High-dimensional eigen-analysis, spiked models, and regularization

Debashis Paul

(University of California at Davis, Dept. of Statistics)

NCAR workshop

May 07, 2007
Agenda

• High dimensional multivariate data: examples

• Review of standard PCA
  – Objective
  – Drawbacks

• Sparsity
  – Motivation
  – Sparse PCA algorithm
  – Augmented Sparse PCA
  – Simulation study
  – Spatio-temporal extension
ECG data

Goal: Structural analysis
Monthly Average Sea Level Pressure: January

$p = 15 \times 72, n \approx 100$ (NCAR database)

Year 1990

Year 1994

Year 1997

Year 2000

Year 2002

Year 2004

**Goal**: Model diagnostics
Other examples

- Analysis of gene expression array (prediction of survival)
- Speech and image recognition (data compression, denoising)
- Financial data - stock prices (structural analysis)
Common characteristics

• High dimension \( (p) \)
• Comparatively low sample size \( (n) \)
• Noisy observations
A standard approach

- Represent the observation vectors in a “good” basis.
- Preserve the covariance structure.
- Accomplish this (approximately) with fewer basis elements.
Principal component analysis

Data: \( X \) (\( p \times n \) matrix) where

\[
X_i, \; i = 1, \ldots, n, \quad \text{i.i.d.} \; (\mu, \Sigma)
\]

- **Strategy**: Find orthonormal basis vectors \( u_1, u_2, \ldots \) such that \( \text{Variance}(u_\nu^t X) \) is maximized subject to \( u_\nu \perp u_1, \ldots, u_{\nu-1} \).

- **Solution**: \( u_1, u_2, \ldots \) are eigenvectors of \( \Sigma \) corresponding to eigenvalues \( \ell_1 \geq \ell_2 \geq \ldots \geq \ell_p \geq 0 \)

\[
\Sigma u_\nu = \ell_\nu u_\nu, \quad \nu = 1, 2, \ldots, \text{rank}(\Sigma)
\]
Principal component analysis : illustration
Standard PCA

- Define *sample covariance matrix*

\[ S := \frac{1}{n}(X - \hat{\mu}1^t)(X - \hat{\mu}1^t)^t, \quad \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i \]

- Estimate eigenvectors and eigenvalues of \( \Sigma \) by eigenvectors and eigenvalues of \( S \).
Model I: High dimensional i.i.d. data

- **Large dimension**: e.g. \( p \) (dimension) and \( n \) (sample size) are such that \( \frac{p}{n} \approx \gamma > 0 \).

- **Model**: \( X_1, \ldots, X_n \) i.i.d. \( N(\mu, \Sigma) \), where \( \Sigma = \Sigma_0 + \sigma^2 I \).

- **Components of variation**:

\[
X_i = \mu + \sum_{\nu=1}^{M} \sqrt{\lambda_{\nu}} v_{\nu i} u_{\nu} + \sigma Z_i, \quad i = 1, \ldots, n
\]

where \( \lambda_1 \geq \ldots \geq \lambda_M > 0 \) are the eigenvalues of \( \Sigma_0 \), and \( u_1, \ldots, u_M \) are corresponding eigenvectors.

\( \{v_{\nu i} : 1 \leq \nu \leq M\} \) i.i.d. \( N(0, 1) \) independent of \( Z_i \sim N_p(0, I_p) \)

- **Interpretation as Random Signal** (\( M \) dimensional) + **Isotropic Noise**
Drawback of standard PCA

- Noisy estimates for large dimension.
- High computational cost.
A toy example

Some sample curves: $M = 1$, $\lambda = 7$, $p = 512$
Standard PCA estimate

True eigenvector

Standard PCA estimate

$M = 1$, $p = 512$, $n = 500$, $\lambda_1 = 7$, $\sigma = 1$
Notion of consistency

An estimator $\hat{u}_\nu$ is consistent for true eigenvector $u_\nu$ if

$$\angle(u_\nu, \hat{u}_\nu) \to 0, \text{ in probability as } n \to \infty$$
Results on sample eigenvalues
(Paul (2005), Baik and Silverstein (2005), Onatski (2005))

Suppose eigenvalues of $\Sigma$ ($p \times p$) are $\ell_1 > \ldots > \ell_M > 1 = \ldots = 1$.
(Note, $\ell_{\nu} = \lambda_{\nu} + \sigma^2$ and $\sigma = 1$). Let $\frac{p}{n} \to \gamma \in (0, 1)$.

