Published in final edited form as:
J Am Stat Assoc. 2009 June 1; 104(486): 682–693. doi:10.1198/jasa.2009.0121.

On Consistency and Sparsity for Principal Components Analysis in High Dimensions

Iain M. Johnstone and **Arthur Yu Lu**
Iain M. Johnstone is Professor of Statistics and Biostatistics, Stanford University, Department of Statistics, 390 Serra Mall, Stanford, CA 94305 (imj@stanford.edu). Arthur Yu Lu is Principal, Renaissance Technologies LLC, 600 Route 25A, East Setauket, NY 11733.

Abstract

Principal components analysis (PCA) is a classic method for the reduction of dimensionality of data in the form of n observations (or cases) of a vector with p variables. Contemporary datasets often have p comparable with or even much larger than n. Our main assertions, in such settings, are (a) that some initial reduction in dimensionality is desirable before applying any PCA-type search for principal modes, and (b) the initial reduction in dimensionality is best achieved by working in a basis in which the signals have a sparse representation. We describe a simple asymptotic model in which the estimate of the leading principal component vector via standard PCA is consistent if and only if $p(n)/n \to 0$. We provide a simple algorithm for selecting a subset of coordinates with largest sample variances, and show that if PCA is done on the selected subset, then consistency is recovered, even if $p(n) \gg n$.

Keywords

Eigenvector estimation; Reduction of dimension; Regularization, Thresholding; Variable selection

1. INTRODUCTION

Suppose $\{\mathbf{x}_i, i = 1, \ldots, n\}$ is a dataset of n observations on p variables. Standard principal components analysis (PCA) looks for vectors ξ that maximize

$$\mathrm{var}\left(\xi^T \mathbf{x}_\ell\right) / \| \xi\|^2. \tag{1}$$

If $\xi_1, \ldots, \xi_k$ have already been found by this optimization, then the maximum defining ξ_{k+1} is taken over vectors ξ orthogonal to $\xi_1, \ldots, \xi_k$.

Our interest lies in situations in which each $\mathbf{x}_i$ is a realization of a possibly high-dimensional signal, so that p is comparable in magnitude with n, or may even be larger. In addition, we have in mind settings in which the signals $\mathbf{x}_i$ contain localized features, so that the principal modes of variation sought by PCA may be localized as well.

These issues are familiar in signal and image processing application areas in which each sample has many variablesâpixels, frequencies, genes, stocks, and so forth. In applications, it is common to combine the use of transform domains and feature selection to achieve an effective reduction of dimensionality. For example, one might transform the data into a suitable orthogonal basis (e.g., wavelets), select coordinates with highest variance, and then do PCA on the reduced set of variables.

A notable example occurs in the work of Wickerhauser (1994a, b), in which the orthobasis itself was chosen from a library of (wavelet packet) bases. Applications to face and fingerprint classification were given. A selection of later examples (by no means exhaustive) would include Feng, Yuen, and Dai (2000) in face recognition; and Kaewpijit, Le Moigne, and El-Ghazawi (2002) and Du and Fowler (2008) for hyper-spectral images. For some further discussion, see Cherkassky and Mulier (1998). A recent approach to variable selection followed by dimensionality reduction that emphasizes sparsity is described by Wolf and Shashua (2005) and Wolf and Bileschi (2005).

The purpose of this article is to contribute some theoretical analysis of PCA in these burgeoning high-dimensional settings. In a simple class of models of factor analysis type, we (a) describe inconsistency results to emphasize that when p is comparable with n, some reduction in dimensionality is desirable before applying any PCA-type search for principal modes; and (b) establish consistency results to illustrate that the reduction in dimensionality can be effected by working in a basis in which the signals have a sparse representation. We will support these assertions with arguments based on statistical performance and computational cost.

We begin, however, with an illustration of our results on a simple constructed example. Consider a single component (or single factor) model, in which, when viewed as p-dimensional column vectors

$$\mathbf{x}_i = \upsilon_i \rho + \sigma z_i, \; i=1, \ldots, n \tag{2}$$

in which $\rho \in \mathbb{R}^p$ is the single component to be estimated, $\upsilon_i \sim N(0, 1)$ are iid Gaussian random effects, and $\boldsymbol{z_i} \sim N_p(0, I)$ are independent noise vectors.

Figure 1a shows an example of $\boldsymbol{\rho}$ with $p = 2{,}048$ and the vector $\rho_l = f(l/p)$, $l \in \{1, \ldots, p\}$, where $f(t)$ is a mixture of beta densities on [0, 1], scaled so that $\|\rho\| = \left(\sum_1^p \rho_l^2\right)^{1/2} = 10$. Figure 1b shows a sample case from model (2): The random effect $\upsilon_i \boldsymbol{\rho}$ is hard to discern in individual cases. Figure 1c shows the result of standard PCA applied to $n = 1{,}024$ observations from (2) with $\sigma = 1$, normalized to the same length as $\|\boldsymbol{\rho}\|$. The effect of the noise remains clearly visible in the estimated principal eigenvector.

For functional data of this type, a regularized approach to PCA has been proposed by Rice and Silverman (1991) and Silverman (1996), (see also Ramsay and Silverman (1997) and references therein). Although smoothing can be incorporated in various ways, we illustrate the method discussed also in Ramsay and Silverman (1997, chap. 7), which replaces (1) with

$$\operatorname{var}\left(\xi^T \mathbf{x}_i\right) / \left| \| \xi\|^2 + \lambda \| D^2 \xi\|^2 \right|, \tag{3}$$

where $D^2\xi$ is the $(p - 2) \times 1$ vector of second differences of ξ, and $\lambda \in (0, \infty)$ is the regularization parameter.

Figure 1d shows the estimated first principal component vector found by maximizing (3) with $\lambda = 10^{-12}$ and $\lambda = 10^{-6}$, respectively. Neither is really satisfactory as an estimate. The first recovers the original peak heights, but fails fully to suppress the remaining baseline noise, whereas the second grossly oversmooths the peaks in an effort to remove all trace of noise. Further investigation with other choices of λ confirms the impression already conveyed here: No single choice of λ succeeds both in preserving peak heights and in removing baseline noise.

Figures 1e and f show the result of the adaptive sparse PCA algorithm to be introduced later, respectively without and with a final thresholding step. Both goals are accomplished quite satisfactorily after thresholding in this example.

This article is organized as follows. Section 2 reviews the inconsistency result Theorem 1. Section 3 sets out the sparsity assumptions and the consistency result (Theorem 2). Section 4 gives an illustrative algorithm, demonstrated on simulated and real data in Section 5. Proofs and their preliminaries are deferred to Section 6 and the Appendix.

2. INCONSISTENCY OF CLASSIC PCA

A basic element of our sparse PCA proposal is initial selection of a relatively small subset of the initial p variables before any PCA is attempted. In this section, we formulate some (in)consistency results that motivate this initial step.

Consider first the single component model (2). The presence of noise means that the sample covariance matrix $S=n^{-1}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^T$ will typically have min(n, p) nonzero eigenvalues. Let $\widehat{\rho}$ be the eigenvector associated with the largest sample eigenvalue, with probability one it is uniquely determined up to sign.

One natural measure of the closeness of $\widehat{\rho}$ to $\boldsymbol{\rho}$ uses the “overlap” $R(\widehat{\rho},\rho)$, defined as the inner product between the vectors after normalization to unit length:

$$R(\widehat{\rho},\rho)=\langle\widehat{\rho}/\parallel\widehat{\rho}\parallel,\rho/\parallel\rho\parallel\rangle=\widehat{\rho}^T\rho/\parallel\widehat{\rho}\parallel\quad\parallel\rho\parallel.$$

Equivalently, $R(\widehat{\rho},\rho)$ is the cosine of the angle between $\widehat{\rho}$ and $\boldsymbol{\rho}$

$$R(\widehat{\rho},\rho)=\cos\quad\alpha(\widehat{\rho},\rho),\tag{4}$$

and we may also write this in terms of a distance metric

$$d(\widehat{\rho},\rho)=\sin\quad\alpha(\widehat{\rho},\rho).\tag{5}$$

For asymptotic results, we will assume that there is a sequence of models (2) indexed by n. Thus, we allow p_n and $\boldsymbol{\rho}_n$ to depend on n, although the dependence will usually not be shown explicitly. (Of course, σ might also be allowed to vary with n, but for simplicity it is assumed fixed.)

Our first interest is whether the estimate $\widehat{\rho}$ is consistent as $n \to \infty$. This turns out to depend crucially on the limiting value

$$\lim_{n\to\infty}p_n/n=c.\tag{6}$$

One setting in which this last assumption may be reasonable is when p_n grows by adding finer scale wavelet coefficients of a fixed function as n increases.

We will also assume that the limiting “signal-to-noise ratio”

$$\lim_{n\to\infty} \| \rho_{\boldsymbol{n}}\|^2/\sigma^2=\omega>0. \tag{7}$$

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

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

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

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

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

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

In short, $\widehat{\rho}$ is a consistent estimate of $\boldsymbol{\rho}$ if and only if $p/n \to 0$. The noise does not average out if there are too many dimensions p relative to sample size n.

Theorem 1 is stated without proof, and we include some bibliographic remarks about its history. The inconsistency phenomenon was first observed in a number of papers in the learning theory literature in physics (see, for example, Biehl and Mietzner 1994; Watkin and Nadal 1994; Reimann, Van den Broeck, and Bex 1996; Hoyle and Rattray 2004). The limiting overlap formula of Theorem 1, derived in a related classification setting, appears in Biehl and Mietzner (1994). These works all use the nonrigorous “replica method” of statistical mechanics.

The first rigorous proof of inconsistency was given, in model (2), by the second author in his Ph.D. dissertation (Lu 2002) and in the initial version of this article (Johnstone and Lu 2004), available at *arxiv.org*. While this article was in publication review, subsequent rigorous proofs were published, along with other results, by Paul (2007) and Nadler (2008).

