On Consistency and Sparsity for PCA in High Dimensions

Ganchao Wei

October 20, 2021

Overview

Introduction

In regular PCA, we assume $n \gg p$. However, in many situations, p is comparable in magnitude with n or even $n < p$ (high-dimensional settings). In this paper, they:

- describe inconsistency results to emphasize that when p is comparable with n, we need to reduce dimension.
- establish consistency results to illustrate that the reduction in dimensionality ca be effected working in a basis in which the signals have a sparse representation.

Setting: single factor model

$$\boldsymbol{x}_i = \nu_i \boldsymbol{\rho} + \sigma \boldsymbol{z}_i, \ \ i = 1, \ldots, n$$

, where $\boldsymbol{x}_i \in \mathbb{R}^p$, $\boldsymbol{\rho}_i \in \mathbb{R}^p$, $\nu_i \sim N(0, 1)$ and $\boldsymbol{z}_i \sim N_p(0, I)$

Introduction: Motivating Example

$$\boldsymbol{x}_i = \nu_i \boldsymbol{\rho} + \sigma \boldsymbol{z}_i, \ i = 1, \ldots, n$$

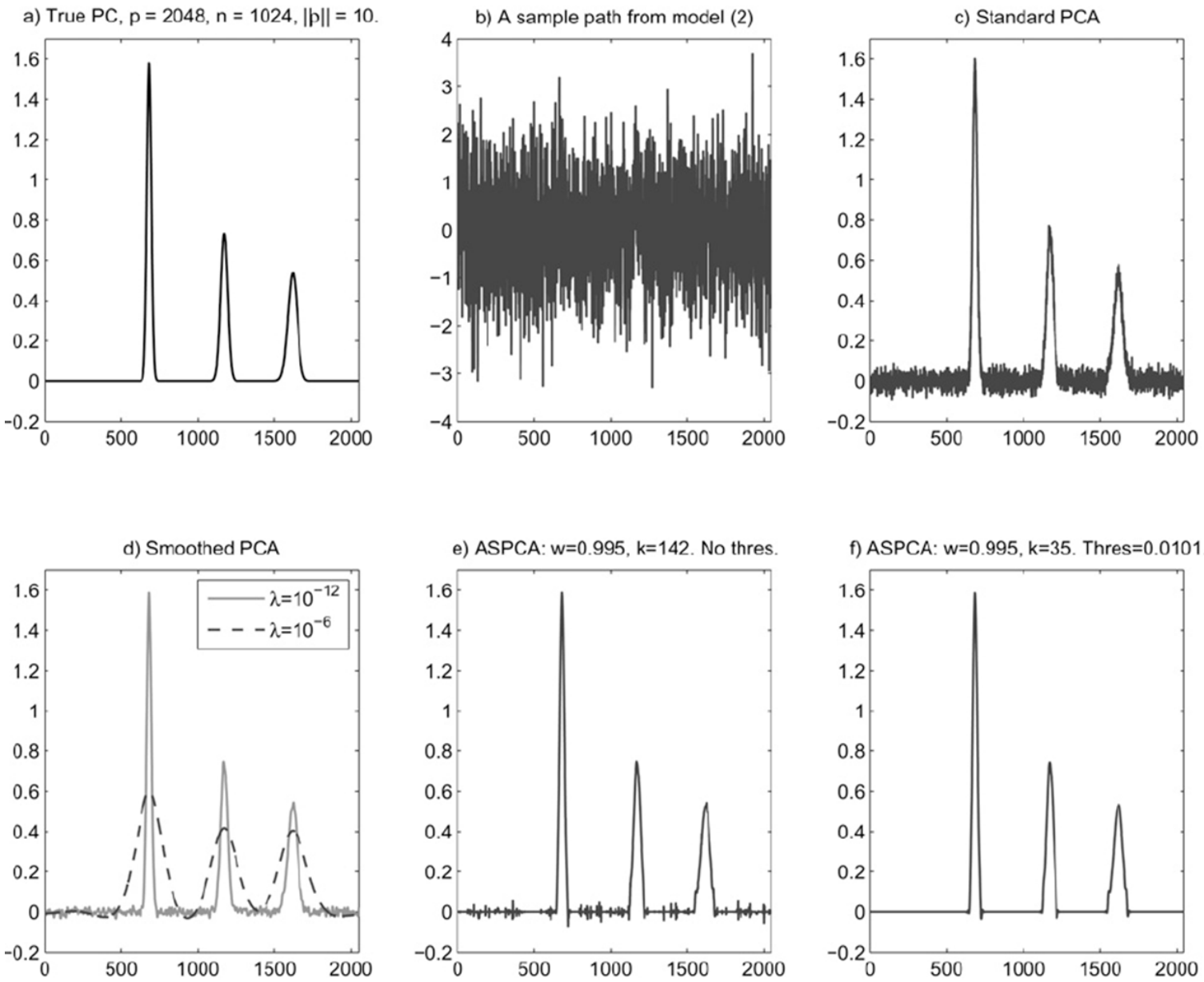
,where $p = 2048, n = 1024$. The vector $\rho_l = f(l/p)$ for $l \in 1, \ldots, p$, and $f(t)$ is a mixture of beta desnities on $[0, 1]$, scaled so that $||\boldsymbol{\rho}||_2 = 10$. Specifically,
$f(t) = C0.7Beta(1500, 3000) + 0.5Beta(1200, 900) + 0.5Beta(600, 160)$.
The $\sigma = 1$. Here, they analyzed the data by 4 different methods (results in the next slide):

