Sparse PCA: Concepts, Theory, and Algorithms

Jing Lei, *Department of Statistics, CMU*

U. Pitt. Biostat Department Seminar, Nov 7, 2013

Collaborators

Vince Q. Vu, Stat. Dept. OSU

Karl Rohe, Stat. Dept. U.W. Madison

Juhee Cho, Stat. Dept. U.W. Madison

Outline

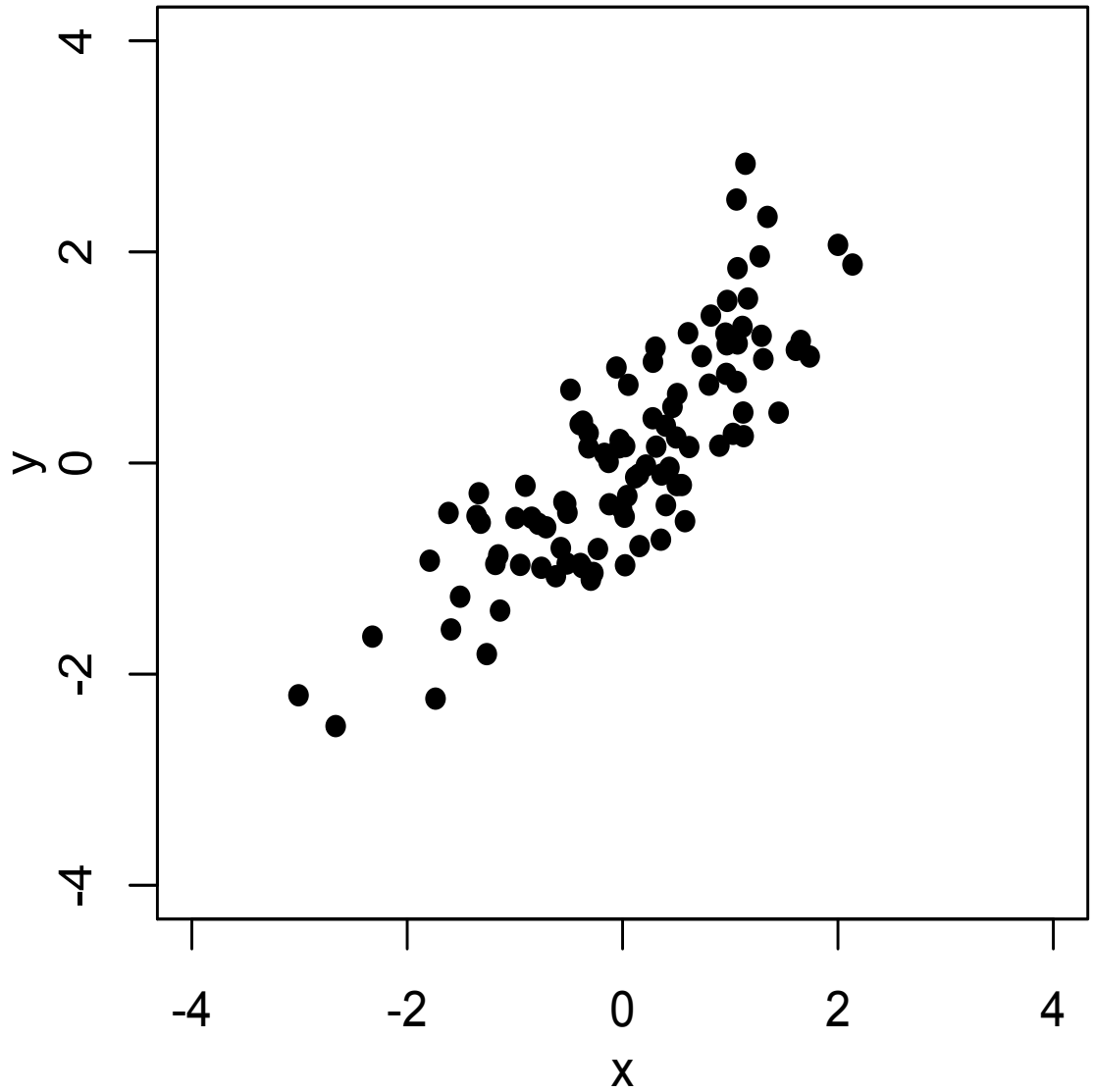
- What is sparse PCA?
- How does sparsity help?
- Can we estimate it in polynomial time?
- Is there a good algorithm?
- How does it work?

"In many physical, statistical, and biological investigations it is desirable to represent a system of points in ... higher dimensioned space by the 'best fitting' straight line or plane."