Remark. Paul (2007) also includes extensions to a considerably more general “multicomponent” or “spiked” covariance model that has attracted interest in the literature. Assume that we have n data vectors $\mathbf{x}_i$, observed at p time points. Viewed as p-dimensional column vectors, this model assumes that

$$\mathbf{x}_i=\mu+\sum_{j=1}^{m} \upsilon_i^j \rho^j+\sigma z_i, \quad i=1,\ldots,n. \tag{8}$$

The mean function $\boldsymbol{\mu}$ is assumed known. The vectors $\boldsymbol{\rho}^j$, $j = 1, \ldots, m \le p$ are unknown and mutually orthogonal, with norms $\rho_j(n) = \|\boldsymbol{\rho}^j\|$ assumed decreasing: $\|\boldsymbol{\rho}^1\| > \|\boldsymbol{\rho}^2\| \ge \ldots \ge \|\boldsymbol{\rho}^m\|$. The multipliers $\upsilon_i^j N(0,1)$ are all independent over $j = 1, \ldots, m$ and $i = 1, \ldots, n$, and the noise vectors $\boldsymbol{z_i} \sim N_p(0, \mathbf{I})$ are independent among themselves and also of the random effects $\left\{\upsilon_i^j\right\}$. The population covariance matrix of the iid vectors ($\mathbf{x}_i$) is given by $\sum = \sum_{j=1}^{m} \rho^j \rho^{jT} + \sigma^2 \boldsymbol{I}$. The vectors ρ^j are the ordered principal component eigenvectors of the population covariance Σ. The asymptotics assume p_n, m_n, and ρ_n^j to be functions of n, and as $n \to \infty$,

$$\varrho_n = \left(\rho_n^1, \ldots, \rho_n^{m_n}, 0, \ldots, 0\right) \to \varrho = \left(\varrho_1, \ldots, \varrho_j, \ldots\right).$$

Paul (2007) shows that it continues to be true in the multicomponent model that $\widehat{\rho}^1$ is consistent if and only if $p_n/n \to 0$. Of course, the inconsistency extends to cases with $p_n/n \to \infty$ because these models are even larger.

3. SPARSITY, SELECTION, AND CONSISTENCY

3.1 Sparsity

The inconsistency Theorem 1 asserts that ordinary PCA becomes confused in the presence of too many variables each with equal independent noise. Ideally, we might wish to reduce the dimensionality from p to a smaller number of variables k before beginning PCA. For this to succeed, the population principal components—**ρ**, in our model—should be essentially concentrated in the smaller number of dimensions, in a manner that can be discovered from the data.

To quantify this, assume that the data and the population principal components are represented, perhaps after transformation, in a fixed orthonormal basis $\{\boldsymbol{e}_\mathbf{v}\}$:

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

The index ν will always indicate the transform domain. In many situations, including all examples in this article, the data $\mathbf{x}_i$ are initially collected in the time domain—for example in [0, 1], with $\mathbf{x}_i = \{x_i(t_l)\}$, where $t_l = l/p$, $l = 1,\ldots, p$. In such cases, the basis vectors $\boldsymbol{e}_\mathbf{v}$ are also time domain functions $\boldsymbol{e}_\nu(t_l)$.

The idea of concentration in a small number of variables can be captured by considering the ordered coefficient magnitudes $|\rho|_{(1)} \geq |\rho|_{(2)} \geq \cdots$. The intuitive idea of sparse representation is that, for relatively small k, the “energy” in the largest k coordinates $\sum_{i=1}^{k} \rho_{(i)}^2$ is close to the total energy $\|\rho\|^2 = \sum_{\nu=1}^{p} \rho_\nu^2$. This can only be true if the magnitudes $|\rho|\nu$ decay rather quickly. Thus, we assume for some $0 < q < 2$ and $C > 0$ that

$$|\rho|_{(\nu)} \quad \leq \quad C\nu^{-1/q}, \quad \nu=1, 2, \ldots. \tag{9}$$

The condition $q < 2$ forces rapid decay—clearly, the more rapid if q is smaller. This notion of “weak ℓ_q decay” is actually equivalent to the concentration of energy in the sums $\sum_{i=1}^{k} \rho_{(i)}^2$ just mentioned (see Donoho (1993) or Johnstone (2003, chap. 15)), but is more convenient for the results given here.

The choice of orthonormal basis $\{\boldsymbol{e}_\nu\}$ for sparse representation will depend on the dataset and problem domain, and thus is beyond the scope of this article. We remark, however, that for certain signal processing settings, wavelet bases can be natural for uncovering sparsity. When one-dimensional signals are smooth or have isolated discontinuities (either in the signal or its derivatives), then it can be shown (e.g., Mallat 1999) that the wavelet coefficients decay rapidly with frequency octave away from the discontinuities. In such cases, assumptions (9) are natural, as is shown in detail, for example, in the references cited earlier. We have therefore used wavelet bases for the examples in this article, but hasten to emphasize that our results apply equally to representations in other bases that might be better suited to, say, economic or genomic data.

3.2 Consistency

If the principal components have a sparse representation in basis $\{\boldsymbol{e}_\nu\}$, then selection of an appropriate subset of variables should overcome the inconsistency problem described by Theorem 1. In this direction, we establish a consistency result for sparse PCA. For simplicity, we use the single component model (2), and assume σ^2 is known—although this latter assumption could be removed by estimating σ^2 using (14), presented later.

We again assume a sequence of models (2) indexed by n. The unknown principal components $\boldsymbol{\rho} = \boldsymbol{\rho}_n$ should satisfy a “uniform sparsity condition”: For some $q \in (0, 2)$ and $C < \infty$ independent of n, each $\boldsymbol{\rho}_n$ satisfies decay condition (9). In addition, as in Theorem 1, the signal strength should stabilize: $\|\boldsymbol{\rho}_\mathbf{n}\| \to \varrho > 0$.

On the assumption of model (2), the sample variances

$$\widehat{\sigma}_\nu^2=n^{-1}\sum_{i=1}^{n} x_{i\nu}^2 \sim \left(\sigma^2+\rho_\nu^2\right)\chi_{(n)}^2/n. \tag{10}$$

Consequently, components ν with large values of ρ_ν will typically have large sample variances. We use here a simple selection rule

$$\widehat{I}=\left\{\nu \quad : \quad \widehat{\sigma}_\nu^2 \ge \sigma^2\,(1+\alpha_n)\right\}, \tag{11}$$

with $\alpha_n = \alpha(n^{-1} \log(n \vee p))^{1/2}$ and α a sufficiently large positive constant—for example, $\alpha > \sqrt{12}$ would work for the proof. (By definition, $n \vee p = \max(p, n)$.)

Let $\boldsymbol{S}_\mathbf{I} = (S_{\nu\nu'} : \nu$ and $\nu' \in \widehat{I})$ denote the sample covariance matrix of the selected variables. Applying PCA to $\boldsymbol{S}_\mathbf{I}$ yields a principal eigenvecto $(\widehat{\rho}_\nu, \nu \in \widehat{I})$. Let $\widehat{\rho}_I$ denote the corresponding vector in the full p-dimensional space:

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

The sparsity assumption implies that $\widehat{\rho}_I$ is a consistent estimator of $\boldsymbol{\rho}$.

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 $n \to \infty$ Assum for some $q \in (0, 2)$ and $C < \infty$, that for each n, $\boldsymbol{\rho}_\mathbf{n}$ satisfies the sparsity condition (9). Then the estimated principal eigenvector $\widehat{\rho}_I$ obtained via subset selection rule (11) is consistent:

$$\alpha\left(\widehat{\rho},\rho\right)\xrightarrow{a.s.}0.$$

Here, α is the angle between $\widehat{\rho}_I$ and **ρ** as in (4). Converting to an estimate $\widehat{\rho}(t)$ in the time domain (as in Step 5 of Section 4, an equivalent statement of the result is that $\|\widehat{\rho}/\|\widehat{\rho}\| - \widehat{s}\rho/\|\rho\|\| \to 0$ in Euclidean norm, where $\widehat{s}=\mathrm{sign}\left(\langle\widehat{\rho},\rho\rangle\right)$.

The proof shows that consistency holds even under weaker assumption $p = o(e^n)$. The result and proof could also be extended to multicomponent systems (8).

Armed with Theorems 1 and 2, let us return briefly to Figure 1. Based on Figure 1c, one might ask if simple thresholding of the standard PCA estimate—either in the original or wavelet domain—might suffice. Although this may work for high signal-to-noise settings, Theorem 1 suggests that such an approach is doomed if $\|\rho\| < p/n$, because $\widehat{\rho}$ is asymptotically orthogonal to ρ. No such constraint applies in Theorem 2—as long as lim $\|\rho_n\| > 0$—as a result of the preliminary reduction of variables.

Bibliographic Remarks—Significant extensions of these results have been obtained since this article was first written. For example, working with the multicomponent model, Paul (2007) and Paul and Johnstone (2004) have derived lower bounds to the possible quality of estimation, and optimal estimators that attain the bounds (at the level of rates of convergence) in certain cases.

In a large *p, n* setting of partial least squares regression, Nadler and Coifman (2005) noted the importance of prior dimension reduction and suggested the use of wavelets to exploit sparsity.

Alternative methods for "sparsifying" PCA have also been proposed, based on connections with LASSO and ℓ_1 penalized regression (Jolliffe, Trendafilov, and Uddin 2003; Zou, Hastie, and Tibshirani 2006; d'Aspremont et al. 2007), and compressive sensing (Fowler 2008). The study of consistency properties for these methods and comparison with those of this article is a natural topic for further research, with significant progress recently reported in Amini and Wainwright (2009).

3.3 Correct Selection Properties

A basic issue raised by the sparse PCA algorithm is whether the selected subset $\widehat{I}$ in fact correctly contains the largest population variances, and only those. We formulate a result, based on large deviations of chi-squared variables, to address this issue. The considerations of this section hold for coefficients in any orthogonal basis.

For this section, assume that the diagonal elements of the sample covariance matrix $S=n^{-1}\sum_1^n \mathbf{x}_i\mathbf{x}_i^T$ have marginal chi-squared distributions—in other words,

$$\widehat{\sigma}_\nu^2=S_{\nu\nu}\sim\widehat{\sigma}_\nu^2\chi_{(n)}^2/n,\qquad \nu=1,\ldots,p. \tag{12}$$

We will not require any assumptions on the joint distribution of $\left\{\widehat{\sigma}_\nu^2\right\}$ The use of the index ν emphasizes the fact that we work in the transform domain.

Denote the ordered population coordinate variances by $\sigma^2_{(1)} \geq \sigma^2_{(2)} \geq \ldots$, and the ordered sample coordinate variances by $\sigma^2_{(1)} \geq \sigma^2_{(2)} \geq \ldots$. A desirable property is that $\widehat{I}$ should, for fixed k and for suitable positive α_n small, (i) include all indices l in $I_{in} = \left\{l{:}\sigma^2_l \geq \sigma^2_{(k)}(1+\alpha_n)\right\}$ and (ii) exclude all indices l in $I_{out} = \left\{l{:}\sigma^2_l \leq \sigma^2_{(k)}(1-\alpha_n)\right\}$. We will show that this, in fact, occurs with high probability if $\alpha_n = \alpha\sqrt{n^{-1}\log n}$, for appropriate $\alpha > 0$.

