Sparse Principal Components and Subspaces

Concepts, Theory, and Computation

Vincent Q. Vu

Department of Statistics
The Ohio State University

October 2, 2013

This talk is based on joint work with...

Jing Lei
Carnegie Mellon U.

Juhee Cho
U. Wisconsin-Madison

Karl Rohe
U. Wisconsin-Madison

Outline

- Background on PCA and high-dimensions
- Sparsity of the leading eigenvector
- Consistent estimation and minimax theory
- Sparse principal subspaces
- Computationally tractable estimation

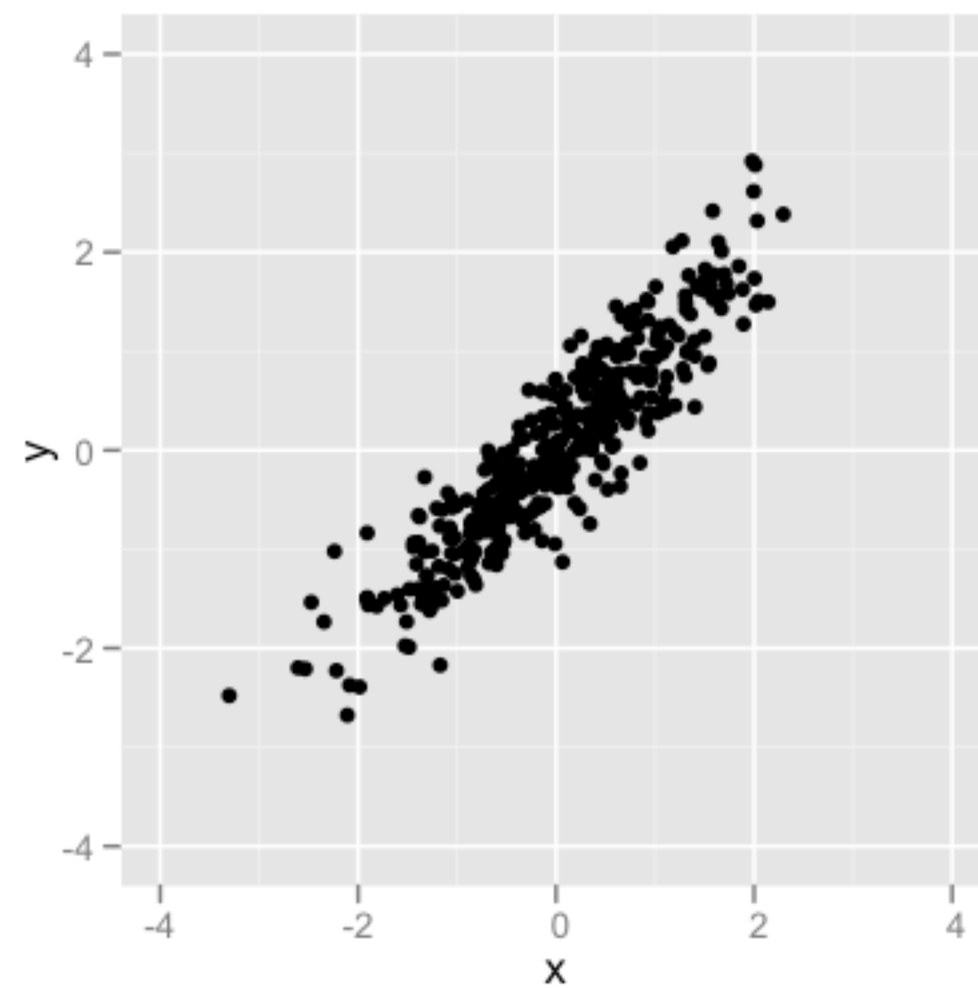
High-Dimensional PCA

"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."

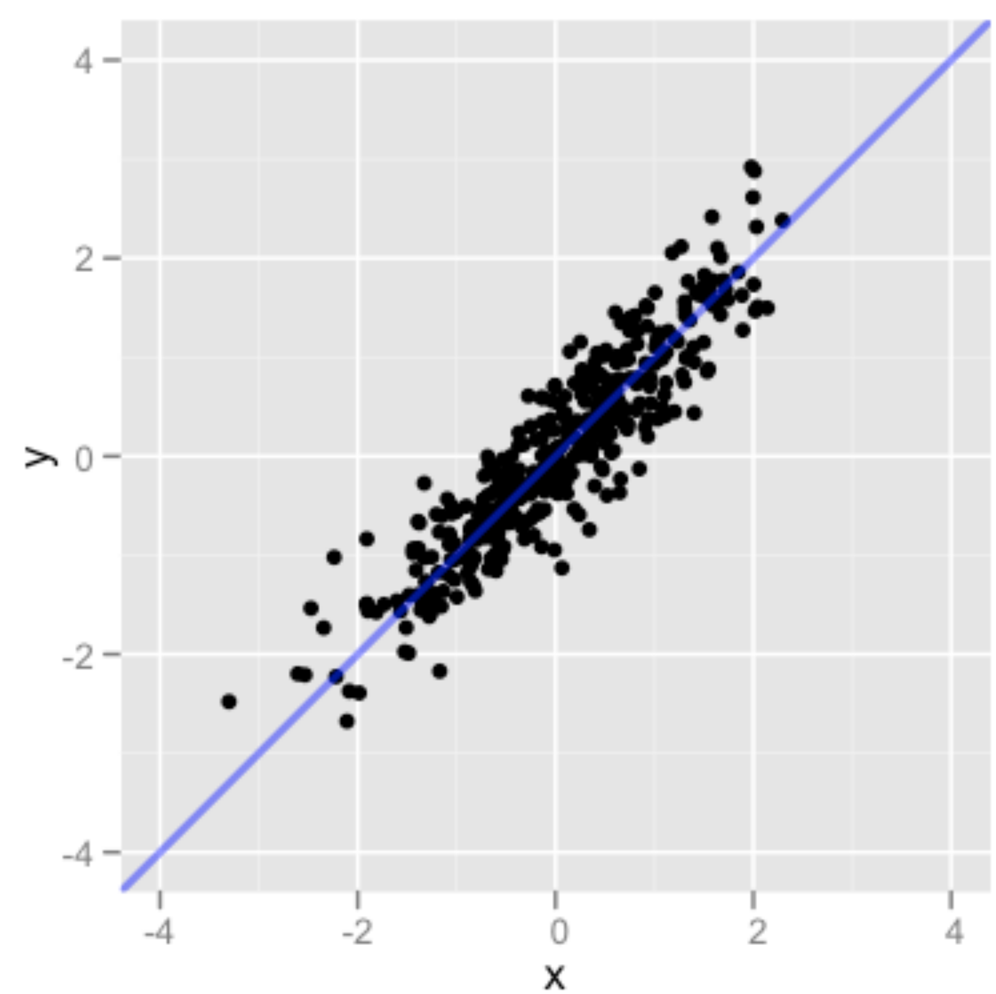
– **Karl Pearson** (1901)

