Sparse Principal Components Analysis

Iain Johnstone, Statistics, Stanford & UC Berkeley

Arthur Yu Lu, Stanford and Renaissance Technologies

FDA Workshop, Gainesville, Jan 10-11,2003

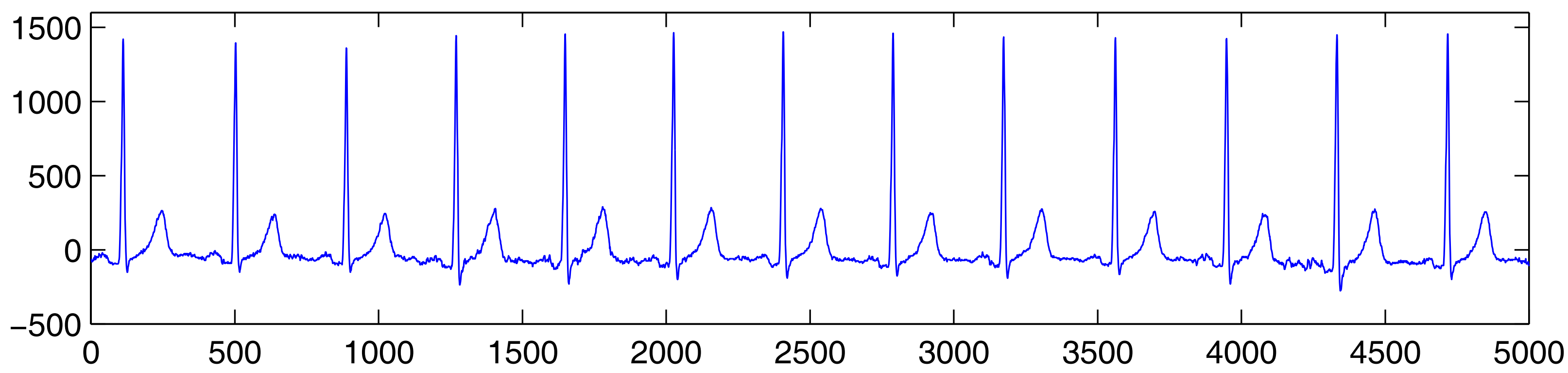
Support thanks: NIH

"Multiscale" Functional Data

Functional data: $x_i = x_i(t_j)$ $\quad j = 1, \ldots, p,$ time points,
$i = 1, \ldots, n,$ cases.

- Focus on PCA: $\longrightarrow$ principal modes of variation
- 'high-dimensional' : $p = O(n)$ or larger
- Signals x_i contain localized features, perhaps on different scales —visible in p.c.'s also?
- Example: ECG signals

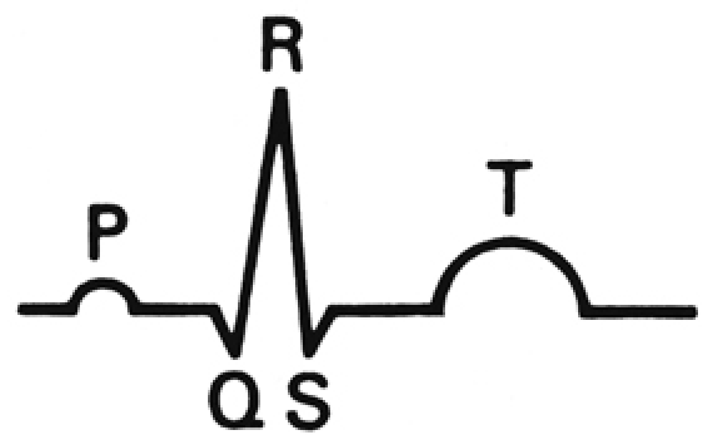
ElectroCardioGram traces



Average beat

vs.

beat to beat variation



Outline

$\Rightarrow$ Multiscale Functional Data $p \asymp n$

- Need for Dimension Reduction: Inconsistency
- Sparse PCA
  - Some sampling properties, incl. consistency
- Examples, incl. ECG

Main Themes

- initial dimension reduction before PCA
  - otherwise, inconsistency!
- use basis with sparse representation
  - so that little is lost in initial dim reduction

Background Theme: role of “Random Matrices”

- Small perturbations of symmetric matrices
- a.s. bounds for extreme eigenvalues of large matrices
- other talks here: large stochastic data matrices: (e.g. Marron)

$\Rightarrow$ more generally, broader role for RMT tools in analysis of FDA methods?

Outline

- Multiscale Functional Data $p \asymp n$

$\Rightarrow$ Need for Dimension Reduction: Inconsistency

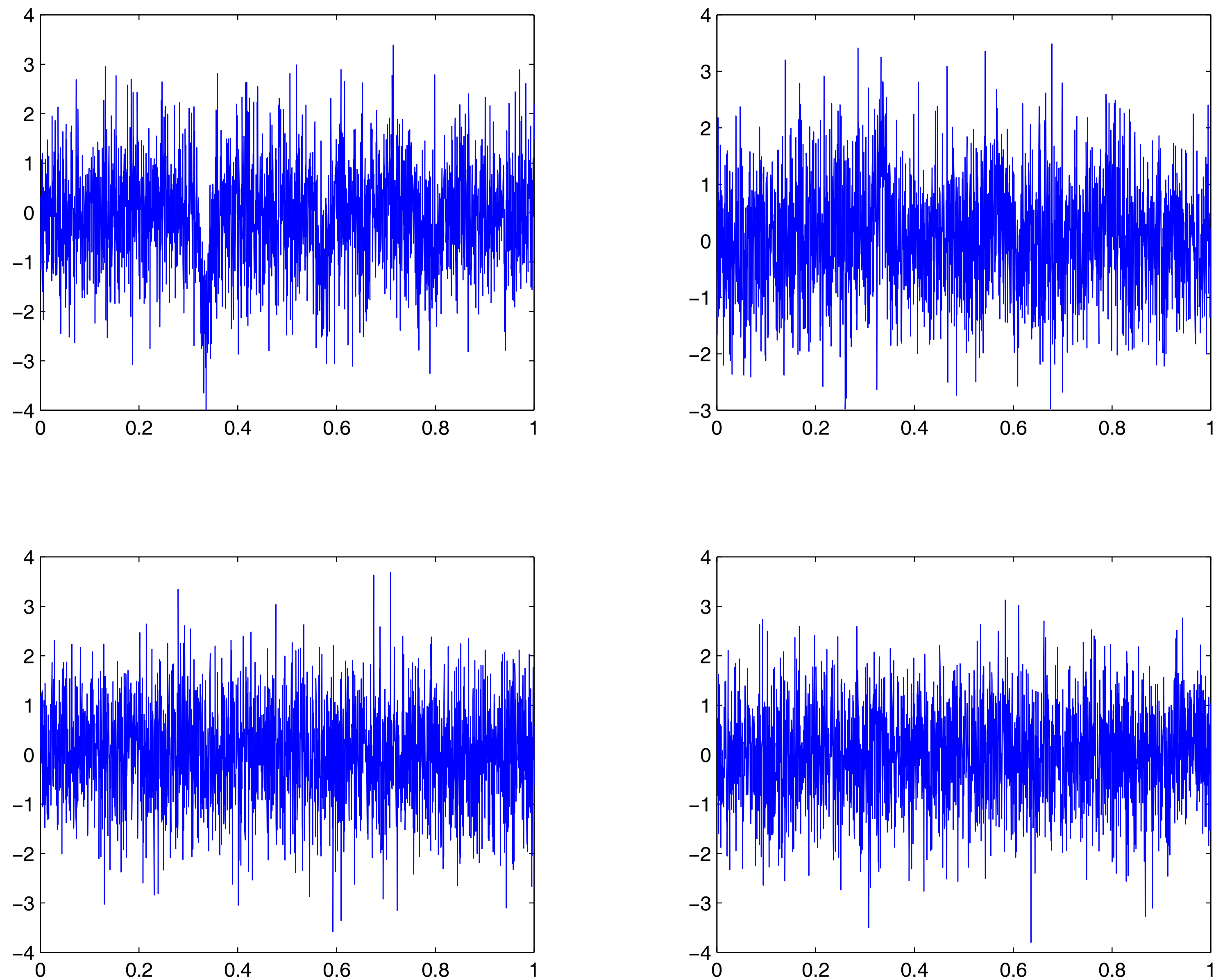
- Sparse PCA
  - Some sampling properties, incl. consistency
- Examples, incl. ECG