We say that a "false exclusion" (FE) occurs if any variable in I_{in} is missed:

$$FE = \bigcup_{l \in I_{in}} \left\{\widehat{\sigma}^2_l < \widehat{\sigma}^2_{(k)}\right\},$$

whereas a "false inclusion" (FI) happens if any variable in I_{out} is spuriously selected:

$$FI = \bigcup_{l \in I_{out}} \left\{\widehat{\sigma}^2_l \geq \widehat{\sigma}^2_{(k)}\right\}.$$

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

The proof is in the Appendix. As an example, if $\alpha = 9$, then $b(\alpha) \doteq 4.36$. As a numerical illustration based on (A.4) (seen in the Appendix), if the subset size $k = 50$, while $p = n = 1{,}000$, then the chance of an inclusion error corresponding to a 25% difference in standard deviations (i.e., $\sqrt{1+\alpha_n} = 1.25$ when $\alpha = 9$) is less than 5%.

4. AN ILLUSTRATIVE ALGORITHM

4.1 An Algorithm

The inconsistency results summarized in Section 2 emphasize the importance of reducing the number of variables before embarking on PCA. The results of Section 3 show that the existence of a sparse representation allows consistency to be recovered. The proof of Theorem 2 relies on two key steps: (i) sparsity allows $\boldsymbol{\rho}$ to be approximated using a relatively small number of coefficients, and (ii) these smaller number of coefficients can be estimated by a reduced PCA.

We use these remarks as the basis for the sparse PCA algorithm to be described in general terms here. Note that the algorithm per se does not require the specification of a particular model, such as the single component system (2) or the multicomponent version (8). Given the widespread use of transform domain and feature selection techniques in the signal processing literature, as described in Section 1, we make no claims for originality; this section is included primarily to illustrate results of previous sections.

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

$$x_i(t1)=\sum_{\nu=1}^{p} x_{i\nu} e_\nu(t1), \quad i=1,\ldots,n; \quad t1=1,\ldots,p.$$

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

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

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

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

Discussion: Steps 2 and 3—An important computational point that is implicit in Steps 2 and 3 is that we only compute the variances $S_{\nu,\nu}$ for the *p* transform domain variables. Off-diagonal covariance elements $S_{\nu,\nu'}$ are only computed for ν, ν′ in the reduced set $\widehat{I}$ of size *k*. The reduced PCA size *k* may be specified in advance or chosen based on the data (see Section 4.2).

Discussion: Step 4—Although not formally studied in the theory in the preceding section, the thresholding step is found in our examples to yield a useful further filtering of noise. For a scalar value *y*, hard thresholding is given, as usual, by $\eta_H(y, \delta) = y\boldsymbol{I}\{|y| \geq \delta\}$. An alternative is soft thresholding $\eta_S(y, \delta) = \mathrm{sgn}(y) \times \max(|y| - \delta, 0)$, but hard thresholding has been used here because it preserves the magnitude of retained signals.

There is considerable freedom in the choice of thresholds δ_j. Trial and error is always possible, of course. Further, more formal choices are suggested by analogy with the signal in Gaussian noise setting $\delta_j=\widehat{\tau}_j\sqrt{2\log k}$ (compare with, for example, Donoho et al. (1995)), and, for this article, we use this choice of δ_j. Here, $\widehat{\tau}_j$ is an estimate of the noise level in $\{\widehat{\rho}_\nu^j, \nu \in \widehat{I}\}$—in this article, estimate (16) is used. Another possibility is to set $\widehat{\tau}_j=MAD\left\{\widehat{\rho}_\nu^j, \nu \in \widehat{I}\right\}/0.6745$, where *MAD* denotes "median absolute deviation."

The consistency result Theorem 2 applies to this algorithm, with subset selection rule (11), and without the thresholding Step 4. Although the thresholding step helps in the examples to follow, theoretical analysis to elucidate its specific advantages is beyond the scope of this article.

Terminology—We will refer to the general procedure specified by Steps 1 through 5 as "sparse PCA." With the specific data-based choice of *k* proposed as method (b) in Section 4.2, with $w = 0.995$, we use the term "adaptive sparse PCA" (ASPCA).

In the rest of this section, we amplify and illustrate various aspects of this algorithm. Given eigenvalue and eigenvector routines, it is not difficult to code. For example, a MATLAB package ASPCALab that includes the algorithms and files that produce the figures in this article is available at www-stat.stanford.edu/~imj/. To exploit wavelet bases, it makes use of the open-source library WaveLab available at www-stat.stanford.edu/~wavelab/.

4.2 Data-Based Choice of *k*

When the size *k* of the set of selected variables $\widehat{I}$ is itself determined from the data, we write $\widehat{k}$ for $|\widehat{I}|$. Here are two possibilities, both based on the sample variances $\widehat{\sigma}_\nu^2$ of Step 2 presented earlier:

(a) Choose coordinates with variance exceeding the estimated noise level by a specified fraction α_n:

$$\widehat{I}=\left\{\nu \quad : \quad \widehat{\sigma}_\nu^2 \geq \widehat{\sigma}^2 (1+\alpha_n)\right\}.$$

This choice was considered in Section 3.

(b) As motivation for the second method, recall that we hope that the selected set of variables $\widehat{I}$ is both small in cardinality and also captures most of the variance of the population principal components, in the sense that the ratio

$$\sum_{\nu\in\widehat{I}}\rho_\nu^2 / \sum_\nu \rho_\nu^2 \tag{13}$$

is close to one for each of the leading population principal components in {$\boldsymbol{\rho}^1$, …, $\boldsymbol{\rho}^m$}. Now let $\chi^2_{(n),\alpha}$ denote the upper α percentile of the $\chi^2_{(n)}$ distribution. If all coordinates were pure noise, one might expect the ordered sample variances $\widehat{\sigma}^2_{(\nu)}$ to be close to $(n-1)^{-1}\widehat{\sigma}^2\chi^2_{(n-1),\nu/(p+1)}$. Define the excess over these percentiles by

$$\eta^2_{(\nu)}=\max\left\{\widehat{\sigma}^2_{(\nu)} - (n-1)^{-1}\widehat{\sigma}^2\chi^2_{(n-1),\nu/(p+1)}, 0\right\},$$

and for a specified fraction *w*(*n*) ∈ (0, 1), set

$$\widehat{I}=\left\{\nu : \sum_{\nu=1}^{\widehat{k}}\eta^2_{(\nu)} \geq w(n)\sum_\nu \eta^2_{(\nu)}\right\},$$

where $\widehat{k}$ is the smallest index *k* for which the inequality holds, and the somewhat sloppy notation refers to the indices ν that contribute to the left-hand sum of excesses $\eta^2_{(\nu)}$. This second method has been used for the figures in this article, typically with *w*(*n*) = 0.995.

Estimation of σ—If the population principal components ρ^j have a sparse representation in basis {$\boldsymbol{e}_\mathbf{v}$}, then we may expect that in most coordinates, ν, {$x_{i\nu}$} will consist largely of noise. This suggests a simple estimate of the noise level on the assumption that the noise level is the same in all coordinates—namely,

$$\widehat{\sigma}^2 = \text{median}\left(\widehat{\sigma}_\nu^2\right). \tag{14}$$

4.3 Computational Complexity

One estimates the cost of sparse PCA by examining its main steps:

1. This depends on the choice of basis. In the wavelet case, no more than $O(np \log p)$ operations are needed. (see, for example, Mallat (1999)).
2. Evaluate and sort the sample variances and select $\widehat{I}: O(np+p \log p)$.
3. Compute a $k \times k$ matrix and its eigendecomposition: $O(k^3+k^2n)$.
4. Apply thresholding to each vector in $\widehat{I}: O\left(k^2\right)$, and estimate $\widehat{\sigma}^2$ and $\|\widehat{\rho}\|^2: O(p)$.
5. Reconstruct eigenvectors in the original sample space: $O(k^2p)$.

Hence, the total cost of sparse PCA is $O(np \log p + k^2 \max(p, n))$. Both standard and smoothed PCA need at least $O((\min(p, n))^3)$ operations. Therefore, if we can find a sparse basis such that $k/p \to 0$, then under the assumption that $p/n \to c$ as $n \to \infty$, the total cost of sparse PCA is $o(p^3)$. We will see in the examples that follow that the savings can be substantial.

5. EXAMPLES

5.1 Simulated Examples

The two examples in this section are both motivated by functional data with localized features. The first is a three-peak principal component depicted in Figure 1, and already discussed in Section 1. The second example, Figure 2, has an underlying first principal component composed of step functions. For both examples, the dimension of data vectors is p = 2,048, the number of observations n = 1,024, and the noise level $\sigma = 1$. However, the amplitudes of $\boldsymbol{\rho}$ differ, with $\|\boldsymbol{\rho}\| = 10$ for the "three-peak" function and $\|\boldsymbol{\rho}\| = 25$ for the "step" function. The corresponding square root signal-to-noise ratios $\omega = \varrho/\sigma$ (Theorem 1) are 10 and ~25 respectively.

Figure 1c and Figure 2c, respectively, show the sample principal components obtained by using standard PCA. Although standard PCA does capture the peaks and steps, it retains significant noise in the flat regions of the function. Corresponding Figure 1d and Figure 2d show results from smooth PCA with the indicated values of the smoothing parameter. Just as for the three-peak curve discussed earlier, in the case of the step function, none of the three estimates simultaneously captures both jumps and flat regions well.

Figures 1e, f and Figures 2e, f present the principal components obtained by sparse PCA without and with the thresholding step, respectively. The WaveLab wavelet bases Symmlet and Haar are used for the "three-peak" and "step" functions respectively. Using method (b) of Section 4.2 with w = 99.5%, the subset step selects k = 142 and 361 for the "three-peak" curve and "step" function, respectively. The sample principal component in Figure 1f is clearly superior to the other sample principal components in Figure 1. Although the principal component function in the step case appears to be only slightly better than the solid, red, smooth PCA estimate, we will see shortly that its squared error is reduced by about 75%.

Table 1 compares the accuracy of the three PCA algorithms, using average squared error (ASE) defined as $\mathrm{ASE}=p^{-1}\|\widehat{\rho}-\rho\|$. ($\widehat{\rho}$ was first normalized to have length $\|\boldsymbol{\rho}\|$ before computing ASE.) The running time is average CPU time over 50 iterations, used by MATLAB on an Intel Core Duo CPU T2400 at 1.83 GHz.

Figure 3 presents boxplots of ASE for the 50 iterations. Overall, sparse PCA with thresholding always gives the best result for the "step" curve. For the "three-peak" function, in only a few iterations (~15%) does sparse PCA generate larger error than smoothed PCA with $\lambda = 10^{-12}$. On average, ASE using sparse PCA is superior to the other methods by a large margin. Overall, Table 1 and Figure 3 show that sparse PCA leads to the most accurate principal component (within the techniques considered) while using much less CPU time than other PCA algorithms.

