Fantope Projection and Selection:

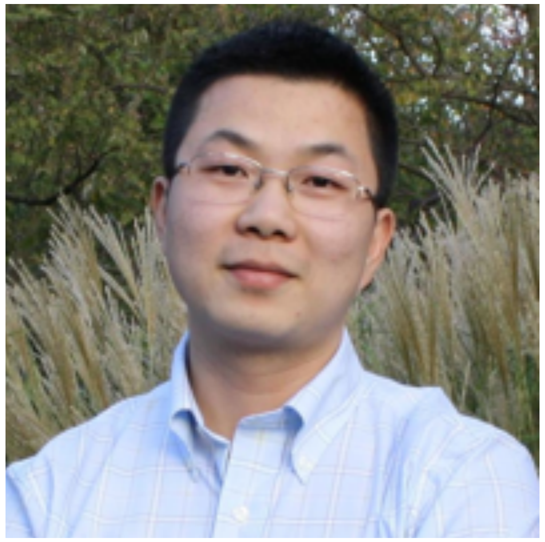
Near-optimal convex relaxation of Sparse PCA

Vincent Q. Vu

Department of Statistics
The Ohio State University

with J. **Lei** (CMU), J. **Cho** (Wisc), K. **Rohe** (Wisc)

This talk is based on work with...

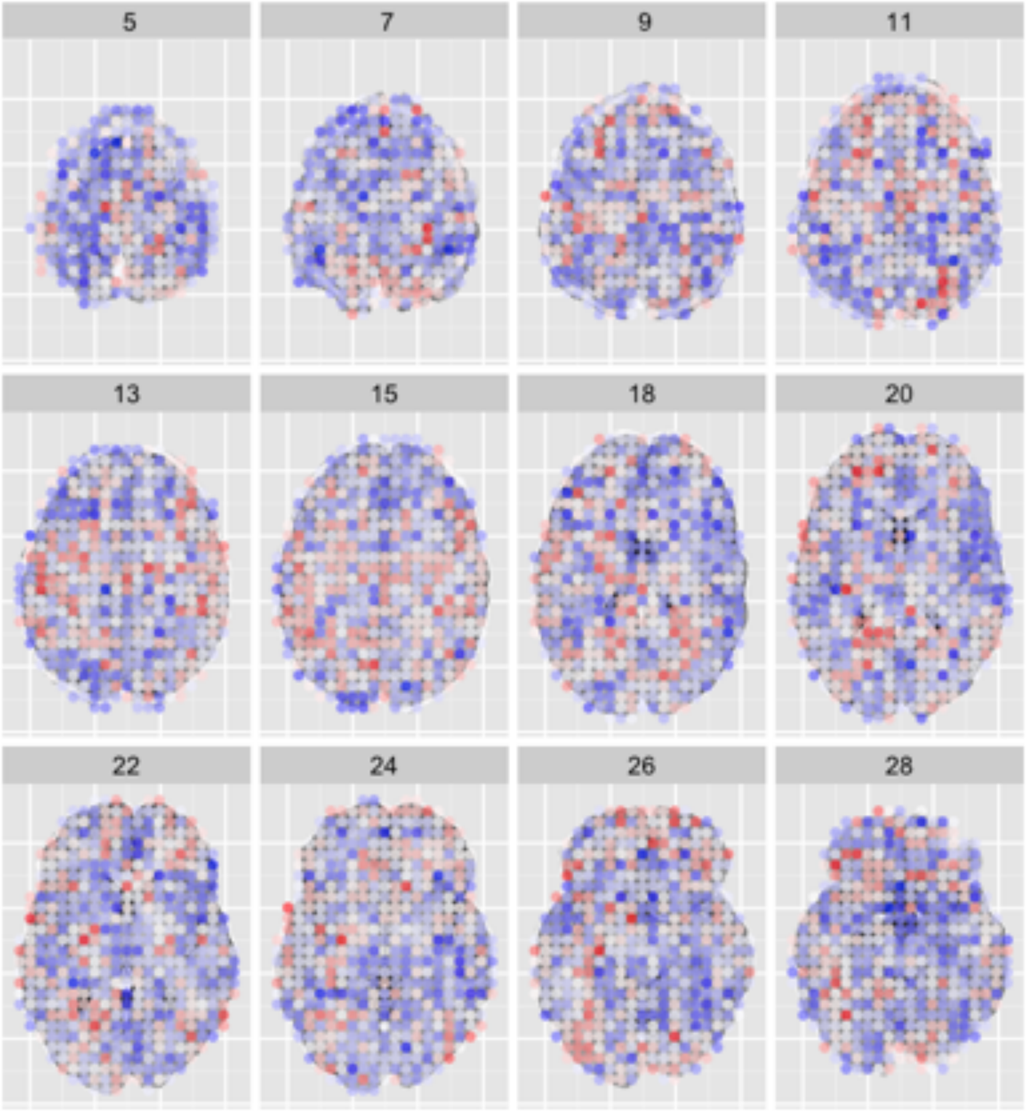


Jing Lei
Carnegie Mellon U.

