Discussion

Boaz NADLER

1. INTRODUCTION

I commend Johnstone and Lu for publishing this important article, which has motivated quite a lot of recent work on sparsity and statistical inference in high-dimensional settings. In their article, Johnstone and Lu present two main results. First, in the presence of considerable noise in the x variables, with a number of samples n not significantly larger than the number of variables p, the sample eigenvectors computed by standard principal component analysis (PCA) may be poor approximations to their population analogs. Second, if the sample observations are known to be sparse in some a priori known basis, then it is possible, via a thresholding procedure, to obtain both improved eigenvector estimates (provably consistent in an appropriate limit) as well as substantial computational savings.

Because PCA is an unsupervised method, one question that a reader may ask is whether this curse of dimensionality, leading to large reconstruction errors in high dimensions, is the result of the lack of supervision. In other words, should we worry about similar problems in supervised settings, such as classification or regression, where for each sample x a response variable y is also given?

Regretfully, the short answer to this question is yes. This curse of dimensionality also affects the supervised scenario. A few years ago, independent of the work of Johnstone and Lu, Ronald Coifman and myself (Nadler and Coifman 2005) considered a regression problem in a setting similar to the one considered in the article by Johnstone and Lu. Because the errors are in the x variables, this is an error-in-variables regression problem. Rather than analyzing the joint limit as both $p, n \to \infty$, in Nadler and Coifman (2005) we kept the number of variables p and the number of samples n as fixed, but viewed the noise strength σ as a small parameter, and expanded the estimated regression vector and resulting mean squared error as a function of σ. We showed that, similar to the findings of Johnstone and Lu, large prediction errors may occur in high-dimensional settings. In particular, for various regression methods, such as classic least squares and partial least squares, we derived the following formula for the mean squared error of prediction:

$$\mathbb{E}[(y-\hat{y})^2] = \frac{\sigma^2}{\| \mathbf{b} \|^2} \times \left(1 + \frac{c_1}{n} + \frac{\sigma^2}{\| \mathbf{b} \|^2}\frac{p^2}{n^2}(c_2 + o(1)) + O(\sigma^4)\right) \quad (1)$$

where σ is the noise level, $\mathbf{b}$ is the true regression vector, and c_1, c_2 are constants independent of n, p. Hence, in the $p \gg n$ setting with substantial noise, the $(p/n)^2$ term inside the brackets may be larger than unity and may dominate the prediction error. Furthermore, in Nadler and Coifman (2005), motivated by problems in chemometrics and spectroscopy, in which the signals are known to be smooth and hence sparse in a wavelet basis, we suggested thresholding the signals by representing them with only a few wavelet coefficients, computed by a joint best basis approach.

A similar phenomenon, of large errors when $p \gg n$ was also shown in a classification setting a decade earlier by Buckheit and Donoho (1995), who also, not surprisingly, suggested thresholding of the wavelet coefficients. Neither of these works presented a consistency proof of the performance of a thresholding procedure, as described by Johnstone and Lu.

2. SOME THEORETICAL CALCULATIONS

There are two main issues raised in the work of Johnstone and Lu: the first is the accuracy in reconstruction of the underlying eigenvectors, and the second is the speed or computational complexity of a suggested algorithm—both under the assumption that the signals are sparse in an a priori known basis.

Let us first consider the issue of accuracy. In contrast to the approach taken by Johnstone and Lu of analyzing consistency in the joint limit $p, n \to \infty$, I shall keep p, n finite and fixed. Consider, then, a one-component system, denote by $\mathbf{v}$ the (unit norm) population eigenvector corresponding to the largest eigenvalue and by $\hat{\mathbf{v}}$ the corresponding eigenvector of sample PCA. From Nadler (2008, corollary 2), it follows that the error in eigenvector reconstruction is given by

$$\sin\theta = \frac{\sigma}{\lambda}\sqrt{\frac{p-1}{n}}\left(1 + O(\sigma) + O(1/\sqrt{n})\right),$$

where θ is the angle between the two vectors $\mathbf{v}, \hat{\mathbf{v}}$, and λ is the largest eigenvalue in the absence of noise (in the notation of Johnstone and Lu, $\lambda = \|\rho\|^2$). Hence,

$$\| \mathbf{v} - \hat{\mathbf{v}} \|^2 = 2(1 - \cos\theta) \approx \frac{\sigma^2}{\lambda}\frac{p-1}{n}$$

and so the accuracy of signal reconstruction is

$$ASE = \frac{1}{p}\| \rho - \hat{\rho} \| = \frac{\sqrt{\lambda}}{p}\| \mathbf{v} - \hat{\mathbf{v}} \| \approx \sigma\sqrt{\frac{1}{pn}}. \quad (2)$$

Note that the approximate expression in Equation (2) is independent of the signal-to-noise ratio. For $p = 2{,}048$, $n = 1{,}024$, and $\sigma = 1$, Equation (2) gives that $(pn)^{-1/2} \approx 6.9 \times 10^{-4}$ in agreement with table 1 in Johnstone and Lu.