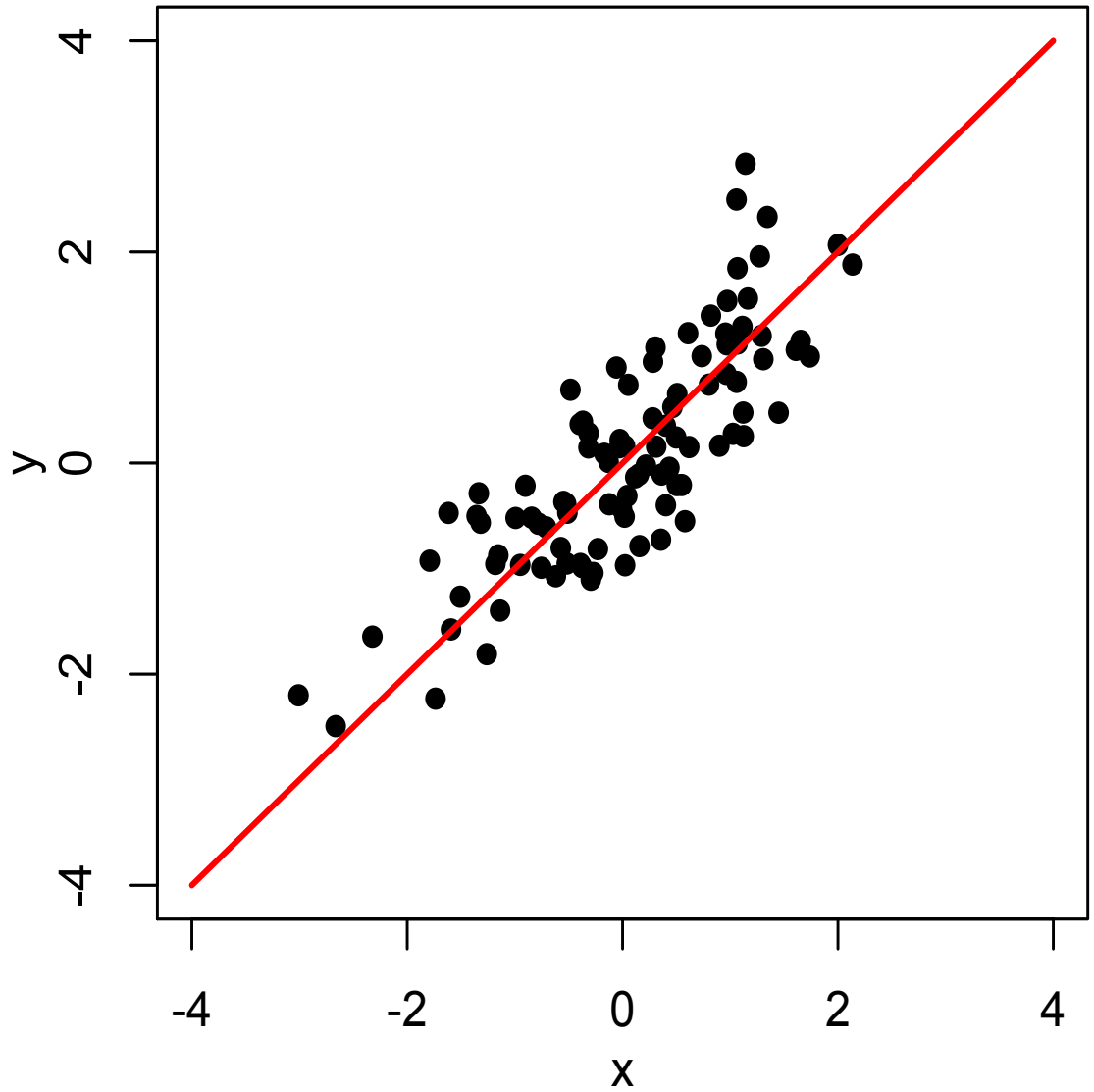
– **Karl Pearson** (1901)

On lines and planes of closest fit to systems of points in space

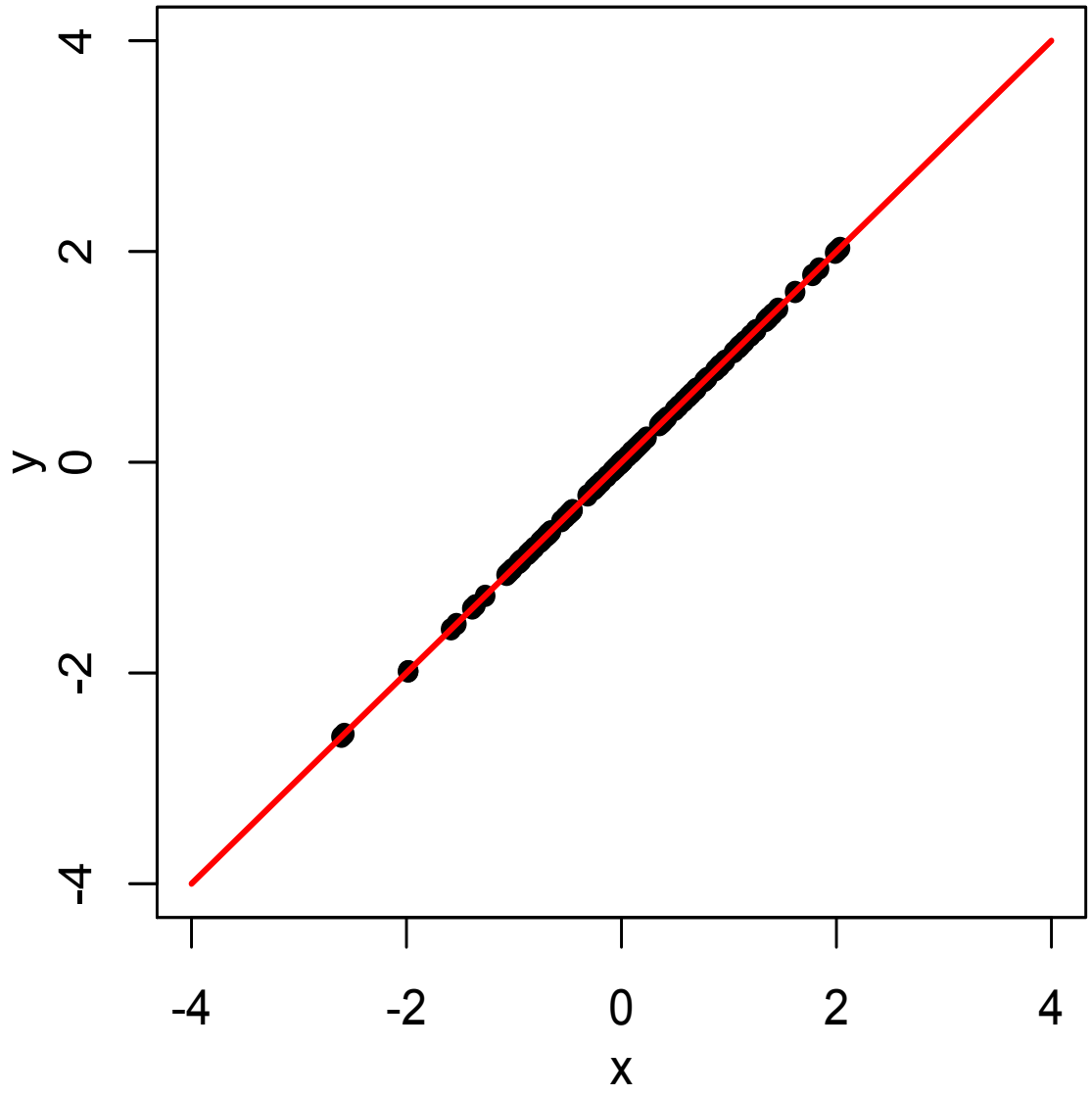
Principal Components Analysis



Principal Components Analysis



Principal Components Analysis



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: $X \to \Pi_d X$
  $\Pi_d = V_d V_d^T$ (d-dimensional projection matrix),
  $V_d = (v_1, \ldots, v_d)$.

