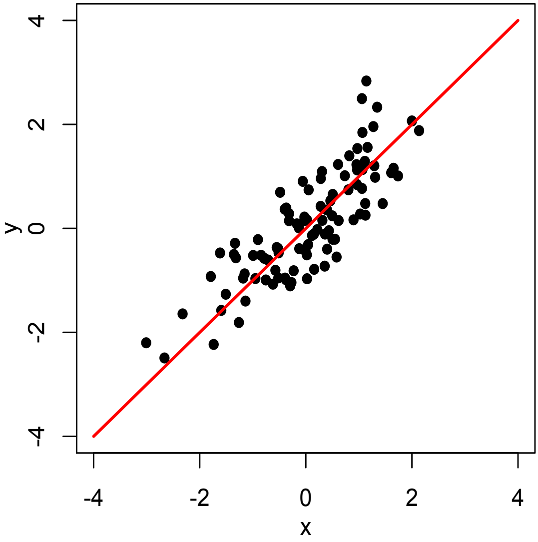
# *Principal Components Analysis*

- 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.

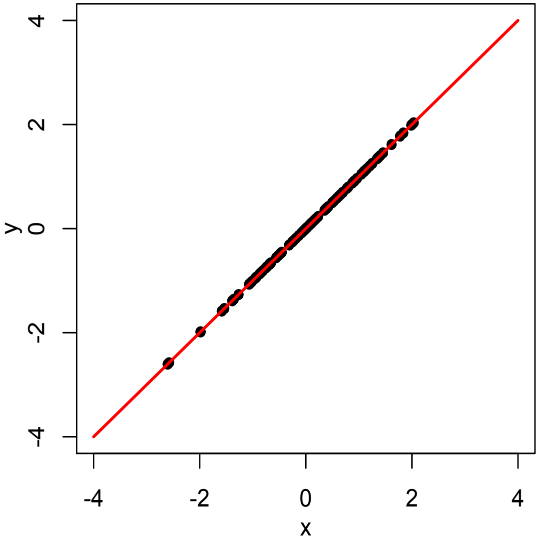
# *Principal Components Analysis*



# *Principal Components Analysis*



# *Principal Components Analysis*



# *Principal Components Analysis*

- $\Sigma = \mathbb{E}(XX^T)$ is the population covariance matrix (say $\mathbb{E}X = 0$).
- Eigen-decomposition
  $\Sigma = VDV^T = \lambda_1 v_1 v_1^T + \lambda_2 v_2 v_2^T + \ldots + \lambda_p v_p v_p^T$
  $D = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ (eigenvalues)
  $VV^T = I_p$, $V = (v_1, v_2, \ldots, v_p)$ (eigenvectors)
- "Optimal" $d$-dimensional projection: $X \to \Pi_d X$
  $\Pi_d = V_d V_d^T$ ($d$-dimensional projection matrix),
  $V_d = (v_1, \ldots, v_d)$.

# *Classical Estimator*

- Sample covariance matrix: $\hat{\Sigma} = n^{-1}(X_1X_1^T + \ldots + X_nX_n^T)$.
- Estimate $(\hat{\lambda}_j, \hat{v}_j)$ by eigen-decomposition of $\hat{\Sigma}$.
  $\hat{V}_d = (\hat{v}_1, \ldots, \hat{v}_d)$, $\hat{\Pi}_d = \hat{V}_d\hat{V}_d^T$.
- Standard theory for $p$ fixed and $n \to \infty$:
  $\hat{\Pi}_d \to \Pi_d$ a.s. if $\lambda_j - \lambda_{j+1} > 0$.

# *High-Dimensional PCA: Challenges*

- Estimation accuracy. Classical theory fails when $p/n \to c > 0$: $\hat{\lambda}_1 \to c' > 1$, and $\hat{v}_1^T v_1 \approx 0$ under a simple model (Johnstone & Lu 2009).
- Interpretability. $\hat{\Pi}_d X$ may be hard to interpret when it involves linear combination of many variables.
- Sparsity is a possible solution.