- standard PCA
- smoothed PCA: similar to LASSO, add penalty terms, i.e. maximize

$$var(\boldsymbol{\xi}' \boldsymbol{x}_i)/[||\boldsymbol{\xi}||^2 + \lambda ||D^2 \boldsymbol{\xi}||^2]$$

- adaptive sparse PCA, without thresholding (this paper)
- adaptive sparse PCA, with thresholding (this paper)

Introduction: Motivating Example



Inconsistency of Classic PCA

First, some notations & definitions:

- $\boldsymbol{S}$ = sample covariance, $\hat{\boldsymbol{\rho}}$ = eigenvectors with largest eigenvalue
- overlap between 2 vectors: $R(\hat{\boldsymbol{\rho}}, \boldsymbol{\rho}) = \cos \angle(\hat{\boldsymbol{\rho}}, \boldsymbol{\rho})$
- distance: $d(\hat{\boldsymbol{\rho}}, \boldsymbol{\rho}) = \sin \angle(\hat{\boldsymbol{\rho}}, \boldsymbol{\rho})$

We also need 2 more assumptions:

- dimension growth: $\lim_{n\to\infty} p_n/n = c$
- limiting SNR: $\lim_{n\to\infty} ||\boldsymbol{\rho}_n||^2/\sigma^2 = \omega > 0$

Inconsistency of Classic PCA

This leads to the limiting results (Theorem 1)

Theorem 1. Assume that there are n observations drawn from the p-dimensional model (2). Assume that $p_n/n \to c$ and that $\|\boldsymbol{\rho}_n\|^2/\sigma^2 \to \omega > 0$. Then almost surely

$$\lim_n R^2(\hat{\boldsymbol{\rho}}, \boldsymbol{\rho}) = R^2_\infty(\omega, c) = \frac{(\omega^2 - c)_+}{\omega^2 + c\omega}.$$

In particular, $R_\infty(\omega, c) < 1$ if and only if $c > 0$, and so $\hat{\boldsymbol{\rho}}$ is a consistent estimator of $\boldsymbol{\rho}$ if and only if $p/n \to 0$.

The situation is even worse if $\omega^2 \leq c$—that is, if

$$\lim \frac{p}{n} \frac{\sigma^4}{||\boldsymbol{\rho}||^4} \geq 1,$$

because $\hat{\boldsymbol{\rho}}$ and $\boldsymbol{\rho}$ are asymptotically orthogonal, and $\hat{\boldsymbol{\rho}}$ ultimately contains no information at all regarding $\boldsymbol{\rho}$.