Classical Estimator

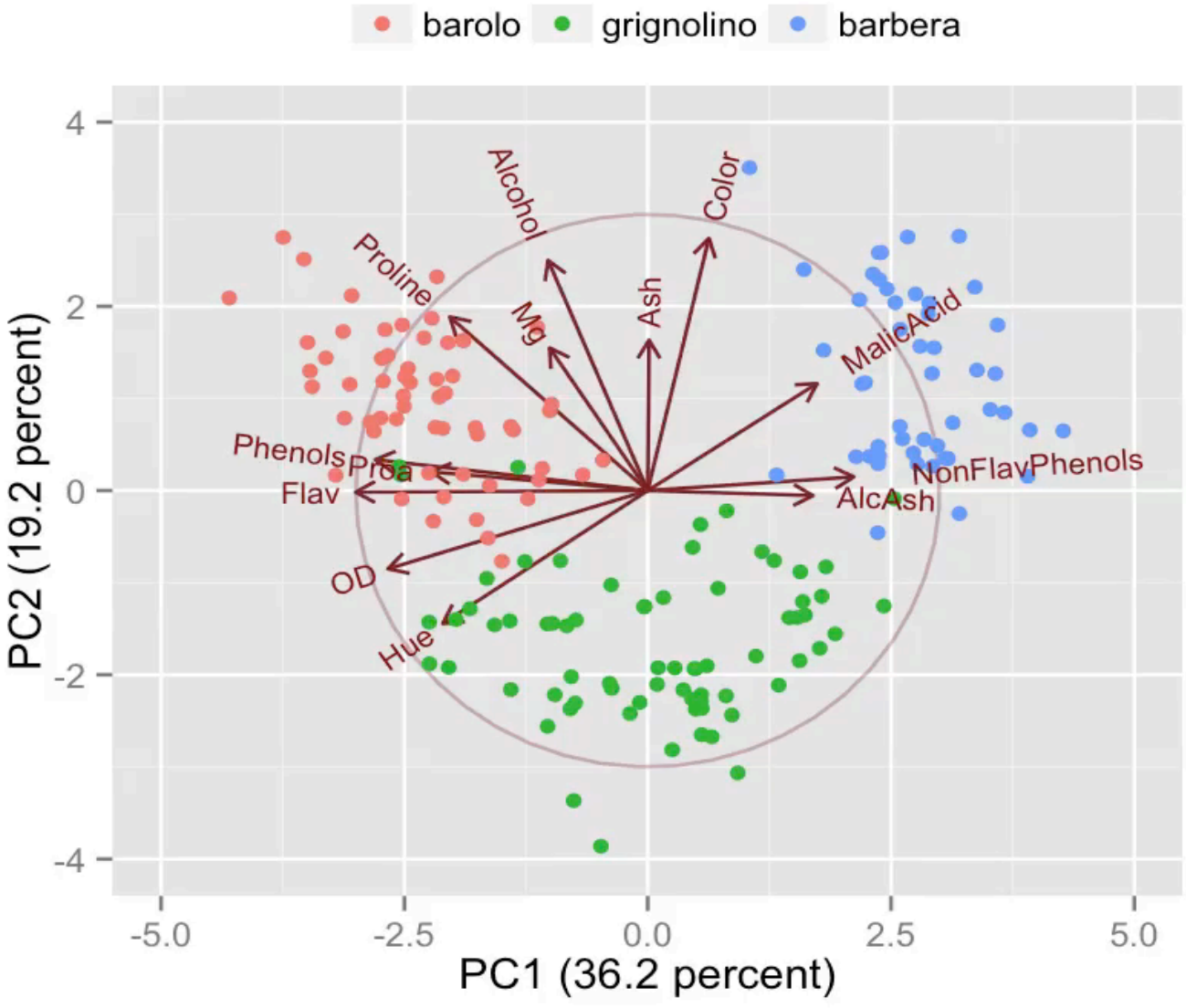
- Sample covariance matrix: $\hat{\Sigma} = n^{-1}(X_1 X_1^T + ... + X_n X_n^T)$.
- Estimate $(\hat{\lambda}_j, \hat{v}_j)$ by eigen-decomposition of $\hat{\Sigma}$.
  $\hat{V}_d = (\hat{v}_1, ..., \hat{v}_d)$, $\hat{\Pi}_d = \hat{V}_d \hat{V}_d^T$.
- These are consistent and asymptotically normal when p is fixed and $n \to \infty$.

What is sparse PCA?

