

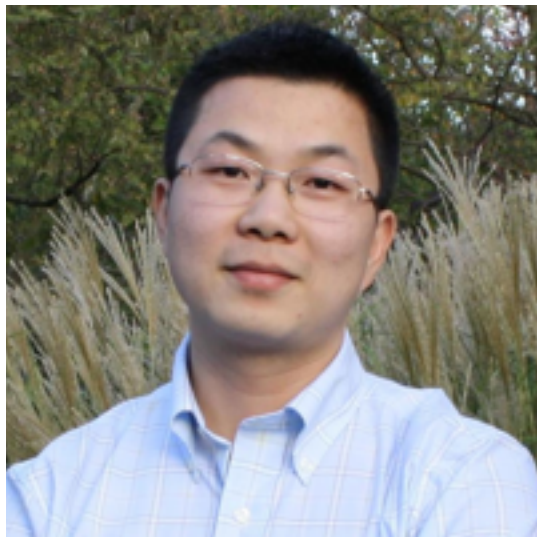
*Fantope Projection and Selection:*  
Near-optimal convex relaxation of  
Sparse PCA

**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.

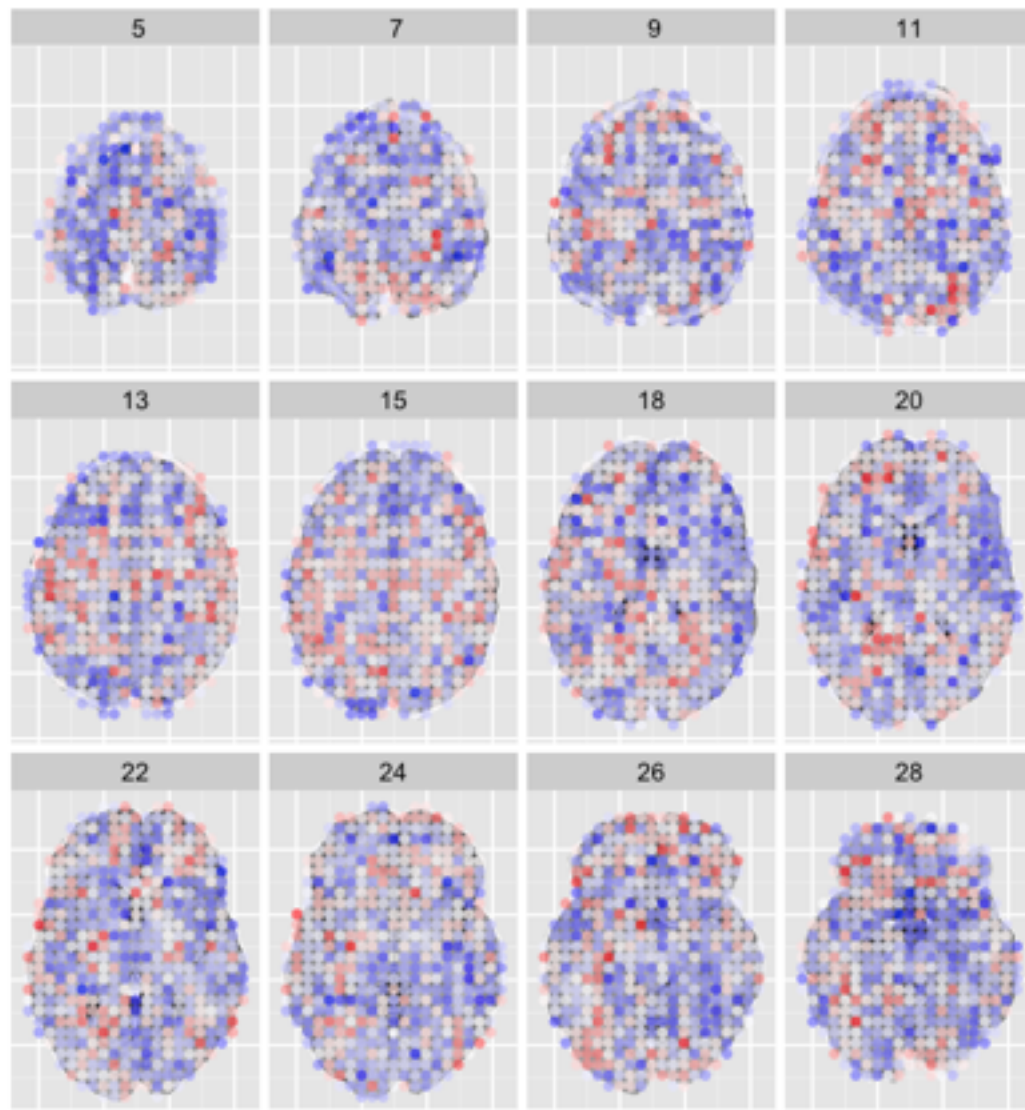