5.2 Noise Level in the Single Component Model

Anderson (1963) obtained the asymptotic distribution of $\sqrt{n}\,(\widehat{\rho}-\rho)$ for fixed p—in particular,

$$\operatorname{var}\left\{\sqrt{n}\,(\widehat{\rho}_{\nu}-\rho_{\nu})\right\}\rightarrow\left(\|\rho\|^{2}+\sigma^{2}\right)\frac{\sigma^{2}}{\|\rho\|^{4}}\left(1-\rho_{\nu}^{2}\right), \tag{15}$$

as $n\rightarrow\infty$. Here, p increases with n, but one can nevertheless use (15) as a heuristic for estimating the variance $\widehat{\tau}$ needed for thresholding. Because the effect of thresholding is to remove noise in small coefficients, setting ρ_{ν} to 0 in (15) suggests

$$\widehat{\tau}\approx\frac{1}{\sqrt{n}}\frac{\sigma\sqrt{\|\rho\|^{2}+\sigma^{2}}}{\|\rho\|^{2}}. \tag{16}$$

Neither $\|\boldsymbol{\rho}\|^2$ nor σ^2 in (16) is known, but they can be estimated by using the information contained in the sample covariance matrix S of (10). Hence, in the single component model,

$$\|\rho\|^{2}=\sum_{1}^{p}\rho_{\nu}^{2}=\sum_{1}^{p}\left\{\mathbf{E}\left(\widehat{\sigma}_{\nu}^{2}\right)-\sigma^{2}\right\}.$$

If ρ_{ν} is a sparse representation of $\boldsymbol{\rho}$, then most coefficients will be small, suggesting the estimate (14) for σ^2. In turn, this suggests as an estimate

$$\widehat{\|\rho\|}^{2}=\max\left\{\sum_{1}^{p}\left[S_{\nu}^{2}-\operatorname{median}\left(S_{\nu}^{2}\right)\right],\quad \operatorname{median}\left(S_{\nu}^{2}\right)\sqrt{p/n}\right\}. \tag{17}$$

5.3 ECG Example

We offer a brief illustration of sparse PCA as applied to some electrocardiogram (ECG) data kindly provided by Jeffrey Froning and Victor Froelicher in the cardiology group at Palo Alto Veterans Affairs Hospital. Beat sequences—typically about 60 cycles in length—were obtained from some 15 healthy patients; we selected two for the preliminary illustrations here. Individual beats are notable for features such as the sharp spike ("QRS complex") and the subsequent lower peak ("T wave"), seen, for example, in Figures 5a and d. The presence of these local features, of differing spatial scales, suggests the use of wavelet bases for efficient representation. Traditional ECG analysis focuses on averages of a series of beats. If one were to look instead at beat-to-beat "variation," one might expect these local features to play a significant role in the principal component eigenvectors.

Considerable preprocessing is routinely done on ECG signals before the beat averages are produced for physician use. Here we summarize certain steps taken with our data, in collaboration with Jeff Froning, preparatory to the PCA analysis. The most important feature of an ECG signal is the Q-R-S complex: The maximum occurs at the R wave, as seen, for example, in Figure 5a. Therefore, we define the length of one cycle as the gap between two adjacent maxima of R waves.

i. "Baseline wander" is observed in many ECG datasets (compare with Figure 4, with a caption that summarizes the adjustment used).

ii. Because pulse rates vary even on short timescales, the duration of each heartbeat cycle may vary as well. We use linear interpolation to equalize the duration of each cycle, and for convenience in using wavelet software, discretize to $512 = 2^9$ sample points in each cycle.

iii. Because of the importance of the R wave, the horizontal positions of the maxima are registered at the 150th position in each cycle.

iv. The ECG data vector is converted into an $n \times p$ data matrix, where n is the number of observed cycles and $p = 512$.

5.4 PCA Analysis

Figures 5a and d show the mean curves for two ECG samples in blue. The number of observations n (i.e., number of heartbeats recorded) are 66 and 61, respectively. The first sample principal components for these two sample sets ("patients") are plotted in Figures 5c and f, with the upper/red curves from standard PCA and the lower/blue curves from sparse PCA, with thresholds chosen subjectively. In both cases, there are two sharp peaks in the vicinity of the QRS complex. The first peak occurs shortly before the 150th position, where all the maxima of R waves are aligned, and the second peak, which has an opposite sign, occurs shortly thereafter. (The lower/blue curves have been offset vertically by −0.4 and −0.2 in Figures 5c and f, respectively, for legibility in monochrome.)

The standard PCA curve in Figure 5c (upper/red) is less noisy than that in Figure 5f (upper/red), even allowing for the difference in vertical scales. Using (14), $\widehat{\sigma}_1^2=24.59$ and $\widehat{\sigma}_2^2=80.77$, whereas the magnitudes of the two mean sample curves are very similar.

The sparse PCA curves (lower/blue) are smoother than the standard PCA ones (upper/red), especially in Figure 5f, where the signal-to-noise ratio is lower. On the other hand, the upper/red and lower/blue curves match quite well at the two main peaks. Sparse PCA has reduced noise in the sample principal component in the baseline while keeping the main features.

There is a notable difference between the estimated principal components for the two patients. In the first case, the principal component is concentrated around the R-wave maximum, and the effect is to accelerate or decelerate the rise (and fall) of this peak from baseline in a given cycle. This is more easily seen by comparing plots of $\bar{\boldsymbol{x}}+2\widehat{\rho}$ (green) with $\bar{\boldsymbol{x}}-2\widehat{\rho}$ (red), shown over a magnified part of the cycle in Figure 5b. In the second patient, the bulk of the energy of the principal component is concentrated in a level shift in the part of the cycle starting with the ST segment. This can be interpreted as beat-to-beat fluctuation in baseline; because each beat is anchored at 0 at the onset point, there is less fluctuation on the left side of the peak. This is particularly evident in Figure 5e. There is, again, a slight acceleration/deceleration in the rise to the R-wave peak—less pronounced in the first case, and also less evident in the fall.

Obvious questions raised by this illustrative example include the nature of effects that may have been introduced by the preprocessing steps—notably, the baseline removal anchored at onset points and the alignment of R-wave maxima. Clearly, some conventions must be adopted to create rectangular data matrices for principal component analysis, but detailed analysis of these issues must await future work.

To summarize, sparse PCA has reduced noise in the sample principal component in the baseline while keeping the main features, and in addition, sparse PCA uses less than 10% of the computing time used by standard PCA in these examples.

6. PROOF OF THEOREM 2

We first establish some notation and recall some pertinent matrix results (e.g., Golub and Van Loan 1996). Norms on vectors are always Euclidean 2-norms: $\|\boldsymbol{x}\|=\left(\sum_{\nu} x_{\nu}^{2}\right)^{1/2}$. Define the 2-norm of a rectangular matrix by $\|\boldsymbol{A}\| = \sup\{\|\boldsymbol{Ax}\| : \|\boldsymbol{x}\| = 1$. If A is real and symmetric, then $\|\boldsymbol{A}\| = \lambda_{max}(\boldsymbol{A})$. If $\boldsymbol{A}_{p\times p}$ is partitioned,

$$\boldsymbol{A}=\begin{pmatrix} a & \boldsymbol{b}^T \\ \boldsymbol{b} & \boldsymbol{C} \end{pmatrix},$$

where $\boldsymbol{b}$ is (p − 1) × 1, then by setting $\boldsymbol{x} = (1, \boldsymbol{0}^T)^T$, one finds that $\|\boldsymbol{b}\| \le \|\boldsymbol{A}\|$. The matrix $\boldsymbol{B} = \boldsymbol{\rho} u^{\mathrm{T}} + u\boldsymbol{\rho}^{\mathrm{T}}$ has at most two nonzero eigenvalues, given by

$$\lambda=(\tau \pm 1)\,\|\rho\|\|\boldsymbol{u}\|,\ \tau=\rho^T\boldsymbol{u}/\|\rho\|\|\boldsymbol{u}\|. \qquad (18)$$

6.1 Perturbation Bounds

Suppose that a symmetric matrix $\boldsymbol{A}_{p\times p}$ has unit eigenvector $\boldsymbol{q}_1$. We wish to bound the effect of a "symmetric" perturbation $\boldsymbol{E}_{p\times p}$ on $\boldsymbol{q}_1$. The following result (Golub and Van Loan 1996, theorem 8.1.10) constructs a unit eigenvector $\widehat{\boldsymbol{q}_1}$ of $\boldsymbol{A} + \boldsymbol{E}$ and bounds its distance from $\boldsymbol{q}_1$ in terms of $\|\boldsymbol{E}\|$.

Let $\boldsymbol{Q}_{p\times p} = [\boldsymbol{q}_1\ \boldsymbol{Q}_2]$ be an orthogonal matrix containing $\boldsymbol{q}_1$ in the first column, and partition conformally

$$\boldsymbol{Q}^T\boldsymbol{A}\boldsymbol{Q}=\begin{pmatrix} \lambda & 0 \\ 0 & \boldsymbol{D}_{22} \end{pmatrix},\ \boldsymbol{Q}^T\boldsymbol{E}\boldsymbol{Q}=\begin{pmatrix} \epsilon & \boldsymbol{e}^T \\ \boldsymbol{e} & \boldsymbol{E}_{22} \end{pmatrix},$$

where $\boldsymbol{D}_{22}$ and $\boldsymbol{E}_{22}$ are both (p − 1) × (p − 1).

Suppose that λ is separated from the set of eigenvalues of $\boldsymbol{D}_{22}$, denoted $\lambda(\boldsymbol{D}_{22})$; set

$$\delta=\min_{\mu\epsilon\lambda(D_{22})}|\lambda-\mu|.$$

If ‖E‖ ≤ δ/5, then there exists $\boldsymbol{r} \in \mathbb{R}^{p-1}$ satisfying

$$\|\boldsymbol{r}\| \le (4/\delta)\,\|e\| \le (4/\delta)\,\|\boldsymbol{E}\| \tag{19}$$

such that $\widehat{\boldsymbol{q}}_1 = \left(1+\boldsymbol{r}^T\boldsymbol{r}\right)^{-1/2}(\boldsymbol{q}_1+\boldsymbol{Q}_2\boldsymbol{r})$ is a unit eigenvector of$\boldsymbol{A}$ + $\boldsymbol{E}$. Moreover, with d as in (5),

$$d\left(\widehat{q}_1, \widehat{q}_1\right) \le (4/\delta)\,\|\mathbf{e}\|.$$

Let us remark that because $\|\boldsymbol{e}\| \le \|\boldsymbol{E}\|$, we have $\|\boldsymbol{r}\| \le 1$ and

$$\boldsymbol{q}_1^T\widehat{\boldsymbol{q}}_1 = \left(1+\|\boldsymbol{r}\|^2\right)^{-1/2} \ge 1/\sqrt{2}. \tag{20}$$

Suppose now that $\boldsymbol{q}_1$ is the eigenvector of $\boldsymbol{A}$ associated with the "principal" eigenvalue $\lambda_1(\boldsymbol{A})$. Here and later, $\lambda_i(\boldsymbol{A})$ denotes the ith largest eigenvalue of $\boldsymbol{A}$. We verify that, using the preceding conditions, $\widehat{\boldsymbol{q}}_1$ is also the principal eigenvector of $\boldsymbol{A} + \boldsymbol{E}$. In other words, if $(\boldsymbol{A}+\boldsymbol{E})\,\widehat{\boldsymbol{q}}_1 = \lambda * \widehat{\boldsymbol{q}}_1$, then, in fact, $\lambda^* = \lambda_1(\boldsymbol{A} + \boldsymbol{E})$.

