Sparse Principal Components Analysis

Iain Johnstone, Statistics, Stanford & UC Berkeley
Arthur Yu Lu, Stanford and Renaissance Technologies

Support thanks: NIH

FD Workshop, Gainesville, Jan 10-11, 2003
Multiscale Functional Data

Functional data: $x_i(t_j)_{j=1,...,p}$, time points, $i=1,...,n$, cases.

• Focus on PCA: principal modes of variation

Signals $x_i$ contain localized features, perhaps on different scales—visible in p.c.'s also?

Example: ECG signals

"high-dimensional": $p = O(n)$ or larger

---

Functional data: $x_i I = I$, time points, $I = \mathcal{I}$, $x \in \mathcal{I}_x$
ElectroCardioGram traces

Average beat vs. beat-to-beat variation

Electrocardiogram traces
Multiscale Functional Data

• Need for Dimension Reduction: Inconsistency

Sparse PCA

• Examples, incl. ECG

– Some sampling properties, incl. consistency

$\mathcal{N}$ Multiscale Functional Data $p \prec n$
Main Themes

- Initial dimension reduction before PCA
- So that little is lost in initial dim reduction
- Use basis with sparse representation
- Otherwise, inconsistency!

- Initial dimension reduction before PCA
Background Theme: Role of "Random Matrices"

- Small perturbations of symmetric matrices
- a.s. bounds for extreme eigenvalues of large matrices
- more generally, broader role for RMT tools in analysis of FDA methods?
• Multiscale Functional Data

\[ p \approx n \]

⇒ Need for Dimension Reduction: Inconsistency

• Sparse PCA

- Some sampling properties, incl. consistency

Examples, incl. ECG

\[ u \prec \mathbf{p} \]

- Multiscale Functional Data
Single Component Model

(\text{Gaussian noise (e.g. ø \cdot ρ)} \sim \mathcal{N}(I, 0)^d) \bullet

(\text{Random effects (e.g. ø \cdot u = 1024)} \sim \mathcal{N}(I, 0)^p) \bullet

\text{Single component to be estimated} \quad d \subseteq d \bullet

\text{[To illustrate, e.g., need for dimension reduction]}

\textbf{Single Component Model}
True PC, $p = 2048, n = 1024, \|p\| = 10.$

$\lambda = 10^{-12}, \lambda = 10^{-6}.$

ASPCA, $w = 99.5\%, k = 372.$

Standard PCA.
Smooth functional PCA

- Seek smoothed p.c.’s $\xi$ by maximizing

$$\frac{\var(\xi^T x)}{\var(\xi)} + \lambda \|D^2 \xi\|^2$$

- Our example: no choice of $\lambda$ can both

  - remove baseline noise
  - preserve peak heights

- Ssilverman (1997)

Smoothing Functional PCA
\[
\text{standard PCA, } w = 99.5\%, k = 372. \\
\text{smoothed PCA, } \lambda = 10^{-12}, \lambda = 10^{-6}. \\
\text{True PC, } p = 2048, n = 1024, \|p\| = 10. \\
\text{ASPCA, } w = 99.5\%, k = 372. \\
\]
Consistency

Single component model:

\[ x_i = v_i \rho_i + \sigma z_i, i = 1, \ldots, n \]

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

Then a.s.

\[ \lim_{n \to \infty} \sin \angle (\hat{\rho}, \rho) \leq 6 \sigma \sqrt{c / \varrho}. \]

In particular:

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

In particular:

\[ \hat{\partial} / \sqrt{c} \geq (d, \varphi) \sin \angle \lim_{n \to \infty} u \]

Suppose \( 0 < \partial \left\| (u \partial) \right\|, c \left\langle u / (u \partial) \right\rangle \)

Single component model:

Consistency
Suppose $A$, $E$ are symmetric, $q_1$, $\hat{q}_1$ are the principal eigenvectors of $A$, $A + E$. Then

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

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

Then

$\exists \gamma \in \mathbb{R}, \forall \lambda_1 \forall \lambda_2 \forall \lambda \forall (V \hat{q}_1 \geq (V \hat{q}_1 - (V \hat{q}_1^T V \lambda \hat{q}_1) \geq (V \hat{q}_1^T V \hat{q}_1$ – $\gamma V$)

$\exists \gamma, \hat{q}_1, \hat{q}_1'$ are the principal eigenvectors of $A$, $A + E$, $\hat{q}_1, \hat{q}_1'$ are symmetric.