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

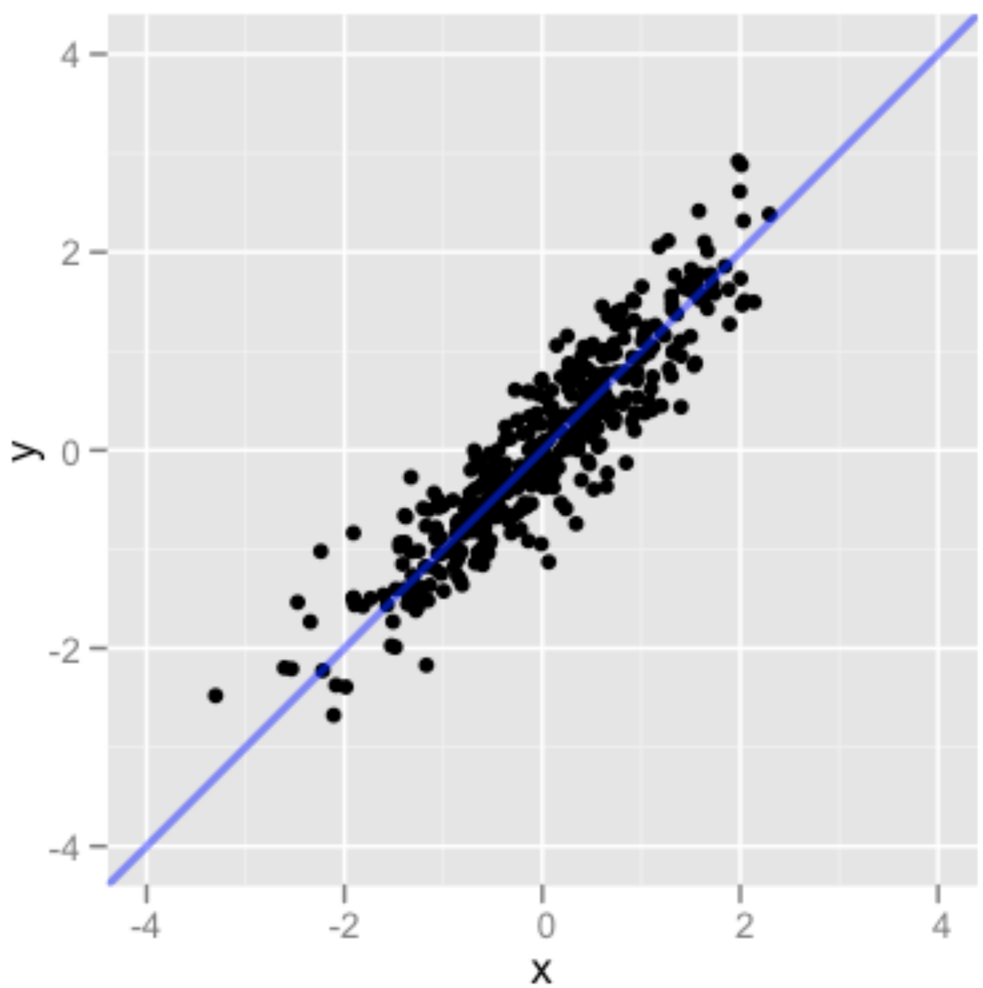
Principal Components Analysis



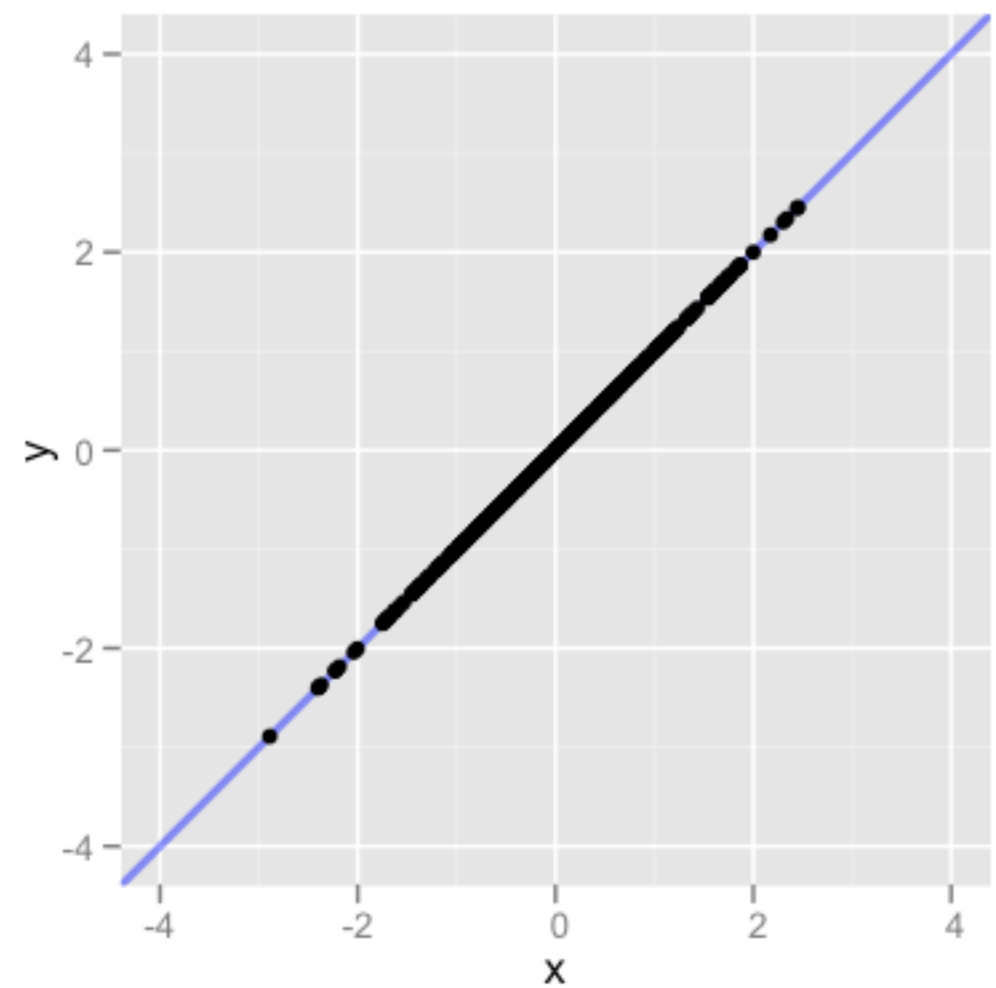
Principal Components Analysis



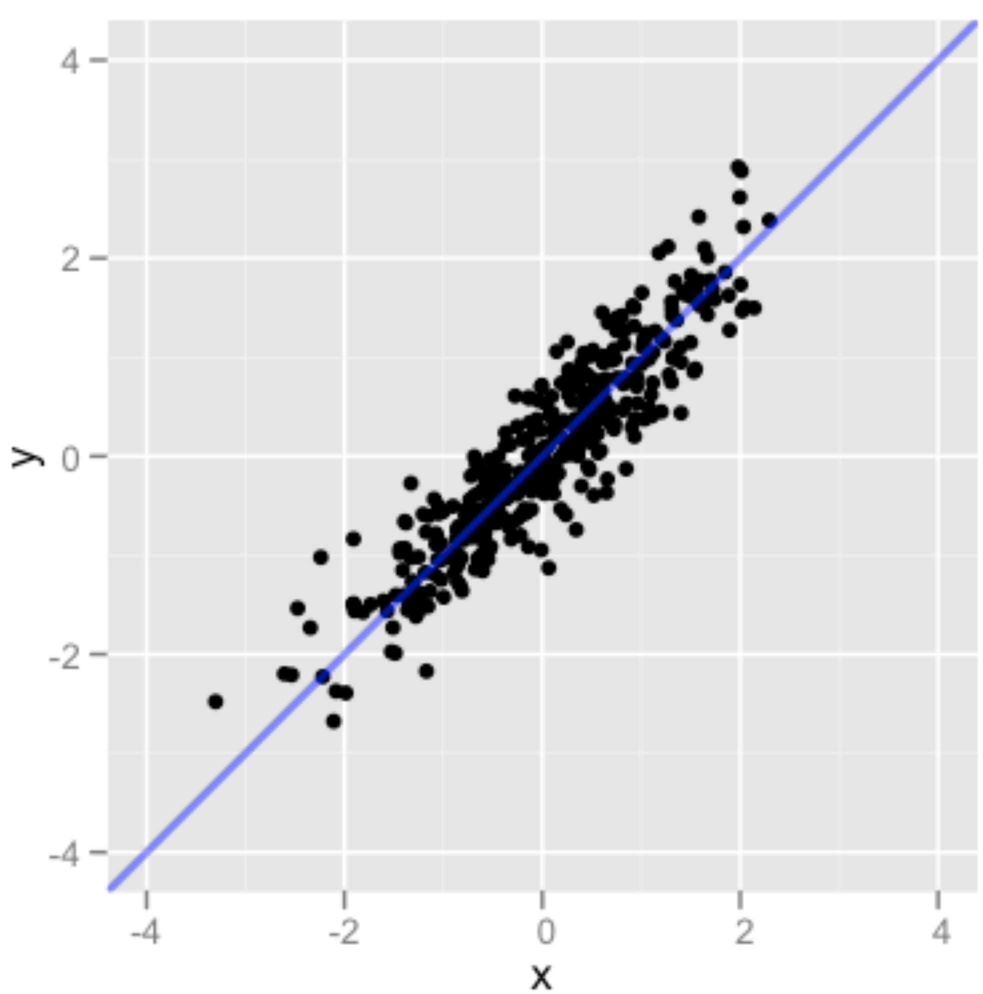
Principal Components Analysis



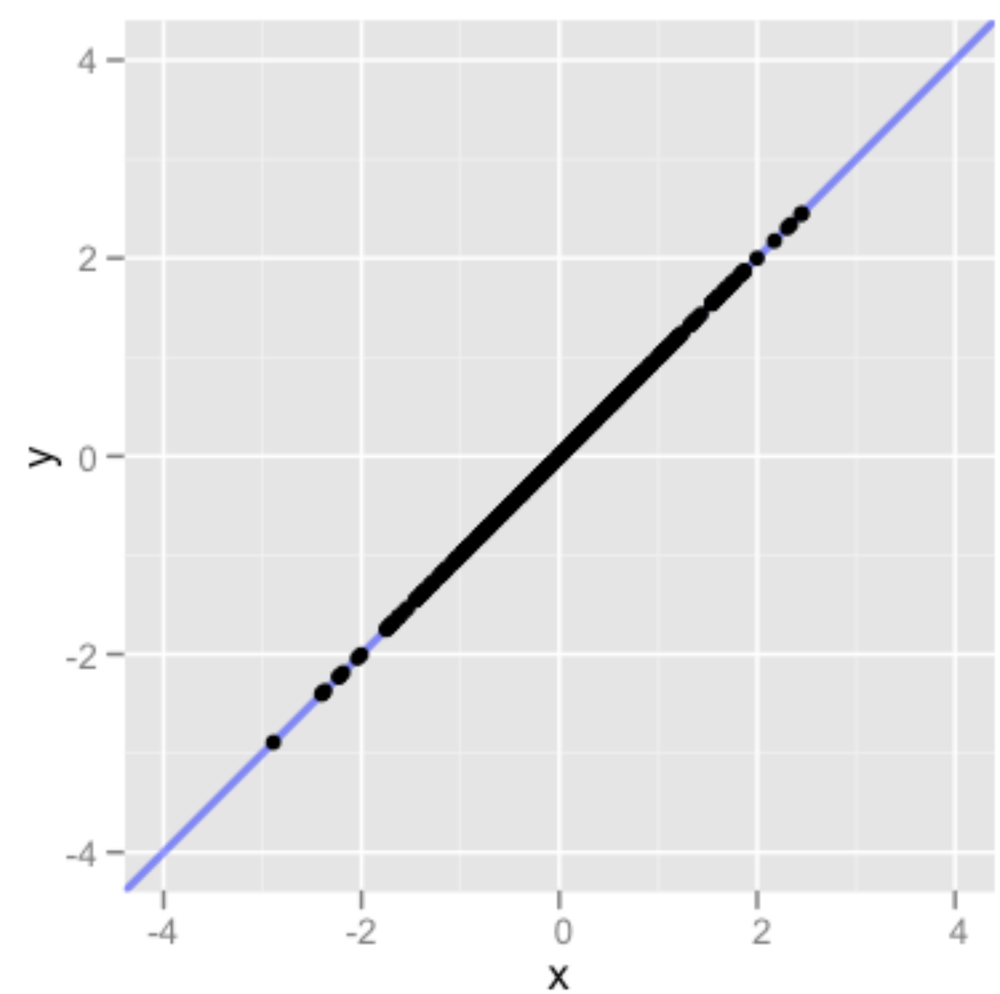
≈



Principal Components Analysis



≈



original data

lower-dimensional projection

Principal Components Analysis

- Suppose $\{ X_1, X_2, ..., X_{\mathbf{n}} \}$ is a dataset of i.i.d. observations on **p** variables
- **p** is ***large***, so **PCA** could be used for dimension reduction

Principal Components Analysis

"Optimal" dimension reduction is determined by **eigenvectors** of the **population covariance matrix**:

$$\Sigma \equiv \mathbb{E}(XX^T)$$

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

Principal Components Analysis

Eigendecomposition

$$\Sigma = V \Lambda V^T = \lambda_1 v_1 v_1^T + \cdots + \lambda_p v_p v_p^T$$

$$\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p),\ \lambda_1 \geq \cdots \geq \lambda_p \geq 0$$ (*eigenvalues*)

$$V = (v_1, \ldots, v_p),\ V^T V = I_p$$ (*eigenvectors*)