The theorem can be easily extended to the multi-component case.

Sparsity

Theorem 1 tells us that regular PCA becomes confused when there are too many variables each with equal independent noise $\Rightarrow$ reduce dimension. Assume the data and population PC's are represented in a fixed orthonormal basis $\boldsymbol{e}_\nu$ (maybe after transformation):

$$\mathbf{x}_i = \sum_{\nu=1}^{p} x_{i,\nu}\boldsymbol{e}_\nu, \ i = 1, \ldots, n, \ \boldsymbol{\rho} = \sum_{\nu=1}^{p} \rho_\nu \boldsymbol{e}_\nu$$

Denote the ordered magnitutes as $|\rho|_{(1)} \geq |\rho|_{(2)} \geq \ldots$ We further need the magnitudes decay rather quickly:

$$|\rho|_\nu \leq C\nu^{-1/q}$$

,where $0 < q < 2$ and $C > 0$.
In other words, we want the "energy" in the largest k coordinates $\sum_{i=1}^{k} \rho_{(i)}^2$ is close to the total energy $||\boldsymbol{\rho}||^2$.

Consistency

To show consistency after selection, we assume: (1) each of unknown $\boldsymbol{\rho} = \boldsymbol{\rho}_n$ decays fast; (2) stable signal strength: $||\boldsymbol{\rho}_n|| \to \varrho > 0$. The sample variances:

$$\hat{\sigma}_\nu^2 = n^{-1} \sum_{i=1}^{n} x_{i\nu}^2 \sim (\sigma^2 + \rho_\nu^2) \chi_{(n)}^2 / n$$

So, larger values of ρ_ν will typically have larger sample variances. This leads to a simple selection rule:

$$\hat{I} = \{\nu : \hat{\sigma}_\nu^2 \geq \sigma^2 (1 + \alpha_n)\}$$

, with $\alpha_n = \alpha(n^{-1} \log(n \vee p))^{1/2}$.

Consistency

Let $\boldsymbol{S}_I = (S_{\nu\nu'} : \nu \text{ and } \nu' \in \hat{I})$ denote the sample covariance of the selected variables. Then apply regular PCA to $\boldsymbol{S}_I$. Let $\hat{\boldsymbol{\rho}}_I$ denote the corresponding vector in the full p-dimensional space:

$$\hat{\boldsymbol{\rho}}_{I,\nu} = \begin{cases} \hat{\boldsymbol{\rho}}_\nu & \nu \in \hat{I} \\ 0 & \nu \notin \hat{I}. \end{cases}$$

Consistency

Under this selection, they show consistency after selection (Theorem 2):

Theorem 2. Assume that the single component model (2) holds with $\log(p \vee n)/n \to 0$ and $\|\boldsymbol{\rho}_n\| \to \varrho > 0$ as $n \to \infty$. Assume for some $q \in (0, 2)$ and $C < \infty$, that for each n, $\boldsymbol{\rho}_n$ satisfies the sparsity condition (9). Then the estimated principal eigenvector $\hat{\boldsymbol{\rho}}_I$ obtained via subset selection rule (11) is consistent:

$$\mathfrak{a}(\hat{\boldsymbol{\rho}}_I, \boldsymbol{\rho}) \xrightarrow{a.s.} 0.$$

Here, $\mathfrak{a}$ is the angle between $\hat{\boldsymbol{\rho}}_I$ and $\boldsymbol{\rho}$ as in (4). Converting to an estimate $\hat{\boldsymbol{\rho}}(t)$ in the time domain (as in Step 5 of Section 4, an equivalent statement of the result is that $\| \hat{\boldsymbol{\rho}} / \| \hat{\boldsymbol{\rho}} \| - \hat{s}\boldsymbol{\rho} / \| \boldsymbol{\rho} \| \| \to 0$ in Euclidean norm, where $\hat{s} = \text{sign}(\langle \hat{\boldsymbol{\rho}}, \boldsymbol{\rho} \rangle)$.