A Matrix Perturbation Theorem.
\[(\cdots, x, \cdots, \top) \overset{\sim}{\leftarrow} (\cdots', \|u\|_F^d\|', \cdots', \|u\|_1^d\|')\]

For asymptotics,

\[\{\imath, \omega\} \text{ noise, independent of } (I, 0)^d \sim \imath \omega \]

\[\text{random effects}\]

\[\{\imath, \omega\} \sim \text{ independent of } \{\imath, \omega\}\]

\[\|u^d\| \lesssim \cdots \lesssim \|\top^d\| \text{ mutually orthogonal, mutually unknown}\]

\[\\begin{align*}
\begin{array}{c}
\|u^{d}_w\| \lesssim \cdots \lesssim \|\top^{d}\| \\
\end{array}
\end{align*}\]

\[u^{d}_w, \cdots, \top^{d} = \imath \omega + \imath d^{\omega}_w \Theta = \imath x\]

A Multicomponent Model
In either single or multicomponent model, Theorem 1: If $\frac{p}{n} \rightarrow c > 0$, then
\[
\lim_{n \to \infty} \sin (\hat{\rho}_1, \rho_1) > 0.
\]

Suggests: Reduce $d \gg k$ before starting PCA relative to $n$. Noise does not average out in PCA if too many dimensions $d$.

\[
\lim_{n \to \infty} \sin \left( \frac{d}{1}, \frac{d}{1} \right) = \lim_{n \to \infty} \frac{u}{n} \left( \frac{d}{1}, \frac{d}{1} \right) < 0.
\]

In either single or multicomponent model, inconsistency.
What goes wrong if $p/n \rightarrow c > 0$?

$$X = \rho v T + \sigma Z S = n$$
$$XX^T = v^T v n \rho \rho + \sigma^2 n ZZ^T$$

Eigenvectors of $n − 1 ZZ^T$ do not approach $1$,

$\lambda_{\max}, \lambda_{\min} \xrightarrow{a.s.} (1 \pm \sqrt{c})^2$ [Geman, Silverstein]

$u = \sigma n^{-1} Z v$ does not vanish:

$\|u\|_2^2 = \sigma^2 p n \chi^2(n) n \chi^2(p) \xrightarrow{a.s.} \sigma^2 c > 0$.

$$B = n L d + L d n + L Z Z \frac{u}{\sqrt{\sigma^2}} + L d d \frac{u}{\sqrt{\sigma^2}} = L X X_{1-u} = S \quad Z \circ + L n d = X$$

What goes wrong if $0 < c \leftarrow u/d$?
Suppose $S = D \pm B \lambda m x$ have principal e-vectors $\hat{\rho} +$ and $\hat{\rho} -$ have same distribution (symmetry) and $\hat{\rho} +$ cannot both be close to $\rho$ and $\hat{\rho} -$ cannot both be close to $\rho$.

But $d(B + D) \perp d B$ and $0 < \|d B\|$ implies $S \subset S$.

Why $\epsilon < 0$ forces inconsistency.
Outline

• Multiscale Functional Data

\[ n \lesssim p \]

• Need for Dimension Reduction: Inconsistency

– Some sampling properties, incl. consistency

• Examples, incl. ECG

Sparse PCA

\[ \Leftarrow \]
Sparse PCA Algorithm

Basis \( x_i(t) = \sum_{\nu} x_{i\nu} e_{\nu}(t) \), \( i = 1, \ldots, n \)

Thresholding

Reduced PCA

Subset

Reconstruction

\( \hat{\rho}_{j\nu} = \eta H(\hat{\rho}_{j\nu}, \delta) \)

\( \hat{x}(t) = \sum_{\nu} \hat{\rho}_{j\nu} e_{\nu}(t) \)
Sparse PCA - Choice of Basis