Single Component Model

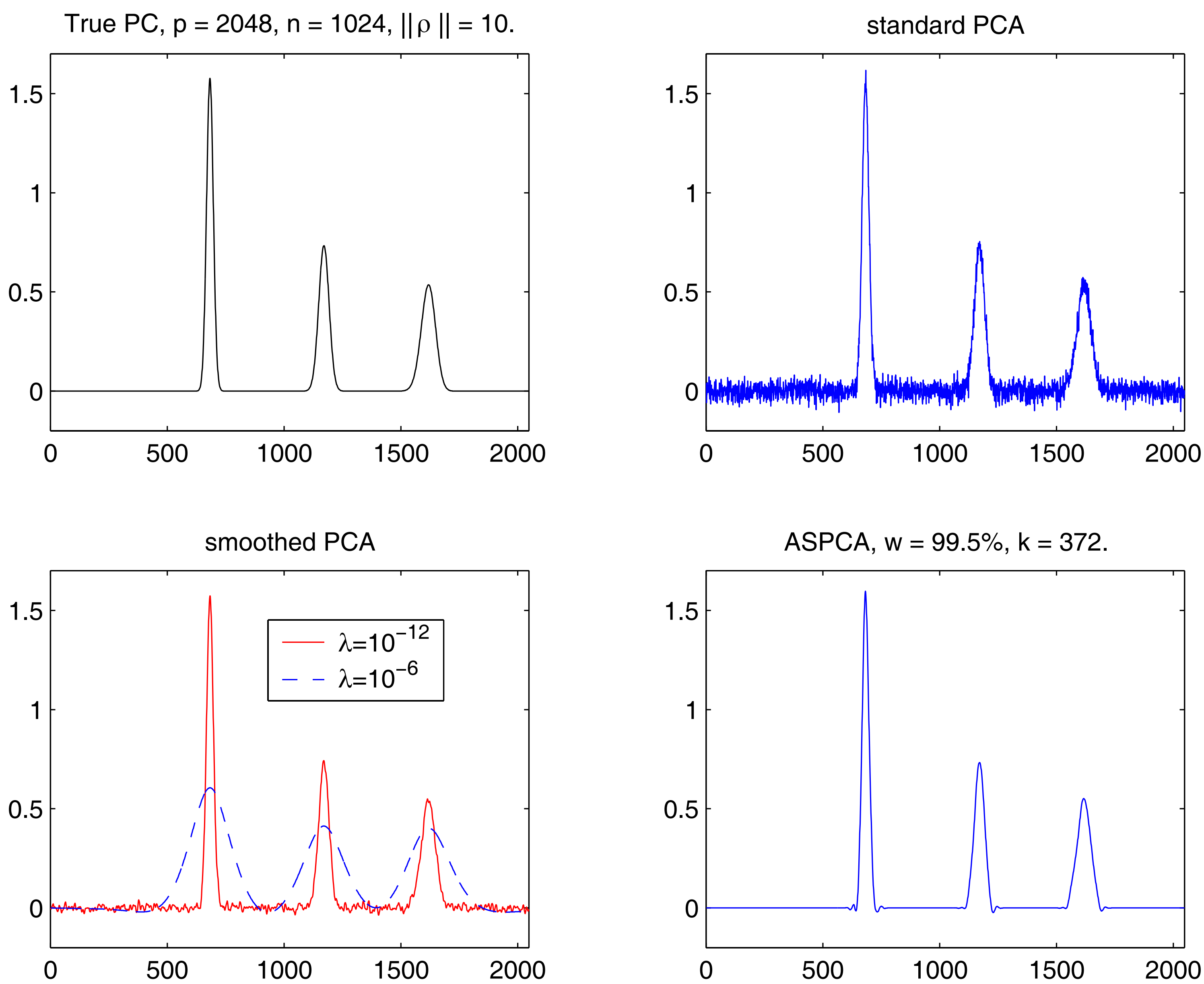
[to illustrate, e.g., need for dimension reduction]

$$x_i = v_i \rho + \sigma z_i \qquad i = 1, \ldots, n$$

- $\rho \in \mathbb{R}^p$, single component to be estimated
  (e.g. $\|\rho\| = 10, p = 2048$)
- $v_i \overset{i.i.d.}{\sim} N(0, 1)$ random effects (e.g. $n = 1024$)
- $z_i \overset{i.i.d.}{\sim} N_p(0, I)$ Gaussian noise (e.g. $\sigma = 1$)