**Juhee Cho**  
U. Wisconsin-Madison



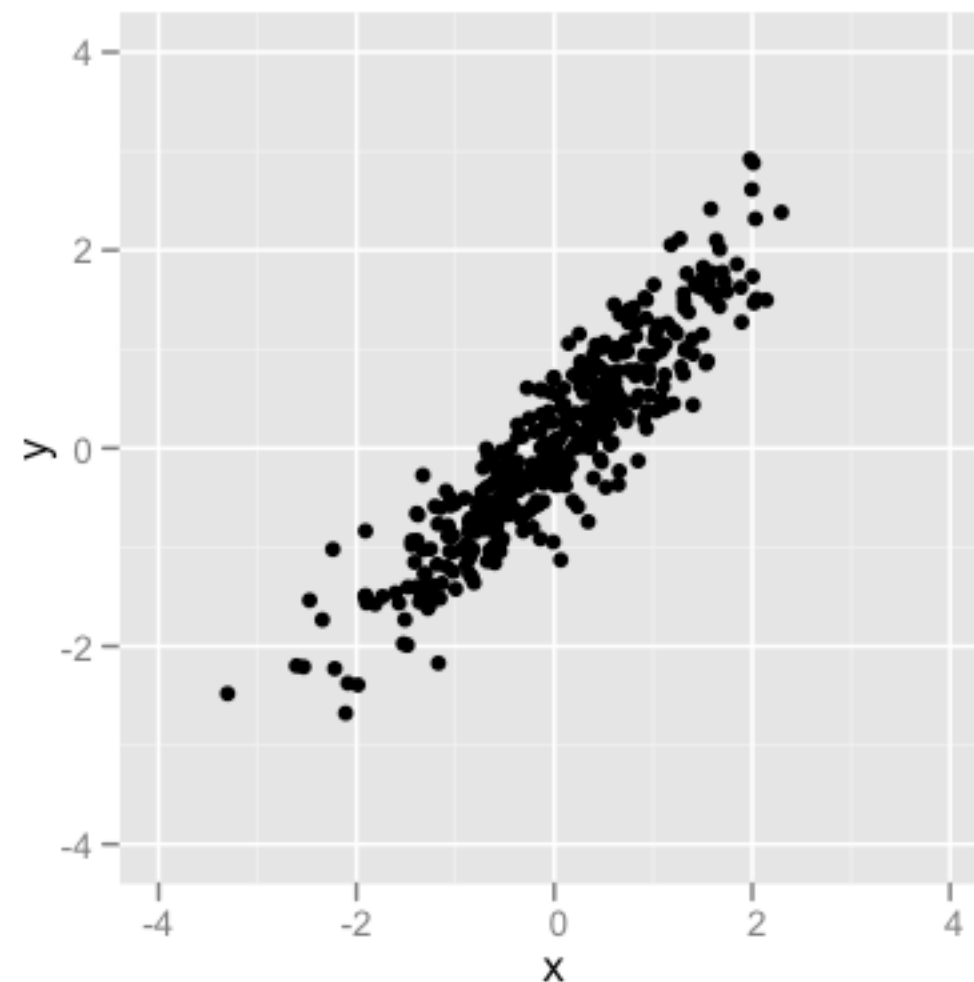
**Karl Rohe**  
U. Wisconsin-Madison

# Example: fMRI

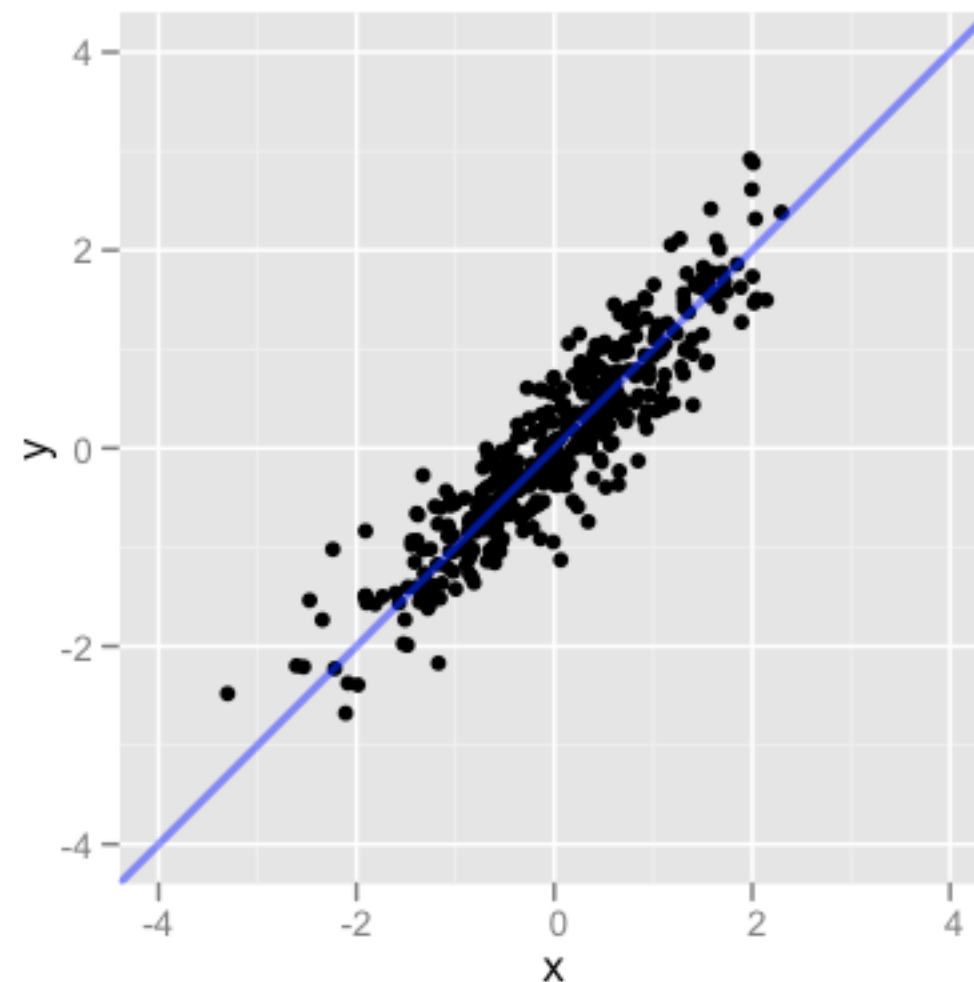


- $n \approx 10^2 \sim 10^3$  images
- $p \approx 10^5 \sim 10^6$  voxels
- Scientists interested in joint modeling of voxels
- But... challenging because of high-dimensionality
- **Dimension reduction** can be beneficial