To show this, we verify that $\lambda^* > \lambda_2(\boldsymbol{A} + \boldsymbol{E})$. Take inner products with $\boldsymbol{q}_1$ in the eigenequation for $\widehat{\boldsymbol{q}}_1$:

$$\lambda^*\boldsymbol{q}_1^T\widehat{\boldsymbol{q}}_1 = \boldsymbol{q}_1^T\boldsymbol{A}\widehat{\boldsymbol{q}}_1 + \boldsymbol{q}_1^T\boldsymbol{E}\widehat{\boldsymbol{q}}_1. \tag{21}$$

Because $\boldsymbol{A}$ is symmetric, $\boldsymbol{q}_1^T\boldsymbol{A} = \lambda_1(\boldsymbol{A})\,\boldsymbol{q}_1^T$. Trivially, we have $\boldsymbol{q}_1^T\boldsymbol{E}\widehat{\boldsymbol{q}}_1 \ge -\|\boldsymbol{E}\|$. Combine these remarks with (20) to get $\lambda* \ge \lambda_1(\boldsymbol{A}) - \sqrt{2}\|\boldsymbol{E}\|$.

Now $\delta = \lambda_1(\boldsymbol{A}) - \lambda_2(\boldsymbol{A})$ and, because from the minimax characterization of eigenvalues (e.g., Golub and Van Loan 1996, p. 396), $\lambda_2(\boldsymbol{A} + \boldsymbol{E}) \le \lambda_2(\boldsymbol{A}) + \|\boldsymbol{E}\|$, we have

$$\begin{aligned} \lambda^* - \quad & \lambda_2(\boldsymbol{A}+\boldsymbol{E}) \ge \delta - \left(1+\sqrt{2}\right)\|\boldsymbol{E}\| \\ & \ge \delta\left[1-\left(1+\sqrt{2}\right)/5\right] > 0. \end{aligned}$$

6.2 Some Limit Theorems

Collect the noise vectors into a matrix $\boldsymbol{Z}_{p\times n} = [\mathrm{z}_1 \ldots \mathrm{z}_n]$. We turn to properties of the noise matrix $\boldsymbol{Z}$. The cross-products matrix $\boldsymbol{Z}\boldsymbol{Z}^T$ has a standard p-dimensional Wishart $W_p(n, \boldsymbol{I})$ distribution with n degrees of freedom and identity covariance matrix (see, for example, Muirhead 1982, p. 82). Thus, the matrix $\boldsymbol{C} = \sigma^2(n^{-1}\boldsymbol{Z}\boldsymbol{Z}^T - \boldsymbol{I}_p)$ is simply a scaled and recentered Wishart matrix.

Geman (1980) and Silverstein (1985), respectively, established almost sure limits for the largest and smallest eigenvalues of a $W_p(n, \boldsymbol{I})$ matrix as $p/n \to c \in [0, \infty)$, from which follows:

$$\lambda_1(\boldsymbol{C}), \lambda_p(\boldsymbol{C}) \to \sigma^2\left(c \pm 2\sqrt{c}\right). \tag{22}$$

(Although the results in the articles cited are for $c \in (0, \infty)$, the results are easily extended to $c = 0$ by simple coupling arguments.)

Suppose, in addition, that υ is an $n \times 1$ vector with independent $N(0, 1)$ entries, which are also independent of $\mathbf{Z}$. Conditioned on υ, the vector $\mathbf{Z}\upsilon$ is distributed as $N_p(\mathbf{0}, \|\upsilon\|^2\boldsymbol{I})$. Because $\boldsymbol{Z}$ is independent of υ, we conclude that

$$\boldsymbol{Z}\boldsymbol{v} \overset{\mathcal{D}}{=} \chi_{(n)}\chi_{(p)}\boldsymbol{U}_p, \tag{23}$$

where $\chi^2_{(n)}$ and $\chi^2_{(p)}$ denote chi-squared variables the $\boldsymbol{U}_p$ a vector uniform on the surface of the $(p - 1)$-dimensional unit sphere S^{p-1} in $\mathbb{R}^p$, and all three variables $\chi_{(n)}$, $\chi_{(p)}$, and $\boldsymbol{U}_p$ are independent.

Now let $\boldsymbol{u} = \sigma n^{-1}\mathbf{Z}\upsilon$. From (23) we have, as $p/n \to c \in [\mathbf{0}, \infty)$,

$$\|\boldsymbol{u}\|^2 \overset{\mathcal{D}}{=} \sigma^2 n^{-2}\chi^2_{(n)}\chi^2_{(p)} \xrightarrow{a.s.} \sigma^2 c. \tag{24}$$

6.3 Proof of Theorem 2

Outline—Recall from (11) that, given $\alpha_n = \alpha(n^{-1}\log(n \vee p))^{1/2}$, the selected subset of variables $\hat{I}$ is defined by $\hat{I} = \left\{\nu : \hat{\sigma}^2_\nu \ge \sigma^2(1+\alpha_n)\right\}$ and that the estimated principal eigenvector based on $\hat{I}$ written $\hat{\rho}_I$. We define a vector $\boldsymbol{\rho}_I$ with coordinates $(\rho_{I,\nu})$ by selecting coordinates from $\boldsymbol{\rho} = (\rho_\nu)$ according to membership in $\hat{I}$:

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

We will use the triangle inequality $d(\hat{\rho}_I, \rho) \le d(\hat{\rho}_I, \rho_I) + d(\rho_I, \rho)$ to show that $\hat{\rho}_I \to \rho$. There are three main steps.

(i) Construct deterministic sets of indices $I_n^{\pm} = \left\{\nu : \rho_\nu^2 \ge \sigma^2 a_{\mp}\alpha_n\right\}$ with constants $a\mp$ to be determined later, which bracket $\hat{I}$ almost surely as $n \to \infty$:

$$I_n^- \subset \hat{I} \subset I_n^+ \quad w.p.1. \tag{25}$$

(ii) The uniform sparsity, combined with $\hat{I}^c \subset (I_n^-)^c$, is used to show that $d(\rho_I, \rho) \xrightarrow{a.s.} 0$.

(iii) The containment $\hat{I} \subset I_n^+$, along with $|I_n^+| = o(n)$ shows that $d(\rho_I, \rho_I) \xrightarrow{a.s.} 0$.

Details

Step (i): We first obtain a bound on the cardinality of $I_n^{\pm}$ using the sparsity condition (9). Using (9), $|\rho|_{(\nu)} \le \boldsymbol{C}\mathbf{v}^{-1/q}$, and so

$$|I_n^{\pm}| \le |\left\{\nu : C^2\nu^{-2/q} \ge \sigma^2 a \mp \alpha_n\right\}| \le C^q/\left(\sigma^2 a \mp \alpha_n\right)^{q/2} = o\left(n^{1/2}\right).$$

Let $\sigma_\nu^2 = \sigma^2 + \rho_\nu^2$. Turning to the bracketing relations (25), we first remark that $\widehat{\sigma}_\nu^2 \overset{\mathcal{D}}{=} \sigma_\nu^2 \chi_{(n)}^2/n$, and when $\nu \in I_n^{\pm}$,

$$\sigma_\nu^2 = \sigma^2\left(1 + \rho_\nu^2/\sigma^2\right) \ge \sigma^2\,(1 + a \mp \alpha_n).$$

Using the definitions of $\widehat{I}$ and writing $\bar{M}_n$ for a random variable with the distribution of $\chi_{(n)}^2/n$ we have

$$P_n^- = P\left(I_n^- \not\subseteq \widehat{I}\right) \le \sum_{\nu \in I_n^-} \mathbf{P}\left\{\widehat{\sigma}_\nu^2 < \sigma^2\,(1+\alpha_n)\right\} \le |I_n^-|\mathbf{P}\left\{\bar{M}_n < (1+\alpha_n)\,/\,(1+a+\alpha_n)\right\}.$$

We apply (A.2) from the Appendix with $\epsilon_n = (a_+ - 1)\alpha_n/(1 + a_+\alpha_n)$ and for n large and α' slightly smaller than $\alpha^2, n\epsilon_n^2 > (a_+ - 1)^2\alpha'\log(n \vee p)$, so that

$$P_n^- \le cn^{1/2}\exp\left\{-n\epsilon_n^2/4\right\} \le cn^{1/2-\alpha_+''},$$

with $\alpha_+'' = (\alpha_+ - 1)^2\alpha'/4$. If $\alpha \ge \sqrt{12}$, then $\alpha_+'' \ge 3$ for suitable $a_+ > 2$.

The argument for the other inclusion is analogous, using (A.3) in place of (A.2):

$$P_n^+ = \mathbf{P}\left(\widehat{I} \not\subseteq I_n^+\right) \le \sum_{\nu \notin I_n^+} \mathbf{P}\left\{\widehat{\sigma}_\nu^2 \ge \sigma^2\,(1+\alpha_n)\right\} \le p\mathbf{P}\left\{\bar{M}_n \ge (1+\alpha_n)\,/\,(1+a-\alpha_n)\right\} \le p(n \vee p)^{-\alpha_-''},$$

with $\alpha_-'' = 3(1 - a_-)^2\alpha'/16$ so long as α' is now slightly *less* than α^2 and n is large enough. If $\alpha \ge \sqrt{12}$, then $\alpha_-'' > 2$ for suitable $a_- < 1 - \sqrt{8/9}$.

By a Borel-Cantelli argument, (25) follows from the bounds on P_n^- and P_n^+.

Step (ii): We first remark that one may easily show that $d(\boldsymbol{\rho} + \boldsymbol{u}, \boldsymbol{\rho}) \le \|\boldsymbol{u}\|/(\|\boldsymbol{\rho}\| - \|\boldsymbol{u}\|)$, so that norm convergence implies d-convergence. So, for $n > n(\omega)$ we have $I_n^- \subset \widehat{I}$ and so

$$\|\rho_I - \rho\|^2 = \sum_{\nu \notin \widehat{I}} \rho_\nu^2 \le \sum_{\nu \notin I_n^-} \rho_\nu^2.$$

When $\nu \in (I_n^-)^c$, we have by definition

$$\rho_\nu^2(n) < \sigma^2 a + \alpha\sqrt{n^{-1}\log(n \vee p)} := \epsilon_n^2,$$

say, while the uniform sparsity condition entails $|\rho|_{(\nu)}^2 \le C^2\nu^{-2/q}$.