4
3
2
1
0
−1
−2
−3
−4
0
0.2
0.4
0.6
0.8
1

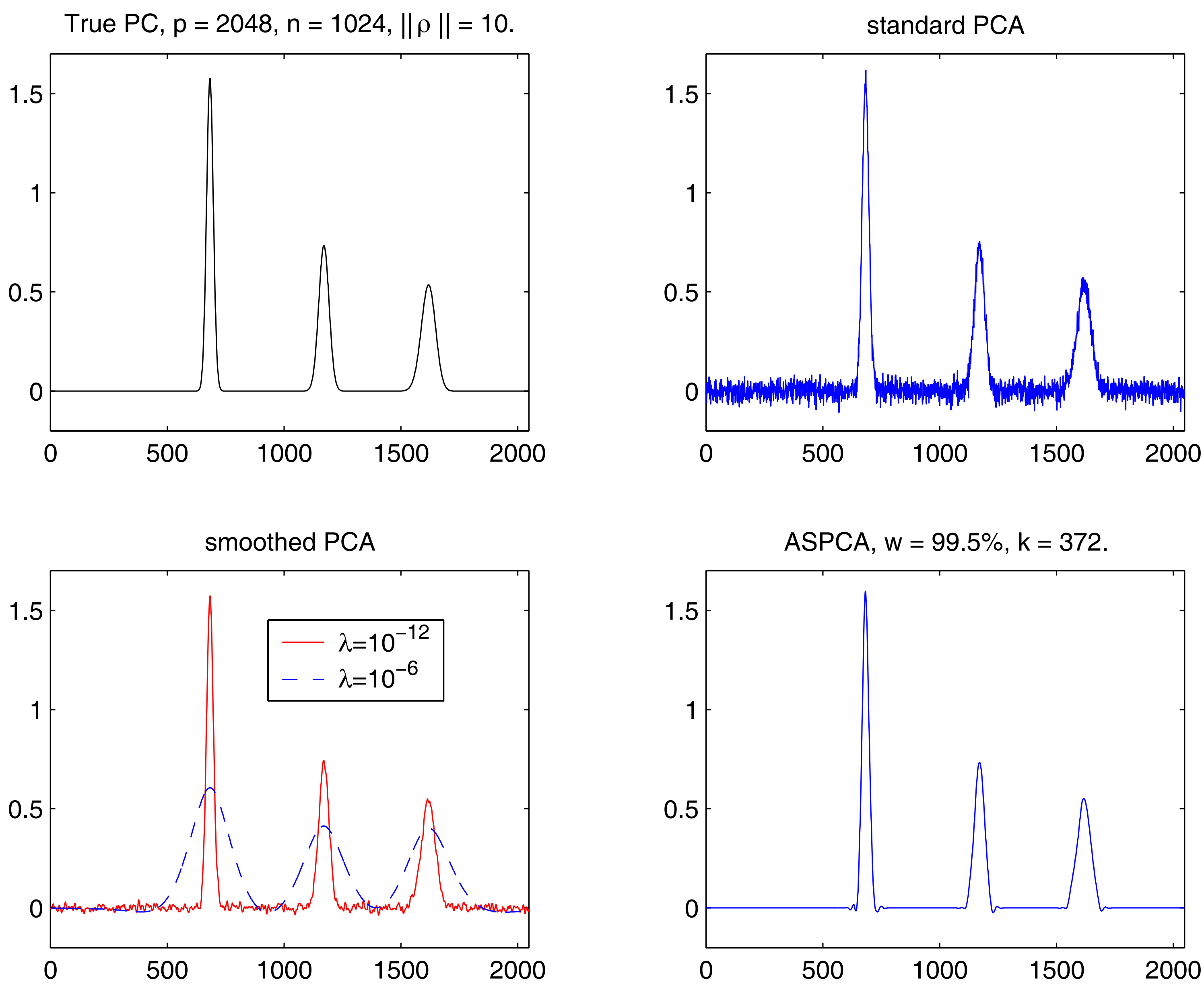

True PC, p = 2048, n = 1024, || ρ || = 10.
standard PCA
smoothed PCA
λ=10^−12
λ=10^−6
ASPCA, w = 99.5%, k = 372.
0
0.5
1
1.5
0
500
1000
1500
2000

Smooth functional PCA

- Rice & Silverman (1991), Silverman (1996), Ramsay & Silverman (1997)
- Seek smoothed p.c.'s ξ by maximizing

$$\frac{\text{Var}\,(\xi^T x_i)}{\|\xi\|^2 + \lambda\|D^2\xi\|^2},$$

- Our example: no choice of λ can *both*
  - preserve peak heights, *and*
  - remove baseline noise


True PC, p = 2048, n = 1024, || ρ || = 10.
standard PCA
smoothed PCA
λ=10^−12
λ=10^−6
ASPCA, w = 99.5%, k = 372.
0
0.5
1
1.5
0
500
1000
1500
2000

Consistency

Single component model: $x_i = v_i \rho + \sigma z_i, \quad i = 1, \ldots, n$

Suppose $p(n)/n \to c, \quad \|\rho(n)\| \to \varrho > 0$.

Then a.s.

$$\overline{\lim}_{n\to\infty} \sin \angle(\hat{\rho}, \rho) \leq 6\sigma\sqrt{c}/\varrho.$$

In particular:

- **consistent** if $p/n \to c = 0$
- correct rate if p *£xed*: $O(\sqrt{c}) = O(1/\sqrt{n})$.
- *but* positive if $c > 0$.

A Matrix Perturbation Theorem

Suppose

- A, E are *symmetric*
- $q_1, \hat{q}_1$ are the *principal* eigenvectors of A, $A + E$
- $\lambda_1(A) - \delta \geq \lambda_2(A) \geq \lambda_\nu(A), \;\; \nu \geq 2$.

Then

$$\sin \angle(\hat{q}_1, q_1) \;\leq\; (4/\delta)\|E\|_2$$

consequence of more general result for invariant subspaces (Stewart, Stewart - Sun).

A Multicomponent Model

$$x_i = \sum_{j=1}^{m} v_i^j \rho^j + \sigma z_i, \qquad i = 1, \ldots, n$$

- ρ^j are unknown, mutually orthogonal, $\|\rho^1\| \geq \cdots \geq \|\rho^m\|$.
- $v_i^j \overset{i.i.d.}{\sim} N(0,1)$ random effects
- $z_i \sim N_p(0, I)$ noise, independent of $\{v_i^j\}$.