Juhee Cho
U. Wisconsin-Madison

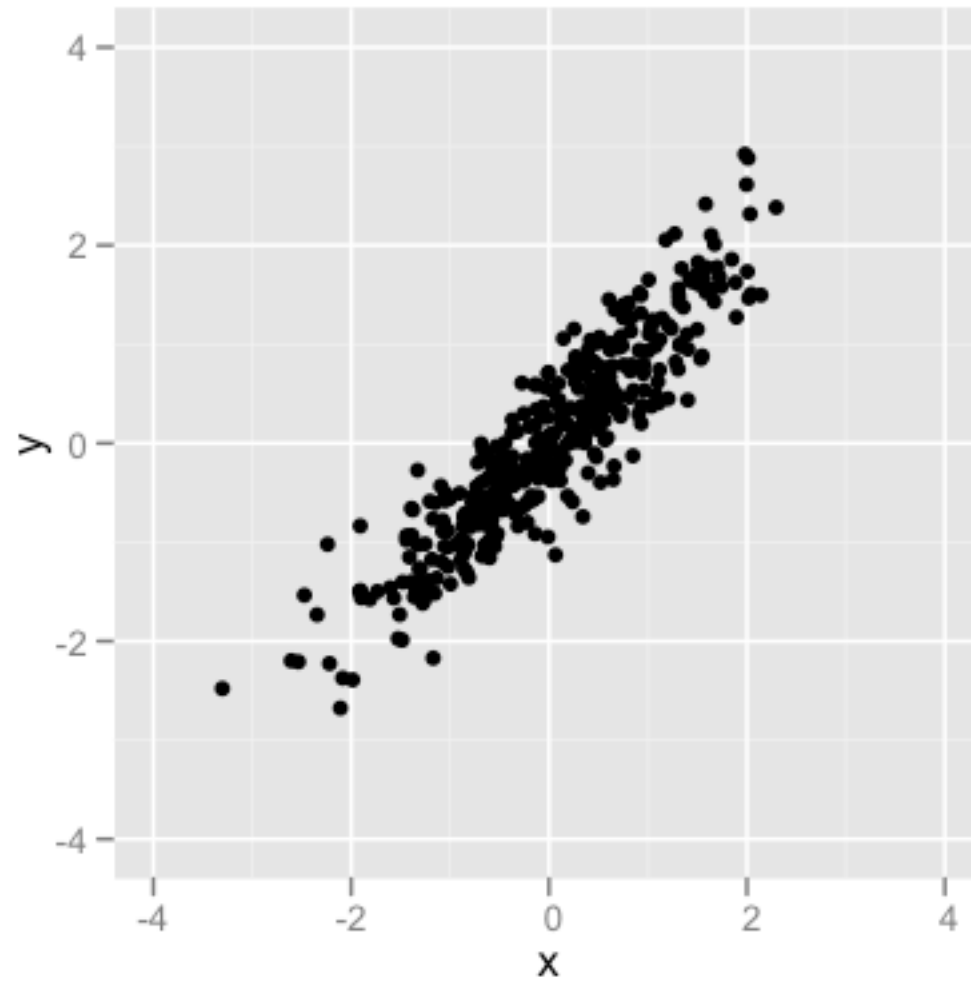
Karl Rohe
U. Wisconsin-Madison

Example: fMRI



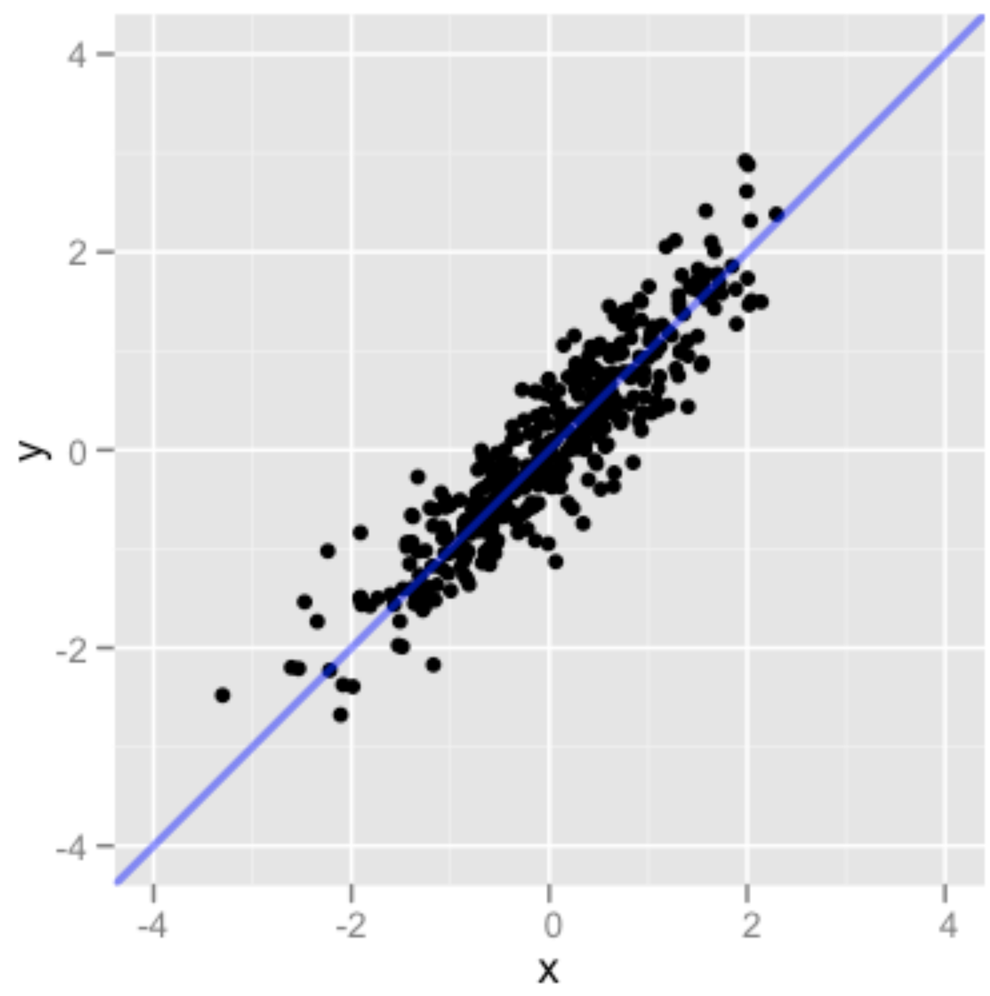
- **n** ≈ 10^2 ~ 10^3 images
- **p** ≈ 10^5 ~ 10^6 voxels
- Scientists interested in joint modeling of voxels
- But… challenging because of high-dimensionality
- **Dimension reduction** can be beneficial