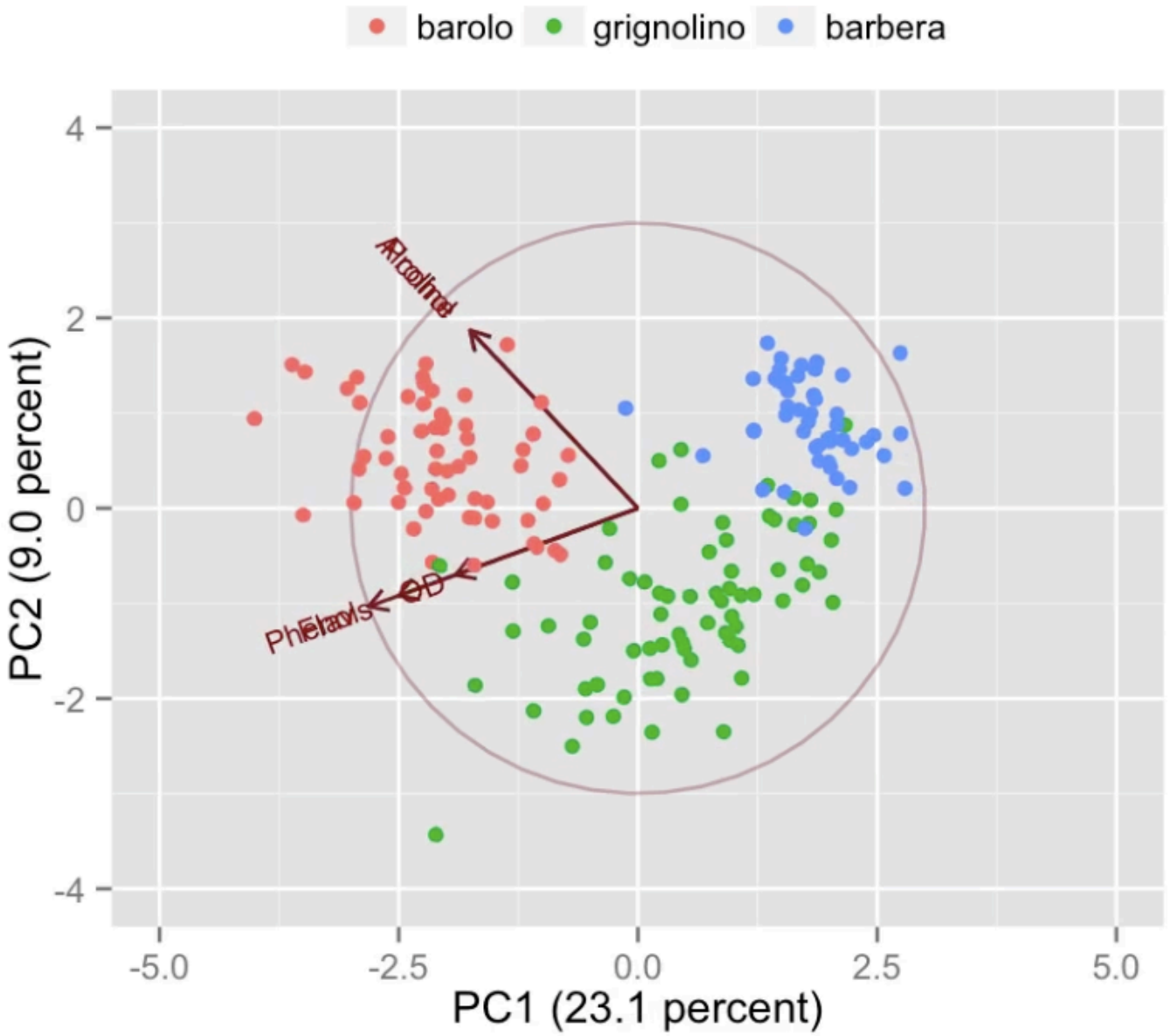
High-Dimensional PCA: Challenges

- When p is large, PCA can be inconsistent (Johnstone & Lu 2009 JASA), and/or hard to interpret.
- Sparse PCA offers simultaneous dimension reduction and variable selection.

UCI wine *Data* ($n = 178, p = 13$), *PCA*



UCI wine *Data* ($n = 178, p = 13$), *sPCA*



The Sparse PCA Model

- A general model:

$$\Sigma = \underbrace{\lambda_1 v_1 v_1^T + \ldots + \lambda_d v_d v_d^T}_{\text{signal}} + \underbrace{\lambda_{d+1}\Sigma'}_{\text{noise}}$$

where $\lambda_d > \lambda_{d+1}$; $\Sigma' \succeq 0$; $\|\Sigma'\| = 1$; $\Sigma' v_j = 0$, $\forall 1 \leq j \leq d$,

and $v_1, \ldots, v_d$ are sparse.

- Other authors (Johnstone & Lu 09, Paul & Johnstone 12, Ma 13, Cai et al 13) consider the Spiked model, with the additional constraint of spherical noise

$$\Sigma' = I_p.$$

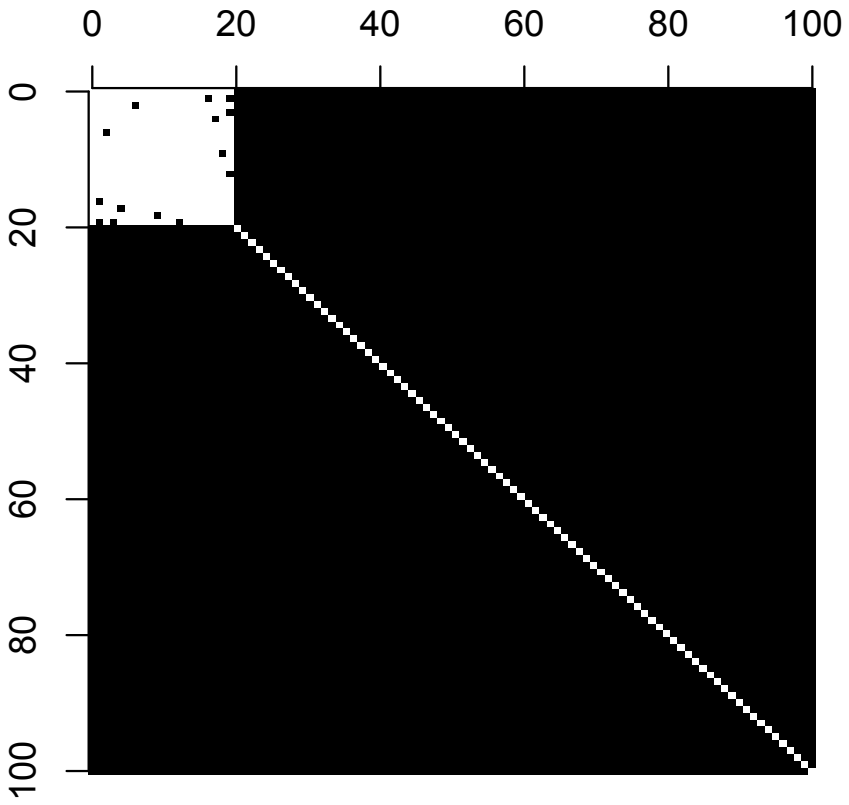
Subspace Sparsity [Vu & L 2013]

- Identifiability. If $\lambda_1 = \lambda_2 = ... = \lambda_d$, 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: $\|V_d\|_{2,0} \leq R_0 \ll p$. $V_d = (v_1, v_2, ..., v_d)$.
- Loss function: $\|\hat{\Pi}_d - \Pi_d\|_F^2$ ($\|\cdot\|_F$: the Frobenius norm).
  Recall: $\Pi_d = V_d V_d^T$, $\hat{\Pi}_d = \hat{V}_d \hat{V}_d^T$.
  Π is uniquely defined as long as $\lambda_d > \lambda_{d+1}$.