For asymptotics,

$$(\|\rho^1(n)\|, \ldots, \|\rho^j(n)\|, \ldots) \xrightarrow{\ell_1} (\varrho_1, \ldots, \varrho_j, \ldots).$$

Inconsistency

In either single or multicomponent model,

Theorem If $p/n \to c > 0$, then

$$\liminf_{n\to\infty} E \sin\angle(\hat{\rho}^1, \rho^1) > 0.$$

- Noise does not average out in PCA if too many dimensions p relative to n.
- Suggests: reduce p to $k \ll p$ before starting PCA

What goes wrong if $p/n \to c > 0$

$$X = \rho v^T + \sigma Z \qquad S = n^{-1} X X^T$$

$$= \underbrace{\frac{v^T v}{n}\rho\rho^T + \frac{\sigma^2}{n} Z Z^T}_{D} + \underbrace{u\rho^T + \rho^T u}_{B}$$

$$= D + B$$

- eigenvalues of $n^{-1} Z Z^T$ do not approach 1:
  - $\lambda_{max}, \lambda_{min} \stackrel{a.s.}{\to} (1 \pm \sqrt{c})^2$ [Geman, Silverstein]
- $u = \sigma n^{-1} Z v$ does not vanish:
  - $\|u\|^2 \stackrel{\mathcal{D}}{=} \sigma^2 \frac{p}{n} \frac{\chi^2_{(n)}}{n} \frac{\chi^2_{(p)}}{p} \stackrel{a.s.}{\to} \sigma^2 c > 0.$

Why $c > 0$ forces inconsistency

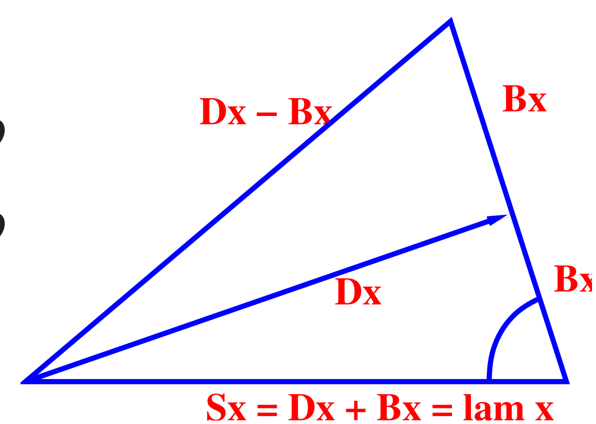
Suppose $S_{\pm} = D \pm B$ have principal e-vectors $\hat{\rho}_{\pm}$

$\hat{\rho}_{+}, \hat{\rho}_{-}$ have same distribution (symmetry)

But $\|B\rho\| > 0$ and $B\rho \sim\perp (D+B)\rho$

$\Rightarrow$ $\hat{\rho}_{+}$ and $\hat{\rho}_{-}$ cannot *both* be close to ρ

$\Rightarrow$ both are inconsistent.



Outline

- Multiscale Functional Data $p \asymp n$
- Need for Dimension Reduction: Inconsistency

$\Rightarrow$ Sparse PCA

  - Some sampling properties, incl. consistency

- Examples, incl. ECG

Sparse PCA Algorithm

Basis	$x_i(t) = \sum_\nu x_{i\nu} e_\nu(t)$, $i = 1, \ldots, n$	$O(np \log p)$
Subset	$\hat{\sigma}_\nu^2 =$ Var$\{x_{i\nu}, i = 1, \ldots, n\}$; $\hat{I}_k = \{\nu \leftrightarrow$ largest k variances $\}$	$O(p \log p)$
Reduced PCA	on $\{x_{i\nu} : \nu \in \hat{I}_k\}$, $\longrightarrow$ eigenvectors $\hat{\rho}_j$	$O(k^3)$
Thresholding	$\hat{\rho}_{j\nu}^* = \eta_H(\hat{\rho}_{j\nu}, \delta)$	$O(k)$
Reconstruct	$\hat{\rho}_j(t) = \sum_\nu \hat{\rho}_{j\nu}^* e_\nu(t)$.	$O(k^2 p)$

Sparse PCA - Choice of Basis

In basis $\{e_\nu(t)\}$, a population p.c. $\{\rho\}$ has coef£cents $\{\rho_\nu\}$:

$$\rho(t) = \sum_{\nu=1}^{p} \rho_\nu e_\nu(t).$$

Sparsity and weak ℓ_p Say $\rho \in w\ell_p(C)$ if

$$|\rho_{(\nu)}| \leq C\nu^{-1/p}, \qquad \nu = 1, 2, \ldots$$

- p small $\Rightarrow$ rapid decay of ordered coef£cients
- choose basis to exploit sparsity

Wavelet Bases and Sparsity

- Expand ρ in wavelet basis $\{\psi_{jk}\}$: $\rho = \sum_{jk} \rho_{jk}\psi_{jk}$
- order coef£cients $\rho_\nu = \nu$-th largest $|\rho_{jk}|$
- **Fact**: smoothness (even non-homogeneous) $\Rightarrow$
  sparse wavelet representation:

$$\rho \in B^\alpha_{p,q} \qquad \Rightarrow \qquad (\rho_\nu) \in w\ell_p \qquad p = 2/(2\alpha + 1).$$

- Hence use wavelet basis here, but algorithm *could* use others..

Sparse PCA Algorithm

Basis	$x_i(t) = \sum_\nu x_{i\nu} e_\nu(t)$,	
	$i = 1, \ldots, n$	$O(np \log p)$
Subset	$\hat{\sigma}_\nu^2 =$ Var$\{x_{i\nu}, i = 1, \ldots, n\}$	
	$\hat{I}_k = \{\nu \leftrightarrow$ largest k variances $\}$	$O(p \log p)$
Reduced PCA	on $\{x_{i\nu} : \nu \in \hat{I}_k\}$,	
	$\longrightarrow$ eigenvectors $\hat{\rho}_j$	$O(k^3)$
Thresholding	$\hat{\rho}^*_{j\nu} = \eta_H(\hat{\rho}_{j\nu}, \delta)$	
		$O(k)$
Reconstruct	$\hat{\rho}_j(t) = \sum_\nu \hat{\rho}^*_{j\nu} e_\nu(t)$.	
		$O(k^2 p)$

Choosing subset size k from data

Aim: Choose $\hat{I}$ to capture most of population p.c.'s ρ variance:

$$\sum_{\nu \in \hat{I}} \rho_\nu^2 = w(n) \sum \rho_\nu^2, \qquad w(n) \nearrow 1.$$

Possibilities: a) $\hat{I} = \{\nu : \hat{\sigma}^2_{(\nu)} \geq \hat{\sigma}^2(1 + L_n)\}$, or

b) de£ne excess over noise using percentiles of $\chi^2_{(n)}$:

$$\hat{\tau}^2_{(\nu)} = \hat{\sigma}^2_{(\nu)} \chi^2_{(n),\nu/n}/n, \qquad \text{and,}$$

$$\hat{I} = \{\nu : \sum_{\nu=1}^{\hat{k}} \hat{\tau}^2_{(\nu)} \geq w(n) \sum_{\nu} \hat{\tau}^2_{(\nu)}\}.$$

Sparse PCA Algorithm

Basis	$x_i(t) = \sum_\nu x_{i\nu} e_\nu(t)$,	
	$i = 1, \ldots, n$	$O(np \log p)$
Subset	$\hat{\sigma}_\nu^2 =$ Var$\{x_{i\nu}, i = 1, \ldots, n\}$	
	$\hat{I}_k = \{\nu \leftrightarrow$ largest k variances $\}$	$O(p \log p)$
Reduced PCA	on $\{x_{i\nu} : \nu \in \hat{I}_k\}$,	
	$\longrightarrow$ eigenvectors $\hat{\rho}_j$	$O(k^3)$
Thresholding	$\hat{\rho}_{j\nu}^* = \eta_H(\hat{\rho}_{j\nu}, \delta)$	
		$O(k)$
Reconstruct	$\hat{\rho}_j(t) = \sum_\nu \hat{\rho}_{j\nu}^* e_\nu(t)$.	
		$O(k^2 p)$

Computational Complexity

For standard PCA on $n \times p$ data set X, running time is

$$O((p \wedge n)^3).$$

For $k = o(p) \asymp n$, reduced PCA is $O(k^3)$.