Principal Components Analysis

Eigendecomposition

$$\Sigma = V\Lambda V^T = \lambda_1 v_1 v_1^T + \cdots + \lambda_p v_p v_p^T$$

$$\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p),\ \lambda_1 \geq \cdots \geq \lambda_p \geq 0$$ (*eigenvalues*)

$$V = (v_1, \ldots, v_p),\ V^T V = I_p$$ (*eigenvectors*)

Optimal projector

$$\Pi_1 = v_1 v_1^T$$ (*rank-1 projector*)

$$\Pi_d = V_d V_d^T,\ V_d = (v_1, \ldots, v_d)$$ (*rank-d projector*)

Classical PCA

Estimate eigenvectors by eigendecomposition of **sample covariance matrix**:

$$\widehat{\Sigma} \equiv \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T$$

Classical PCA

Estimate eigenvectors by eigendecomposition of **sample covariance matrix**:

$$\widehat{\Sigma} \equiv \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T$$

Standard PCA estimator:

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

Classical PCA

Estimate eigenvectors by eigendecomposition of **sample covariance matrix**:

$$\widehat{\Sigma} \equiv \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T$$

Standard PCA estimator:

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

Standard theory for **p** <u>fixed</u> and **n**→∞:

$$\widehat{\Pi}_d \to \Pi_d \text{ a.s. if } \lambda_d - \lambda_{d+1} > 0$$

High-Dimensional PCA: Challenges

High-Dimensional PCA: Challenges

- In modern applications, *e.g.* neuroimaging, genetics: $\mathbf{p \approx n}$ and often $\mathbf{p >> n}$

High-Dimensional PCA: Challenges

- In modern applications, e.g. neuroimaging, genetics: **p≈n** and often **p>>n**
- ***Accuracy***: Standard PCA estimator can be inconsistent (Johnstone & Lu 2009):

$$\hat{v}_1^T v_1 \approx 0 \quad (\text{when } p/n \to c > 0, \lambda_1 - \lambda_2 \to c' > 0)$$

High-Dimensional PCA: Challenges

- In modern applications, *e.g.* neuroimaging, genetics: **p≈n** and often **p>>n**
- ***Accuracy***: Standard PCA estimator can be inconsistent (Johnstone & Lu 2009):

$$\hat{v}_1^T v_1 \approx 0 \quad (\text{when } p/n \to c > 0, \lambda_1 - \lambda_2 \to c' > 0)$$

- ***Interpretability***: PCA difficult to interpret when estimated projector depends on many variables

PCA in High-Dimensions

- Unconstrained estimation generally inconsistent
- Need additional **structural constraints** to have consistency
- What structural constraints make sense?

Sparsity

Sparsity

- Few variables have large effects – most others negligible

Sparsity

- Few variables have large effects – most others negligible
- Sometimes appropriate after *sparsifying* transformation

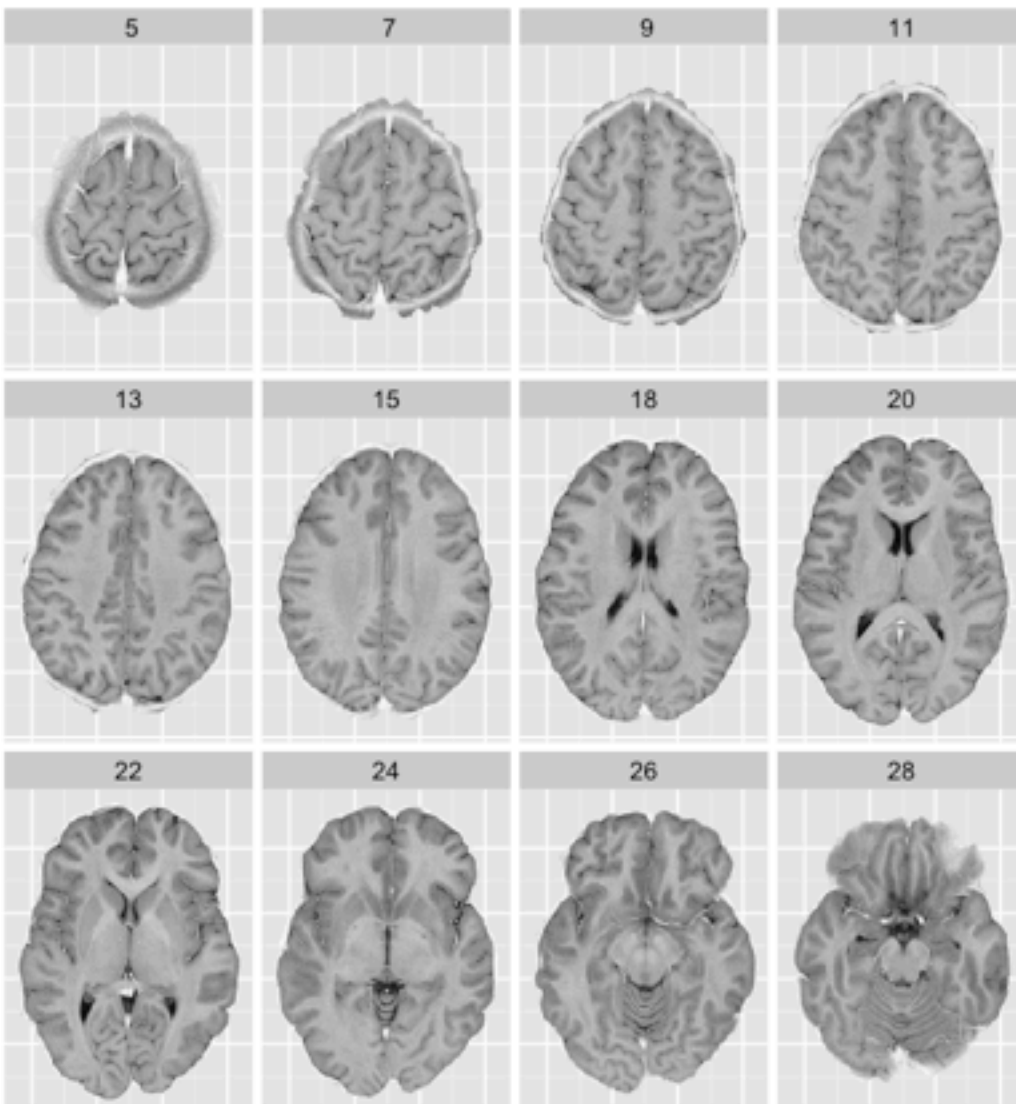
  – smoothness, localization, or periodicity correspond to sparsity in known bases: e.g. wavelets or Fourier

Sparsity

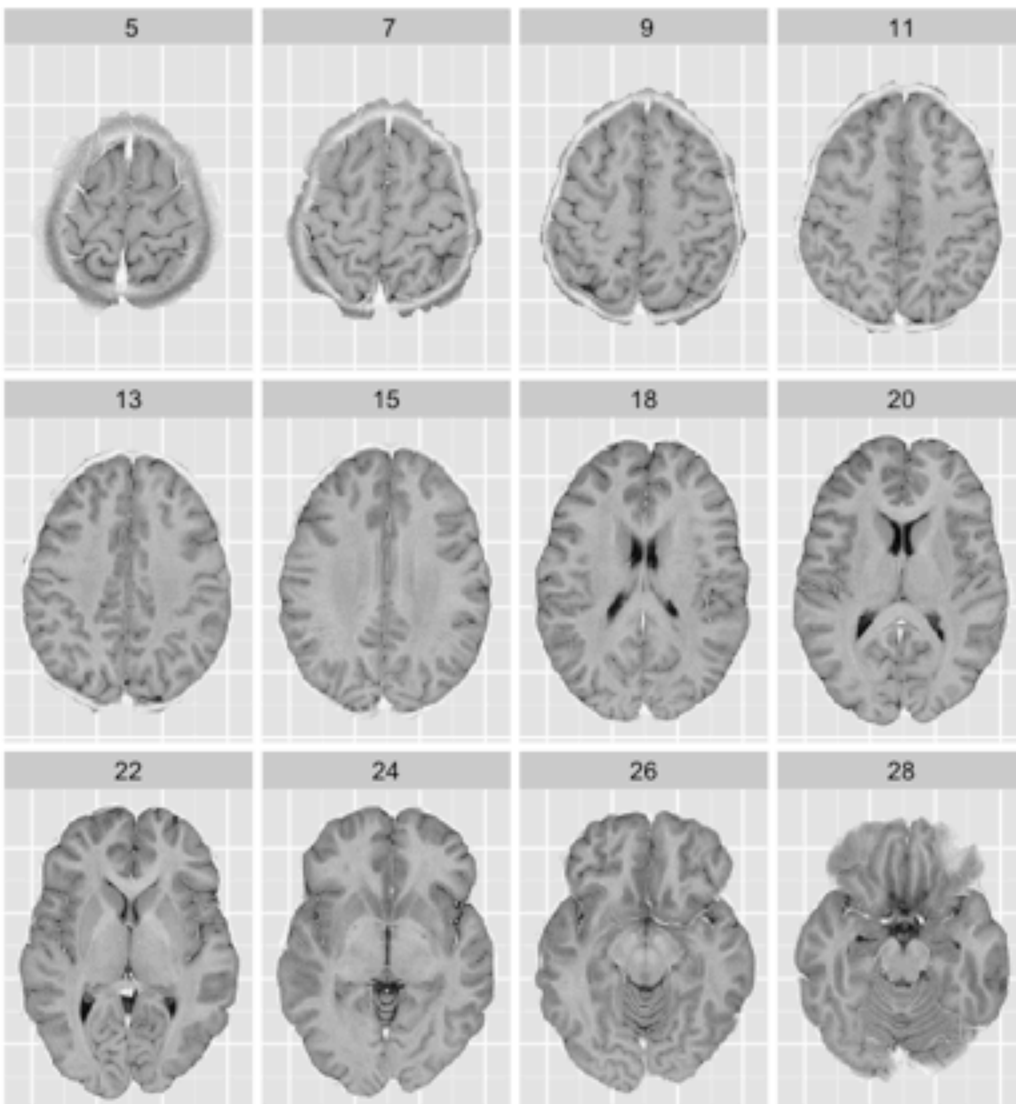
- Few variables have large effects – most others negligible
- Sometimes appropriate after *sparsifying* transformation

  – smoothness, localization, or periodicity correspond to sparsity in known bases: e.g. wavelets or Fourier
- Can make estimation feasible ***and*** enhance interpretability