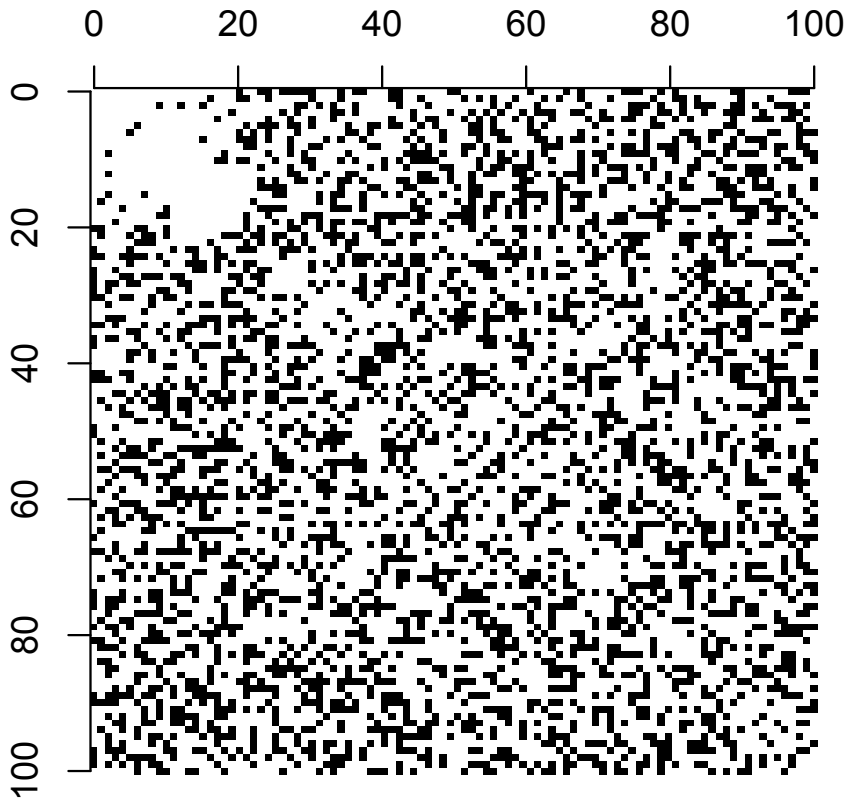
Spiked Model is a Special Case of General Model

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

In spiked model, all black cells outside the upper 20×20 are 0.



Covariance Pattern of Spiked Model



Covariance Pattern of General Model

How does sparsity help?

How Does Sparsity Help?

- Question: how does sparsity help PCA?
  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. λ_{d+1} is close to 0.
  4. $\lambda_d - \lambda_{d+1}$ is away from 0.
  5. R_0 is small.

Answer: Sparsity Determines the Error Rate

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

Under the general model, assuming V_d is R_0-sparse, then the optimal error rate of estimating Π_d is

$$\|\hat{\Pi}_d - \Pi_d\|_F^2 \asymp R_0 \frac{\lambda_1 \lambda_{d+1}}{(\lambda_d - \lambda_{d+1})^2} \frac{d + \log p}{n},$$

and can be achieved by

$$\hat{V}_d = \arg \max_{U^T U = I_d, \|U\|_{2,0} \leq R_0} \text{Trace}(U^T \hat{\Sigma} U).$$

About This Result

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

Can we estimate it in polynomial time?

Plan: formulate it as a convex optimization.

Some Linear Algebra

- Ordinary PCA (Ky Fan's Theorem):

$$\max_{U \in \mathbb{R}^{p \times d}} \text{Trace}(U^T \hat{\Sigma} U), \text{ s.t. } U^T U = I_d .$$

- Change to a linear problem by considering $Z = UU^T$:

$$\max_{Z,U} \text{Trace}(\hat{\Sigma} Z), \text{ s.t. } Z = UU^T, \; U^T U = I_d .$$

or equivalently

$$\min_Z \; -\text{Trace}(\hat{\Sigma} Z), \text{ s.t. } Z \text{ is a } d\text{-dim projection matrix}.$$

Convex Relaxation of Sparse PCA

- Add sparsity

$$\min_Z \; -\mathrm{Trace}(\hat{\Sigma} Z) + \lambda \|Z\|_1, \quad \text{s.t. } Z \text{ is a } d\text{-dim projection matrix}$$

$\|Z\|_1$: entry-wise ℓ_1 norm; λ: tuning parameter.

- The tightest convex relaxation would be

$$\min_Z \; -\mathrm{Trace}(\hat{\Sigma} Z) + \lambda \|Z\|_1,$$
$$\text{s.t. } Z \in \mathsf{Convhull}(\text{all } d\text{-dim projection matrices}).$$

The Fantope

Theorem *(Fillmore & Williams 71)*

$$\begin{aligned}
&\mathsf{Convhull}(\text{all } d\text{-dim projection matrices}) \\
=&\{Z : Z = Z^T \ , 0 \preceq Z \preceq I_p, \ \mathrm{Trace}(Z) = d\} \\
:=&\mathscr{F}_{p,d} \quad (\text{the Fantope})
\end{aligned}$$

Ky Fan (1914–2010)

Convex Relaxation of Sparse PCA

Fantope Projection and Selection (FPS) [VCLR13]

$$\min_Z \quad -\mathrm{Trace}(\hat{\Sigma} Z) + \lambda \|Z\|_1, \quad s.t.\ Z \in \mathscr{F}_{p,d}\,.$$

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

Under the general PCA model, with R_0-sparsity on V_d, the global optimizer $\hat{Z}$ satisfies (w.h.p)

$$\|\hat{Z} - \Pi_d\|_F^2 \lesssim R_0^2 \frac{\lambda_1 \lambda_{d+1}}{(\lambda_d - \lambda_{d+1})^2} \frac{\log p}{n}\,.$$

When d is small, this has an extra factor of R_0 (compare to minimax rate), which may be unavoidable for polynomial time algorithms.

How to solve FPS?

A Hard Problem at First Glance

$$\min_Z \quad -\text{Trace}(\hat{\Sigma} Z) + \lambda \|Z\|_1, \quad s.t.\ Z \in \mathscr{F}_{p,d}\,.$$

- FPS minimizes a linear function of Z over the intersection of two convex sets:
  1. The ℓ_1 ball;
  2. The Fantope $\mathscr{F}_{p,d}$.
- Similar to the ℓ_1-penalized inverse covariance matrix estimation (graphical Lasso)
- Traditional algorithms may be slow.
- There are powerful tools designed for problems like these.

Alternating Direction Method of Multipliers

Equivalent formulation using convex indicator $\mathbf{1}_{\mathscr{F}_{p,d}}(\cdot)$:

$$\min_Z \quad -\mathrm{Trace}(\hat{\Sigma}Z) + \mathbf{1}_{\mathscr{F}_{p,d}}(Z) + \lambda \|Z\|_1 .$$

Z appears in three terms. We split these into Z and Y

$$\min_{Z,Y} \quad -\mathrm{Trace}(\hat{\Sigma}Z) + \mathbf{1}_{\mathscr{F}_{p,d}}(Z) + \lambda \|Y\|_1, \;\; s.t. \;\; Z = Y .$$

Augmented Lagrangian with dual variable W

$$\begin{aligned} \min_{Z,Y,W} \quad & -\mathrm{Trace}(\hat{\Sigma}Z) + \mathbf{1}_{\mathscr{F}_{p,d}}(Z) + \lambda \|Y\|_1 \\ & + \mathrm{Trace}((Z-Y)W) + \frac{\rho}{2}\|Z-Y\|_F^2 . \end{aligned}$$

ρ is a penalty parameter (like the step size).