- Phase transition:

$$
\hat{\ell}_\nu \xrightarrow{a.s.} \begin{cases} 
(1 + \sqrt{\gamma})^2 & \text{if } \ell_{\nu} \in (1, 1 + \sqrt{\gamma}] \\
\rho_{\nu} := \ell_{\nu} \left(1 + \frac{\gamma}{\ell_{\nu} - 1}\right) & \text{if } \ell_{\nu} > 1 + \sqrt{\gamma}
\end{cases}
$$

- Asymptotic distribution: If $\ell_{\nu} > 1 + \sqrt{\gamma}$ and of multiplicity 1, and $|\frac{p}{n} - \gamma| = o(n^{-1/2})$,

$$
\sqrt{n}(\hat{\ell}_\nu - \rho_{\nu}) \Rightarrow N(0, \tilde{\sigma}_{\nu}^2), \text{ where } \tilde{\sigma}_{\nu}^2 = 2\ell_{\nu}^2 \left(1 - \frac{\gamma}{(\ell_{\nu} - 1)^2}\right)
$$
Phase transition of eigenvalues
Asymptotics of sample eigenvectors
(Paul (2005), Hoyle and Rattray (2004), Onatski (2005))

Eigenvector associated with $\ell_\nu > 1$ of multiplicity one.

True: $u_\nu$ Estimated: $\hat{u}_\nu$

- Phase transition:

$$|\langle u_\nu, \hat{u}_\nu \rangle| \xrightarrow{a.s.} \begin{cases} 0 & \text{if } \ell_\nu \in (1, 1 + \sqrt{\gamma}] \\ \sqrt{\frac{1 - \frac{\gamma}{(\ell_\nu - 1)^2}}{1 + \frac{\gamma}{\ell_\nu - 1}}} & \text{if } \ell_\nu > 1 + \sqrt{\gamma} \end{cases}$$
Issues

• How to reduce noise from the estimates of eigenvectors, and achieve consistency?
• How many components are to be used to represent the data?
• Can these be done at a low computational cost?
Smoothing based approach

• Assume that the observations are noisy versions of smooth curves

• While estimating the principal components, add a penalty to ensure that the estimate is reasonable smooth.

• Boils down to essentially applying a linear smoothing of the curves and then a further smoothing of eigenvectors of the covariance matrix of the smoothed curves.

• Needs specification of a smoothing parameter.

Sparsity based approach

• Assume that the observations are **noisy versions of some random curves**

• Assume that the noiseless curves can be **sparsely represented** in some known orthogonal basis

• Data first transformed in an **orthogonal basis**

• Sparseness of the observation vectors in the transformed basis used for reducing noise

Sparsity : an illustration

\[ \sum |\text{wavelet coefficients}| = 3.732, \quad \sum |\text{Fourier coefficients}| = 6.348 \]
Suitable basis \equiv \text{sparse representation}

- **Heuristics:**
  If the eigenvectors $u_\nu$ are \textit{sparse} (in an appropriate basis), \textit{i.e.} \textit{have few significantly large coefficients}, then a smaller submatrix of $\Sigma$ contains most of the useful information.

- **Motivation:**
  - Observations are smooth random functions $\implies$ eigenvectors are smooth and can be sparsely represented in a wavelet/spline basis.
  - In many applications (e.g. microarray experiments) very few coordinates (genes) may carry any useful information.
Sparse PCA: Idea

With $\Sigma = \Sigma_0 + \sigma^2 I$,

$$\Sigma_{kk} = \sum_{\nu=1}^{M} \lambda_{\nu} u_{\nu k}^2 + \sigma^2 = \Sigma_{0,kk} + \sigma^2$$

The sequence $(\Sigma_{0,kk}, k = 1, \ldots, p)$ is sparse if and only if the eigenvectors $\{u_{\nu}, \nu = 1, \ldots, M\}$ are sparse.
Sparse PCA Algorithm (Lu, 2002)

Hereafter we assume that we already have a *sparse representation* of the observations in some *known basis*.