Overall, if say $p \geq n$, reduce from

$$O(p^3) \qquad \longrightarrow \qquad O(k^2 p + np \log p).$$

Sparse PCA Algorithm

Basis	$x_i(t) = \sum_\nu x_{i\nu} e_\nu(t)$,	
	$i = 1, \ldots, n$	$O(np \log p)$
Subset	$\hat{\sigma}_\nu^2 =$ Var$\{x_{i\nu}, i = 1, \ldots, n\}$	
	$\hat{I}_k = \{\nu \leftrightarrow$ largest k variances $\}$	$O(p \log p)$
Reduced PCA	on $\{x_{i\nu} : \nu \in \hat{I}_k\}$,	
	$\longrightarrow$ eigenvectors $\hat{\rho}_j$	$O(k^3)$
Thresholding	$\hat{\rho}_{j\nu}^* = \eta_H(\hat{\rho}_{j\nu}, \delta)$	
		$O(k)$
Reconstruct	$\hat{\rho}_j(t) = \sum_\nu \hat{\rho}_{j\nu}^* e_\nu(t)$.	
		$O(k^2 p)$

Thresholding subset eigenvectors

Reason: estimated e-vectors on reduced variable set still noisy.

By analogy with wavelet shrinkage in regression: *keep* large coef£cients, *kill* noise.

Here, usual *hard thresholding:*

$$\eta_H(y, \delta) = \begin{cases} y & |y| \geq \delta \\ 0 & \text{otherwise} \end{cases}$$

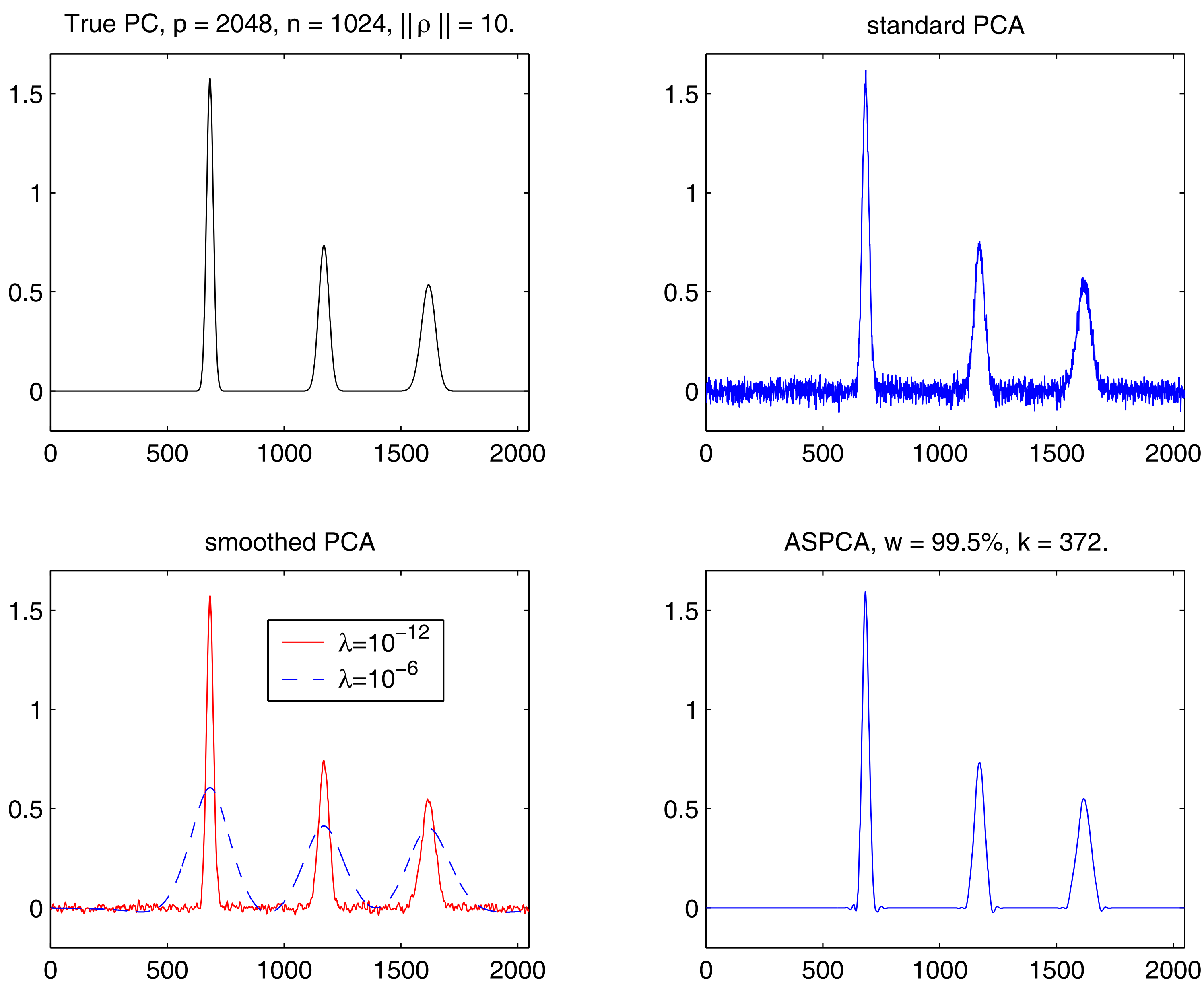
Other variants: soft, SCAD ...