Principal Components Analysis



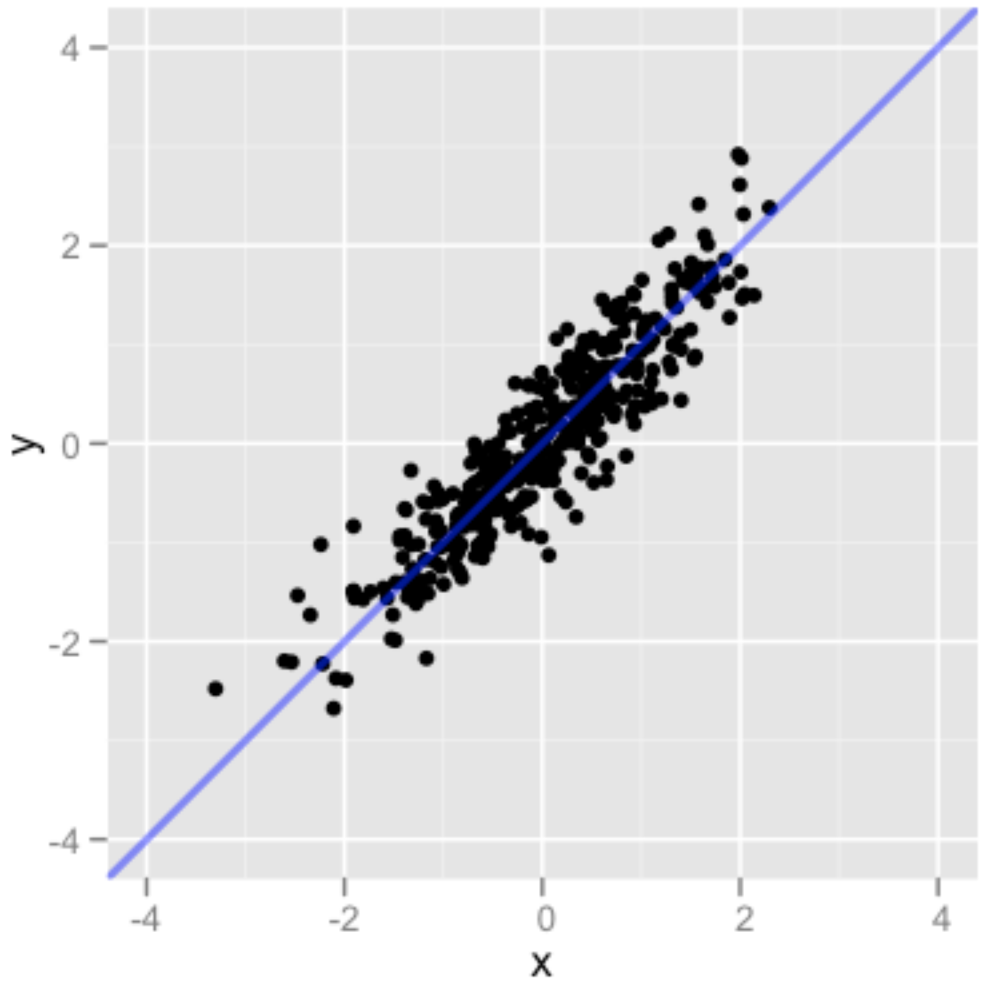
Pearson (1901)