Update Scheme is Simple

From current state $(Z^{old}, Y^{old}, W^{old})$, the variables are updated by iteratively optimizing the Lagrangian over Z and Y.

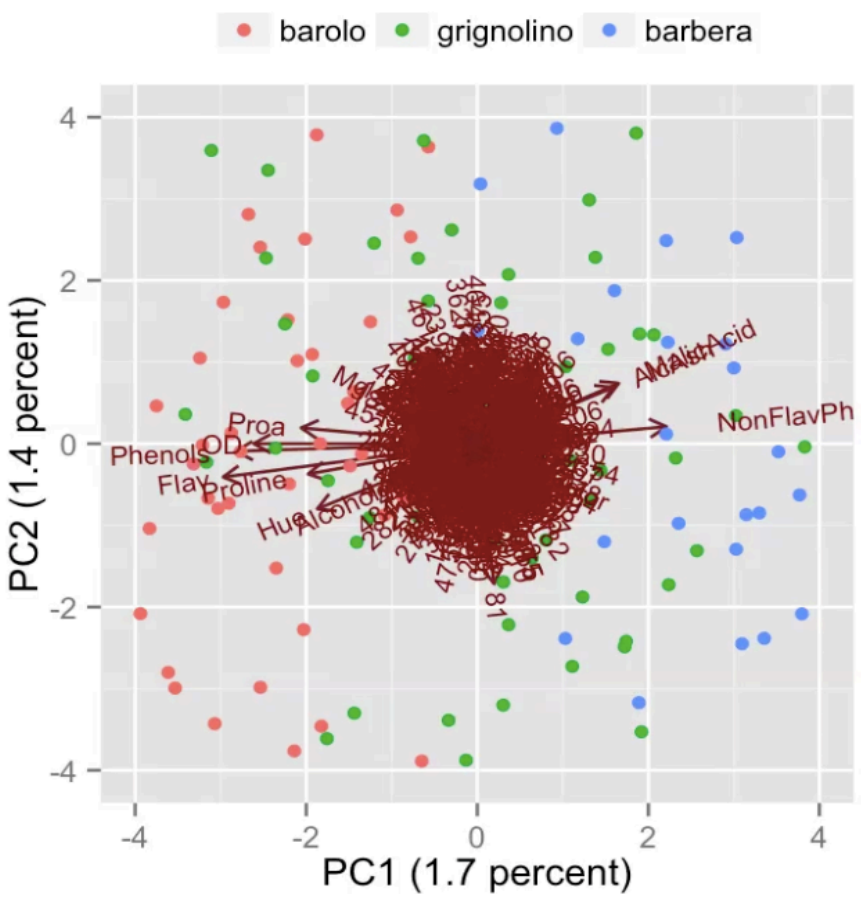
$$
\begin{aligned}
Z^{new} &= \mathscr{P}_{\mathscr{F}_{p,d}}(Y^{old} + (\hat{\Sigma} - W^{old})/\rho), \quad \text{Fantope projection} \\
Y^{new} &= ST_{\lambda/\rho}(Z^{new} + W^{old}/\rho), \quad \text{soft thresholding} \\
W^{new} &= W^{old} + \rho(Z^{new} - Y^{new}), \quad \text{dual update}
\end{aligned}
$$

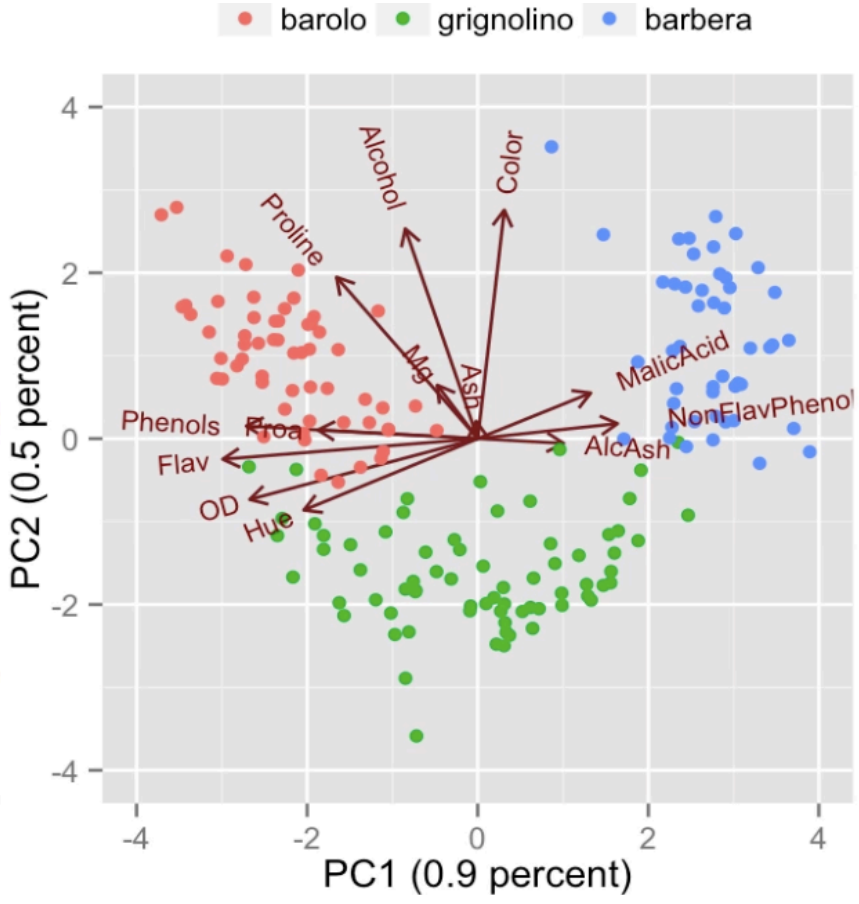
Then $(Z^{old}, Y^{old}, W^{old}) \leftarrow (Z^{new}, Y^{new}, W^{new})$ and repeat until convergence is observed.