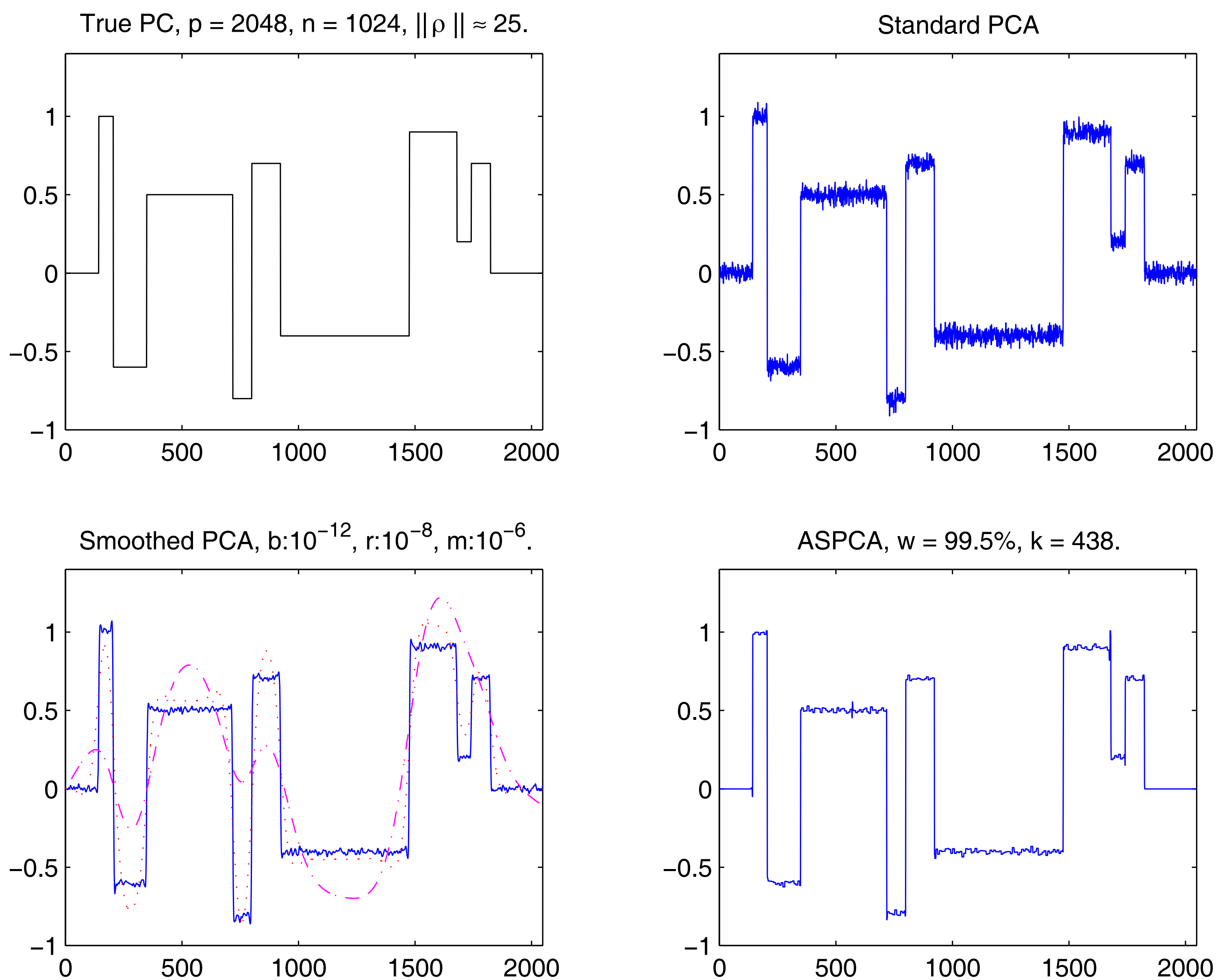
Choice of δ.   *For now*, use usual “universal” threshold

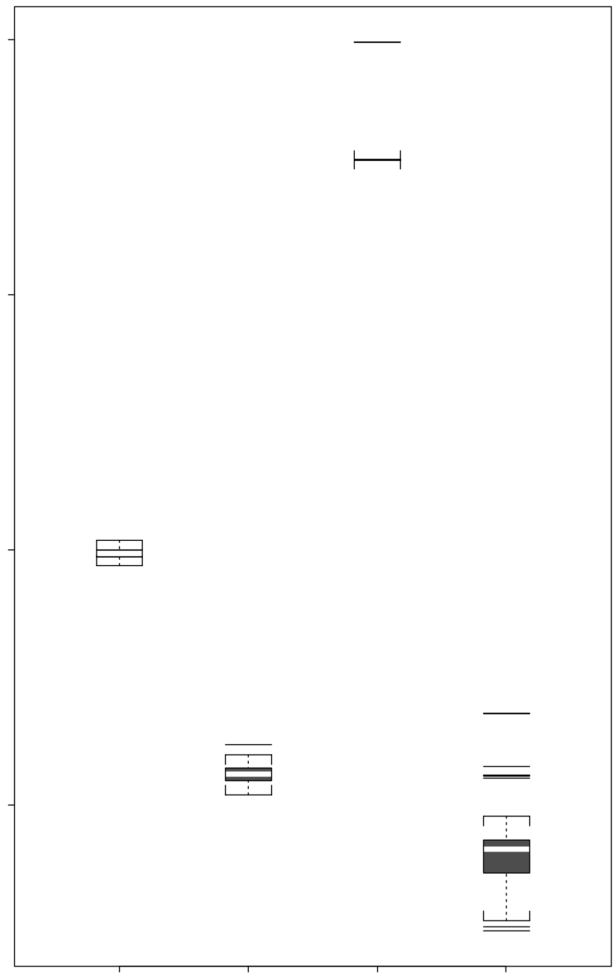
$$\delta = \hat{\tau}_j \sqrt{2 \log k}, \qquad \hat{\tau}_j = MAD\{\hat{\rho}_{j,\nu}, \nu = 1, \ldots, k\}/0.6745.$$