Principal Components Analysis

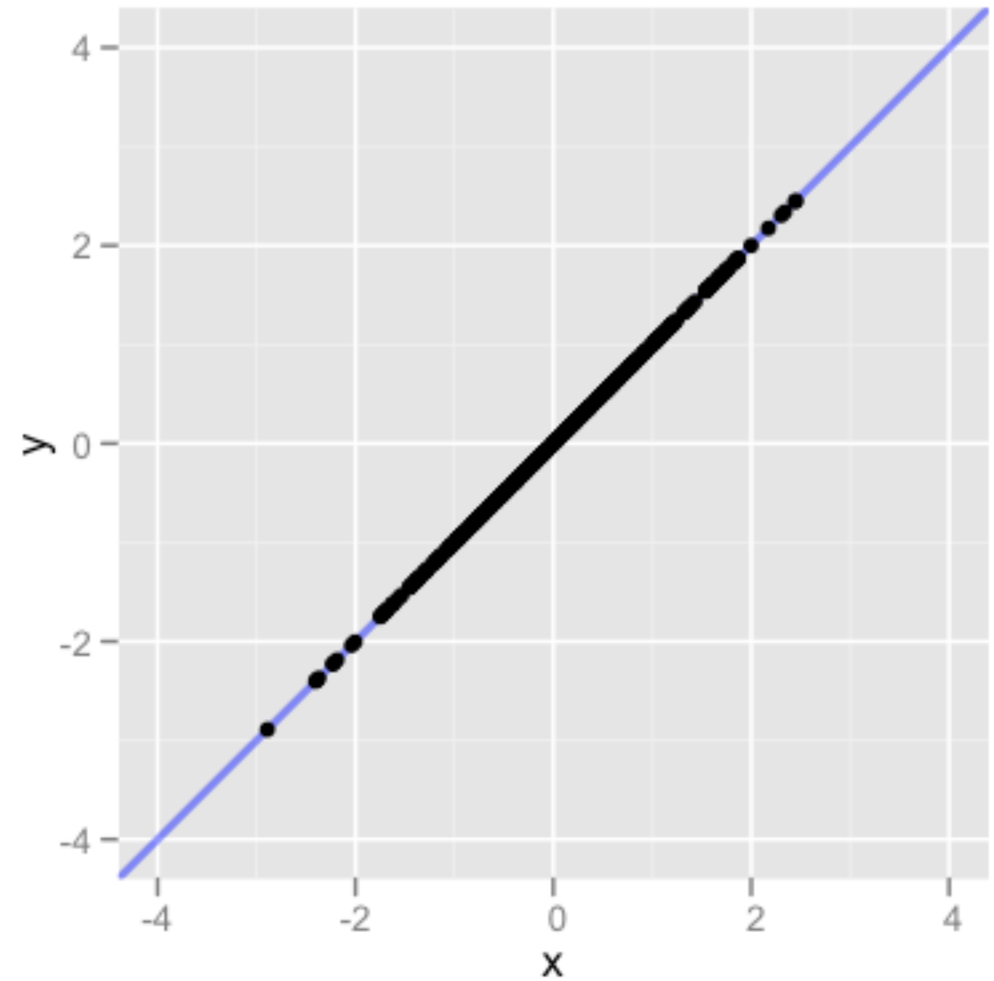


Pearson (1901)

Principal Components Analysis



≈



original data

lower-dimensional projection

Principal Components Analysis

- Suppose $\{ X_1, X_2, ..., X_n \}$ is a dataset of i.i.d. observations on **p** variables
- **p** is ***large***, so want to use **PCA** for dimension reduction

PCA

- Population covariance* and its eigendecomposition

$$\begin{aligned}\Sigma :=& \mathbb{E}(XX^T) \\ =& \lambda_1 v_1 v_1^T + \cdots + \lambda_p v_p v_p^T\end{aligned}$$

eigenvalues $\lambda_1 \geq \ldots \geq \lambda_p$ and eigenvectors $v_1, \ldots, v_p$

- "Optimal" d-dimensional projection

$$\Pi_k = V_k V_k^T, \quad V_k = (v_1, \ldots, v_k)$$

(*assume $\mathbb{E}X = 0$ to simplify presentation)

Classic PCA estimate

- Sample covariance and its eigendecomposition

$$\begin{aligned}\widehat{\Sigma} &= n^{-1}(X_1 X_1^T + \cdots + X_n X_n^T) \\ &= \hat{\lambda}_1 \hat{v}_1 \hat{v}_1^T + \cdots + \hat{\lambda}_p \hat{v}_p \hat{v}_p^T\end{aligned}$$

- PCA estimate of d-dimensional projection

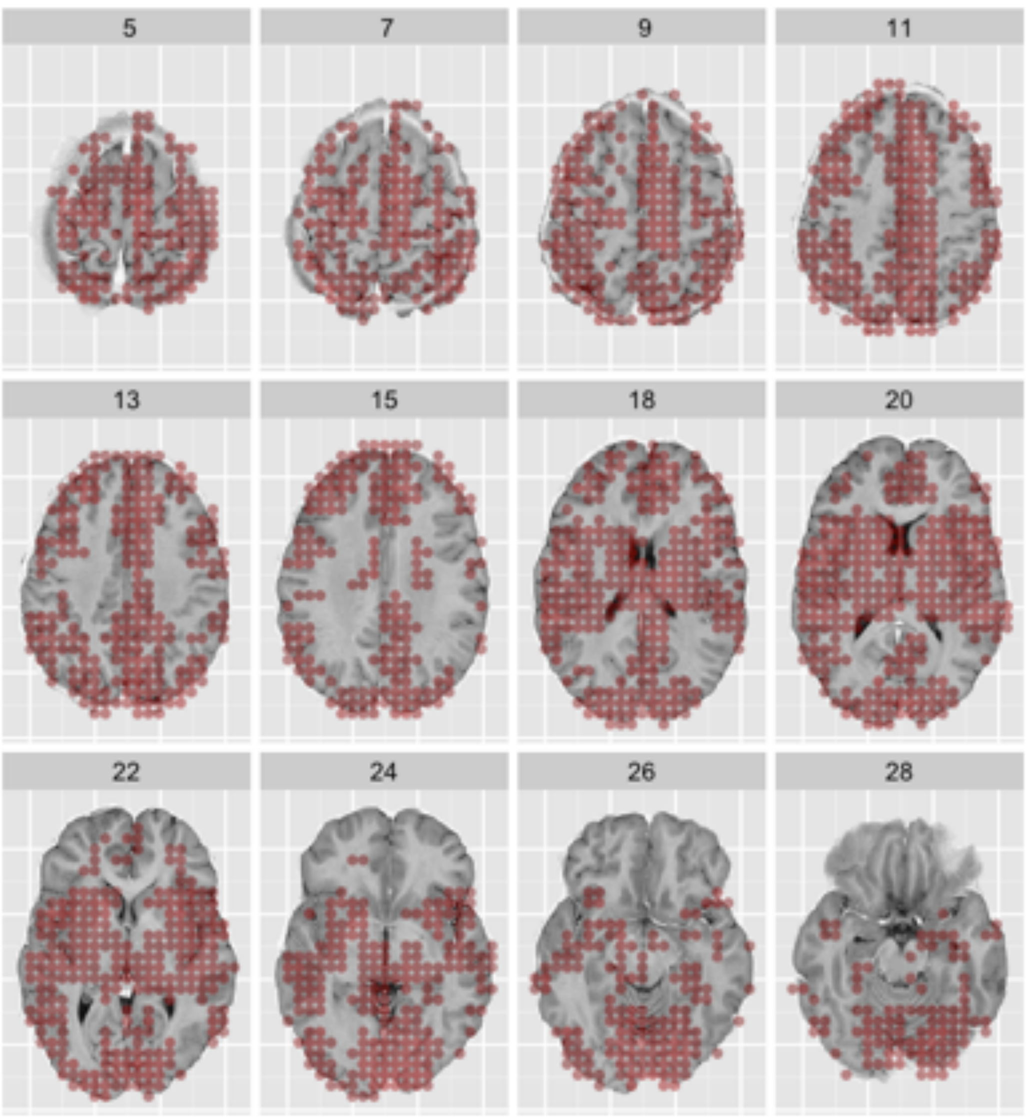
$$\widehat{\Pi}_k = \widehat{V}_k \widehat{V}_k^T \,, \quad \widehat{V}_k = (\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_k)$$