Putting these together, and defining $s_1 = s_1(n)$ as the solution of the equation $\boldsymbol{C}\boldsymbol{s}^{-1/q} = \epsilon_n$, and writing $a \wedge b$ for min(a, b), we obtain, as $n \to \infty$,

$$\begin{aligned}\sum_{\nu \notin I_n^-}\rho_\nu^2 &\le \sum_\nu \epsilon_n^2 \wedge \rho_\nu^2 \le \sum_\nu \epsilon_n^2 \wedge |\rho|_{(\nu)}^2 \\ &\le \sum_\nu \epsilon_n^2 \wedge C^2\nu^{-2/q} \le \int_0^\infty \epsilon_n^2 \wedge C^2 s^{-2/q} ds \\ &= s_1\epsilon_n^2 + q(2-q)^{-1}C^2 s_1^{1-2/q} = [2/(2-q)]\, C^q \epsilon_n^{2-q} \to 0.\end{aligned}$$

<u>**Step (iii):**</u> We adopt the abbreviations

$$\begin{aligned}\boldsymbol{u}_I &= \left(u_\nu : \nu \in \widehat{I}\right), \boldsymbol{Z}_I = \left(z_{\nu i} : \nu \in \widehat{I}, i=1,\ldots,n\right), \\ \boldsymbol{s}_I &= \left(S_{\nu\nu'} : \nu, \nu' \in \widehat{I}\right),\end{aligned}$$

and similarly for $\boldsymbol{E}_I$. We consider $\boldsymbol{S}_I^* = \boldsymbol{S}_I - \sigma^2 \boldsymbol{I}_{\widehat{k}} = \rho_I\rho_I^T + \boldsymbol{E}_I$ and note that the perturbation term has the decomposition

$$\boldsymbol{E}_I = \upsilon_s \rho_I \rho_I^T + \rho_I \boldsymbol{u}_I^T + \boldsymbol{u}_I \rho_I^T + \sigma^2\left(n^{-1}\boldsymbol{Z}_I\boldsymbol{Z}_I^T - \boldsymbol{I}\right),$$

so that

$$\|\boldsymbol{E}_I\| \le \upsilon_s\|\rho\|^2 + 2\|\rho_I\|\|\boldsymbol{u}_I\| + \sigma^2\left|\lambda_{max}\left(n^{-1}\boldsymbol{Z}_I\boldsymbol{Z}_I^T\right) - 1\right|.$$

Consider the first term on the right side. Because $\|\rho_I - \rho\| \xrightarrow{a.s.} 0$ from Step (ii), it follows that $\|\rho_I\| \xrightarrow{a.s.} \|\rho\|$. Because $\upsilon_s \xrightarrow{a.s.} 0$, the first term is asymptotically negligible.

Let $\boldsymbol{Z}_{I^+} = (z_{\nu i} : \nu \in I_n^+, i=1,\ldots,n)$ and $\boldsymbol{u}_{I^+} = (u_\nu : \nu \in I_n^+)$. On the event $\Omega_n = \left\{\widehat{I} \subset I_n^+\right\}$, we have $\| \boldsymbol{u}_I \| \le \| \boldsymbol{u}_{I^+} \|$ and setting $k_+ = |I_n^+|$, by the same arguments as led to (24), we have

$$\|\boldsymbol{u}_{I^-}\|^2 \overset{\mathcal{D}}{=} \sigma^2 (k_+/n)\left(\chi_{(n)}^2/n\right)\left(\chi_{(k_+)}^2/k_+\right) \xrightarrow{a.s.} 0,$$

because $k_+ = o(n)$ from Step (i).

Finally, because on the event Ω_n, the matrix $\boldsymbol{Z}_{I^+}$ contains $\boldsymbol{Z}_I$, along with some additional rows, it follows that $\lambda_{max}\left(n^{-1}\boldsymbol{Z}_I\boldsymbol{Z}_I^T - \boldsymbol{I}\right) \le \lambda_{max}\left(n^{-1}\boldsymbol{Z}_{I^+}\boldsymbol{Z}_{I^+}^T - \boldsymbol{I}\right) \xrightarrow{a.s.} 0$ by (22), again because $k_+ = o(n)$. Combining previous bounds, we conclude that $\|\boldsymbol{E}_I\| \to 0$.

The separation $\delta_n = \|\boldsymbol{\rho}_I\|^2 \to \|\boldsymbol{\rho}\|^2 > 0$ and so, by the perturbation bound (19),

$$d\left(\widehat{\rho}_l,\rho_l\right)\le(4/\delta_n)\,\|\boldsymbol{E}_l\|\xrightarrow{a.s.}0.$$

7. CONCLUSIONS

In models with observational noise such as (2), in which the number of variables *p* grows with the number of cases *n,* we have reviewed results that show that standard PCA yields consistent estimates of the principal eigenvectors if and only if *p/n* → 0.

If the leading population principal eigenvector has a sparse representation in a given basis, Theorem 2 shows that it can be consistently estimated by selecting a subset of variables with variances above a threshold and then by restricting the PCA to this selected set. Incorporating a threshold is found empirically to be helpful. Future theoretical work might explore the tradeoff between variable selection and thresholding.

In summary, sparse PCA as described here may be of practical benefit in high-dimensional settings with substantial observational noise in which variation between individuals resides mainly in a subset of the coordinates in which the data are represented (perhaps after transformation).

Acknowledgments

This work was supported in part by grants NSF DMS 0072661, 0505303, and NIH EB R01 EB001988.

APPENDIX

A.1 Large Deviation Inequalities

If $\bar{X}=n^{-1}\sum_1^n X_i$ is the average of iid variates with moment-generating function exp{∧(λ)} = **E** exp{λX_1}, then Cramer's theorem (see, for example, Dembo and Zeitouni 1993, sections 2.2.2 and 2.2.12) says that for $x > \mathbf{E}X_1$,

$$\mathbf{P}\left\{\bar{X}>x\right\}\le\exp\left\{-n\Lambda*(x)\right\},\tag{A.1}$$

where the conjugate function $\wedge^*(x) = \sup_\lambda\{\lambda x - \wedge(\lambda)\}$. The same bound holds for $\mathbf{P}\left\{\bar{X}<x\right\}$ when $x < \mathbf{E}X_1$. When applied to the $\chi^2_{(n)}$ distribution, with $X_1 = z^2$ and $z \sim N(0, 1)$, the moment-generating function

$$\Lambda(\lambda)=-\frac{1}{2}\log(1-2\lambda)$$

and the conjugate function

$$\Lambda*(x)=\frac{1}{2}\left[x-1-\log x\right].$$

The bounds

$$\log(1+\epsilon) \le \begin{cases} \epsilon - \epsilon^2/2 & -1<\epsilon<0, \\ \epsilon - 3\epsilon^2/8 & 0 \le \epsilon<\frac{1}{2}, \end{cases}$$

(the latter following, for example, from (47) in Johnstone (2001)) yield

$$\mathbf{P}\left\{\chi^2_{(n)} \le n(1-\epsilon)\right\} \le \exp\left\{-n\epsilon^2/4\right\}, 0 \le \epsilon<1, \tag{A.2}$$

$$\mathbf{P}\left\{\chi^2_{(n)} \ge n(1+\epsilon)\right\} \le \exp\left\{-3n\epsilon^2/16\right\}, 0 \le \epsilon<\frac{1}{2}. \tag{A.3}$$

A.2 Proof of Theorem 3

Assume, without loss of generality, that $\sigma_1^2 \ge \sigma_2^2 \ge \ldots \ge \sigma_p^2$.

False Inclusion

For any fixed constant t, and $l \in \boldsymbol{I}_{out}$,

$$\widehat{\sigma}_i^2 \ge t \quad \text{for} \quad i=1,\ldots,k \quad \text{and} \quad \widehat{\sigma}_l^2<t \Rightarrow \widehat{\sigma}_l^2<\widehat{\sigma}_{(k)}^2.$$

This threshold device leads to bounds on error probabilities using only marginal distributions. For example, consider false inclusion of variable l:

$$\mathbf{P}\left\{\widehat{\sigma}_l^2 \ge \widehat{\sigma}_{(k)}^2\right\} \le \Sigma_{i=1}^k \mathbf{P}\left\{\widehat{\sigma}_i^2<t\right\} + \mathbf{P}\left\{\widehat{\sigma}_l^2 \ge t\right\}.$$

Write $\bar{M}_n$ for a $\chi^2_{(n)}/n$ variate, and note from (12) that $\widehat{\sigma}_\nu^2 \sigma_\nu^2 \bar{M}_n$. Set $t=\sigma_k^2(1-\epsilon_n)$ for a value of ϵ_n to be determined. Because $\sigma_i^2 \ge \sigma_k^2$ and $\sigma_l^2 \le \sigma_k^2(1-\alpha_n)$, we arrive at

$$\begin{aligned} \mathbf{P}\left\{\widehat{\sigma}_l^2 \ge \widehat{\sigma}_{(k)}^2\right\} & \le k\mathbf{P}\left\{\bar{M}_n<1-\epsilon_n\right\} + \mathbf{P}\left\{\bar{M}_n \ge \tfrac{1-\epsilon_n}{1-\alpha_n}\right\} \\ & \le k\exp \left\{-\tfrac{n\epsilon_n^2}{4}\right\} + \exp\left\{-\tfrac{3n}{16}\left(\tfrac{\alpha_n-\epsilon_n}{1-\alpha_n}\right)^2\right\} \end{aligned}$$

using large deviation bound (A.2). With the choice $\epsilon_n = \sqrt{3}\alpha_n/\left(2+\sqrt{3}\right)$ both exponents are bounded above by $-b(\alpha)\log(n \vee p)$, and so $\mathbf{P}\{FI\} \le p(k+1)(n \vee p)^{-b(\alpha)}$.

False Exclusion

The argument is similar, starting with the remark that for any fixed t and $l \in \boldsymbol{I}_{in}$,

$$\widehat{\sigma}_i^2 \le t \quad \text{for} \quad i \ge k, i \ne l \quad \text{and} \quad \widehat{\sigma}_l^2 \ge t \Rightarrow \widehat{\sigma}_l^2 \ge \widehat{\sigma}_{(k)}^2.$$

Consequently, if we set $t=\sigma_k^2(1+\varepsilon_n)$ and use $\sigma_1^2 \ge \sigma_k^2(1+\alpha_n)$, we get

$$\begin{aligned} \mathbf{P}\left\{\widehat{\sigma}_l^2<\widehat{\sigma}_{(k)}^2\right\} & \leq \Sigma_{i\geq k}\mathbf{P}\left\{\widehat{\sigma}_i^2>t\right\}+\mathbf{P}\left\{\widehat{\sigma}_l^2<t\right\} \\ & \leq (p-1)\,\mathbf{P}\left\{\bar{M}_n>1+\epsilon_n\right\} \\ & +\mathbf{P}\left\{\bar{M}_n>\frac{1+\epsilon_n}{1+\alpha_n}\right\} \\ & \leq (p-1)\exp\left\{-\frac{3n\epsilon_n^2}{16}\right\} \\ & +\exp\left\{-\frac{n}{4}\left(\frac{\alpha_n-\epsilon_n}{1+\alpha_n}\right)^2\right\}, \end{aligned}$$

this time using (A.3). The bound $\mathbf{P}\{\mathrm{FE}\} \leq pk(n \vee p)^{-b(\alpha)}\, ke^{-\mathrm{b}(\alpha)(1\,-\,2\mathrm{a}_n)\log(n\vee p)}$ follows on setting $\epsilon_n=2\alpha_n/\left(2+\sqrt{3}\right)$ and noting that $(1 + \alpha_n)^{-2} \geq 1 - 2\alpha_n$.