- Set a threshold $t_n > \sigma^2$. Select coordinate $k$ if $S_{kk} > t_n$. Call this set $B_n$.
- Perform PCA on the data $\{X_{li} : l \in B_n, i = 1, \ldots, n\}$.
- Extend the eigenvectors to full dimension by adding zeros to coordinates not in $B_n$. 

Consistency

Result (Johnstone and Lu, 2004): 

Under assumption of Gaussianity, and if $\| u_\nu \|_q \leq C_\nu < \infty$, for some $0 < q < 2$, then taking $t_n = \sigma^2 (1 + c_0 \sqrt{\frac{\log (p \wedge n)}{n}})$, the estimate $\hat{u}_\nu$ obtained from SPCA scheme satisfies

$$\angle (\hat{u}_\nu, u_\nu) \rightarrow 0 \quad \text{a.s. as } n \rightarrow \infty$$

Here $\| u_\nu \|_q \overset{def}{=} [\sum_{k=1}^{p} |u_{\nu k}|^q]^{1/q}$
Drawback of Sparse PCA

- Sparse PCA does not use the covariance structure for selecting coordinates.

- Sparse PCA can only recover coordinates that are *rather large* (of size at least $c_0 \sigma \left( \frac{\log(p\sqrt{n})}{n\lambda_2^2} \right)^{1/4}$).

- Hence, if the eigenvectors are not *very sparse*, Sparse PCA scheme can have *large* bias.
Improved estimation scheme

**Idea:** If we have a “reasonably good” estimator of the eigenvectors $u_{ij}$’s then use that to utilize the **structure of covariance**.

$$M = 1, \; \Sigma_{0,bb} = \lambda_{1} u_{1b}^2, \; \Sigma_{0,ss} = \lambda_{1} u_{1s}^2, \; \Sigma_{0,bs} = \lambda_{1} u_{1b} u_{1s}$$
Structure of covariance

- Three **classes of coordinates** (based on the size of their variance): $B \equiv \text{big}$, $S \equiv \text{small}$, $T \equiv \text{tiny}$.

- $S \cup T = B^c$, $B \cup B^c = \{1, \ldots, p\}$.

- Partition $\Sigma_0$ and $\mathbf{u}_\nu$:

\[
\Sigma_0 = \begin{bmatrix}
\Sigma_{0,BB} & \Sigma_{0,BB^c} \\
\Sigma_{0,B^cB} & \Sigma_{0,B^cB^c}
\end{bmatrix} \quad \mathbf{u}_\nu = \begin{bmatrix}
\mathbf{u}_{\nu,B} \\
\mathbf{u}_{\nu,B^c}
\end{bmatrix} \quad \mathbf{u}_{\nu,B^c} = \begin{bmatrix}
\mathbf{u}_{\nu,S} \\
\mathbf{u}_{\nu,T}
\end{bmatrix}
\]

- Want to retain coordinates that are in $B \cup S$, but don’t want to select coordinates that are in $T$. 
Augmented SPCA

Stage I

- Get preliminary estimates $\hat{u}_\nu$’s by SPCA scheme. $\hat{B} =$ set of selected coordinates. Let $q_n$ be the size of the set $\hat{B}$.

- Estimate $M$ by retaining only those eigenvalues $\hat{\ell}_1 \geq \ldots \geq \hat{\ell}_{\min\{q_n,n\}}$ of $S_{\hat{B}\hat{B}}$ that are above a threshold $\Lambda_n$. 
Stage II

• Compute \( \tilde{u}_{\nu, \tilde{B}_c} := \ell_{\nu}^{-1} s_{\tilde{B}_c \tilde{B}} \tilde{u}_{\nu, \tilde{B}}, \quad \nu = 1, \ldots, \tilde{M} \)

• Select a coordinate \( k \) from \( \tilde{B}_c \) if

\[
\sum_{\nu=1}^{\tilde{M}} \ell_{\nu} \tilde{u}_{\nu, k}^2 > t_n
\]

for some threshold \( t_n \).