Example: fMRI

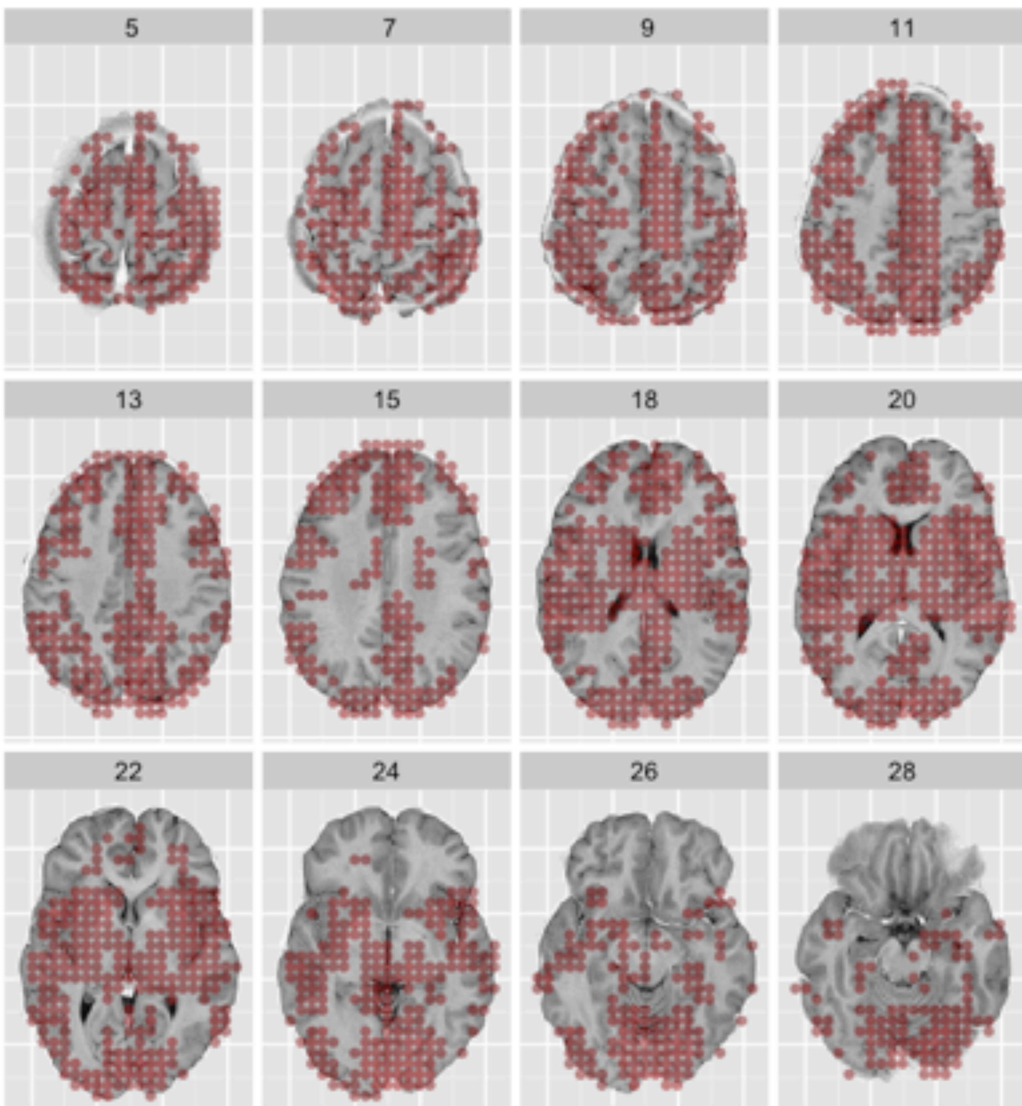


Example: fMRI



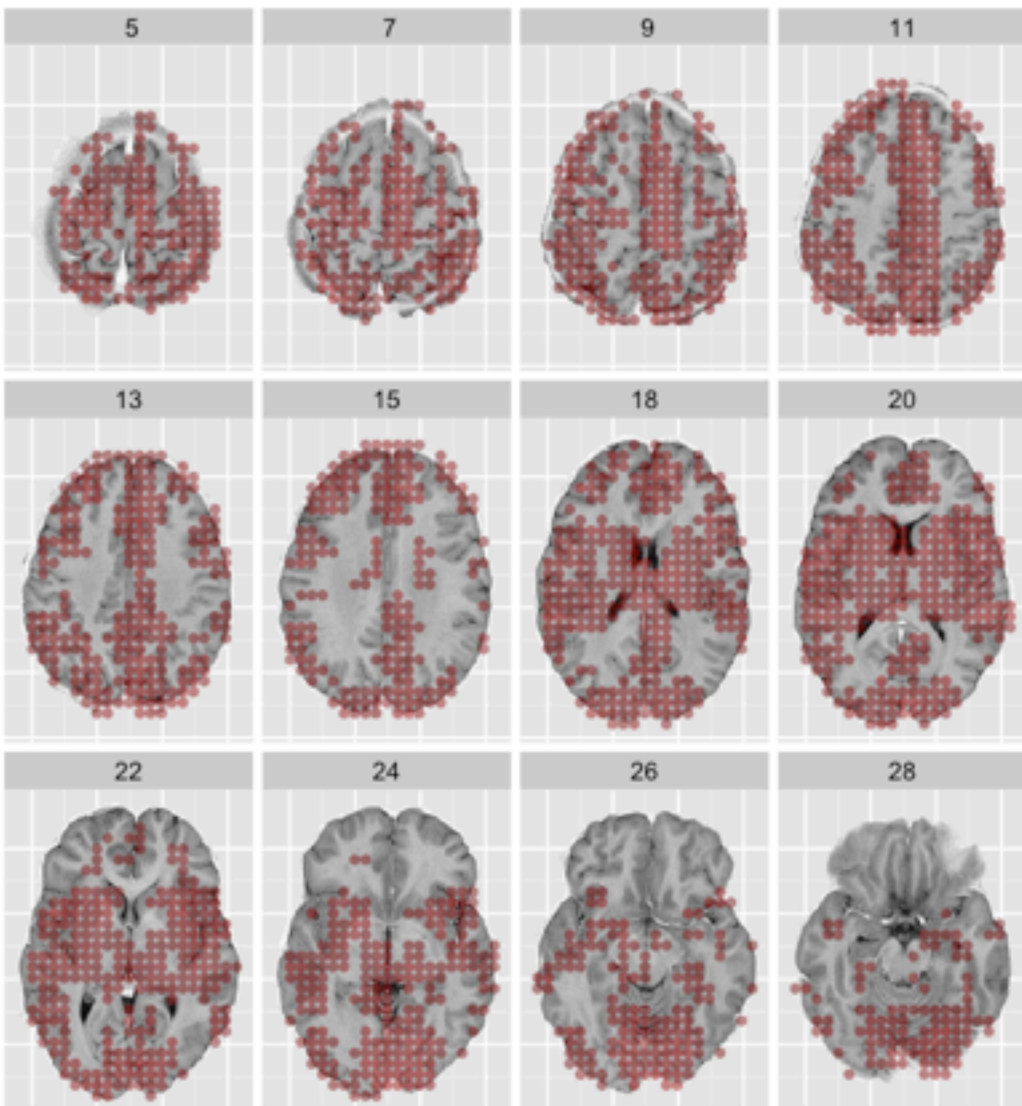
- **p** ≈ 10,000 ~ 40,000

Example: fMRI



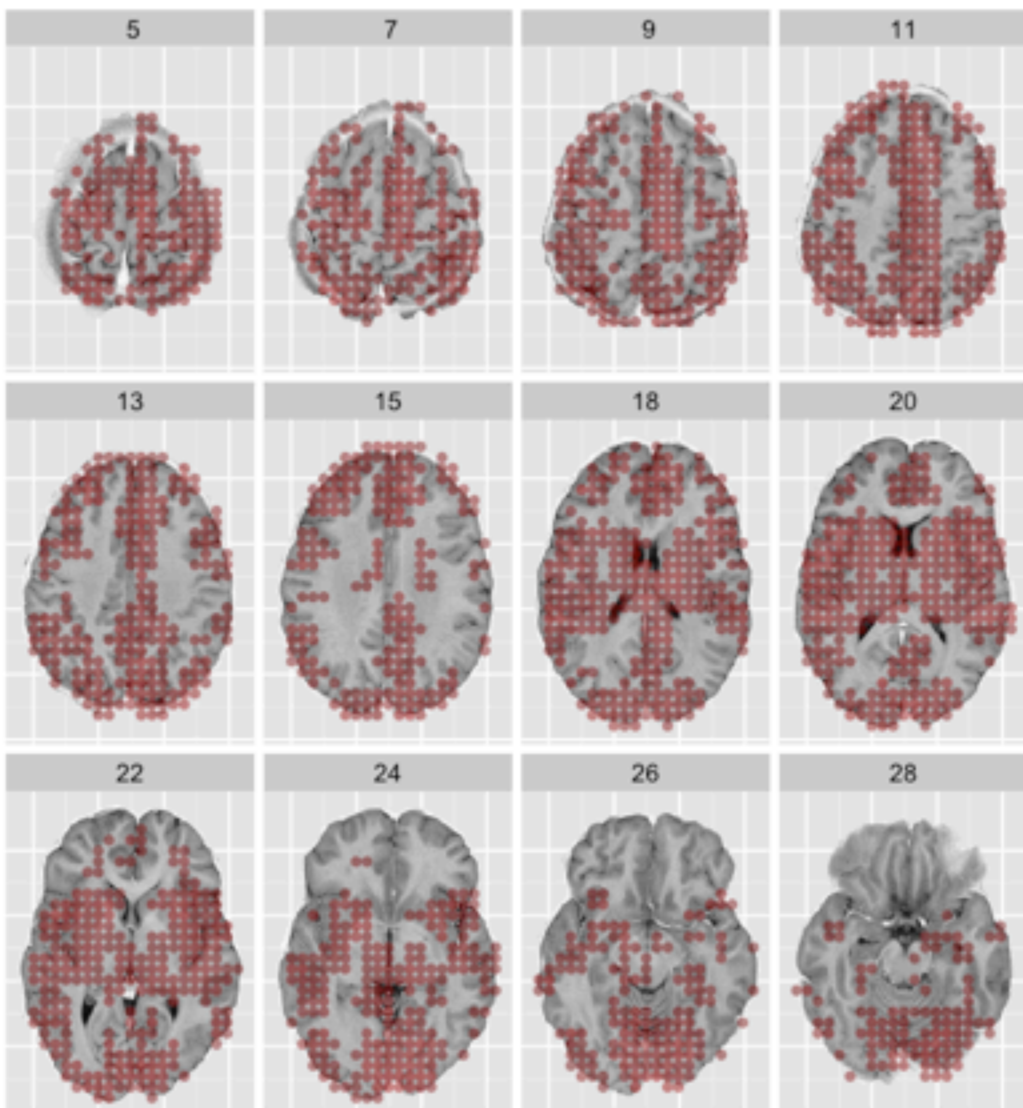
- **p** ≈ 10,000 ~ 40,000
- "Interesting" activity spatially localized