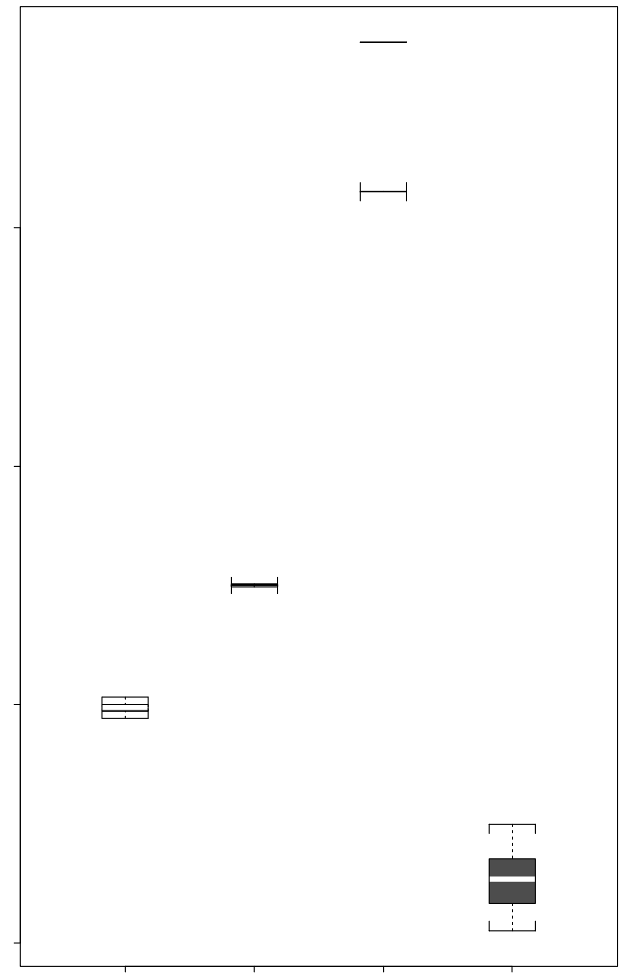
Sparse PCA Algorithm

Basis	$x_i(t) = \sum_\nu x_{i\nu} e_\nu(t)$,	
	$i = 1, \ldots, n$	$O(np \log p)$
Subset	$\hat{\sigma}_\nu^2 =$ Var$\{x_{i\nu}, i = 1, \ldots, n\}$	
	$\hat{I}_k = \{\nu \leftrightarrow$ largest k variances $\}$	$O(p \log p)$
Reduced PCA	on $\{x_{i\nu} : \nu \in \hat{I}_k\}$,	
	$\longrightarrow$ eigenvectors $\hat{\rho}_j$	$O(k^3)$
Thresholding	$\hat{\rho}_{j\nu}^* = \eta_H(\hat{\rho}_{j\nu}, \delta)$	
		$O(k)$
Reconstruct	$\hat{\rho}_j(t) = \sum_\nu \hat{\rho}_{j\nu}^* e_\nu(t)$.	
		$O(k^2 p)$


True PC, p = 2048, n = 1024, || ρ || = 10.
standard PCA
smoothed PCA
ASPCA, w = 99.5%, k = 372.
λ=10^-12
λ=10^-6
0
0.5
1
1.5
0
500
1000
1500
2000


True PC, p = 2048, n = 1024, || ρ || ≈ 25.
Standard PCA
Smoothed PCA, b:10^−12, r:10^−8, m:10^−6.
ASPCA, w = 99.5%, k = 438.
1
0.5
0
−0.5
−1
0
500
1000
1500
2000





Speed and Accuracy Comparison

	Standard PCA	Smoothed $\lambda : 10^{-12}$	Smoothed $\lambda : 10^{-6}$	Sparse PCA
ASE (3-peak)	9.681e-04	1.327e-04	3.627e-2	**7.500e-05**
Time (3-peak)	$\sim$ 12min	$\sim$ 47 min	$\sim$ 43 min	**1 min 15 s**
ASE (step)	9.715e-04	3.174e-3	1.694e-2	**1.947e-04**
Time (step)	$\sim$ 12min	$\sim$ 47 min	$\sim$ 46 min	**1 min 31 s**

Outline

- Multiscale Functional Data $p \asymp n$
- Need for Dimension Reduction: Inconsistency
- Sparse PCA

$\Rightarrow$ Some sampling properties, incl. consistency

- Examples, incl. ECG

Correct Selection Properties

Does $\hat{I}_k \leftrightarrow \hat{\sigma}^2_{(1)}, \ldots, \hat{\sigma}^2_{(k)}$ include the “right” variables?

i.e. include all “large” : $I_{in} = \{\nu : \sigma^2_\nu \geq \sigma^2_{(k)}(1 + \alpha_n)\}$,

and exclude all “small” : $I_{out} = \{\nu : \sigma^2_\nu \leq \sigma^2_{(k)}(1 - \alpha_n)\}$.

De£ne: FE = False Exclusion $= \{I_{in} \subset \hat{I}_k\}^c$

FI = False Inclusion $= \{I_{out} \subset \hat{I}^c_k\}^c$

Theorem Suppose $\hat{\sigma}^2_\nu \sim \sigma^2_\nu \chi^2_{(n)}/n$. If $\alpha = \gamma\sqrt{\log n/n}$, then

$$P\{FE \cup FI\} \leq 3pk\, n^{-b\gamma^2},$$

E.g.: $(k, p, n) = (50, 1000, 1000),\ 1 + \alpha_n = 1.25^2,\ P \leq .02$

Consistency of Sparse PCA

Single component model. Suppose (i) $p/n \to c > 0$,
(ii) $\|\rho(n)\| \to \varrho > 0$.

Assume *Sparsity*: $\rho(n) \in w\ell_p(C)$ uniformly in n

Subset selection rule: $\hat{I} = \{\nu : \hat{\sigma}_\nu^2 > \sigma^2(1 + c\sqrt{2\log p}\sqrt{2/n})\}$

Let $\hat{\rho}$ denote sparse PCA estimate based on $\hat{I}$.

Theorem $\angle(\hat{\rho}, \rho) \stackrel{a.s.}{\longrightarrow} 0.$

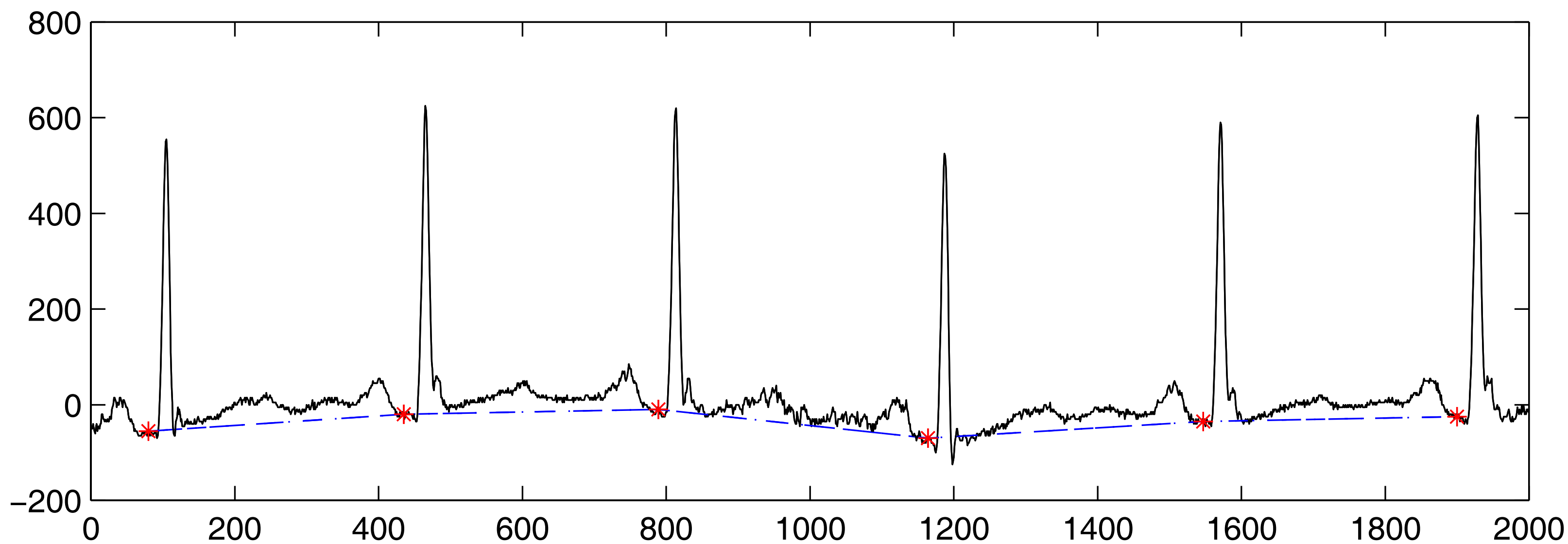
Outline

- Multiscale Functional Data $p \asymp n$
- Need for Dimension Reduction: Inconsistency
- Sparse PCA
  - Some sampling properties, incl. consistency