Detail: Fantope projection $\approx$ soft thresholding singular values.

How does it work?

Wine Data Again. Added 487 Noise Columns.



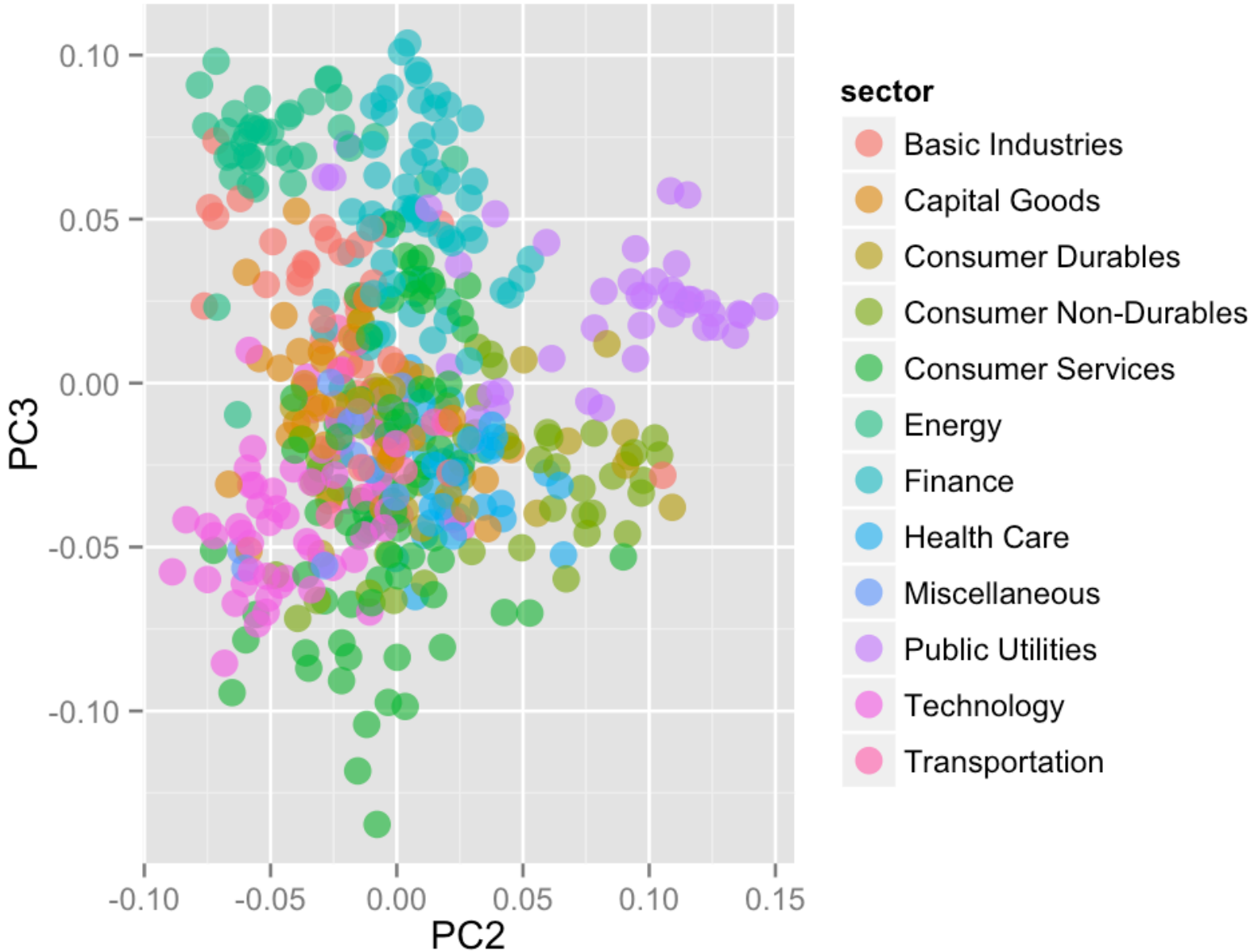


Where else can we use it?

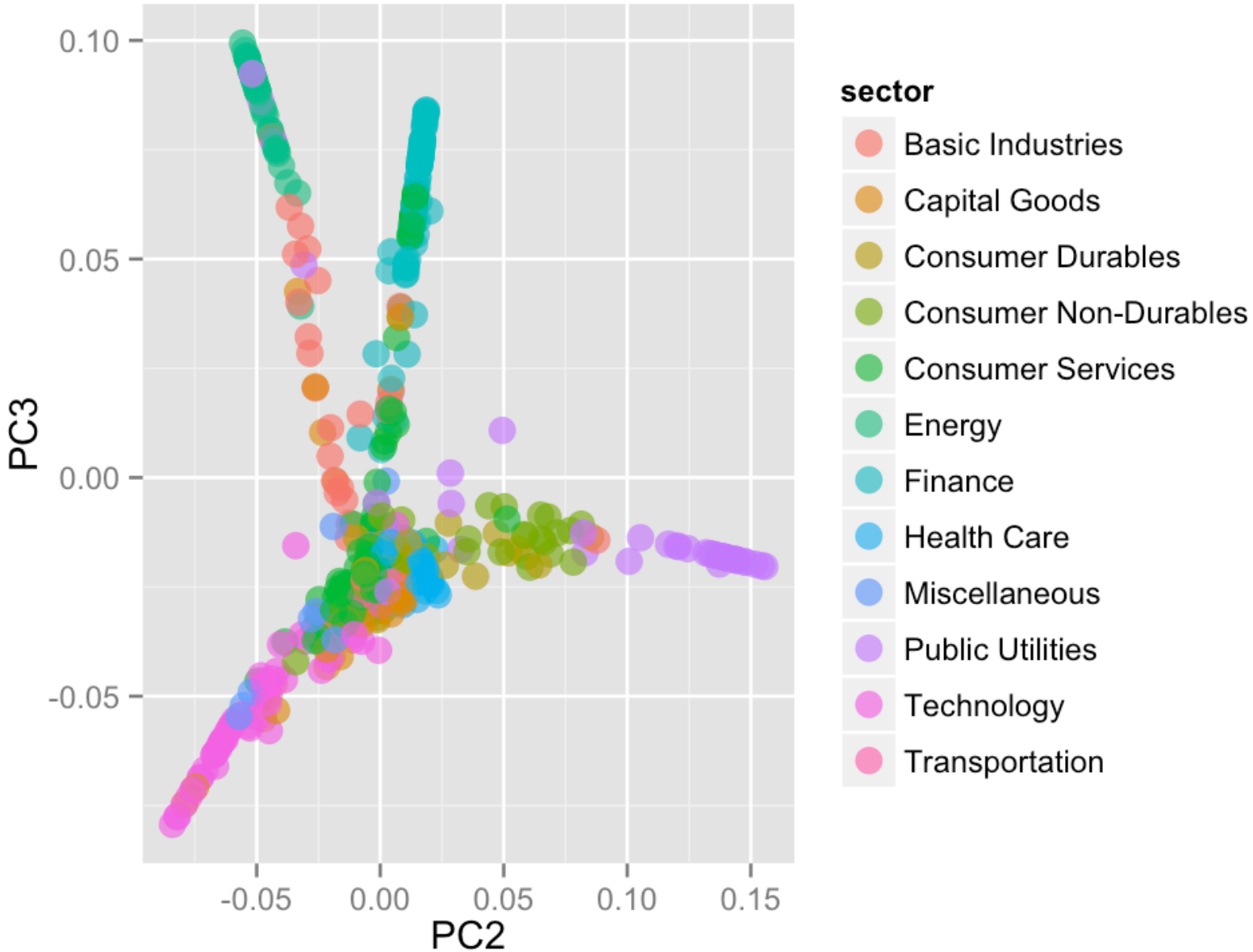
- Variable clustering: finding sets of variables with similar correlation patterns.
- Intuition: These variables will point to similar directions in the bi-plot.
- Example: S&P 500 data.