Correct Selection Properties

Question: Do the selected subset $\hat{I}$ in fact correctly contains the largest **population** variances? And on those (no false inclusion)?
For this section, assume sample (coordinate) variance have marginal χ^2 distribution:

$$\hat{\sigma}_\nu^2 = S_{\nu\nu} \sim \sigma_\nu^2 \chi_{(n)}^2 / n$$

Denote the orderd population & sample coordinate variance as $\sigma_{(1)}^2 \geq \sigma_{(2)}^2 \geq \ldots$ and $\hat{\sigma}_{(1)}^2 \geq \hat{\sigma}_{(2)}^2 \geq \ldots$. And further denote:

- $I_{in} = \{I : \sigma_I^2 \geq \sigma_{(k)}^2(1+\alpha_n)\}$
- $I_{out} = \{I : \sigma_I^2 \leq \sigma_{(k)}^2(1+\alpha_n)\}$
- false exclusion (FE): $FE = \cup_{I \in I_{in}} \{\hat{\sigma}_I^2 < \hat{\sigma}_{(k)}^2\}$
- false inclusion (FI): $FI = \cup_{I \in I_{out}} \{\hat{\sigma}_I^2 \geq \hat{\sigma}_{(k)}^2\}$

Correct Selection Properties

The correct selection properties are shown in Theorem 3:

Theorem 3. Assume that the sample variances satisfy (12) and that a subset of size k of variables is sought. With $\alpha_n = \alpha n^{-1/2}(\log n)^{1/2}$, the probability of an inclusion error of either type is polynomially small:

$$\mathbf{P}\{FE \cup FI\} \leq 2pk(p \vee n)^{-b(\alpha)} + k(p \vee n)^{-(1-2\alpha_n)b(\alpha)},$$

with $b(\alpha) = [\alpha\sqrt{3}/(4 + 2\sqrt{3})]^2$.

An Illustrative Algorithm

1. *Compute Basis Coefficients.* Given a basis $\{\boldsymbol{e_\nu}\}$ for $\mathbb{R}^p$, compute coordinates $x_{i\nu} = (\mathbf{x}_i, \boldsymbol{e_\nu})$ in this basis for each $\mathbf{x}_i$:

$$x_i(t_l) = \sum_{\nu=1}^{p} x_{i\nu} e_\nu(t_l), \quad i = 1, \ldots, n; \quad t_l = 1, \ldots, p.$$

2. *Subset.* Calculate the sample variances $\hat{\sigma}_\nu^2 = \widehat{\mathrm{var}}(x_{i\nu})$. Let $\hat{I} \subset \{1, \ldots, p\}$ denote the set of indices ν corresponding to the largest k variances.
3. *Reduced PCA.* Apply standard PCA to the reduced dataset $\{x_{i\nu}, \nu \in \hat{I}, i = 1, \ldots, n\}$ on the selected k-dimensional subset, obtaining eigenvectors $\hat{\boldsymbol{\rho}}^j = (\hat{\rho}_\nu^j)$, $j = 1, \ldots, k, \nu \in \hat{I}$.
4. *Thresholding.* Filter out noise in the estimated eigenvectors by hard thresholding

$$\tilde{\rho}_\nu^j = \eta_H(\hat{\rho}_\nu^j, \delta_j).$$

5. *Reconstruction.* Return to the original signal domain, using the given basis $\{\boldsymbol{e_\nu}\}$, and set

$$\hat{\rho}_j(t_l) = \sum_{\nu \in \hat{I}} \tilde{\rho}_\nu^j e_\nu(t_l).$$

An Illustrative Algorithm

Some comments about thresholding:

- hard thresholding: $\eta_H(y, \delta) = yI\{|y| \geq \delta\}$
- choose δ_j: by analogy with the signal in Gaussian noies setting $\delta_j = \hat{\tau}_j\sqrt{2\log k}$
- $\hat{\tau}_j$ is an estimate of the noise level in $\{\hat{\rho}_\nu^j, \nu \in \hat{I}\}$
- In this paper, they estimate it as $\hat{\tau} \approx \frac{1}{\sqrt{n}} \frac{\sigma\sqrt{||\boldsymbol{\rho}||^2+\sigma^2}}{||\boldsymbol{\rho}||^2}$. This is derived from the asymptotic distribution of $\hat{\boldsymbol{\rho}}$. $||\boldsymbol{\rho}||^2$ and σ^2 can be estimated by data.
- We can also do things as $\hat{\tau}_j = MAD\{\hat{\rho}_\nu^j, \nu \in \hat{I}\}/0.6745$

Data-based Choice of k and estimation of σ

The paper provides 2 possibilities:

- shown in previous: $\hat{I} = \{\nu : \hat{\sigma}_\nu^2 \geq \sigma^2(1 + \alpha_n)\}$
- Define

$$\eta_{(\nu)}^2 = \max\{\hat{\sigma}_{(\nu)}^2 - (n-1)^{-1}\hat{\sigma}^2\chi_{(n-1),\nu/(p+1)}^2, 0\},$$

and for a specified fraction $w(n) \in (0, 1)$, set

$$\hat{I} = \{\nu : \sum_{\nu=1}^{\hat{k}} \eta_{(\nu)}^2 \geq w(n) \sum_{\nu} \eta_{(\nu)}^2\},$$

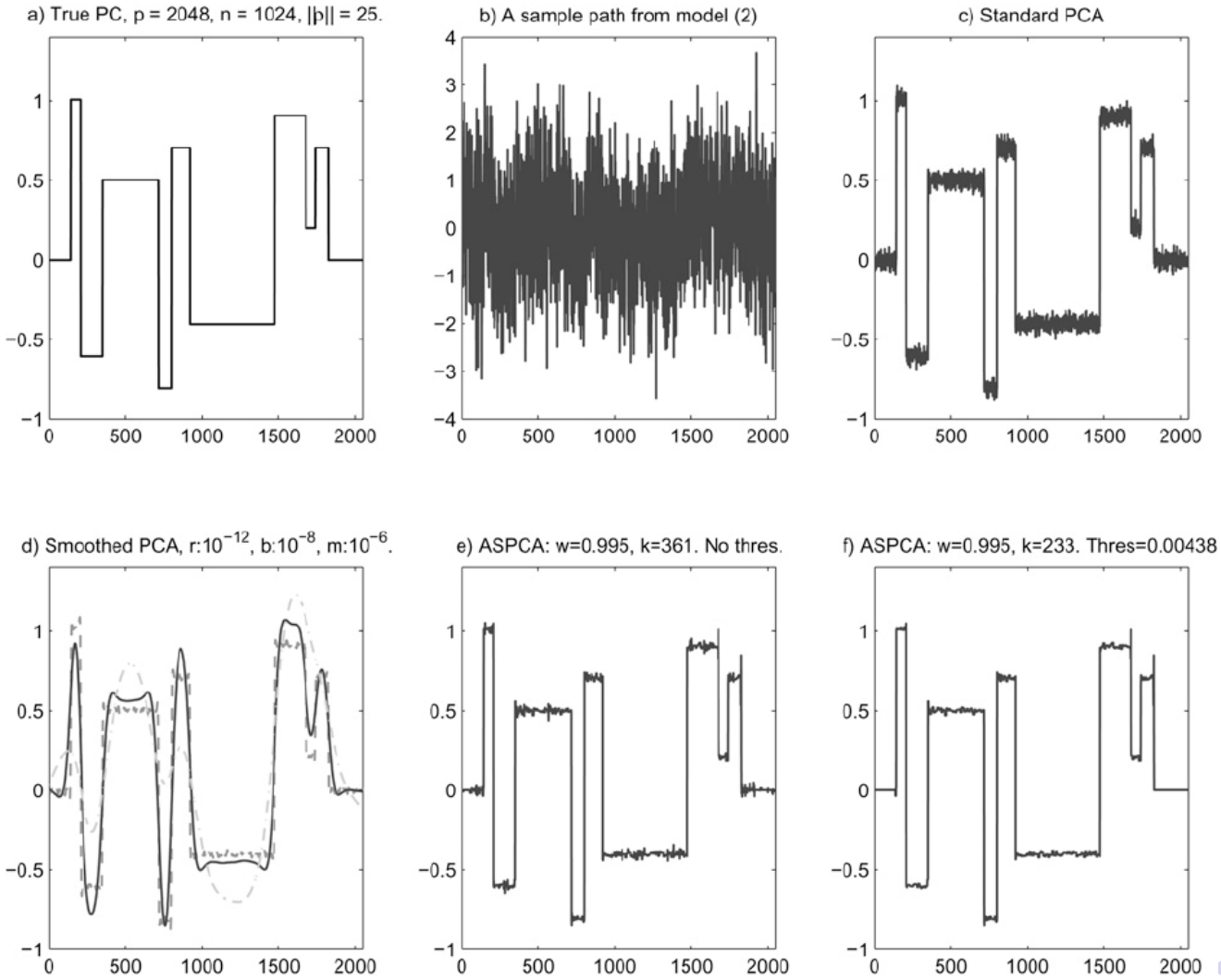
Estimation of σ:

If the population PC have a sparse representation, then in most coordinates, ν, $x_{i\nu}$ will consist largely of noise. Then we can estimate σ as:

$$\hat{\sigma}^2 = median(\hat{\sigma}_\nu^2)$$

Simulations

The first simulation ("3-peak") is shown in the motivation example. Here, they further show another ("step") example.

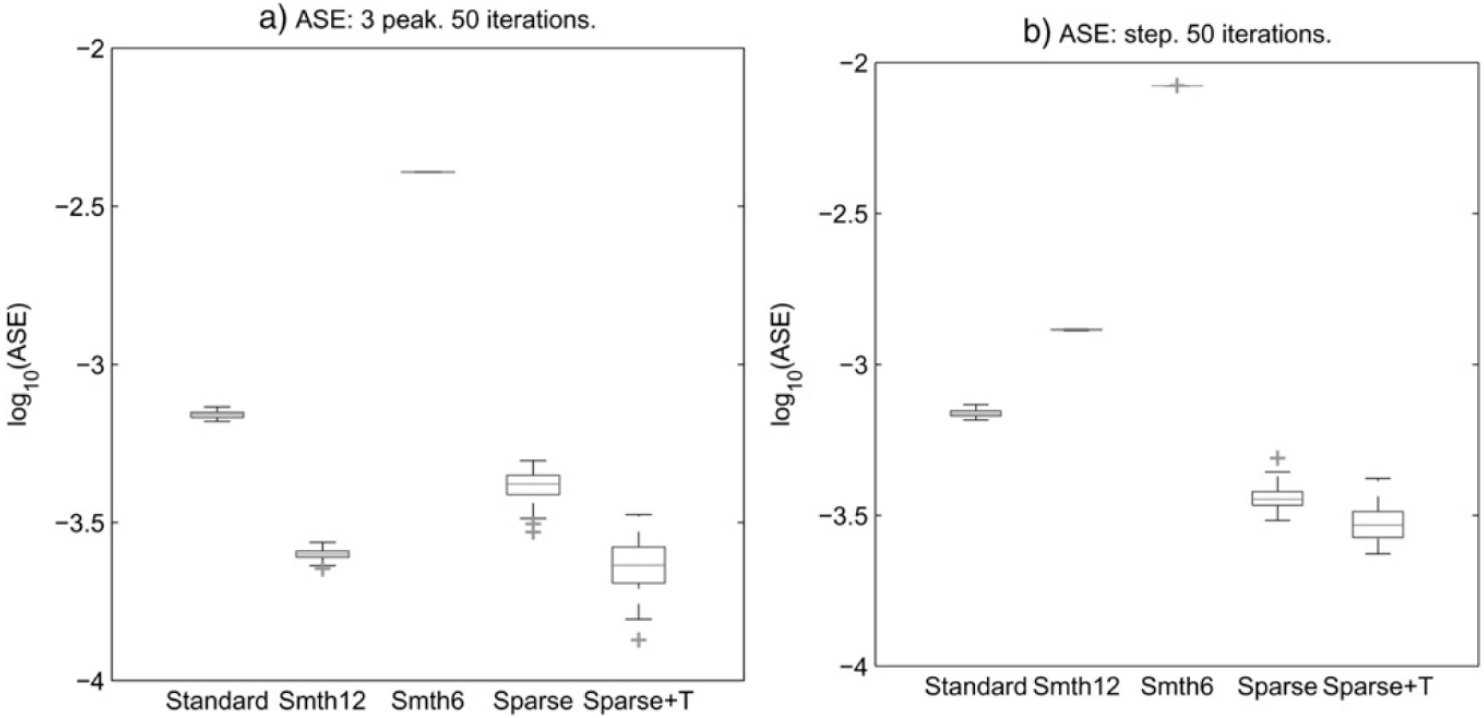


Simulations

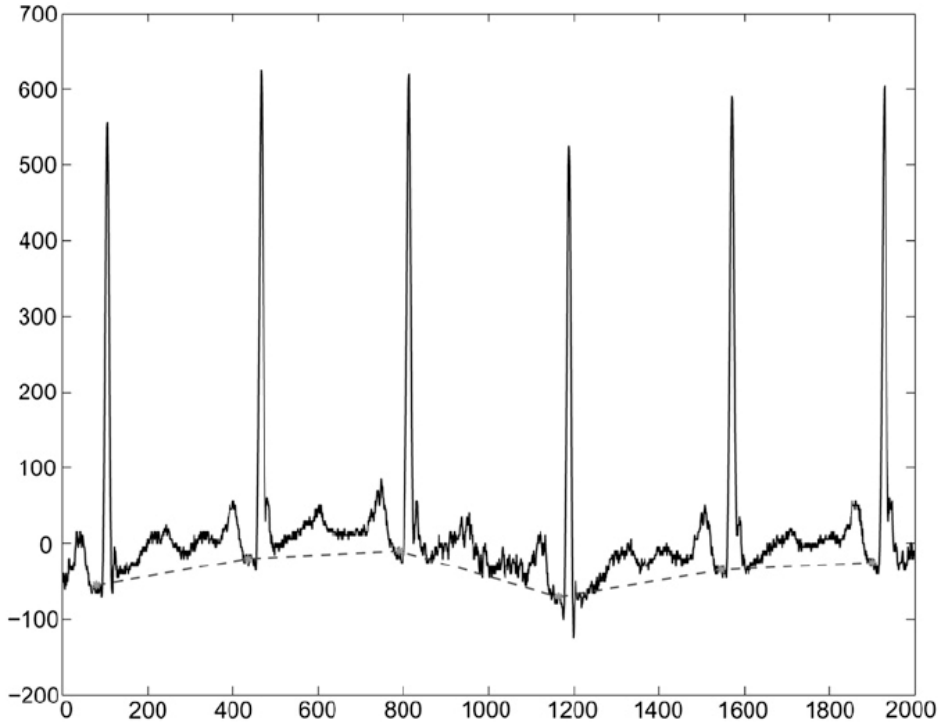
The average squared error (ASE) is defined as $ASE = p^{-1}||\hat{\rho} - \rho||$. The comparisons among different methods:

Table 1. Accuracy and efficiency comparison

	Standard PCA	Smoothed λ: 10^{-12}	Smoothed λ: 10^{-6}	Sparse PCA	Sparse + Threshold PCA
ASE (three-peak)	6.9e-04	2.5e-04	4.1e-3	4.1e-4	2.3e-04
Time (three-peak) (sec)	81.9	42.7	40.8	3.2	3.0
ASE (step)	6.9e-04	1.3e-3	8.4e-3	3.7e-4	3.0e-04
Time (step) (sec)	80.8	42.5	40.8	1.7	1.5