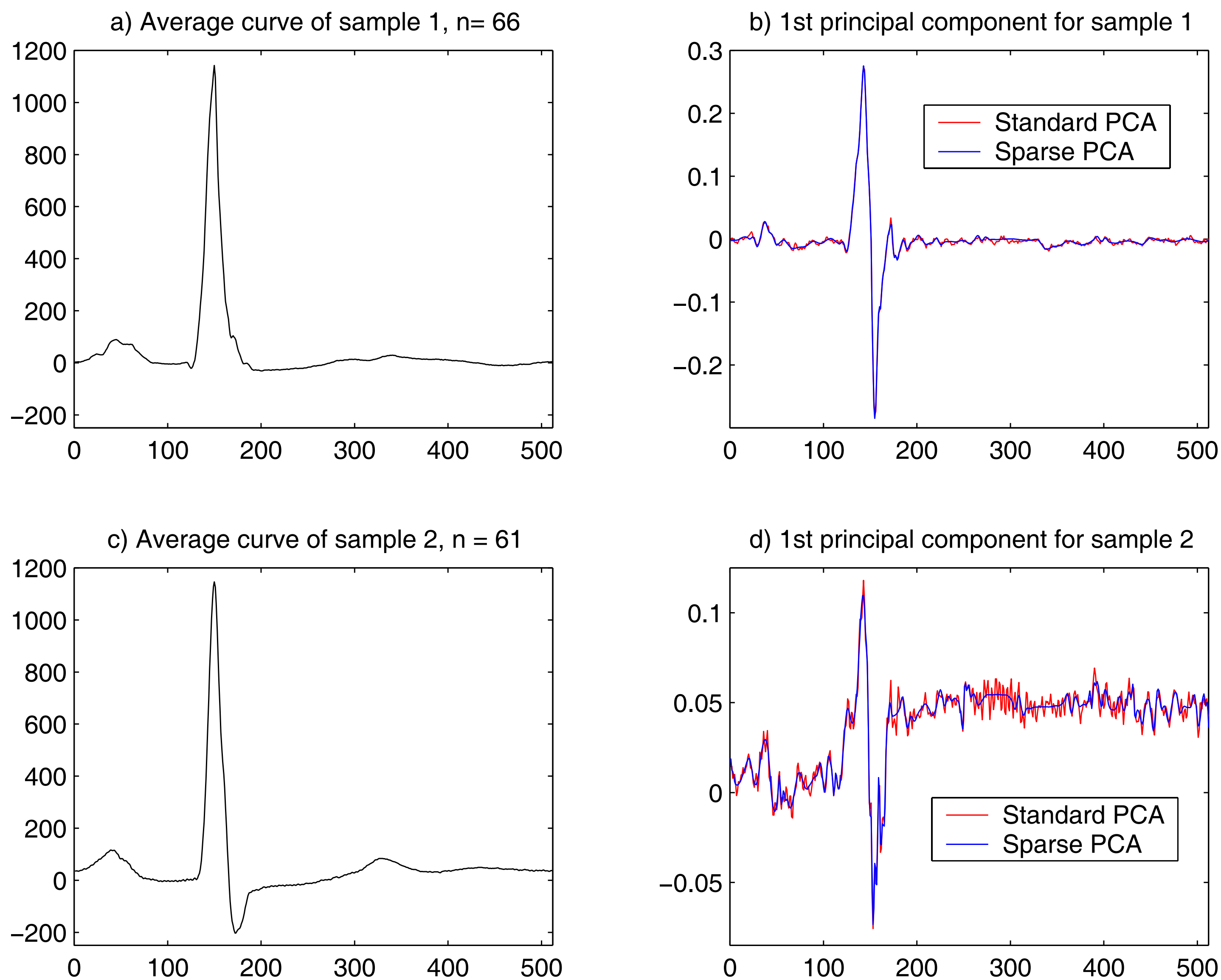
$\Rightarrow$ Examples, incl. ECG

ECG example

Preprocessing: piecewise linear baseline wander removal,
registration at $R-$wave maximum,
interpolation to 512 samples per cycle



Data: V. Froelicher, J. Froning, Palo Alto VA Hospital.


a) Average curve of sample 1, n= 66
1200
1000
800
600
400
200
0
−200
0
100
200
300
400
500
b) 1st principal component for sample 1
0.3
0.2
0.1
0
−0.1
−0.2
Standard PCA
Sparse PCA
c) Average curve of sample 2, n = 61
d) 1st principal component for sample 2
0.1
0.05
−0.05

Remarks

- both p.c.s correspond to change in shape of R-wave peak
- noise is larger in second case: $\hat{\sigma}_1^2 = 24.97, \quad \hat{\sigma}_2^2 = 82.12.$
- Sparse PCA uses $< 10\%$ of computing time for standard PCA.

Much more to do here:

- more work on interpretation with cardiologists
- effect of registration
- thresholding, multiple leads, ...

Conclusion: Main Themes Again

- initial dimension reduction before PCA
  - otherwise, inconsistency!
- use basis with sparse representation
  - so that little is lost in initial dim reduction
- Background role for large random matrices

SAMSI, ‘Random Matrices’ and FDA?

- **S**tatistical and **A**pplied **M**ath **S**ci **I**nsitute, at NISS, NC.
- RMT active area in math, physics and probability
- Possible semester program, Spring 2005: bring together statisticians and applied math people
- Aim: formulate methodologic & theory questions from statistical areas to pro£t from RMT tools
- Possible statistical areas: climatology (EOFs..), document retrieval, Functional Data Analysis
- Array of problems where p is not £xed, or not small....

References

Ramsay, J. O. & Silverman, B. W. (1997), *Functional Data Analysis*, Springer.

Rice, J. A. & Silverman, B. W. (1991), ‘Estimating the mean and covariance structure nonparametrically when the data are curves’, *Journal of the Royal Statistical Society, Series B (Methodological)* **53**, 233–243.

Silverman, B. W. (1996), ‘Smoothed functional principal components analysis by choice of norm’, *Annals of Statistics* **24**(1), 1–24.