- Consistent (converges to truth) when **p** is <u>fixed</u> and **n→∞**

High-dimensional challenges

- In contemporary applications, *e.g.* neuroimaging: **$p \approx n$** and often **$p > n$**
- When **p/n** $\rightarrow c \in (0,\infty]$, classic PCA can be **inconsistent** (Johnstone & Lu '09), (*e.g.* $\hat{v}_1^T v_1 \approx 0$) and/or **difficult to interpret**
- Sparse PCA can help

Example: fMRI



- "Interesting" activity often spatially localized
- Locations not known in advance
- Localization = **sparsity**
- Combine **dimension reduction** *and* **sparsity**?

Outline

- Sparse PCA and subspace estimation
- A convex relaxation and its near-optimality
- Some synthetic examples
- Whither sparisity?

Sparse PCA and Subspace Estimation

Sparse PCA

- Many methods proposed over past 10 years:

  ***Joliffe**, et al. (2003); **Zou**, et al. (2006); **d'Aspremont**, et al. (2007); **Shen** and **Huang** (2008); **Johnstone** and **Lu** (2009); **Witten**, et al. (2009); **Journée** et al. (2010); and many more*

- Mostly algorithmic proposals for **k**=**1**

- Few theoretical guarantees on statistical error and strong assumptions (e.g. spiked covariance model)

Subspace sparsity

- If $\boldsymbol{\lambda_1}=\boldsymbol{\lambda_2}=\ldots=\boldsymbol{\lambda_k}$, then cannot distinguish $\mathbf{V_k}$ and $\mathbf{V_k Q}$ from observed data for any orthogonal $\mathbf{Q}$
- Good notion of sparsity must be rotation invariant
- **Row sparsity** – two equivalent definitions:
  - At most **s** rows of $\mathbf{V_k}$ (and hence Π_k) are nonzero
  - Projection depends on fewer than **s** variables

(VL '13)

General sparse PCA model

$$\Sigma = \underset{\textbf{signal}}{\begin{array}{c c} \scriptstyle s & \scriptstyle p\text{-}s \\ \multicolumn{2}{c}{\begin{bmatrix} UDU^T & 0 \\ 0 & 0 \end{bmatrix}} \end{array}} + \underset{\textbf{noise}}{\begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{bmatrix}}, \quad \Pi_k = \begin{bmatrix} UU^T & 0 \\ 0 & 0 \end{bmatrix}$$

$$\textbf{signal} = \lambda_1 v_1 v_1^T + \cdots + \lambda_k v_k v_k^T$$

$$\textbf{noise} = \lambda_{k+1} v_{k+1} v_{k+1}^T + \cdots + \lambda_p v_p v_p^T$$

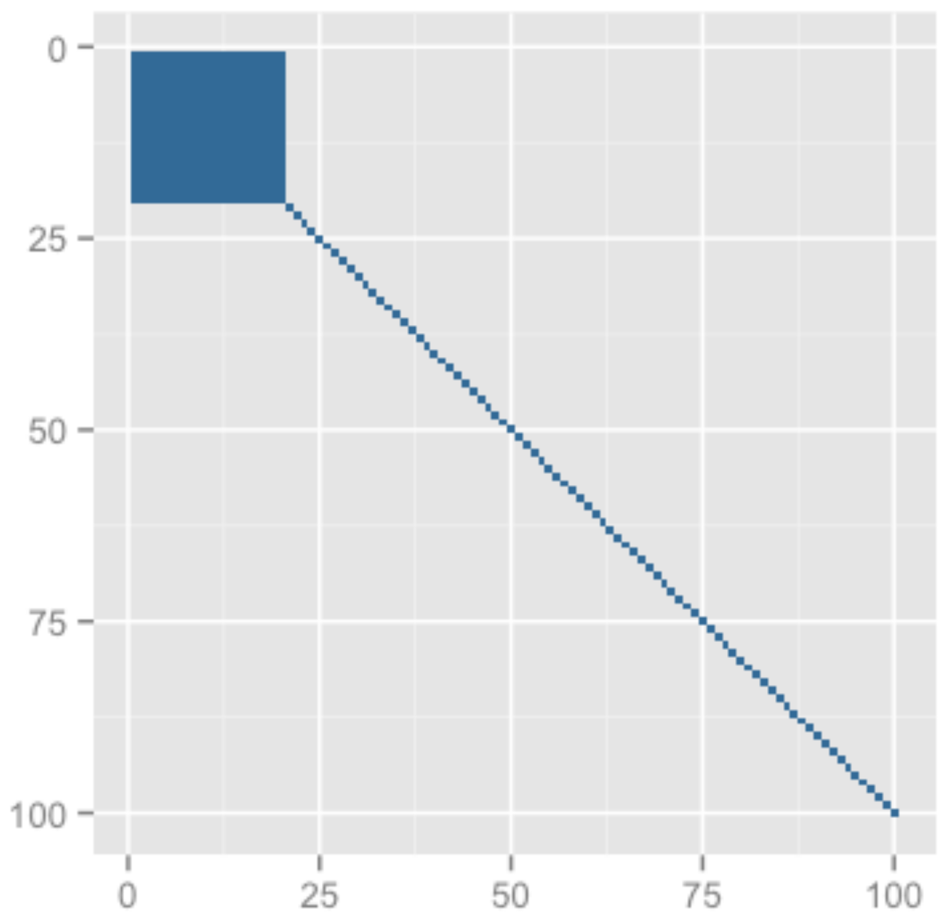
$$D = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$$

$$U = \text{nonzero block of } \mathbf{V_k}$$

Decomposition always exists (for some **s**) and unique when $\lambda_k > \lambda_{k+1}$