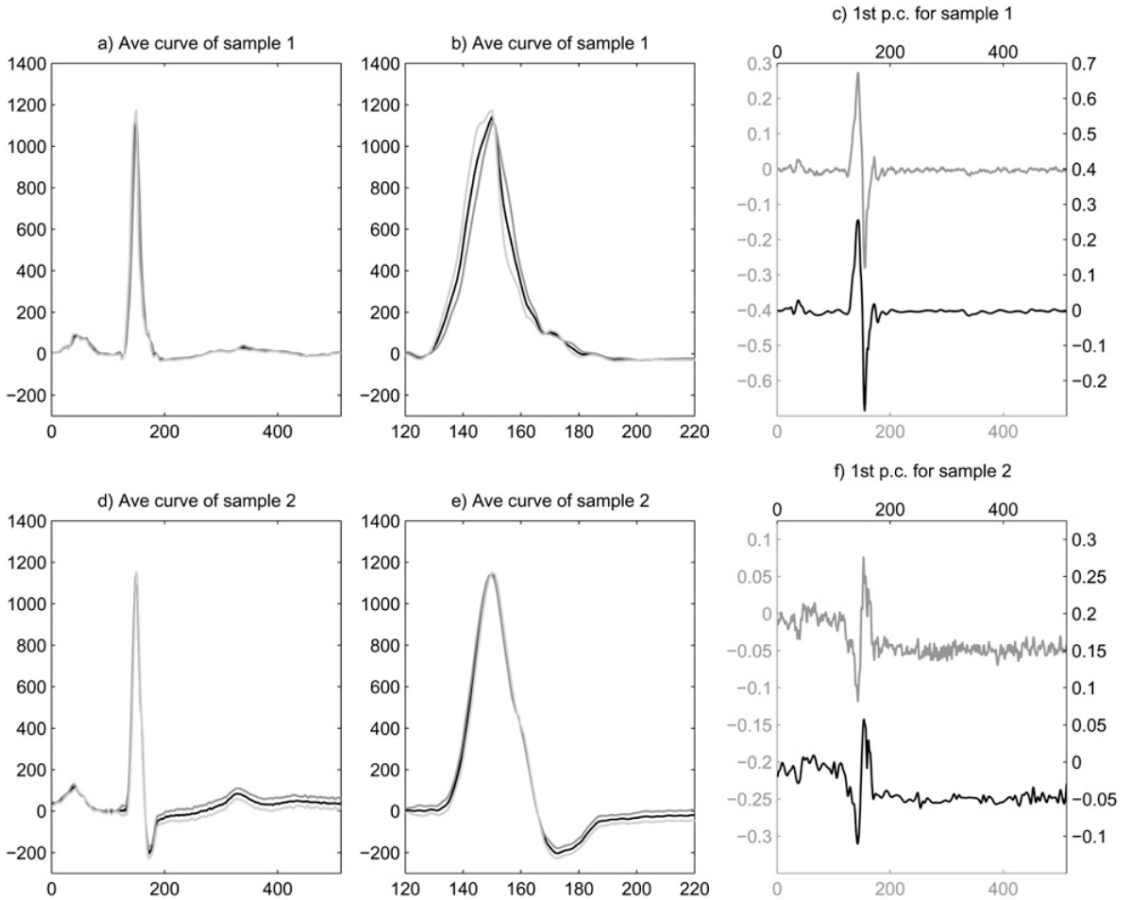
ECG Example



- The ECG data has "baseline wander", and this is adjusted by piece-wise linear baseline.
- Each individual betas are combined by sharp spike ("QRS complex", max at R wave) + lower peak ("T wave")

ECG Example

After prepossessing (e.g. adjust baseline wander and align peaks), they converted the ECG data vector to a $n \times 512$ matrix. n is the number of cycle, $p = 512$ is the duration of the cycle. The wavelet base is used here.



The SPCA is less noisy while keeping the features.