• \( \hat{S} := \) coordinates selected at the second stage. Perform standard PCA on the submatrix of \( S \) (full data) corresp. to \( \hat{C} = \hat{B} \cup \hat{S} \).
Choice of threshold

• Stage I (Same as Sparse PCA):

\[ t_n = \sigma^2 (1 + c_0 \sqrt{\frac{\log(p \vee n)}{n}}) \]

• Threshold for the eigenvalues (for estimating \( M \)):

\[ \Lambda_n = \sigma^2 [(1 + \sqrt{\frac{qn}{n}})^2 + c_1 \sqrt{\frac{\log(p \vee n)qn}{n}}] \]

• Stage II:

\[ \bar{t}_n = \sigma^2 c_2 \frac{\log(p \vee n)}{n} \]
Computational efficiency

- ASPCA requires $O(\min\{nk_n^2, n^2k_n\})$ computations, where $k_n$ is the number of selected coordinate.
- Typically $k_n$ is a lot smaller than $p$ (due to sparsity).
- Much faster than standard PCA or smoothed (regularized) PCA (Ramsay and Silverman, 1997).
Simulation study

Dimension $p = 512$, sample size $n = 500$. $M = 3$, $\lambda_1 = 25$, $\lambda_2 = 15$, $\lambda_3 = 10$, $\sigma = 1$. 
Analysis

**Loss function:**

Squared error type loss

\[
L(u, v) = 2(1 - |\langle u, v \rangle|) = \| u - \text{sign}(\langle u, v \rangle)v \|^2 = 4 \sin^2 \frac{1}{2} \angle(u, v)
\]

for \( u, v \in \mathbb{S}^{p-1} \).

**Assumptions:**

- \( \lambda_1 \rightarrow \lambda_{1,\infty} \in [0, \infty], \frac{\lambda_1}{\lambda_2} \rightarrow r_\nu \) for \( \nu = 1, \ldots, M \) with
  \[
  1 = r_1 > r_2 > \ldots > r_M > 0.
  \]
- \( c_1 \log n \leq \log p \leq c_2 \log n \) for \( c_2 \geq 1 \geq c_1 > 0 \).
Two important functions

\[ h(x) = \frac{x^2}{1+x}, \quad x > 0, \quad g(x, y) = \frac{(x - y)^2}{xy}, \quad x \neq y > 0 \]

(nonparametric component) (nuisance parameters)

Risk behavior: standard PCA

Let \( \Theta = \{u_\nu \text{ orthonormal} : u_\nu \in S^{p-1}, \nu = 1, \ldots, M \} \)

W.l.o.g. take \( \sigma = 1 \).

\[
\sup_{u \in \Theta} \mathbb{E}L(\hat{u}_\nu, u_\nu) \leq \left[ \frac{p}{n h(\lambda_\nu)} + \frac{1}{n} \sum_{\nu \neq \nu'} \frac{1}{g(1 + \lambda_\nu, 1 + \lambda_{\nu'})} \right] (1 + o(1))
\]

whenever RHS converges to 0.
Parameter space

Assume $\Theta_q = \{ u_\nu \in S^{p-1} \text{ orthonormal} : \| u_\nu \|^q \leq C_{\nu}^q, \ \nu = 1, \ldots, M \}$.
Risk behavior: Lower bound

\[
\inf_{\hat{u}_\nu} \sup_{u \in \Theta_q} \mathbb{E}L(\hat{u}_\nu, u_\nu) \geq K \left[ \max\{\delta_n, \frac{1}{n} \max_{\nu \neq \nu'} \frac{1}{g(1 + \lambda_\nu, 1 + \lambda_{\nu'})} \} \wedge 1 \right] \quad (1)
\]

where

\[
\delta_n = \begin{cases} 
  \frac{C^q_{\nu} (\log p)^{1-a/2}}{(nh(\lambda_{\nu}))^{1-a/2}} & \text{if } \frac{C^q_{\nu} (nh(\lambda_{\nu}))^{q/2}}{(\log p)^{q/2}} = O(p^{1-\alpha}), \quad \text{highly sparse} \\
  \frac{C^q_{\nu}}{(nh(\lambda_{\nu}))^{1-q/2}} & \text{if } C^q_{\nu} (nh(\lambda_{\nu}))^{q/2} \leq p, \quad \text{sparse} \\
  \frac{p}{nh(\lambda_{\nu})} & \text{if } C^q_{\nu} (nh(\lambda_{\nu}))^{q/2} \geq p, \quad \text{dense}
\end{cases}
\]