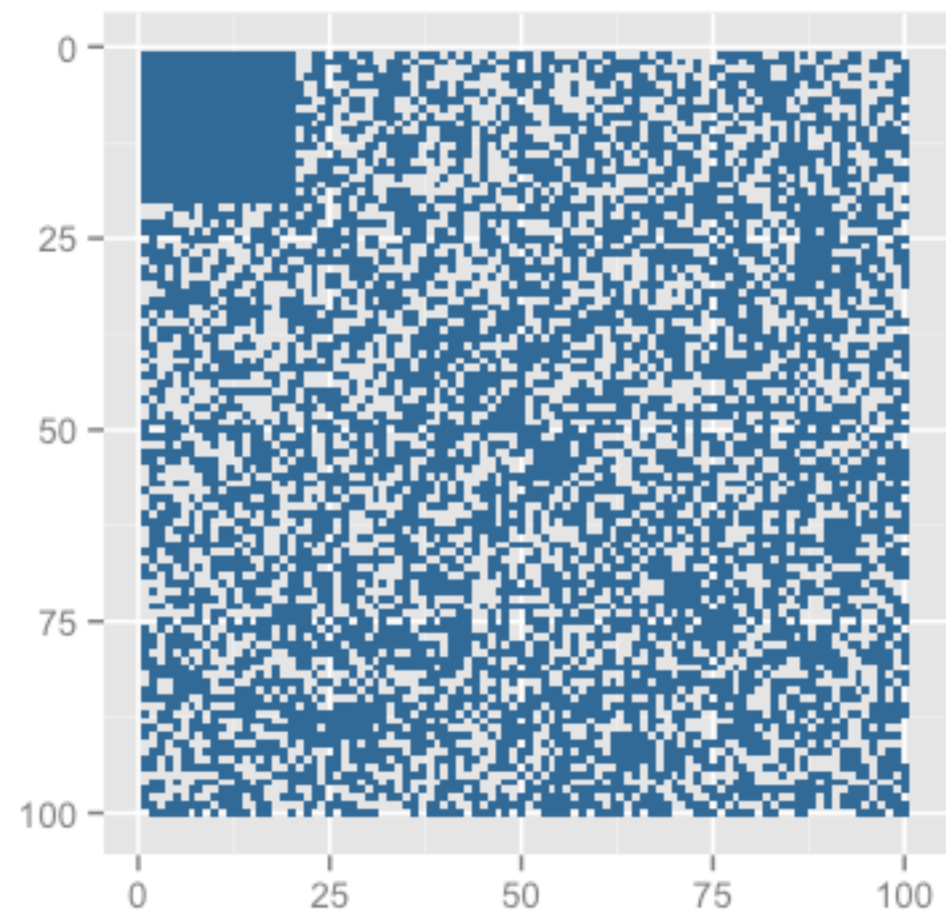
Spiked model vs General model

Locations of large nonzero entries: $|\Sigma(i,j)| \geq 0.01$



Spiked model

$(\Gamma = I)$



General model

Sparsity enables estimation

Theorem (VL '13)
Under the sparse PCA model*, the optimal error rate of estimating Π_k is

$$\min_{\widehat{\Pi}_k} \max_{\Sigma} \mathbb{E}\|\widehat{\Pi}_k - \Pi_k\|_F^2 \asymp s \cdot \frac{\lambda_1 \lambda_{k+1}}{(\lambda_k - \lambda_{k+1})^2} \cdot \frac{k + \log p}{n}$$

and can be achieved by

$$\widehat{\Pi}_k = \arg\max_{\Pi} \ \operatorname{trace}(\widehat{\Sigma}\Pi)$$

where the max is over ***s***-sparse, rank-***k*** projection matrices.

** with sub-Gaussian X_i*

Computation?

- Theorem gives optimal dependence on $(n, p, s, k, \lambda_1, \lambda_k, \lambda_{k+1})$
- No additional assumptions on noise Γ (e.g., spiked covariance model not necessary)
- But constructed minimax optimal estimator is impractical to compute :-(

Convex Relaxation

Convex relaxation of sparse PCA

Fantope Projection and Selection (VLCR '13)

$$\max_{H} \; \operatorname{trace}(\widehat{\Sigma} H) - \rho \sum_{ij} |H_{ij}| \quad \text{subject to} \quad \begin{cases} 0 \preceq H \preceq I \\ \operatorname{trace}(H) = k \end{cases}$$

PCA — **sparsity** $(\rho \geq 0)$ — **convex hull of rank-k projection matrices**

Constraint set called **Fantope** (*Fillmore & Williams '71*, *Overton & Womersly '02*) — named after Ky Fan

Reduces to: classic PCA when **ρ**=0; DSPCA (d'Aspremont et al. '07) when **k**=1

FPS

- Solved efficiently by alternating direction method of multipliers (ADMM) with two main steps:
  - Projection onto Fantope (≈ same difficulty as SVD)
  - Entry-wise soft-thresholding (L_1 proximal operator)
- Iteration complexity $O(p^3)$ – but typically $O(kp^2)$ and dependent on choice of tuning parameter $\boldsymbol{\rho}$

FPS is near-optimal

Theorem (VLCR '13)

Under the sparse PCA model*, if

$$\rho \sim \sqrt{\log p / n}$$

then any solution $\widehat{H}$ of **FPS** satisfies (with high probability)

$$\|\widehat{H} - \Pi_k\|_F^2 \lesssim s^2 \cdot \frac{\lambda_1 \lambda_{k+1}}{(\lambda_k - \lambda_{k+1})^2} \cdot \frac{\log p}{n}$$

* *with sub-Gaussian X_i*

Computational barrier?