# *Sparsity for Principal Subspaces [Vu & L 2012b]*

- Identifiability. If $\lambda_1 = \lambda_2 = \ldots = \lambda_d$, then one cannot distinguish $V_d$ and $V_d Q$ from observed data for any orthogonal $Q$.
- Intuition: a good notion of sparsity must be rotation invariant.
- Matrix $(2,0)$ norm: for any matrix $V \in \mathbb{R}^{p \times d}$,
  $\|V\|_{2,0} = \#$ of non-zero rows in $V$
- Row sparsity: $\|V_d\|_{2,0} \leq R_0 \ll p$. $V_d = (v_1, v_2, \ldots, v_d)$.
- Loss function: $\|\hat{\Pi}_d - \Pi_d\|_F^2$ ($\|\cdot\|_F$: the Frobenius norm).
  Recall: $\hat{\Pi}_d = V_d V_d^T$, $\hat{\Pi}_d = \hat{V}_d \hat{V}_d^T$.

# *Two Sparse PCA Models*

1. Spiked model:

$$\Sigma = (\lambda_1 - \lambda_{d+1}) v_1 v_1^T + \ldots + (\lambda_d - \lambda_{d+1}) v_d v_d^T + \lambda_{d+1} I_p \,.$$

2. General model:

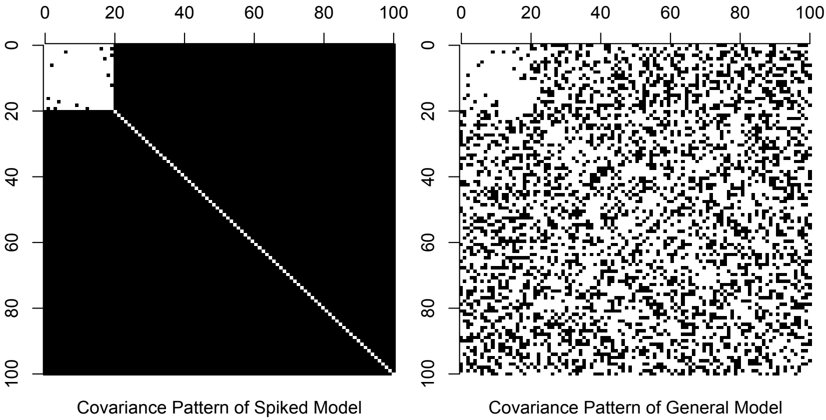
$$\Sigma = \lambda_1 v_1 v_1^T + \ldots + \lambda_d v_d v_d^T + \lambda_{d+1} \Sigma'$$

$$\text{where } \; \Sigma' \succeq 0, \;\; \|\Sigma'\| = 1, \;\; \Sigma' v_j = 0, \; \forall 1 \le j \le d \,.$$

## *Spiked Model is a Special Case of General Model*

Black cell: $|\Sigma(i,j)| \leq 0.01$, White cell: $|\Sigma(i,j)| > 0.01$

In spiked model, all black cells outside the upper $20\times 20$ are 0.



## *How Does Sparsity Help?*

- Question: how does sparsity help with the estimation?
  1. How well can we do if sparsity is assumed?
  2. How to estimate under sparsity assumption?
- Intuition: Estimation is easy if
  1. $n$ is large.
  2. $p$ is small.
  3. $\lambda_{d+1}$ is close to 0.
  4. $\lambda_d - \lambda_{d+1}$ is away from 0.
  5. $R_0$ is small.
- Under the spiked model, [Johnstone & Lu 2009] gives a consistent estimator of $v_1$ when $p/n \to c > 0$, and others fixed.

# *A Minimax Framework*

Find $f(n,p,R_0,\lambda_1,\lambda_2)$ such that

$$\sup_{\Sigma} \mathbb{E}\|\hat{\Pi}_d - \Pi_d\|_F^2 \gtrsim f(n,p,R_0,\vec{\lambda}), \quad \forall \text{ estimator } \hat{\Pi}_d\,,$$

and a particular estimator $\hat{\Pi}_d$ such that

$$\mathbb{E}\|\hat{\Pi}_d - \Pi_d\|_F^2 \lesssim f(n,p,R_0,\vec{\lambda}), \quad \forall\, \Sigma.$$

$\Sigma$ is taken over all matrices in the sparse PCA model.