Example: fMRI



- **p** $\approx$ 10,000 ~ 40,000
- "Interesting" activity spatially localized
- Locations not known in advance

Example: fMRI



- **p** $\approx$ 10,000 ~ 40,000
- "Interesting" activity spatially localized
- Locations not known in advance
- Sparsity = spatial localization

Sparse PCA

- Many methods proposed over last 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 <u>many more</u>

Sparse PCA

- Many methods proposed over last 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 <u>many more</u>

- Mostly algorithmic proposals

Sparse PCA

- Many methods proposed over last 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 <u>many more</u>

- Mostly algorithmic proposals

- Few theoretical guarantees on statistical error – strong assumptions (spiked covariance model)

Sparse PCA framework

(d=1 case)

Sparse PCA framework
(d=1 case)

- **Goal**: Estimate principal eigenvector v_1

Sparse PCA framework
(d=1 case)

- **Goal**: Estimate principal eigenvector v_1
- Identifiability condition: $\lambda_1 > \lambda_2$

Sparse PCA framework
(d=1 case)

- **Goal**: Estimate principal eigenvector v_1
- Identifiability condition: $\lambda_1 > \lambda_2$
- Sparsity assumption: $\|v_1\|_0 \leq R_0 \ll p$

Sparse PCA framework
(d=1 case)

- **Goal**: Estimate principal eigenvector v_1
- Identifiability condition: $\lambda_1 > \lambda_2$
- Sparsity assumption: $\|v_1\|_0 \leq R_0 \ll p$
- Covariance assumption –
  - Spiked model: $\Sigma = (\lambda_1 - \lambda_2) v_1 v_1^T + \lambda_2 I_p$
  - General model: $\Sigma = \lambda_1 v_1 v_1^T + \cdots + \lambda_p v_p v_p^T$

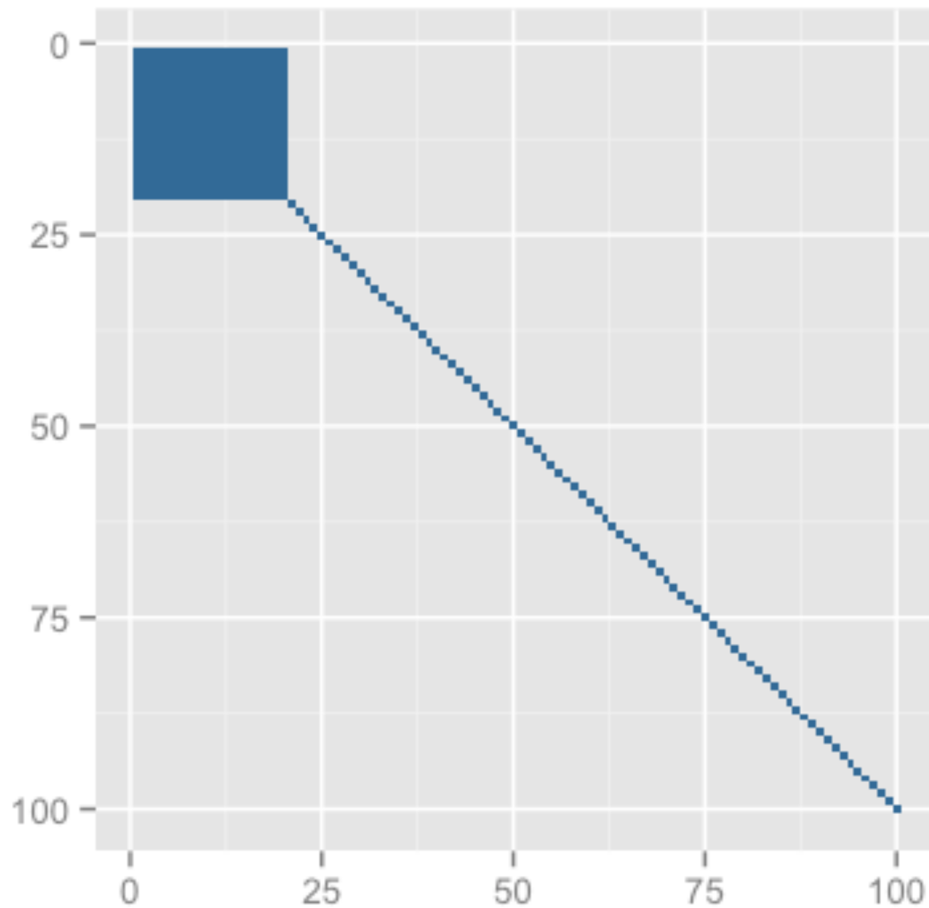
Sparse PCA framework
(d=1 case)

- **Goal**: Estimate principal eigenvector v_1
- Identifiability condition: $\lambda_1 > \lambda_2$
- Sparsity assumption: $\|v_1\|_0 \leq R_0 \ll p$
- Covariance assumption –
    - Spiked model: $\Sigma = (\lambda_1 - \lambda_2) v_1 v_1^T + \lambda_2 I_p$
    - General model: $\Sigma = \lambda_1 v_1 v_1^T + \cdots + \lambda_p v_p v_p^T$
- i.i.d. sub-Gaussian data

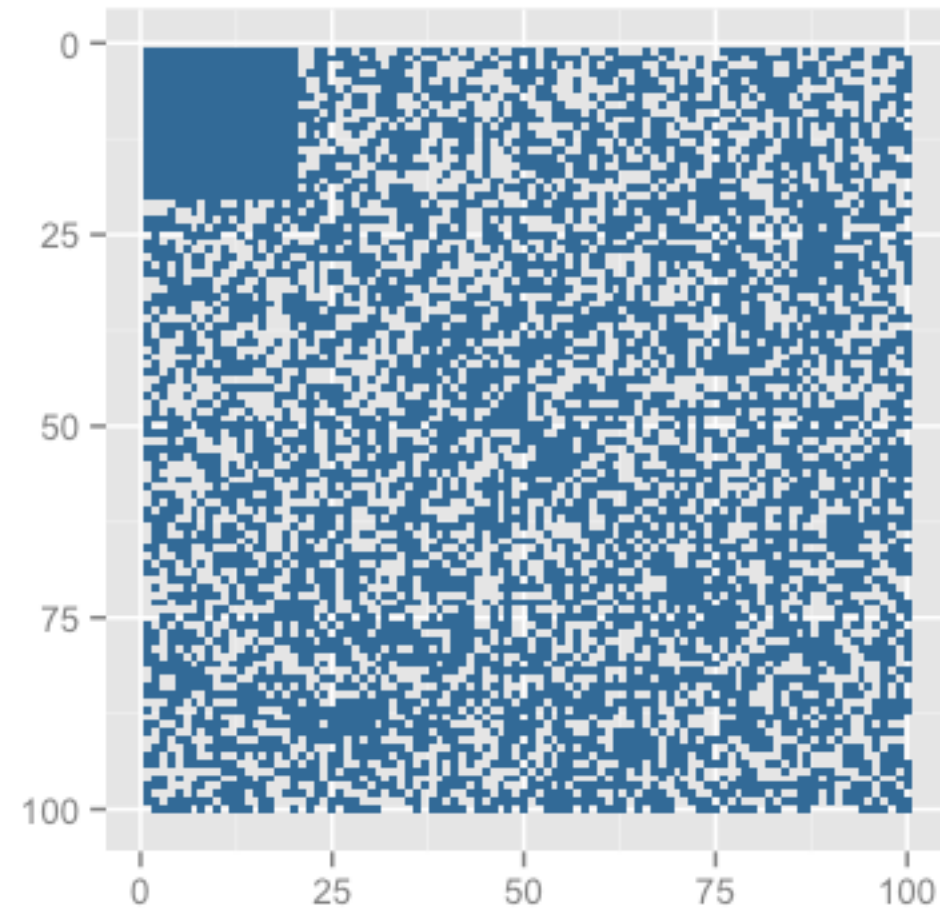
Spiked Model vs ***General Model***

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

$$\|v_1\|_0 = 20$$



Spiked model



General model

How does sparsity help PCA?

How does sparsity help PCA?

- **Questions**
  - How well can we estimate v_1 if sparsity assumed?
  - How do we estimate v_1 if sparsity assumed?

How does sparsity help PCA?

- **Questions**
  - How well can we estimate v_1 if sparsity assumed?
  - How do we estimate v_1 if sparsity assumed?
- **Intuition** – Estimation is easy if
  - R_0 small and λ_1-λ_2 large

How does sparsity help PCA?

- **Questions**
  - How well can we estimate v_1 if sparsity assumed?
  - How do we estimate v_1 if sparsity assumed?
- **Intuition** – Estimation is easy if
  - R_0 small and λ_1-λ_2 large
- Under **spiked model** (Johnstone & Lu 2003/9) give a consistent estimator of v_1 when $p/n \rightarrow c$

Minimax theory

(d = 1 case)

Minimax Framework

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

$$f(n, p, R_0, \lambda_1, \lambda_2) \lesssim \sup_{\Sigma} \mathbb{E}\|\hat{v}_1 - v_1\|_2^2$$

for all estimators $\hat{v}_1$ and a particular estimator $\hat{v}_1$ such that

$$\sup_{\Sigma} \mathbb{E}\|\hat{v}_1 - v_1\|_2^2 \lesssim f(n, p, R_0, \lambda_1, \lambda_2)$$

for all covariance matrices in the sparse PCA model.