- When subspace dimension **k**=1

$$\frac{\text{FPS error rate}}{\text{optimal error rate}} \sim s$$

  - Extra factor **s** maybe unavoidable for polynomial time algorithms (*Berthet & Rigollet '13*)
  - Maybe possible to get tighter rate under stronger assumptions, e.g. spiked covariance?

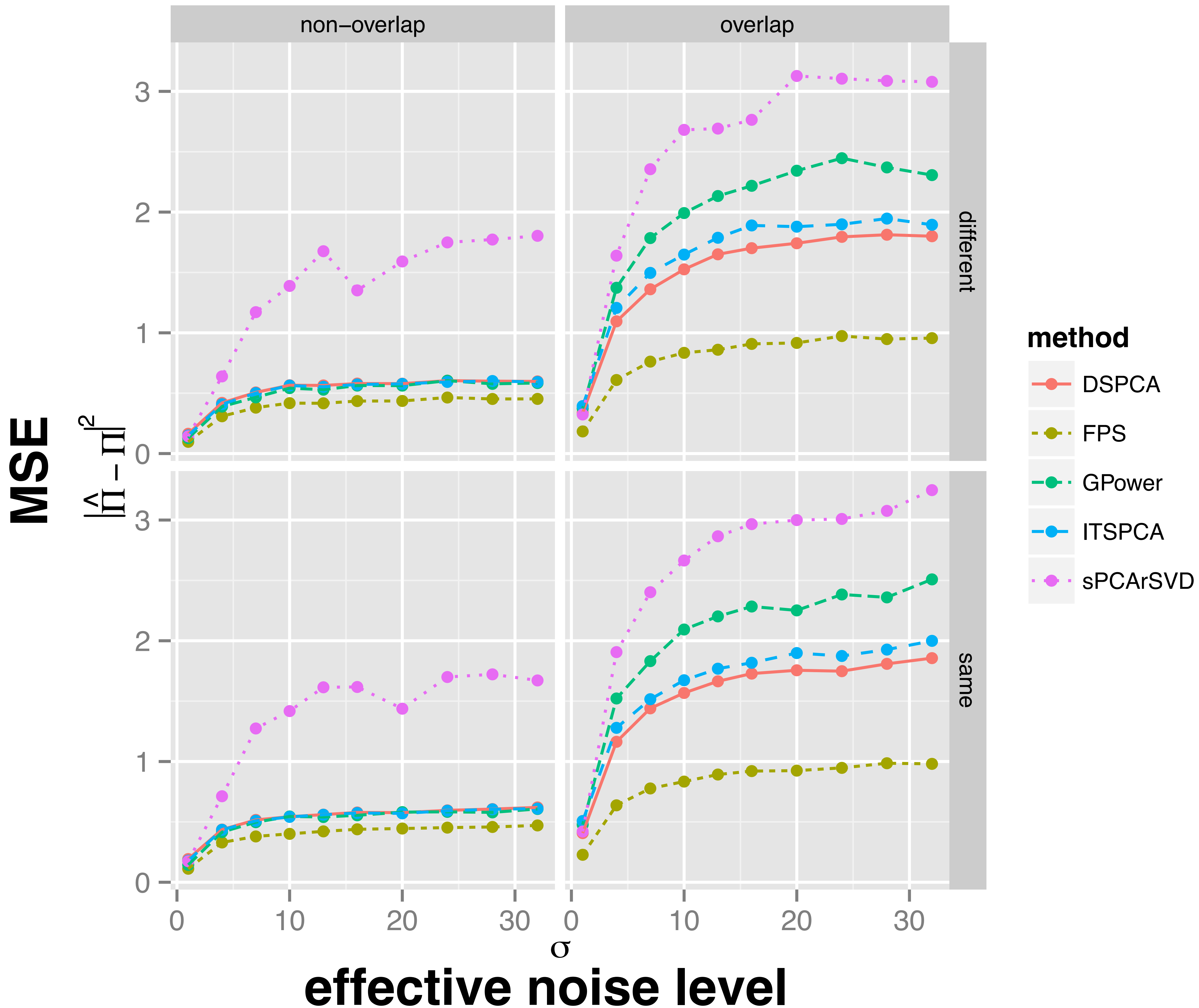
Examples

Simulation

- **n** = **256**, **p** = **512**, **k** = **6**
- True projection matrix: **s** = **30** non-zero rows
- Sparsity pattern: one **overlapping** block or three **non-overlapping** blocks
- Leading eigenvalues: all **same** or **different**
- Effective noise level $\boldsymbol{\sigma^2}$ varied by adjusting spectral gap
- Error criterion: MSE (averaged over 100 simulations)

Methods

- DSPCA (d'Aspremont et al. '07) (same as k=1 FPS)
- Variations on iterative thresholding / truncated power
  - GPower (Journée et al. '10)
  - ITSPCA (Ma '13)
  - sPCA-rSVD (Shen & Huang '08; Witten et al. '09)
- Tuning parameter selected by cross-validation


non−overlap
overlap
different
same
method
DSPCA
FPS
GPower
ITSPCA
sPCArSVD
MSE
$|\hat{\Pi} - \Pi|^2$
σ
effective noise level
0
1
2
3
0
10
20
30

Observations

- All methods about the same when spectral gap is huge (**noise ≈ 0**)
- All methods degrade as spectral gap decreases
- Iterative thresholding methods degrade substantially when sparsity pattern of eigenvectors overlap
- Iterative thresholding methods generally faster, but have larger error