# Principal Components Analysis

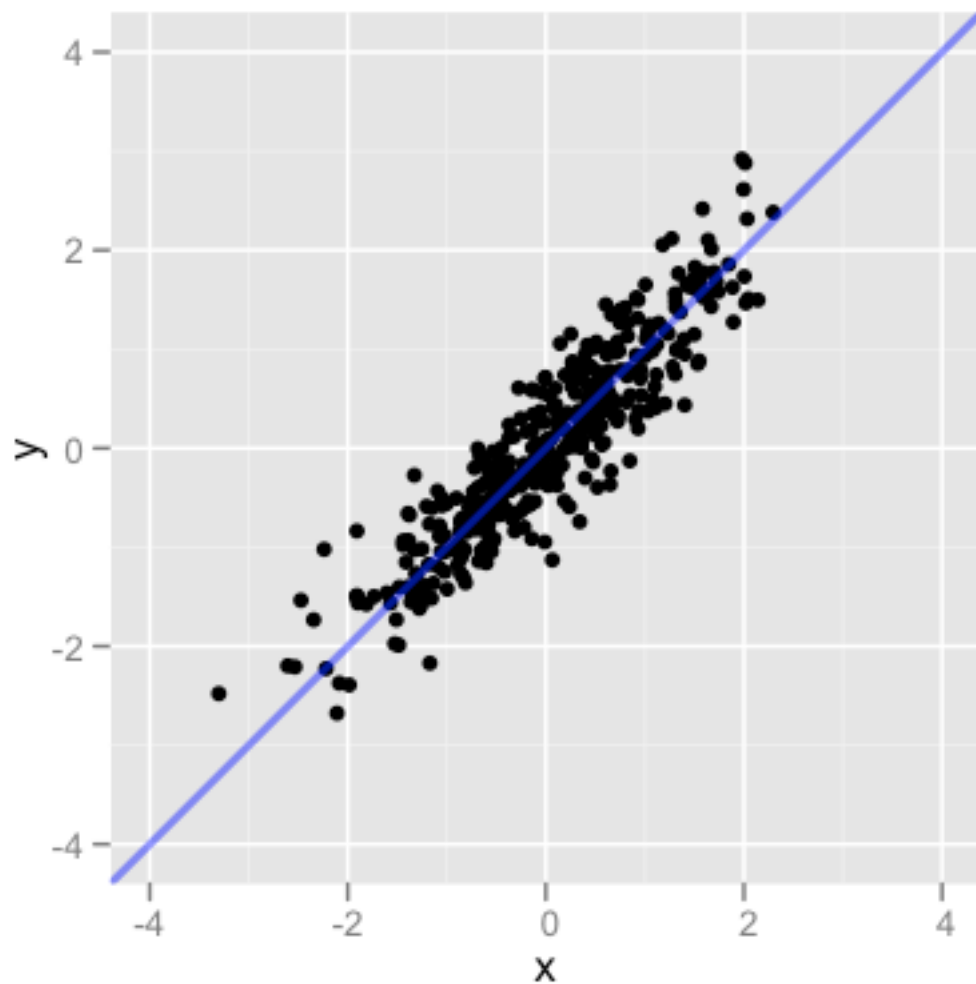


# Principal Components Analysis



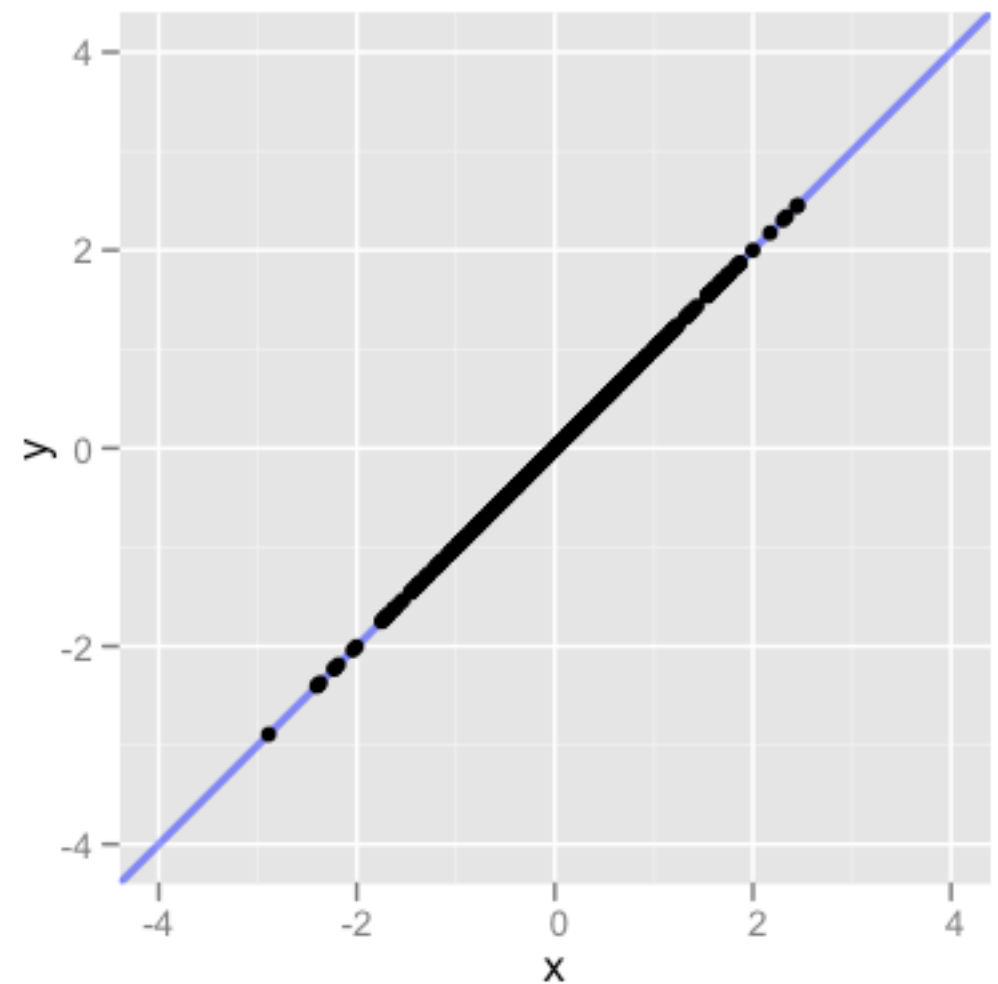
Pearson (1901)

# Principal Components Analysis



original data

$\approx$



lower-dimensional  
projection

# Principal Components Analysis

- Suppose  $\{X_1, X_2, \dots, X_n\}$  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  $\lambda_1 \geq \dots \geq \lambda_p$  and eigenvectors  $v_1, \dots, v_p$