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

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

Sparsity and weak \( \ell_p \) say

\[
|\rho_\nu| \leq C_\nu \frac{1}{p}, \nu = 1, 2, \ldots
\]

\( d \) small \( \Rightarrow \) rapid decay of ordered coefficients

\[
(\mathcal{C})_{d/m}^{a} d \subseteq (\mathcal{C})_{1}^{a} d
\]

\[
\bigcap_{d}^{a} \left( \mathcal{T} \right)^{a} d = (\mathcal{T})^{a} d
\]

\( \left\{ \mathcal{T}^{a} d \right\} \) basis of \( \{ d \} \) has coefficients in basis of choice of basis.
Wavelet Bases and Sparsity

• Expand $\rho$ in wavelet basis $\{\psi_{jk}\}$:

$$\rho = \sum_{jk} \rho_{jk} \psi_{jk}$$

• Order coefficients $\rho_\nu = \nu$-th largest $|\rho_{jk}|$

**Fact**: smoothness (even non-homogeneous) $\implies$ sparse wavelet representation:

$$\rho_\nu \in \mathcal{W}_p \implies \rho \in B_{\alpha,p,q} \implies b^{d/2^{\alpha+1}}$$

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

$$(1 + \alpha/2)/2 = d \quad d \in m \in (\alpha d) \quad \iff \quad b^{d/2^{\alpha+1}} \in d$$

Sparse wavelet representation:

$\iff$ (smoothness (even non-homogeneous))

$\exists \{\psi_i\}$ largest $d = \alpha d$

Order coefficients $\{\psi_i\}$ $\implies d = \sum \{\psi_i\}$

Expand in wavelet basis $d$
Sparse PCA Algorithm

Basis $x_i(t) = \sum_\nu x_i \nu e_\nu(t)$, $i=1,...,n$

Subset $\hat{\sigma}_2 = \text{Var}\{x_i \nu, i=1,...,n\}$

Thresholding $\hat{\rho}^*_j \nu = \eta H(\hat{\rho}^*_j \nu, \delta)$

Reduced PCA on $\{x_i \nu: \nu \in \hat{\mathcal{I}}_k\}$

Reconstruct $\hat{\rho}_j(t) = \sum_\nu \hat{\rho}^*_j \nu e_\nu(t)$.
Choosing subset size $k$ from data

**Aim:**

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

$$\sum_{\nu \in \hat{\text{I}}} \rho^2_\nu = w(n) \sum \rho^2_\nu, w(n) \rightarrow 1.$$ 

**Possibilities:**

a) $\hat{\text{I}} = \{ \nu : \hat{\sigma}^2(\nu) \geq \hat{\sigma}^2(1 + L \ln) \}$, or

b) define excess over noise using percentiles of $\chi^2(n)$:

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

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

\[ \checkmark \text{Aim: Choose } \hat{I} \text{ to capture most of population p.c.'s variance:} \]

Choosing subset size $k$ from data
Sparse PCA Algorithm

Basis

$\sum_{\nu} x_i \nu e_{\nu}(t) = (t)^{\mathcal{I}_p} x$

Thresholding

$\hat{\sigma}^2_{\nu} = \text{Var}\{x_i \nu, i = 1, \ldots, n\}$

Reduced PCA

$\hat{\rho}_{j\nu} = \eta \mathcal{H}(\hat{\rho}_{j\nu}, \delta)$

Reconstruct

$\hat{\rho}_{j}(t) = \sum_{\nu} \hat{\rho}^{\ast}_{j\nu} e_{\nu}(t)$

Subject

$\{u', \ldots, I', \mathcal{I} \} = \mathcal{I}_p \bigwedge \mathcal{I}_p (u, \ldots, I) = \mathcal{I}_p \mathcal{I}_p$
For standard PCA on $n \times p$ data set $X$, running time is

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

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

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

$$O(p^3) \quad \rightarrow \quad O(k^2p + np \log p).$$
Sparse PCA Algorithm

\[ \sum_{i=1}^{n} \nu x_{i\nu} e_{\nu}(t) = (t)^{\nu}x \]