For numerical bounds, setting $L_n = \log(n \vee p)$, we may collect the preceding bounds in the form

$$\begin{aligned} \mathbf{P}\left(FE\cup FI\right)\leq\; & \left[pk+(p-1)(k-1)\right]e^{-b(\alpha)L_n} \\ & +pe^{-b(\alpha)L_n/(1-\alpha_n)^2} \\ & +(k-1)\,e^{-b(\alpha)L_n/(1+\alpha_n)^2}. \end{aligned} \tag{A.4}$$

REFERENCES

Amini, AA.; Wainwright, MJ. The Annals of Statistics. 2009. High-Dimensional Analysis of Semidefinite Relaxations for Sparse Principal Components.

Anderson TW. Asymptotic Theory for Principal Component Analysis. Annals of Mathematical Statistics 1963;34:122–148.

Biehl M, Mietzner A. Statistical Mechanics of Unsupervised Structure Recognition. Journal of Physics A: Mathematical and General 1994;27:1885–1897.

Cherkassky, V.; Mulier, F. Learning from Data. Wiley; New York: 1998.

d'Aspremont A, El Ghaoui L, Jordan M, Lanckriet G. A Direct Formulation for Sparse PCA Using Semidefinite Programming. SIAM Review 2007;49:434–448.

Dembo, A.; Zeitouni, O. Large Deviations Techniques and Applications. Jones and Bartlett; Boston: 1993.

Donoho D. Unconditional Bases Are Optimal Bases for Data Compression and Statistical Estimation. Applied and Computational Harmonic Analysis 1993;1:100–115.

Donoho DL, Johnstone IM, Kerkyacharian G, Picard D. “Wavelet Shrinkage: Asymptopia” (with discussion). Journal of the Royal Statistical Society, Ser. B 1995;57:301–369.

Du Q, Fowler JE. Low-Complexity Principal Component Analysis for Hyperspectral Image Compression. International Journal of High Performance Computing Applications 2008;22:438–448.

Feng GC, Yuen PC, Dai DQ. Human Face Recognition Using PCA on Wavelet Subband. Journal of Electronic Imaging 2000;9:226–233.

Fowler, JE. Compressive–Projection Principal Component Analysis for the Compression of Hyperspectral Signatures. In: Storer, JA.; Marcellin, MW., editors. Data Compression Conference, 2008. DCC 2008. IEEE; Snowbird, UT: 2008. p. 83-92.

Geman S. A Limit Theorem for the Norm of Random Matrices. Annals of Probability 1980;8:252–261.

Golub, GH.; Van Loan, CF. Matrix Computations. 3rd ed. Johns Hopkins University Press; Baltimore: 1996.

Hoyle DC, Rattra y, M. Principal-Component-Analysis Eigenvalue Spectra from Data with Symmetry Breaking Structure. Physical Review E: Statistical, Nonlinear, and Soft Matter Physics 2004;69:026124.

Johnstone, IM. Chi Square Oracle Inequalities. In: de Gunst, M.; Klaassen, C.; van der Waart, A., editors. Festschrift for Willem R. van Zwet (vol. 36 of IMS Lecture Notes—Monographs). Institute of Mathematical Statistics; Beachwood, OH: 2001. p. 399-418.

Johnstone, IM. Function Estimation and Gaussian Sequence Models. Draft of a monograph. 2003. Available at www-stat.stanford.edu/~imj

Johnstone, IM.; Lu, AY. Technical Report. Stanford University, Dept. of Statistics; 2004. Sparse Principal Components Analysis. Available at *arxiv.org* as e-print 0901.4392

Jolliffe IT, Trendafilov NT, Uddin M. A Modified Principal Component Technique Based on the LASSO. Journal of Computational and Graphical Statistics 2003;12:531–547.

Kaewpijit, S.; Le Moigne, J.; El-Ghazawi, T. A Wavelet-Based PCA Reduction for Hyperspectral Imagery. Geoscience and Remote Sensing Symposium, 2002.IGARSS'02. 2002 IEEE International; Washington, DC: IEEE; 2002. p. 2581-2583.

Lu, AY. Ph.D. dissertation. Stanford University, Dept. of Statistics; 2002. Sparse Principal Components Analysis for Functional Data.

Mallat, S. A Wavelet Tour of Signal Processing. 2nd ed.. Academic Press; New York: 1999.

Muirhead, RJ. Aspects of Multivariate Statistical Theory. Wiley; NewYork: 1982.

Nadler B. Finite Sample Approximation Results for Principal Component Analysis: A Matrix Perturbation Approach. The Annals of Statistics 2008;36:2791–2817.

Nadler B, Coifman R. The Prediction Error in CLS and PLS: The Importance of Feature Selection Prior to Multivariate Calibration. Journal of Chemometrics 2005;19:107–118.

Paul D. Asymptotics of Sample Eigenstructure for a Large Dimensional Spiked Covariance Model. Statistica Sinica 2007;17:1617–1642.

Paul, D.; Johnstone, I. Technical Report. Stanford University, Dept. of Statistics; 2004. Estimation of Principal Components through Coordinate Selection.

Ramsay, JO.; Silverman, BW. Functional Data Analysis. Springer; Berlin: 1997.

Reimann P, Van den Broeck C, Bex GJ. A Gaussian Scenario for Unsupervised Learning. Journal of Physics A: Mathematical and General 1996;29:3521–3535.

Rice JA, Silverman BW. Estimating the Mean and Covariance Structure Nonparametrically When the Data Are Curves. Journal of the Royal Statistical Society, Ser. B 1991;53:233–243.

Silverman BW. Smoothed Functional Principal Components Analysis by Choice of Norm. The Annals of Statistics 1996;24:1–24.

Silverstein JW. The Smallest Eigenvalue of a Large Dimensional Wishart Matrix. Annals of Probability 1985;13:1364–1368.

Watkin TLH, Nadal J-P. Optimal Unsupervised Learning. Journal of Physics A: Mathematical and General 1994;27:1899–1915.

Wickerhauser MV. Large-Rank Approximate Principal Component Analysis with Wavelets for Signal Feature Discrimination and the Inversion of Complicated Maps. Journal of Chemical Information and Computer Sciences 1994a;34:1036–1046.

Wickerhauser MV. Two Fast Approximate Wavelet Algorithms for Image Processing, Classification, and Recognition. Optical Engineering 1994b;33:2225–2235. (Special issue on Adapted Wavelet Analysis.).

Wolf, L.; Bileschi, S. Computer Vision and Pattern Recognition. Vol. vol. 2. IEEE Computer Society; Washington, DC: 2005. Combining Variable Selection wit Dimensionality Reduction; p. 801-806.2005

Wolf L, Shashua A. Feature Selection for Unsupervised and Supervised Inference: The Emergence of Sparsity in a Weighted-Based Approach. Journal of Machine Learning Research 2005;6:1855–1887.

Zou H, Hastie T, Tibshirani R. Sparse Principal Component Analysis. Journal of Computational and Graphical Statistics 2006;15:265–286.

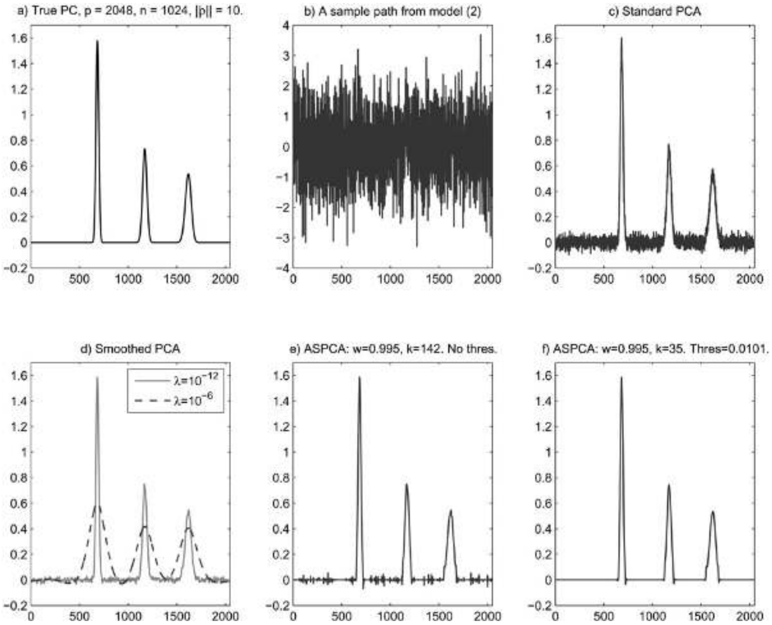

Figure 1.

True principal component, the “three-peak” curve. (a) The single component $\rho_l = f(l/p)$ where $f(t) = C\{0.7B(1{,}500, 3{,}000) + 0.5B(1{,}200, 900) + 0.5B(600, 160)\}$ and $B(a, b)(t) = [\Gamma(a + b)/(\Gamma(a)\Gamma(b))]t^{a\bullet 1}(1 - \mathrm{t})^{b-1}$ denotes the beta density on [0, 1]. (b) A sample case drawn from model (2) with σ = 1, n = 1,024 replications in total, p = 2,048. (c) Sample principal component by standard PCA. (d) Sample principal component by smoothed PCA using $\lambda = 10^{-12}$ and $\lambda = 10^{-6}$. (e, f) Sample principal component by sparse PCA with weighting function w = 99.5%, and k = 142 and 35, respectively, without and with a thresholding step.