- “Optimal” d-dimensional projection

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

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



# Classic PCA estimate

- Sample covariance and its eigendecomposition

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

- PCA estimate of d-dimensional projection

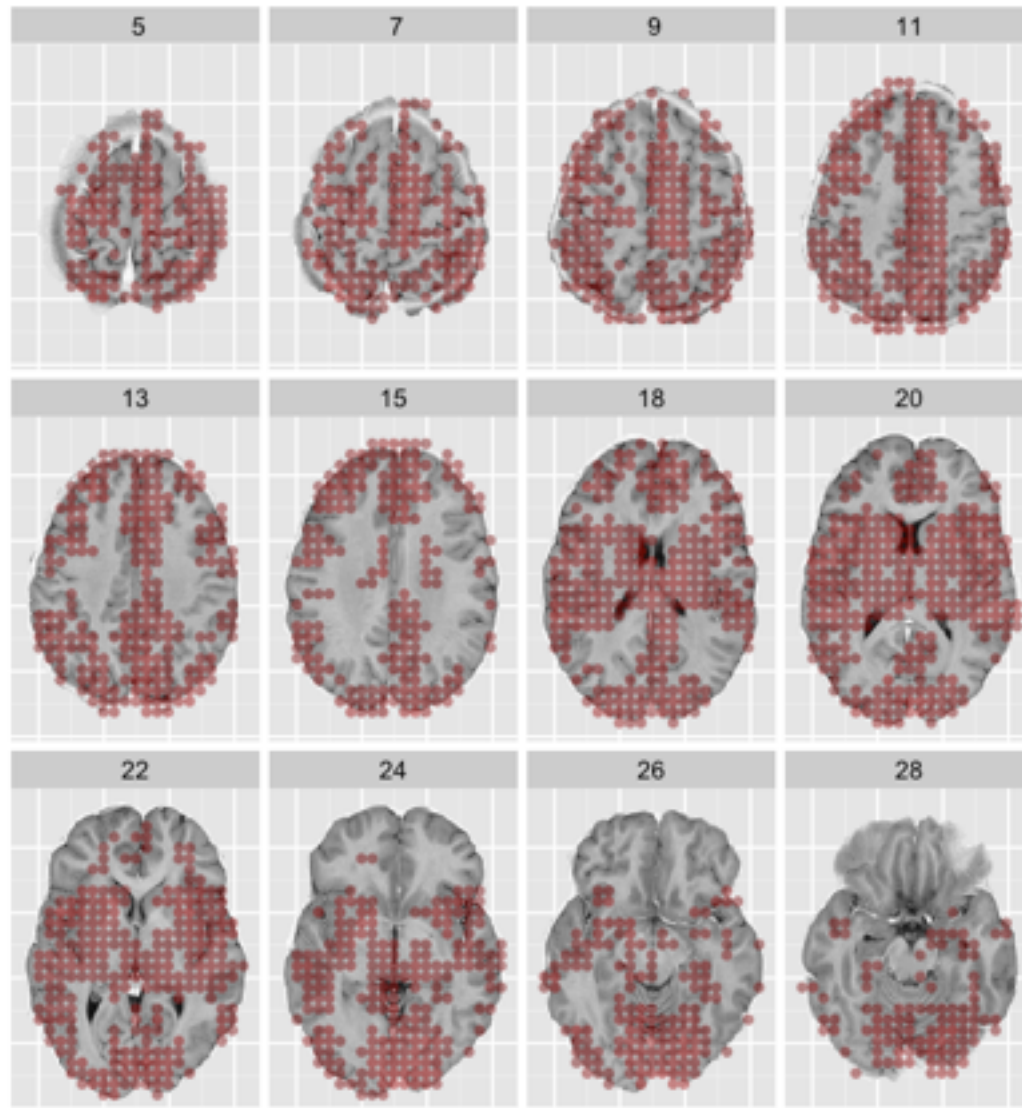
$$\hat{\Pi}_k = \hat{V}_k \hat{V}_k^T, \quad \hat{V}_k = (\hat{v}_1, \hat{v}_2, \dots, \hat{v}_k)$$

- Consistent (converges to truth) when **p** is fixed and **n**  $\rightarrow \infty$

# 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 sparsity?

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

# General sparse PCA model

$$\Sigma = \begin{array}{cc} \begin{matrix} s & p-s \end{matrix} \\ \begin{bmatrix} UDU^T & 0 \\ 0 & 0 \end{bmatrix} & + \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{bmatrix} \end{array}, \quad \Pi_k = \begin{bmatrix} UU^T & 0 \\ 0 & 0 \end{bmatrix}$$