Reconstruct

\[ (\varphi_{\nu}\hat{c}_{\nu}) H u = \nu \hat{c}_{\nu} \]

Thresholding

\[ \hat{\rho}_{j\nu} \leftarrow \text{eigenvectors} \]

Reduced PCA

\[ \{ \nu \in \Theta : \nu \text{largest variances} \} = \vartheta \]

Subset

\[ \{ u^j I, \cdots, I = \nu \text{var} \} = \vartheta \]

Basis

\[ (d \log d) \Omega \]

\[ (d \log d) \Omega \]

\[ (d \Omega \log d) \Omega \]

\[ (d \Omega \log d) \Omega \]
Thresholding subset eigenvectors

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

By analogy with wavelet shrinkage in regression:

keep large coefficients, kill noise.

Here, usual hard thresholding:

For now, use usual "universal" threshold

Other variants: soft, SCAD ...

Choice of $\delta$.

For now, use usual "universal" threshold

$\delta = \hat{\tau}_{j} \sqrt{2 \log k}$,

$\hat{\tau}_{j} = \frac{\text{MAD}\{\hat{\rho}_{j,\nu}, \nu = 1, \ldots, k\}}{0.6745}$. 

Thresholding subset eigenvectors
Sparse PCA Algorithm

Basis

\[ \sum_{\nu} x_{i\nu} e_{\nu}(t) = (t)^{\nu} \]

Subset

\[ \hat{\sigma}_{\nu}^2 = \text{Var}\{x_{i\nu}, i = 1, \ldots, n\} \]

\[ \hat{I}_k = \{ \nu \leftrightarrow \text{largest } k \text{ variances}\} \]

Reduced PCA

\[ \text{on } \{ \nu \in \mathcal{I} : \nu x \} \]

\[ \{u, \ldots, \mathcal{I} = \nu x\} \]

Thresholding

\[ \hat{\rho}_{j\nu}^* = \eta H(\hat{\rho}_{j\nu}, \delta) \]

Reconstruct

\[ \hat{\rho}_j(t) = \sum_{\nu} \hat{\rho}_{j\nu}^* e_{\nu}(t) = (t)^{\nu} x \]

\[ O(k^3) \]

\[ O(p \log p) \]

\[ O(k^2 p) \]
\[
\begin{align*}
\text{True PC}, &\quad p = 2048, n = 1024, \quad \|p\| = 10. \\
\text{standard PCA} &\quad \lambda = 10^{-12} \\
\text{smoothed PCA} &\quad \lambda = 10^{-6} \\
\text{ASPCA}, &\quad w = 99.5\%, k = 372.
\end{align*}
\]
Smoothed PCA, \( b: 10^{-12}, r: 10^{-8}, m: 10^{-6} \).

ASPCA, \( w = 99.5\%, k = 438 \).

Standard PCA.
## Speed and Accuracy Comparison

<table>
<thead>
<tr>
<th>Time (step)</th>
<th>ASE (3-peak)</th>
<th>Time (step)</th>
<th>ASE (3-peak)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 min 31 s</td>
<td>1.947e-04</td>
<td>1 min 46 min</td>
<td>3.174e-03</td>
</tr>
<tr>
<td>1.47e-04</td>
<td>9.715e-04</td>
<td>1 min 47 min</td>
<td>1.694e-2</td>
</tr>
<tr>
<td>1 min 15 s</td>
<td>3.627e-2</td>
<td>1 min 47 min</td>
<td>1.947e-04</td>
</tr>
<tr>
<td>7.50e-05</td>
<td>9.681e-04</td>
<td>1 min 12 min</td>
<td>1.0e-6</td>
</tr>
<tr>
<td>PCA Sparse</td>
<td>PCA Smoothed</td>
<td>PCA Smoothed</td>
<td>PCA Standard</td>
</tr>
</tbody>
</table>

*Note: Times are approximate and may vary.*
Outline

- Multiscale Functional Data $\sim p \gg n$
- Need for Dimension Reduction: Inconsistency
- Sparse PCA
- Some sampling properties, incl. consistency
- Examples, incl. ECG
Correct Selection Properties