# *Answer to the Minimax Question*

***Theorem**: Minimax Error Rate of Estimating $V_d$ (Vu and Lei 2012b)*

Under the general model, the minimax rate of estimating $V_d V_d^T$ is

$$f_d(n, p, R_0, \vec{\lambda}) \asymp R_0 \frac{\lambda_1 \lambda_{d+1}}{(\lambda_d - \lambda_{d+1})^2} \frac{d + \log p}{n},$$

and can be achieved by

$$\hat{V}_d = \arg \max_{V_d^T V_d = I_d, \|V_d\|_{2,0} \le R_0} \mathrm{Tr}(V_d^T \hat{\Sigma} V_d).$$

# *About This Result*

- Good news
  - Exact minimax error rate in $(n, p, d, R_0, \vec{\lambda})$ for general models.
  - First consistency result for $\ell_1$ constrained/penalized PCA (Jolliffe et al 2003, Zou et al 2006).
- Price to pay
  - Finding the global maximizer is computationally demanding.
- Extensions
  - Soft sparsity: $\ell_q$-ball with $q \in [0, 1]$ [Vu & L 2012a,b].
  - Feasible algorithms [Vu, Cho, L, Rohe 2013].

# Related Work

- When $d = 1$, [Birnbaum et al 2012, and Ma 2013] established the minimax rate under the spiked model, where the estimator is obtained by power method and thresholding.
- For subspace estimation, the minimax rate is independently obtained by [Cai et al 2012] under a Gaussian spiked model.

# *Feasible Algorithm Via Convex Relaxation*

- For $d = 1$, the optimal estimator (consider $Z = v_1 v_1^T$) is

$$\hat{Z} = \arg\max_{Z} \ \mathrm{Tr}(\hat{\Sigma} Z) - \lambda \|Z\|_0,$$

$$\text{s.t.} \ \ \mathrm{rank}(Z) = 1, \ Z \succeq 0, \ \mathrm{Tr}(Z) = 1.$$

- [d'Aspremont et al 2004] proposed an SDP relaxation

$$\hat{Z} = \arg\max_{Z} \mathrm{Tr}(\hat{\Sigma} Z) - \lambda \|Z\|_1, \ \ \text{s.t.} \ \ Z \succeq 0, \ \mathrm{Tr}(Z) = 1,$$

- $\hat{Z}$ gives consistent variable selection with optimal rate under a stringent spiked model, provided that $\hat{Z}$ is rank 1 [Amini & Wainwright 2009].

# *Preliminary Results for SDP Relaxation*

***Theorem**: Error Bound for SDP Relaxation [VCLR 2013]*

When $d = 1$ under the general model, assume $\|v_1\|_0 \le R_0$ and choose $\lambda \asymp \frac{\lambda_1}{\lambda_1 - \lambda_2}\sqrt{\log p/n}$ in the SDP relaxation. Then w.h.p the global optimizer $\hat{Z}$ satisfies

$$\|\hat{Z} - v_1 v_1^T\|_2^2 \lesssim R_0^2 \frac{\lambda_1^2}{(\lambda_1 - \lambda_2)^2} \frac{\log p}{n} .$$

# SDP Reslaxation is *Near* Optimal

- Recall the SDP rate and minimax rate ($d = 1$, $q = 0$)

$$R_0^2 \frac{\lambda_1^2}{(\lambda_1 - \lambda_2)^2} \frac{\log p}{n} \quad \text{vs.} \quad R_0 \frac{\lambda_1 \lambda_2}{(\lambda_1 - \lambda_2)^2} \frac{\log p}{n}$$

- These are off by a factor of

$$R_0 \frac{\lambda_1}{\lambda_2} .$$

- The $R_0$ factor is unavoidable for polynomial time algorithms in a hypothesis testing context [Berthet & Rigollet 2013].
- $\lambda_1 / \lambda_2$ factor may be removable using finer analysis.

# *Summary*

- Sparsity helps improve both estimation accuracy and interpretability of PCA in high dimensions.
- Sparsity can be defined for principal subspaces.
- Minimax error rates are established for general covariance models.
- Convex relaxation using SDP is near-optimal.

# *Ongoing Work*

- Statistical properties for SDP relaxation under soft sparsity.
- SDP relaxation for subspaces ($d > 1$).
- Other penalties than $\ell_1$, such as the group lasso penalty.

# *Main References*


1. V. Vu and J. Lei (2012) "Minimax rates of estimation for sparse PCA in high dimensions", *AISTATS'12*
2. Vincent Vu and Jing Lei (2013) "Minimax Sparse Principal Subspace Estimation in High Dimensions", revision submitted.
3. Vincent Q. Vu, Juhee Cho, Jing Lei, and Karl Rohe (2013), ongoing work.