**signal** **noise**

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

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

$$D = \text{diag}(\lambda_1, \dots, \lambda_k)$$

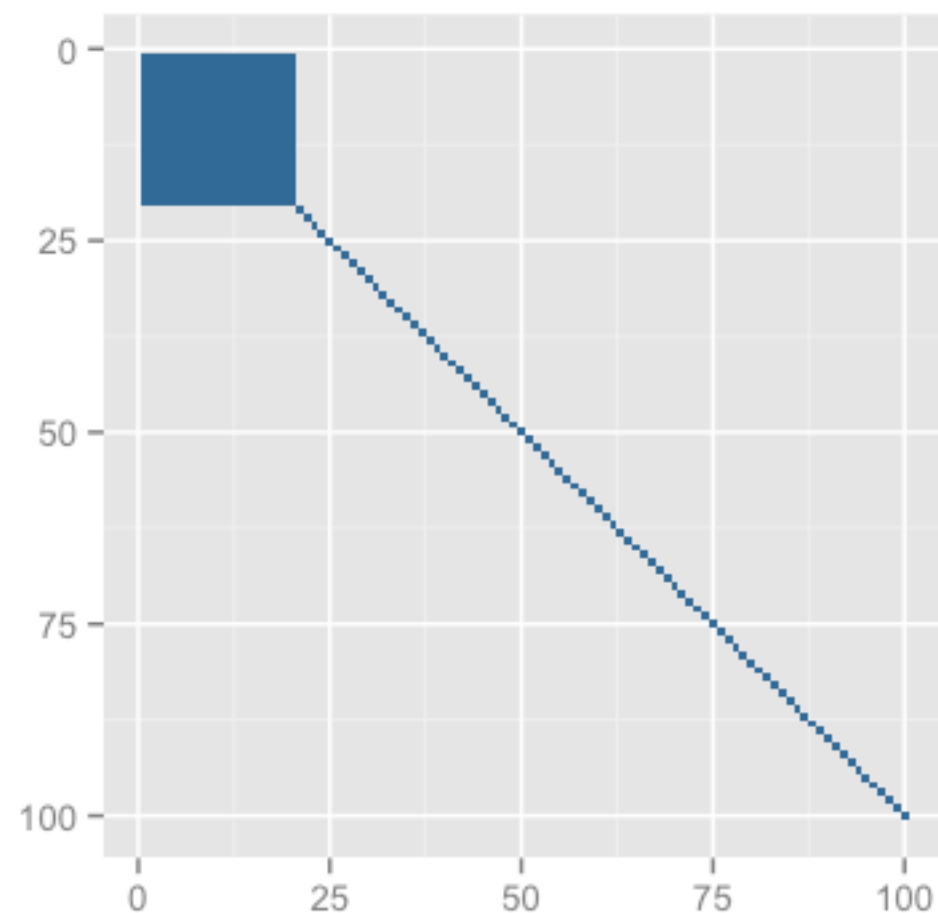
$$U = \text{nonzero block of } \mathbf{V}_k$$

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

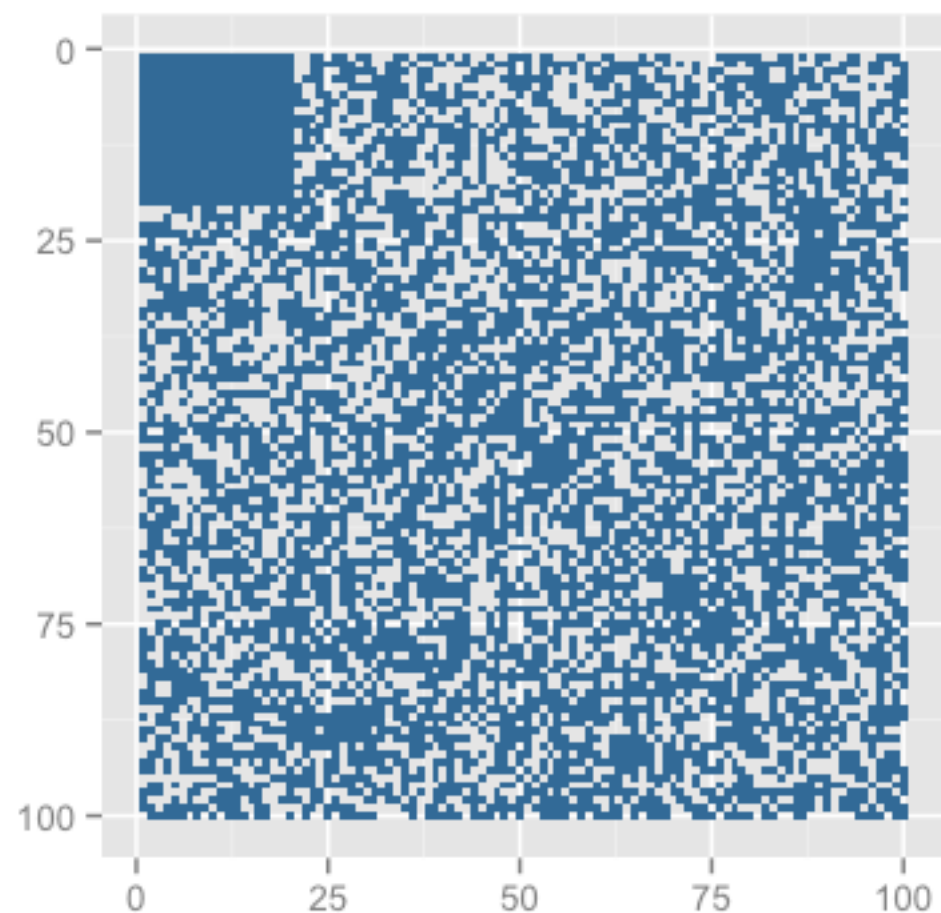


# Spiked model vs General model

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  $\Pi_k$  is

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

and can be achieved by

$$\hat{\Pi}_k = \arg \max_{\Pi} \text{trace}(\hat{\Sigma} \Pi)$$

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  $\Gamma$  (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)