Existing results for d=1

Under the spiked model Birnbaum et al. (2013) and Ma (2013) show that (***roughly***)

$$f(n, p, R_0, \lambda_1, \lambda_2) \sim \frac{R_0}{n} \cdot \frac{\lambda_1 \lambda_2}{(\lambda_1 - \lambda_2)^2} \cdot \log p$$

where the estimator is a thresholded power method (***up to log n factor***)

Existing results for d=1

Under the spiked model Birnbaum et al. (2013) and Ma (2013) show that (***roughly***)

$$f(n, p, R_0, \lambda_1, \lambda_2) \sim \frac{R_0}{n} \cdot \frac{\lambda_1 \lambda_2}{(\lambda_1 - \lambda_2)^2} \cdot \log p$$

where the estimator is a thresholded power method (***up to log n factor***)

Can we close the log term gap?
What about the general model?

Minimax Optimal Rate

Theorem (*V and Lei, 2013*)

Under the **general model**, the minimax error rate of estimating $\mathbf{v_1}$ is

$$f(n, p, R_0, \lambda_1, \lambda_2) \asymp \frac{R_0}{n} \cdot \frac{\lambda_1 \lambda_2}{(\lambda_1 - \lambda_2)^2} \cdot \log p$$

where the **exact rate** is achieved by

$$\hat{v}_1 = \operatorname*{arg\,max}_{\|v\|_2 = 1, \|v\|_0 \leq R_0} v^T \widehat{\Sigma} v$$

Consequences

- **Good news**

Consequences

- **Good news**
  - Exact minimax rate in $(n,p,R_0,\lambda_1, \lambda_2)$ for general model

Consequences

- **Good news**
  - Exact minimax rate in $(n,p,R_0,\lambda_1, \lambda_2)$ for general model
  - Extensions to L_q sparsity provide first consistency result for L_q constrained/penalized PCA (Joliffe et al 2003, Shen and Huang 2008, Witten et al 2009)

Consequences

- **Good news**
  - Exact minimax rate in $(n,p,R_0,\lambda_1, \lambda_2)$ for general model
  - Extensions to L_q sparsity provide first consistency result for L_q constrained/penalized PCA (Joliffe et al 2003, Shen and Huang 2008, Witten et al 2009)
- **Bad news**

Consequences

- **Good news**
  - Exact minimax rate in $(n,p,R_0,\lambda_1, \lambda_2)$ for general model
  - Extensions to L_q sparsity provide first consistency result for L_q constrained/penalized PCA (Joliffe et al 2003, Shen and Huang 2008, Witten et al 2009)
- **Bad news**
  - Estimator is computationally intractable (**NP**-hard in **p**)

Consequences

- **Good news**
  - Exact minimax rate in $(n,p,R_0,\lambda_1, \lambda_2)$ for general model
  - Extensions to L_q sparsity provide first consistency result for L_q constrained/penalized PCA (Joliffe et al 2003, Shen and Huang 2008, Witten et al 2009)
- **Bad news**
  - Estimator is computationally intractable (**NP**-hard in **p**)
  - Estimation not possible if $\lambda_1 \approx \lambda_2$

Multiple Eigenvectors?

- Most sparse PCA methods only estimate single eigenvectors, and are extended to multiple eigenvectors by iterative deflation

Multiple Eigenvectors?

- Most sparse PCA methods only estimate single eigenvectors, and are extended to multiple eigenvectors by iterative deflation
- Iterative deflation methods are heuristic and can be suboptimal (Mackey 2009)

Multiple Eigenvectors?

- Most sparse PCA methods only estimate single eigenvectors, and are extended to multiple eigenvectors by iterative deflation
- Iterative deflation methods are heuristic and can be suboptimal (Mackey 2009)
- If $\lambda_1 \approx \lambda_2$, then it makes less sense to think about *distinct* eigenvectors

Sparse Principal Subspaces
($d \geq 1$ case)

Sparse Principal Subspaces

- **Identifiability** – If $\lambda_1 = \lambda_2 = \ldots = \lambda_d$ then impossible to distinguish V_d and V_dQ from the data for any orthogonal Q.

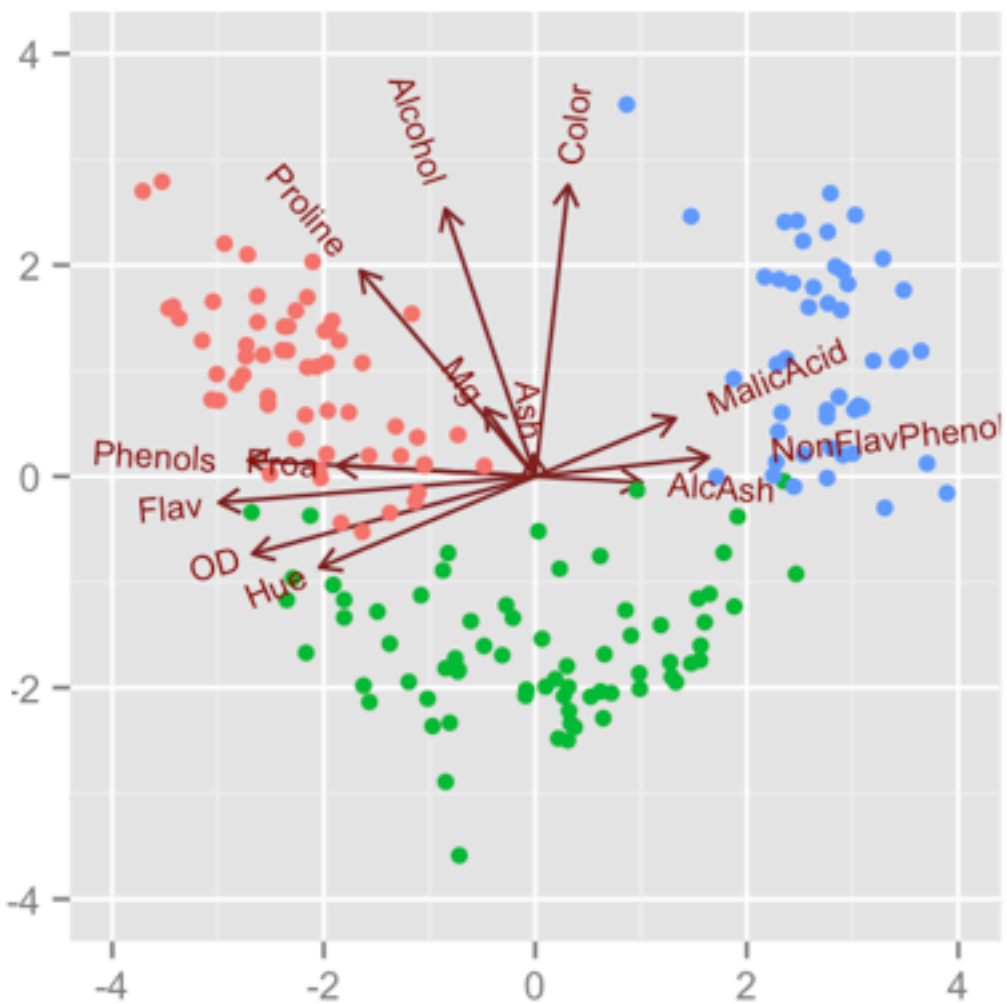
Sparse Principal Subspaces

- **Identifiability** – If $\lambda_1 = \lambda_2 = \ldots = \lambda_d$ then impossible to distinguish V_d and V_dQ from the data for any orthogonal Q.
- **Sparsity** – How to extend sparsity to subspaces? Good notion of sparsity should be **rotation invariant**

Sparse Principal Subspaces

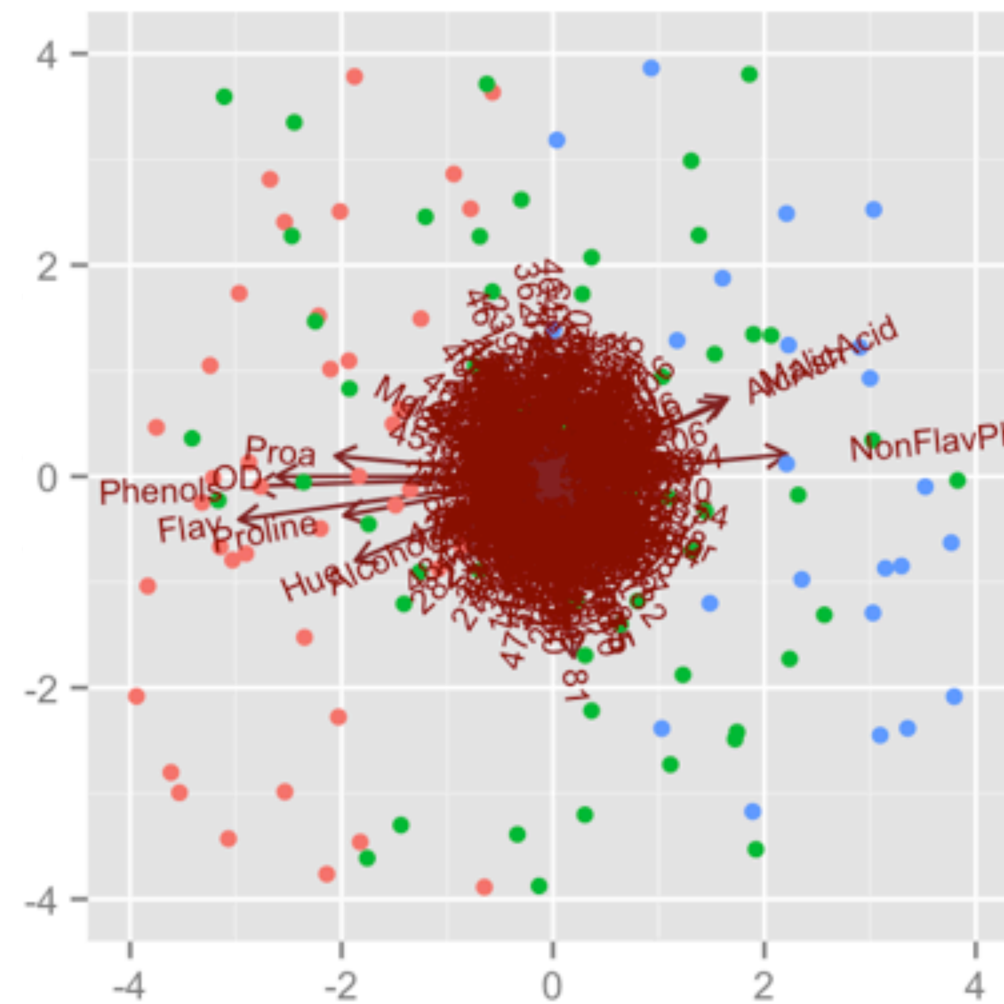
- **Identifiability** – If $\lambda_1 = \lambda_2 = \ldots = \lambda_d$ then impossible to distinguish V_d and V_dQ from the data for any orthogonal Q.
- **Sparsity** – How to extend sparsity to subspaces? Good notion of sparsity should be **rotation invariant**
- **Intuitively** – A subspace is sparse if its projector depends on a small number of variables