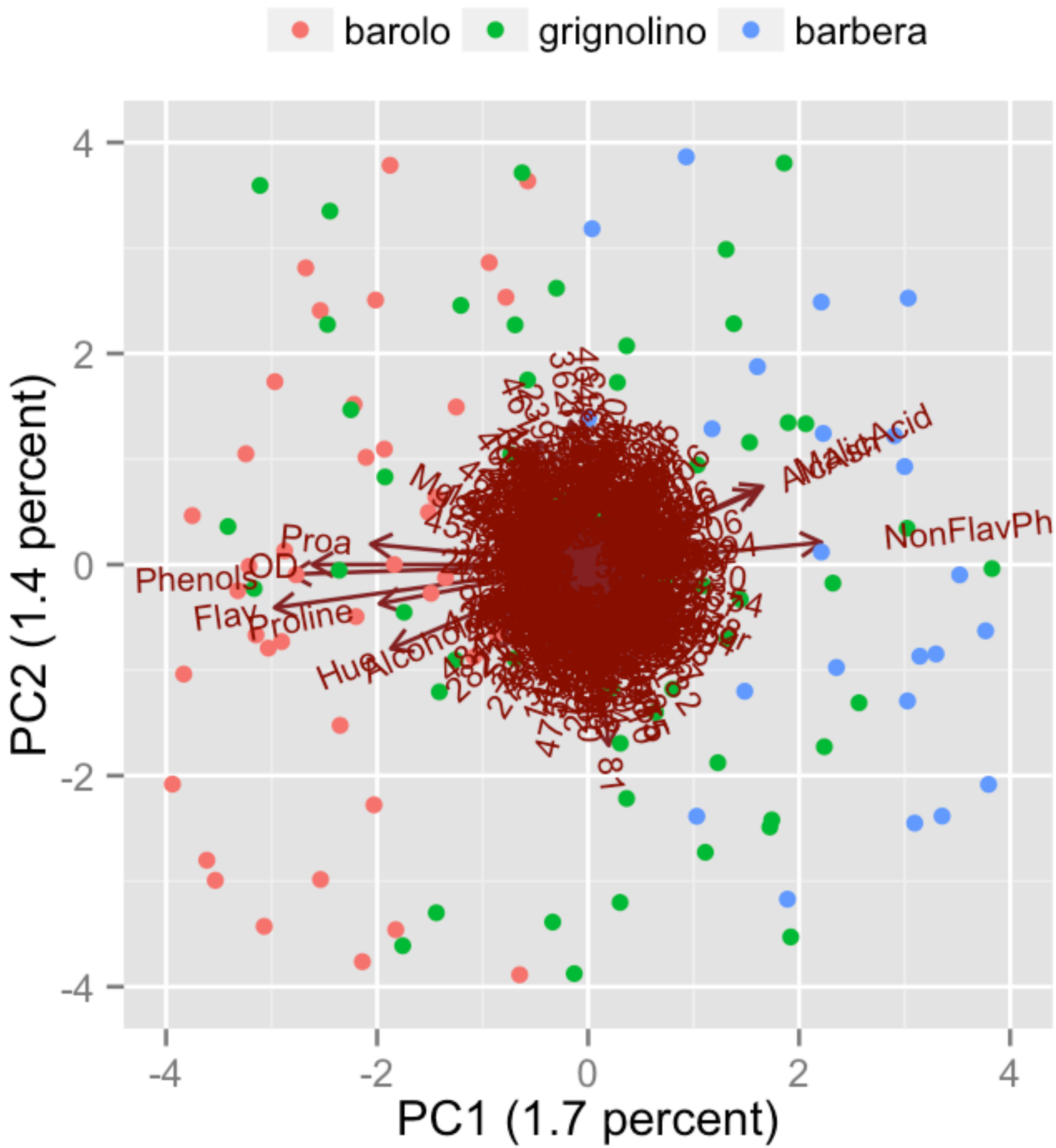
Wine data

- Data on **n=178** wines grown over a decade in the same region of Italy; Measurements on **s=13** constituents

- Divided into 3 different cultivars: Barolo, Grignolino, Barbera

- ***Synthetically enlarged*** by adding 487 **noise variables** by randomly, independently copying and permuting the real variables – resulting **p=500**

- Next movie shows **k=2** and effect of changing tuning parameter **ρ** from min to max (and back)



barolo
grignolino
barbera
PC1 (1.7 percent)
PC2 (1.4 percent)
Proa
OD
Phenols
Flav
Proline
Hue
Alcohol
NonFlavPh
MalicAcid
-4
-2
0
2
4

Sparsity?

Sparsity?

- Sparsity is a strong assumption
- Important questions
  - If sparsity is **true**, can we recover the sparsity pattern?
  - If sparsity is **false**, can we still interpret? **Yes** – see arXiv preprint

FPS is sparsistent

Theorem (LV '14)

Under the sparse PCA model, FPS is **unique** and **correctly selects** the relevant variables with high probability if

$n \gtrsim s^2 \log p$	sample complexity*
$\lVert \Gamma_{21} \rVert_{2 \to 1} \lesssim 1/s$	incoherence
$\min_{j \leq s} \Pi_{jj} \gtrsim s\sqrt{\log p / n}$	signal strength
$\rho \sim \sqrt{\log p / n}$	tuning parameter

(omitted constants depend on eigenvalues)

** minimax lower bound ~ s log p (Amini & Wainwright 2009)*

Summary

- Sparse PCA is an important topic – simultaneous **dimension reduction** and **variable selection**
- Convex relaxation is **nearly statistically optimal** and applicable to **general models** under weak conditions
- Consistent sparsity pattern recovery requires true sparsity and stronger conditions
- But **sparsity not necessary** for sparse PCA to be useful or for its theoretical justification

Thank you!

References

- VL ('12) "Minimax rates of estimation for sparse PCA in high dimensions." *AISTATS*
- VL ('13) "Minimax sparse principal subspace estimation in high dimensions." *Annals of Statistics*
- VCLR ('13) "Fantope projection and selection."*NIPS; extended version in preparation*
- LV ('14) "Sparsistency and agnostic inference in sparse PCA." *arXiv preprint; submitted*