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

**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  $\rho=0$ ; 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 ( $\approx$  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  $\rho$

# FPS is near-optimal

**Theorem** (VLCR '13)

Under the sparse PCA model\*, if

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

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

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

\* 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

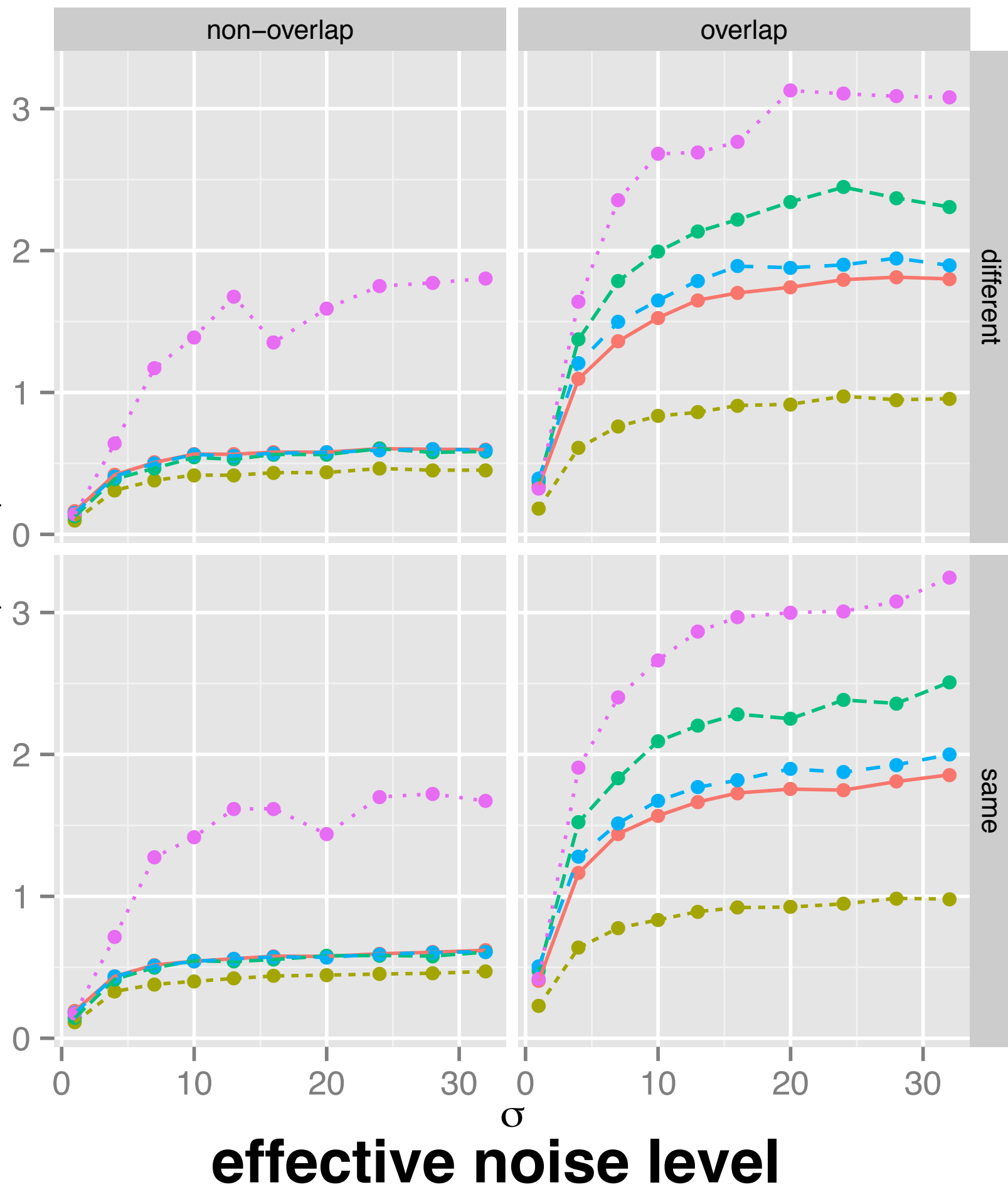
- $\mathbf{n} = 256$ ,  $\mathbf{p} = 512$ ,  $\mathbf{k} = 6$
- True projection matrix:  $\mathbf{s} = 30$  non-zero rows
- Sparsity pattern: one **overlapping** block or three **non-overlapping** blocks
- Leading eigenvalues: all **same** or **different**
- Effective noise level  $\sigma^2$  varied 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