  500 (n) daily returns for 500 (p) stocks.

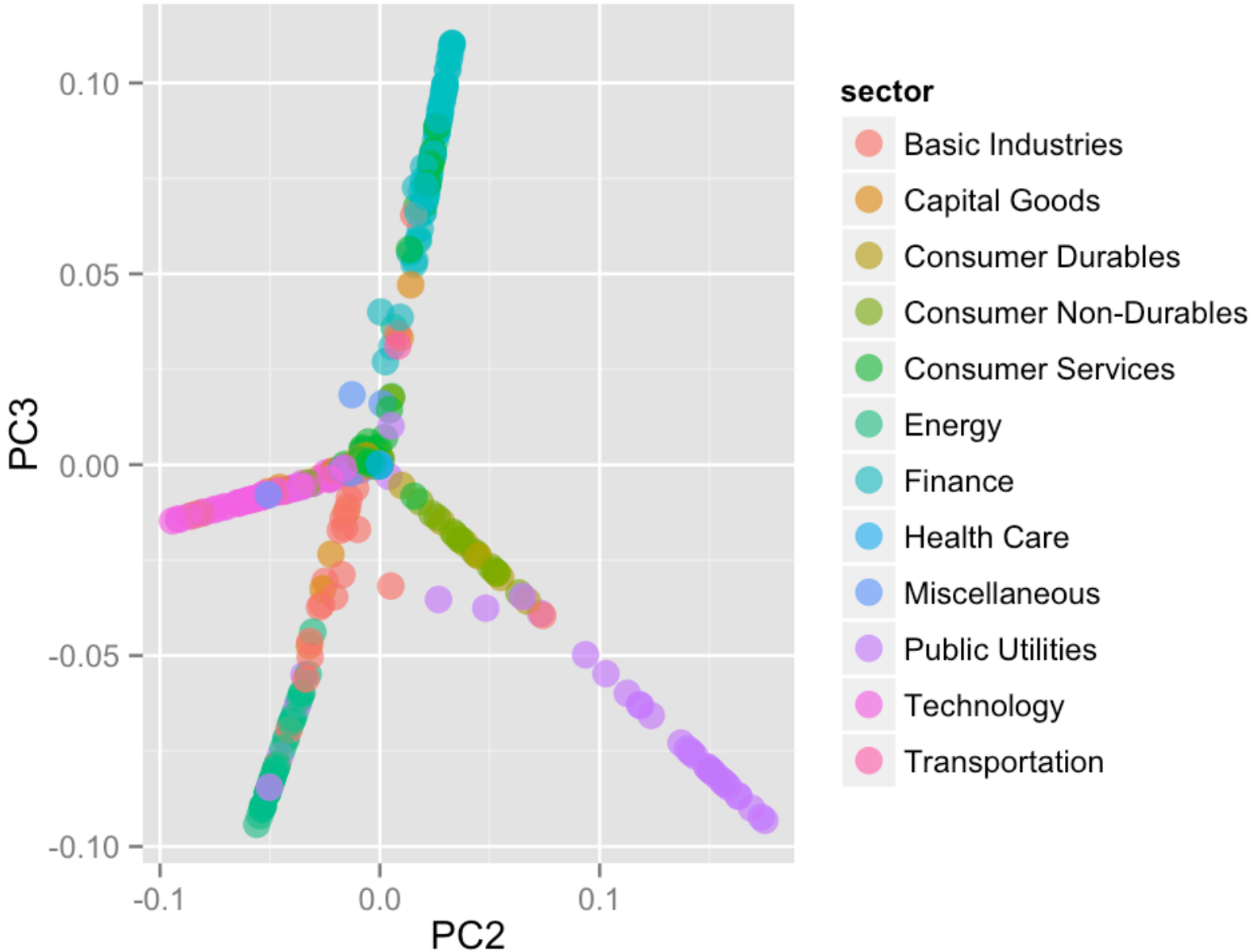
Example: S&P 500 data



Example: S&P 500 data



Example: S&P 500 data



Another Example

- New York Times data shows similar variable clustering effect. Some of the clusters are
  1. “bowl”, “butter”, “chopped”, “cup”, “add”, “tablespoon”, “gram”, “oil”, “pan”, “water”, ...
  2. “fund”, “investment”, “market”, “price”, “stock”, ...
  3. “administration”, “meeting”, “bush”, “congress”, “washington” “white house”, ...

Future & Ongoing Work

- Applications
  1. Linear discriminant analysis
  2. Clustering
  3. Network data
- Theory & methods
  1. Sparsistency
  2. Hypothesis testing
  3. Sparse singular value decomposition

Summary

- Sparsity PCA offers simultaneous dimension reduction and variable selection.
- It makes more sense to focus on the subspace.
- It can be estimated at optimal rate.
- Near optimal practical method: Fantope + ADMM.
- It works, but its behavior needs better understanding.

Thank You!

Questions?