Sparse



Projection depends on 13 variables

Not sparse



Projection depends on 500 variables

Row sparsity

- Matrix (2,0)-norm – for any *p* × *d* matrix

$$\|V\|_{2,0} = \ \# \text{ of nonzero rows in } V$$

- Row sparsity:

$$\|V_d\|_{2,0} \leq R_0 \ll p, \; V_d = (v_1, \ldots, v_d)$$

- Row sparsity is rotation invariant – for any orthogonal Q:

$$\|V_d\|_{2,0} = \|V_d Q\|_{2,0}$$

Subspace distance

- Measure distance between two subspaces with *canonical angles*

$$\|\sin\Theta(\hat{V}_d, V_d)\|_F^2 = \frac{1}{2}\|\hat{V}_d\hat{V}_d^T - V_dV_d^T\|_F^2$$

- Sum of squares of sines of canonical angles
- If d=1, equivalent to squared Euclidean distance

Minimax Optimal Rate (d≥1)

Theorem (*V and Lei, 2013*)

Under the **general model**, the minimax error rate of estimating $\mathbf{V_d}$ is

$$\min_{\hat{V}_d} \max_{\Sigma} \mathbb{E}\|\sin\Theta(\hat{V}_d, V_d)\|_F^2 \asymp R_0 \cdot \frac{\lambda_1 \lambda_{d+1}}{n(\lambda_d - \lambda_{d+1})^2} \cdot (d + \log p)$$

where the **exact rate** is achieved by

$$\hat{V}_d = \underset{V^T V = I_d, \|V\|_{2,0} \le R_0}{\arg\max} \operatorname{trace}(V^T \widehat{\Sigma} V)$$

Rate independently obtained by Cai et al. (2013) for Gaussian spiked model

optimal
minimax
rate

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

optimal minimax rate $\asymp R_0 \cdot \dfrac{\lambda_1 \lambda_{d+1}}{n(\lambda_d - \lambda_{d+1})^2} \cdot (d + \log p)$

# of active variables

optimal minimax rate $\asymp R_0 \cdot \dfrac{\lambda_1 \lambda_{d+1}}{n(\lambda_d - \lambda_{d+1})^2} \cdot (d + \log p)$

# of active variables

effective noise variance

optimal minimax rate $\asymp R_0 \cdot \dfrac{\lambda_1 \lambda_{d+1}}{n(\lambda_d - \lambda_{d+1})^2} \cdot (d + \log p)$

# of active variables

effective noise variance

estimation error

optimal minimax rate $\asymp R_0 \cdot \dfrac{\lambda_1 \lambda_{d+1}}{n(\lambda_d - \lambda_{d+1})^2} \cdot (d + \log p)$

# of active variables

effective noise variance

estimation error

selection error

Consequences

- **Good news**

Consequences

- **Good news**
  - Exact minimax rate for **general model**

Consequences

- **Good news**
  - Exact minimax rate for **general model**
  - **Sparsity enables** estimation in high-dimensions

Consequences

- **Good news**
  - Exact minimax rate for **general model**
  - **Sparsity enables** estimation in high-dimensions
  - Extensions to L_q (weak) row sparsity

Consequences

- **Good news**
  - Exact minimax rate for **general model**
  - **Sparsity enables** estimation in high-dimensions
  - Extensions to L_q (weak) row sparsity
- **Bad news**

Consequences

- **Good news**
  - Exact minimax rate for **general model**
  - **Sparsity enables** estimation in high-dimensions
  - Extensions to L_q (weak) row sparsity
- **Bad news**
  - Estimator is computationally intractable (**NP**-hard)

Computation

Sparse PCA computation

- Almost all formulations of sparse PCA involve:

Sparse PCA computation

- Almost all formulations of sparse PCA involve:
  - Non-convex optimization problems with no statistical guarantees for local optima

Sparse PCA computation

- Almost all formulations of sparse PCA involve:
  - Non-convex optimization problems with no statistical guarantees for local optima
  - Strong assumptions and sensitivity to initial value

Sparse PCA computation

- Almost all formulations of sparse PCA involve:
  - Non-convex optimization problems with no statistical guarantees for local optima
  - Strong assumptions and sensitivity to initial value
- Is there a **polynomial time** method with **strong statistical guarantees** for the general model?

Minimax Optimal Estimator – but NP-Hard

$$\arg\max_{V} \quad \operatorname{trace}(V^T \widehat{\Sigma} V) - \lambda \|V\|_{2,0}$$

$$\text{subject to } V^T V = I_d$$

Minimax Optimal Estimator – but NP-Hard

$$\begin{aligned} &\arg\max_{V} \quad \operatorname{trace}(V^T \widehat{\Sigma} V) - \lambda \|V\|_{2,0} \\ &\text{subject to } V^T V = I_d \end{aligned}$$

or, equivalently,

$$\begin{aligned} &\arg\max_{Z} \quad \operatorname{trace}(\widehat{\Sigma} Z) - \lambda \|Z\|_{2,0} \\ &\text{subject to } Z \in \{Z : Z \text{ is a rank-}d \text{ projector}\} \end{aligned}$$

Difficulties

$$\begin{aligned}&\arg\max_{Z} \quad \operatorname{trace}(\widehat{\Sigma} Z) - \lambda \|Z\|_{2,0} \\ &\text{subject to } Z \in \{Z : Z \text{ is a rank-}d \text{ projector}\}\end{aligned}$$

Difficulties

$$\begin{aligned}&\arg\max_{Z} \quad \operatorname{trace}(\widehat{\Sigma} Z) - \lambda \|Z\|_{2,0} \\ &\text{subject to } Z \in \{Z : Z \text{ is a rank-}d \text{ projector}\}\end{aligned}$$

- Penalty is non-convex

Difficulties

$$\begin{aligned} &\arg\max_{Z} \quad \operatorname{trace}(\widehat{\Sigma} Z) - \lambda \|Z\|_{2,0} \\ &\text{subject to } Z \in \{Z : Z \text{ is a rank-}d \text{ projector}\} \end{aligned}$$

- Penalty is non-convex
- Constraint set is non-convex

Difficulties

$$\underset{Z}{\arg\max} \quad \operatorname{trace}(\widehat{\Sigma} Z) - \lambda \|Z\|_{2,0}$$
$$\text{subject to } Z \in \{Z : Z \text{ is a rank-}d \text{ projector}\}$$

- Penalty is non-convex
- Constraint set is non-convex
- Solution? **Use convex hulls!**

Convex Relaxation

$$\underset{Z}{\arg\max} \quad \operatorname{trace}(\widehat{\Sigma} Z) - \lambda \|Z\|_{2,0}$$

$$\text{subject to } Z \in \{Z : Z \text{ is a rank-}d \text{ projector}\}$$

Convex Relaxation

$$\begin{aligned}&\arg\max_{Z} \quad \operatorname{trace}(\widehat{\Sigma} Z) - \lambda \|Z\|_1 \\ &\text{subject to } Z \in \{Z : Z \text{ is a rank-}d \text{ projector}\}\end{aligned}$$

- Convex penalty function – entrywise L1 norm

Convex Relaxation

$$\begin{aligned} &\arg\max_{Z} \quad \operatorname{trace}(\widehat{\Sigma} Z) - \lambda \|Z\|_1 \\ &\text{subject to } Z \in \{Z : 0 \preceq Z \preceq I \text{ and } \operatorname{trace}(Z) = d\} \end{aligned}$$

- Convex penalty function – entrywise L1 norm
- Convex constraint set – **The Fantope**

Convex Relaxation

$$\begin{aligned} &\underset{Z}{\arg\max} \quad \operatorname{trace}(\widehat{\Sigma} Z) - \lambda \|Z\|_1 \\ &\text{subject to } Z \in \{Z : 0 \preceq Z \preceq I \text{ and } \operatorname{trace}(Z) = d\} \end{aligned}$$

- Convex penalty function – entrywise L1 norm
- Convex constraint set – **The Fantope**
- Amazing fact – Fantope is convex hull of rank-d projectors (see Overton & Womersley, 1992)

Fantope Projection and Selection

$$\begin{aligned} &\operatorname*{arg\,max}_{Z} \quad \operatorname{trace}(\widehat{\Sigma} Z) - \lambda \|Z\|_1 \\ &\text{subject to } Z \in \{Z : 0 \preceq Z \preceq I \text{ and } \operatorname{trace}(Z) = d\} \end{aligned}$$

- Equivalent to a semidefinite program (SDP)
- d=1 case proposed by d'Aspremont et al. 2007
- Avoids orthogonality/deflation issues by ***directly*** estimating projector

Fantope Projection and Selection

- Computable in polynomial time

Fantope Projection and Selection

- Computable in polynomial time
- Alternating Direction Method of Multipliers (ADMM) algorithm has two main steps:

Fantope Projection and Selection

- Computable in polynomial time
- Alternating Direction Method of Multipliers (ADMM) algorithm has two main steps:
  - Fantope Projection – have exact analytic solution

Fantope Projection and Selection

- Computable in polynomial time
- Alternating Direction Method of Multipliers (ADMM) algorithm has two main steps:
  - Fantope Projection – have exact analytic solution
  - Element-wise soft-thresholding (***selection***)

Guarantee for FPS

Theorem (*VLCR 2013*)

Under the **general model**, assume the principal subspace is R_0 row-sparse. If regularization parameter chosen appropriately*, then FPS estimate $\widehat{Z}$ satisfies

$$\|\widehat{Z} - V_d V_d^T\|_F^2 \lesssim R_0^2 \cdot \frac{\lambda_1 \lambda_{d+1}}{n(\lambda_d - \lambda_{d+1})^2} \cdot \log p$$

with high probability, regardless of its rank.