**MSE**

$$|\hat{\Pi} - \Pi|^2$$

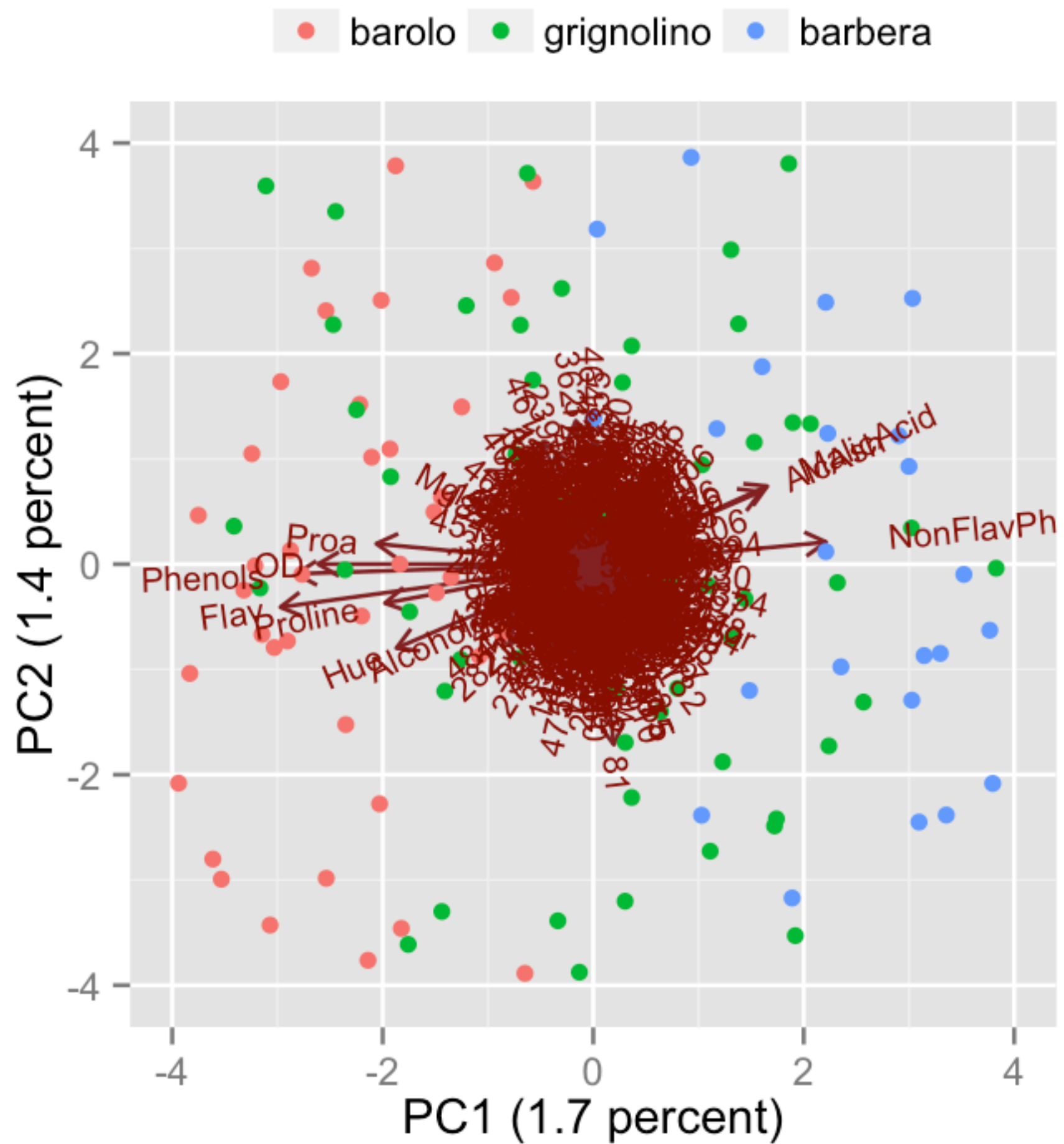


# Observations

- All methods about the same when spectral gap is huge (**noise  $\approx 0$** )
- 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 **p** from min to max (and back)



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*
$\ \Gamma_{21}\ _{2 \rightarrow 1} \lesssim 1/s$	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*  $\sim s \log p$  (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*