where \(\alpha \in (0, 1)\).
Understanding lower bound

• Note that $\delta_n$ is of the form $m_n \tau_n^2$, where

$$m_n = \begin{cases} 
\left[ \frac{C^2 nh(\lambda_\nu)}{\log p} \right]^{q/2} \\
\left[ C^2 nh(\lambda_\nu) \right]^{q/2} \\
p
\end{cases} \quad \tau_n = \begin{cases} 
\frac{\sqrt{\log p}}{\sqrt{nh(\lambda_\nu)}} & \text{highly sparse} \\
\frac{1}{\sqrt{nh(\lambda_\nu)}} & \text{sparse} \\
\frac{1}{\sqrt{nh(\lambda_\nu)}} & \text{dense}
\end{cases}$$

• Interpretation:

$m_n$ = effective dimension

$\tau_n = (\sqrt{\log p}) \times$ asymptotic noise per parameter
Noteworthy features

• Three regimes of sparsity - highly sparse, sparse, dense.

• Dense \equiv \textit{no knowledge} about the eigenvectors \implies no estimator uniformly consistent over $\mathbb{S}_p^{-1}$ when $\frac{p}{nh(\lambda)} \not\to 0$.

• Standard PCA has optimal rate of convergence in the dense case.

• Dependence on the \textit{spacings} between eigenvalues
Risk behavior : Augmented SPCA estimates

\[
\sup_{\mathbf{u} \in \Theta_q} \mathbb{E}L(\hat{\mathbf{u}}_\nu, \mathbf{u}_\nu) \leq K(q) [\bar{\tau}_\nu^{-q/2} \left( \sum_{\nu' = 1}^{M} \bar{\tau}_{\nu'}^{q/2} C^q_{\nu'} \right) (\log p)^{1-q/2} \frac{1}{(nh(\lambda_\nu))^{1-q/2}} + \frac{1}{n} \sum_{\nu \neq \nu'} \frac{1}{g(1 + \lambda_\nu, 1 + \lambda_{\nu'})} ] (1 + o(1))
\]

provided the risk of for SPCA estimates converge to zero. Here \( \bar{\tau}_\nu = h(\lambda_\nu)/h(\lambda_1) \).

- Does not depend on \( p \) except for a \( \log p \) factor.

- In the case \( M = 1 \), upper bound of the form \( K(q) m_n \bar{\tau}_n^2 (1 + o(1)) \),

where

\[
m_n = \left[ \frac{C^2 n h(\lambda_\nu)}{\log p} \right]^{q/2}, \quad \bar{\tau}_n = \frac{\sqrt{\log p}}{\sqrt{nh(\lambda_\nu)}}
\]
Summary of the results

When sparsity is measured over parameter spaces that impose constraint on the $l^q$ norm ($0 < q < 2$) of the eigenvectors,

- There are different regimes of sparsity that give rise to different rates of convergence
- ASPCA achieves the optimal rate in the highly sparse case
- Usual PCA can achieve the optimal rate if there are sufficiently many observations
- Without enough sparsity it may not be possible to estimate the eigenvectors uniformly consistently
Beyond i.i.d. setting

- Separable space-time covariance.
- Dynamic spatio-temporal systems.
Model II: Separable Spatio-temporal process

- Apart from spatial (coordinate-wise) dependence, there is also a “time”-dependence - observations $X_1, \ldots, X_n$ are \textbf{not} independent.

- \textbf{Separable spatio-temporal model}: assume that spatial variability and temporal variability do not depend on each other. i.e., the $N \times n$ data matrix $X_n = [X_1 : \ldots : X_n]$ has covariance $\Sigma_N \otimes \Delta_n$ where $\Sigma_N$ is $N \times N$ and $\Delta_n$ is $n \times n$. 
Separable “Spiked” model

$\Sigma_N$ has eigenvalues $\ell_1 \geq \ldots \geq \ell_M > 1 = \ldots = 1$

Empirical distribution of the eigenvalues of $\Delta_n$ converges to a
distribution $G$ compactly supported on $\mathbb{R}^+$ (and satisfies some
technical condition).