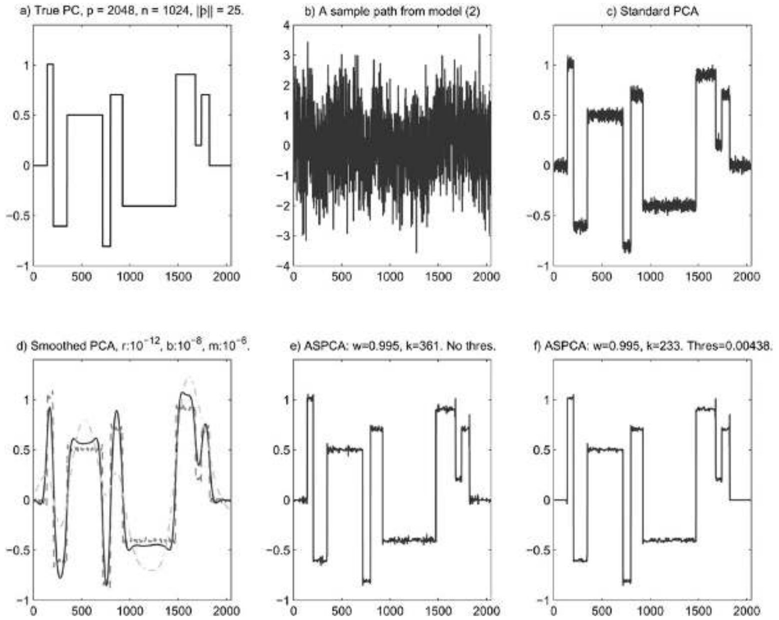

Figure 2.

Comparison of the sample principal components for a step function. (a) True principal component ρ_l. (b) a sample case drawn from model (2) with $\sigma = 1$, $p = 2048$. (c) Sample principal component by standard PCA. (d) Sample principal component by smoothed PCA using $\lambda = 10^{-12}$ and 10^{-6}. (e, f) Sample principal component by sparse PCA with weighting function $w = 99.5\%$, and $k = 361$ and 233 without and with a thresholding step, respectively.

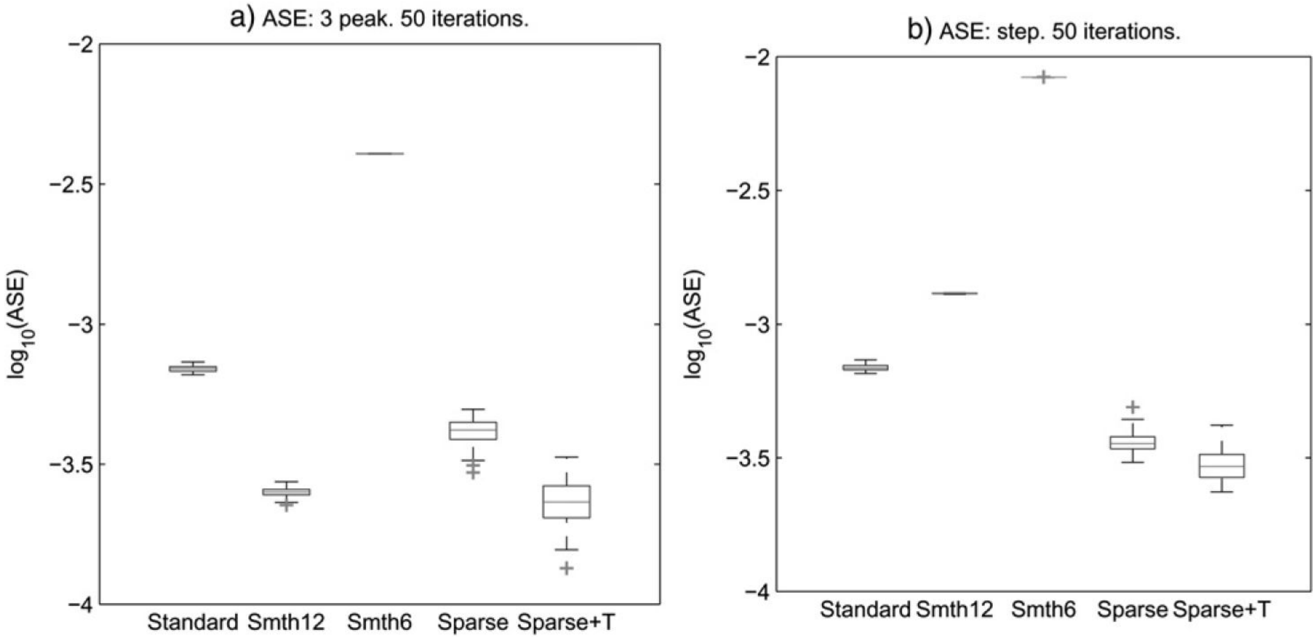

Figure 3.

Side-by-side boxplots of ASE from 50 iterations using different algorithms for the “three-peak” function (a) and for the “step” function (b).

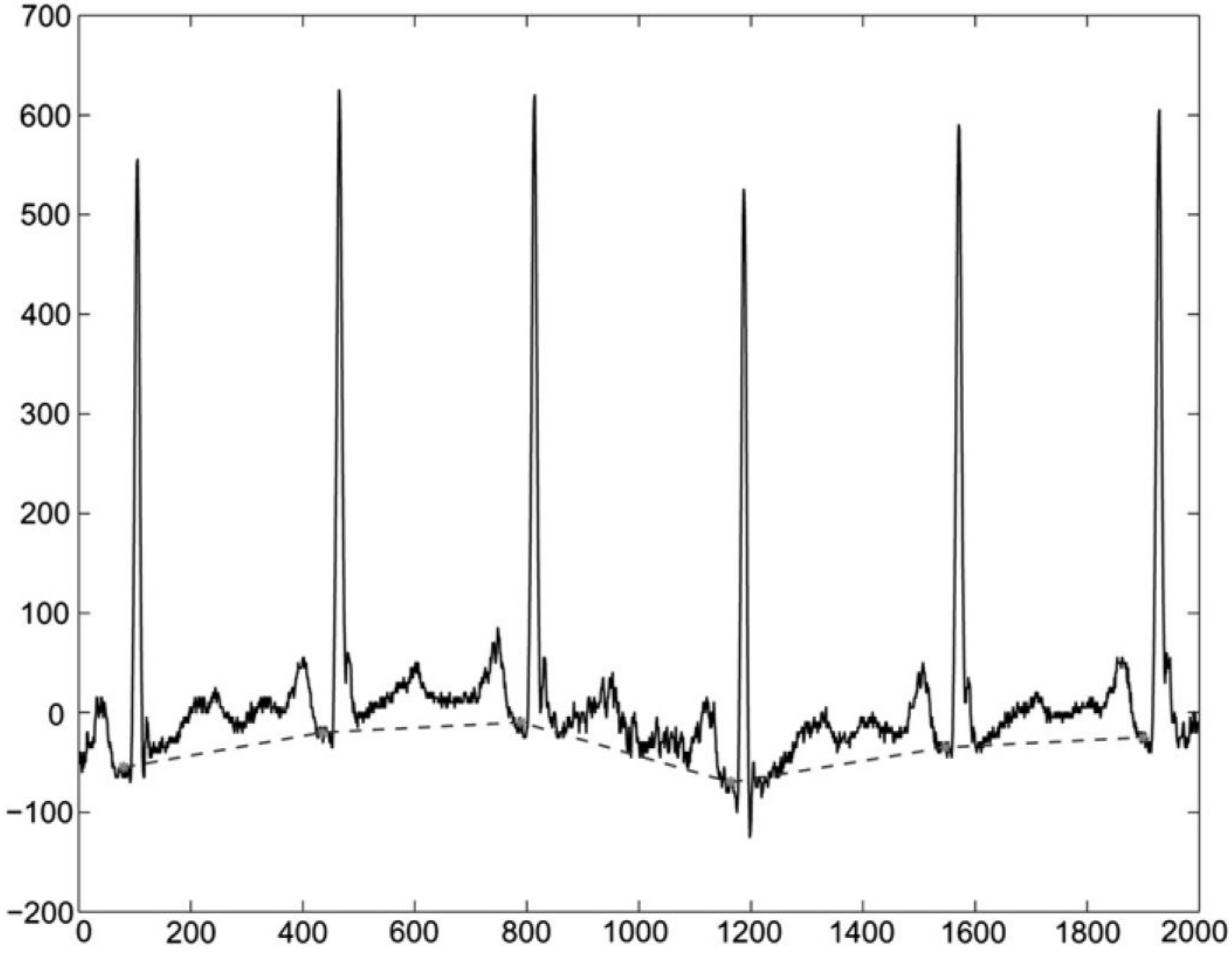

Figure 4.

Baseline wander is observed in many ECG datasets. One common remedy for this problem is to deduct a piecewise linear baseline from the signal, the linear segment (dashed line) between two beats being determined from two adjacent onset points. The onset positions of R waves are shown by asterisks. Their exact locations vary for different patients and, as seen here, even for adjacent R waves. The locations are determined manually in this example. To reduce the effect of noise, the values of onset points are calculated by an average of 5 points close to the onset position.

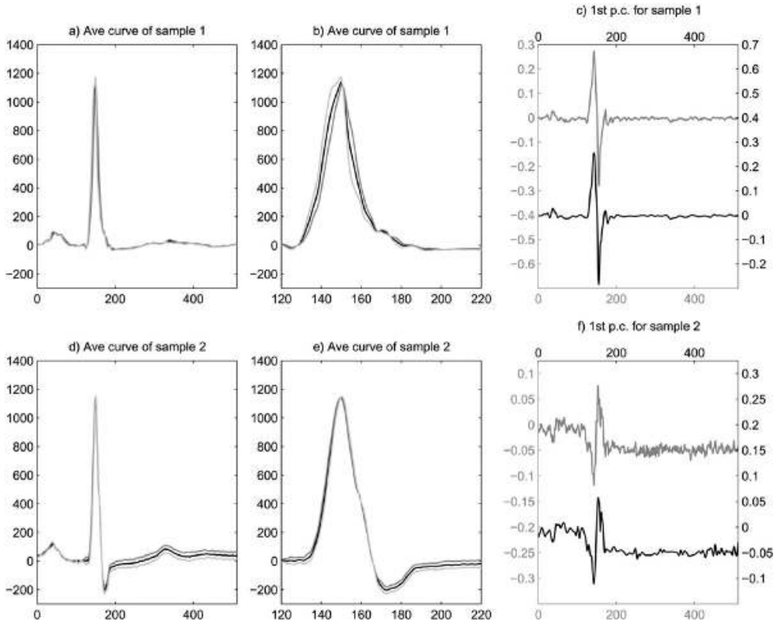

Figure 5.

ECG examples. (Note: Colors refer to online version; “vertical offset” to monochrome print version for clarity.) (a) Mean curve for ECG sample 1, $n = 66$, in blue, along with $\bar{\boldsymbol{x}} + 2\hat{\rho}$ (green) and $\bar{\boldsymbol{x}} - 2\hat{\rho}$ (red), with $\hat{\rho}$ being the estimated first principal component from sparse PCA (see also (c)). (b) Magnified section of (a) over the range 120 to 220. (c) First principal components for sample 1 from standard (upper/left *y*-axis) and sparse PCA (lower/right *y*-axis; vertical offset, −0.4); threshold, 0.0044. (d–f) Corresponding plots for sample 2, $n = 61$. Vertical offset for sparse PCA in (f) is −0.2 and threshold is 0.0075.

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