$* \quad \lambda \asymp \sqrt{(\lambda_1 \lambda_{d+1} \log p)/n}$

FPS is *near*-Optimal

Recall minimax optimal rate (d=1) vs FPS rate:

$$R_0 \cdot \frac{\lambda_1 \lambda_{d+1}}{n(\lambda_d - \lambda_{d+1})^2} \cdot \log p \qquad R_0^2 \cdot \frac{\lambda_1 \lambda_{d+1}}{n(\lambda_d - \lambda_{d+1})^2} \cdot \log p$$

minimax optimal FPS

FPS rate is off by factor of R_0

FPS is *near*-Optimal

Recall minimax optimal rate (d=1) vs FPS rate:

$$R_0 \cdot \frac{\lambda_1 \lambda_{d+1}}{n(\lambda_d - \lambda_{d+1})^2} \cdot \log p \qquad R_0^2 \cdot \frac{\lambda_1 \lambda_{d+1}}{n(\lambda_d - \lambda_{d+1})^2} \cdot \log p$$

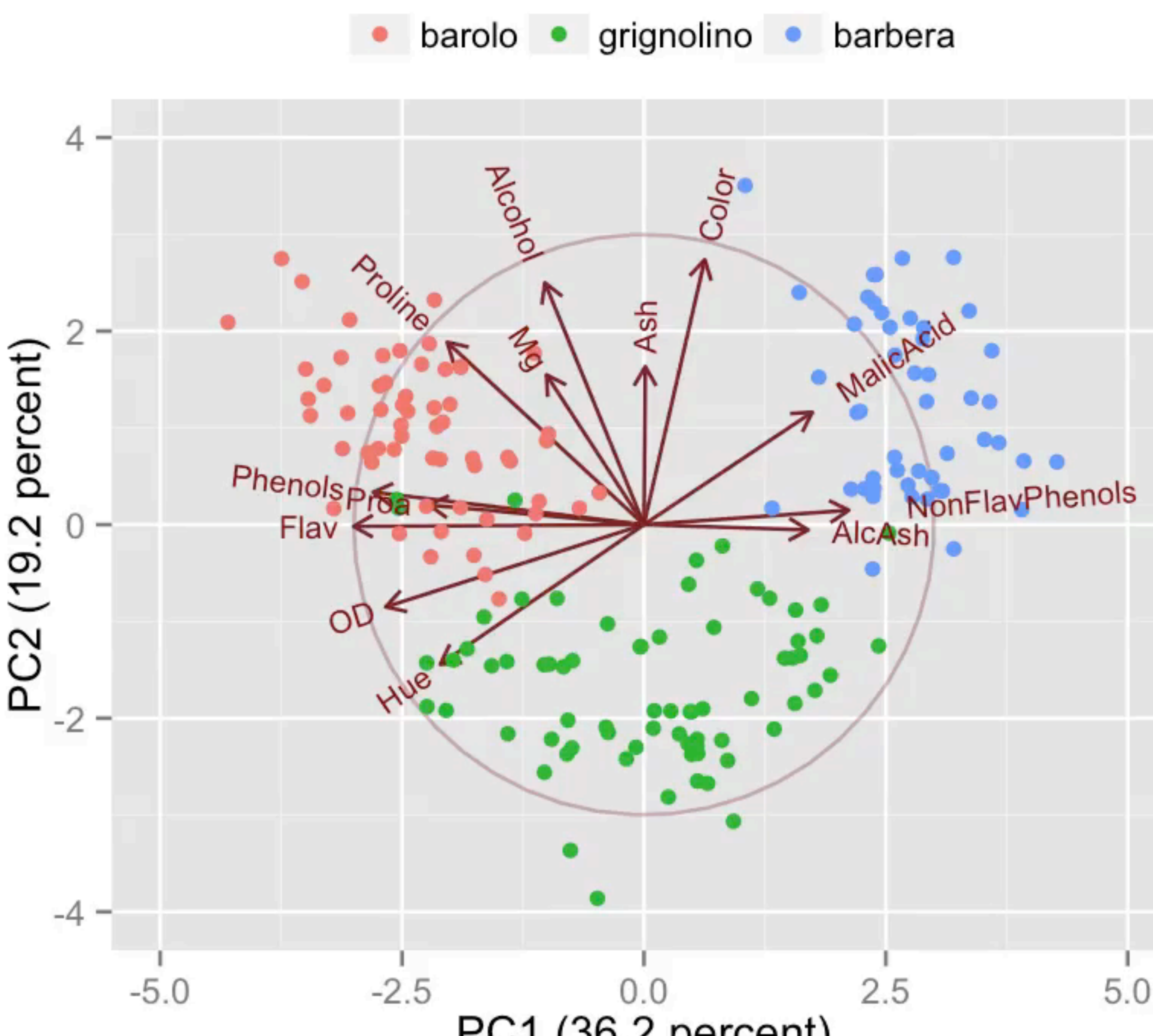
minimax optimal FPS

FPS rate is off by factor of R_0

When d=1, $\mathbf{R_0}$ factor maybe unavoidable for **any polynomial time** algorithm in a hypothesis testing framework (Berthet & Rigollet 2013)

Small illustration

- Data on **n=178** wines grown over a decade in the same region of Italy
- 3 different cultivars: Barolo, Grignolino, Barbera
- Measurements on **p=13** constituents
- Will show **d=2** subspace estimated by FPS over a range of regularization parameter values

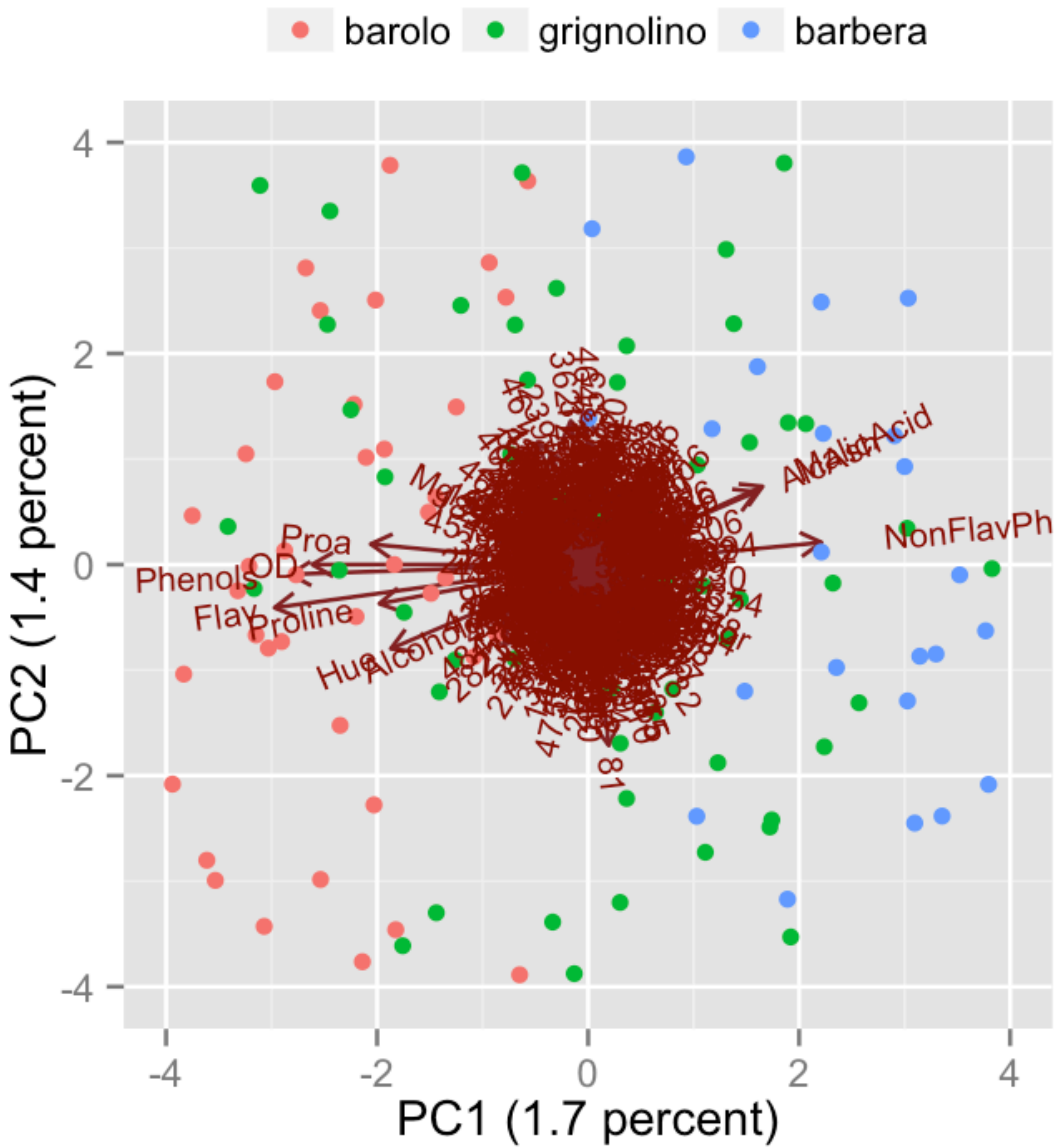


barolo
grignolino
barbera
PC2 (19.2 percent)
PC1 (36.2 percent)
4
2
0
-2
-4
-5.0
-2.5
0.0
2.5
5.0
Alcohol
Color
Proline
Mg
Ash
MalicAcid
Phenols
Proa
Flav
NonFlavPhenols
AlcAsh
OD
Hue

Movie not shown

Synthetic illustration

- Dataset ***synthetically enlarged*** by adding 487 noise variables by randomly, independently copying and permuting the real variables – result **n=178, p=500**
- Does FPS recover the 13 real variables?
- Does FPS projection reveal the 3 clusters?



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

Movie not shown

Summary

- Sparsity helps both estimation **accuracy** and **interpretation** of PCA in high dimensions
- Row sparsity is a rotation invariant notion of **subspace sparsity**
- Minimax rates reveal **gap** between computationally tractable and optimal (NP-hard) procedures
- Convex relaxation (FPS) is **near-optimal**

Ongoing work

- Fast algorithm and computational insights for FPS to enable processing of larger scale data
- Is FPS rate optimal among polynomial time methods?

Thank you!

References

- Vu & Lei (2012) "Minimax rates of estimation for sparse PCA in high dimensions." AISTATS
- Vu & Lei (2013) "Minimax sparse principal subspace estimation in high dimensions." Annals of Statistics, *to appear*
- Vu, Cho, Lei & Rohe (2013) "Fantope projection and selection." *manuscript in preparation; preliminary report to appear in* NIPS