Let $F^\Delta$ denote the limiting ESD of the matrix $\frac{1}{n}Z\Delta_nZ^T$ where $Z$ is
$(N - M) \times n$ with i.i.d. $N(0, 1)$ entries.

**Result**: There is a threshold $\tau_{c,\Delta}$ such that, if $\ell_k > \tau_{c,\Delta}$ then $\hat{\ell}_k \xrightarrow{a.s.} \rho_k$
where $\rho_k$ satisfies

$$\rho_k = \ell_k \int \frac{t}{1 - t\alpha_1(\rho_k, c)}dG(t)$$

where

$$\alpha_j(\rho, c) = c \int \frac{1}{(\rho - x)^j}dF^\Delta(x), \quad j = 1, 2, \ldots.$$ 

for $\rho > \max \ supp(F^\Delta)$. 

48
Asymptotic normality

**Result**: Under the stated assumptions (plus a few regularity conditions), if \( \ell_k \) is of multiplicity one, then

\[
\sqrt{n(\hat{\ell}_k - \rho_k)} \implies N(0, \sigma^2(\ell_k))
\]

where

\[
\sigma^2(\ell_k) = 2\ell_k^2 \left[ \frac{\beta(\rho_k, c)(1 + \alpha_2(\rho_k, c)\beta(\rho_k, c))}{(1 + \ell_k\alpha_2(\rho_k, c)\beta(\rho_k, c))^2} \right]
\]

with

\[
\beta(\rho_k, c) := \int \frac{t^2}{(1 - t\alpha_1(\rho_k, c))^2} dG(t).
\]

Analogous results for sample eigenvectors (a.s. limit, asymptotic normality).
Model III: Autoregressive process

- A Vector Autoregressive Process of order 1 (VAR(1)) in dimension $N$ is given by
  
  $$X_t = AX_{t-1} + \varepsilon_t, \quad t = \ldots, -2, -1, 0, 1, 2, \ldots,$$
  
  where $\varepsilon_t \in \mathbb{R}^N$ are i.i.d. with mean 0 and covariance matrix $\Psi$, $A \in \mathbb{R}^{N \times N}$.

- **Special case:** $A = A^T$ and $\Psi = I_N$. If $\|A\| < 1$, then $\{X_t\}$ is stationary and $\Sigma_N := \text{Var}(X_t) = (I - A^2)^{-1}$.

- **Spiked model:** If $A = A^T$, $\Psi = I_N$, $\|A\| < 1$ and rank$(A) = r$ fixed and < $N$, then $\Sigma$ has eigenvalues $\ell_j = \frac{1}{1-\sigma_j^2(A)}$, $j = 1, \ldots, M$, and $\ell_j = 1$ for $j = M + 1, \ldots, N$. 
Asymptotics for eigenvalues

- We observe $X_1, \ldots, X_n$. Assume that $n = n(N)$ is such that $\frac{N}{n} \to c \in (0, 1)$.

- If $j \in \{1, \ldots, r\}$ is such that $\ell_j > \tau_c$, for some $\tau_c > 0$, then

  $$\hat{\ell}_j \xrightarrow{a.s.} \rho_j \quad \text{with} \quad \rho_j = \frac{\ell_j}{1 - g_c(\rho_j)},$$

  with $g_c(\rho) = c \int \frac{1}{\rho - x} dF^M_P(x)$.

- Asymptotic normality of $j$-th sample eigenvalue and corresponding eigenvector if $\ell_j > \tau_c$ and has multiplicity 1.
Research directions

- Exploration of the relationship between nonparametric regression and estimation of eigenvectors
- Different approach to regularization - e.g. direct penalization
- Confidence balls for eigenvectors for high-dimensional covariances
- Extension to spatio-temporal problems (for example, those arising in geophysical and meteorological applications)
- Application to signal estimation for sensor networks under noisy background
- Application to spectro-temporal problems in chemometrics