$E. 9.$: Suppose $\hat{\sigma}^2 \sim \sigma^2 \chi^2(n)/n$. If $\alpha = \gamma \sqrt{\log n}/n$, then

$$P\{\text{FE} \cup \text{FI}\} \leq 3\sqrt{\gamma} n^{-\beta \gamma^2/4}.$$

E.g.: $(50, 1000, 1000, 1000) = (u, \nu, \hat{\sigma}^2, \hat{\sigma}^2(1)_{\frac{\gamma}{2}}) \to (u, \nu, \hat{\sigma}^2, \hat{\sigma}^2(1)_{\frac{\gamma}{2}}) \to I. 25$.

De£ne:

$$\{\frac{\gamma}{2} I \subseteq \nu_{\text{no}} I\} = \text{False Inclusion} = I \hat{\Sigma}$$

$$\{\frac{\gamma}{2} I \subseteq \nu_{\text{true}} I\} = \text{False Exclusion} = \Sigma \hat{\Sigma}$$

$$\{\frac{\gamma}{2} I \overset{\hat{\sigma}}{\subseteq} \nu_{\text{large}} I\} = \text{Include all “large” variables}$$

$$\{\frac{\gamma}{2} I \overset{\hat{\sigma}}{\subseteq} \nu_{\text{small}} I\} = \text{Exclude all “small” variables}$$

I.e.: Does $I \hat{\Sigma}$ include the “right” variables?
Consistency of Sparse PCA

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

Assume Sparsity: \( \rho(n) \in w^{\ell_p}(C) \) uniformly in \( n \).

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

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

\begin{align*}
\angle (\hat{\rho}, \rho) \overset{a.s.}{\to} 0.
\end{align*}

Theorem

\( n \)

\begin{align*}
\left\{ \left( \frac{u}{\varrho} \right)^{\frac{d}{2}} \left( \frac{d}{2} \right)^{\frac{1}{2}} \log \left( \frac{d}{2} \right) + 1 \right\} \varrho^2 < \frac{a}{\varrho} n \cdot \lambda \} = \hat{I}.
\end{align*}

Subset selection rule: \( \hat{I} \).

Assume Sparsity: \( \mathcal{C}^{d \mathcal{F}_m} \supseteq (u)^d \).

\begin{align*}
0 < \bar{d} \leftarrow \| (u)^d \| \quad \text{(II)}
\end{align*}

Single component model. Suppose (i) \( 0 < c \leftarrow u/d \) (ii) Suppose (i) \( 0 < c \leftarrow u/d \) (ii)
Outline

• Multiscale Functional Data
  \[ n \lesssim p \]

• Need for Dimension Reduction: Inconsistency

• Some sampling properties, incl. consistency

Examples, incl. ECG

Sparse PCA
Preprocessing: piecewise linear baseline wander removal,
registration at R-wave maximum,
interpolation to 512 samples per cycle.
c) Average curve of sample 2, n = 61

b) 1st principal component for sample 1

Standard PCA  
Sparse PCA

a) Average curve of sample 1, n = 66

d) 1st principal component for sample 2

Standard PCA  
Sparse PCA
Remarks

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

- More work on interpretation with cardiologists.
- Effect of registration.
- Thresholding, multiple leads.

Much more to do here:

- More work on interpretation with cardiologists.
- Effect of registration.
- Thresholding, multiple leads.
Background role for large random matrices

- so that little is lost in initial dim reduction

use basis with sparse representation

- otherwise, inconsistency!

Initial dimension reduction before PCA
SAMSI, Random Matrices, and FDA?

Array of problems where $p$ is not fixed, or not small...

Retrieval, Functional Data Analysis

Possible statistical areas: Climatology (EOFs), document areas to profit from RMT tools

Aim: formulate methodological & theory questions from statistical

statistics and applied math people

Possible semester program, Spring 2005: bring together

RMT active area in math, physics and probability

Statistical and Applied Math SCI Institute, at NISS, NC.