Now consider the effect of variable selection (after transformation to an appropriate basis, in which the signals are sparse). Let k denote the number of chosen variables, ρ_k the response vector ρ restricted to these variables, $\rho_k^{\perp} = \rho - \rho_k$ its orthogonal complement, and $\hat{\rho}_k$ the eigenvector of sample PCA

Boaz Nadler is Senior Scientist, Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, 76100, Israel. (E-mail: *boaz.nadler@weizmann.ac.il*)

Journal of the American Statistical Association
June 2009, Vol. 104, No. 486
DOI 10.1198/jasa.2009.0147

Table 1. Accuracy and timing comparison of different algorithms

	PCA	BL	Sparse + Threshold	CORR
ASE (three-peak)	6.9e−4	1.8e−4	2.2e−4	1.5e−4
No. of features	2,048	NR	495	43
Time (sec)	2.5	3.1	1.7	1.7
ASE (step)	6.9e−4	2.4e−4	2.9e−4	2.5e−4
No. of features	2,048	NR	415	240
Time [sec]	2.5	2.8	1.5	1.6

NOTE: ASE, averaged root squared error; BL, Bickel and Levina (2008); CORR, correlation matrix; NR, not relevant.

computed only on these k chosen variables. Then, following the same reasoning as noted earlier, we have a reconstruction error of

$$ASE_k = \frac{1}{p} \| \rho - \hat{\rho}_k \| = \frac{1}{p} \left(\| \rho - \rho_k \|^2 + \| \rho_k - \hat{\rho}_k \|^2 \right)^{1/2} = \frac{1}{p} \left(\| \rho_k^{\perp} \|^2 + \frac{k-1}{n} \sigma^2 \right)^{1/2}. \quad (3)$$

This formula provides insight into the best achievable accuracy for finite p, n. The optimal basis is, of course, one in which the first basis function is simply the vector ρ. Then $k = 1$, $\rho_k^{\perp} = 0$, and $ASE_1 = 0$. Of course, we do not know the vector ρ, but rather assume it has a sparse representation in a given basis. The quality of this basis can be assessed by its minimal achievable (theoretical) error—for example, the value of k for which ASE_k is minimal. Then, the performance of any algorithm for sparse reconstruction can be checked against this optimal error. Another interpretation of Equation (3) is to view variable selection as a simple bias–variance tradeoff. The first term in (3) is the bias resulting from choosing only k variables, whereas the second term is the (smaller) variance of reconstruction in the lower dimensional space.

A second insight from Equation (3) is with respect to the set of features that yield the minimal error. At the optimal value of k, we have that $ASE_{k+1} > ASE_k$, and $ASE_k < ASE_{k-1}$. These conditions give

$$\rho_{(k)}^2 > \frac{\sigma^2}{n} \text{ and } \rho_{(k+1)}^2 < \frac{\sigma^2}{n}, \quad (4)$$

where $\rho_{(\nu)}$ are the coefficients of the signal ρ sorted in decreasing order of magnitude (in absolute value). In other words, for optimal reconstruction we need to find all the features of the signal with energy larger than σ^2/n. The more observations we have, the more features of the signal we should choose. Regretfully, however, a simple thresholding of the empirical variances at a threshold of say $\sigma^2(1 + 1/n)$, in general, will not yield an optimal set of features. The reason is that noise variables themselves have an empirical variance with mean σ^2, but fluctuation on the order of $\sigma^2/\sqrt{n} \gg \sigma^2/n$. Moreover, in high dimensions ($p \gg 1$) where, by the sparsity assumption, most variables are pure noise, even larger signal variables may be difficult to detect, because some of the noise variables will have an empirical variance as large as $\sigma^2(1 + \sqrt{2 \log p/n})$.

Figure 1 presents the theoretical optimal curve for ASE_k as a function of k, assuming an oracle that, for each value of k gives

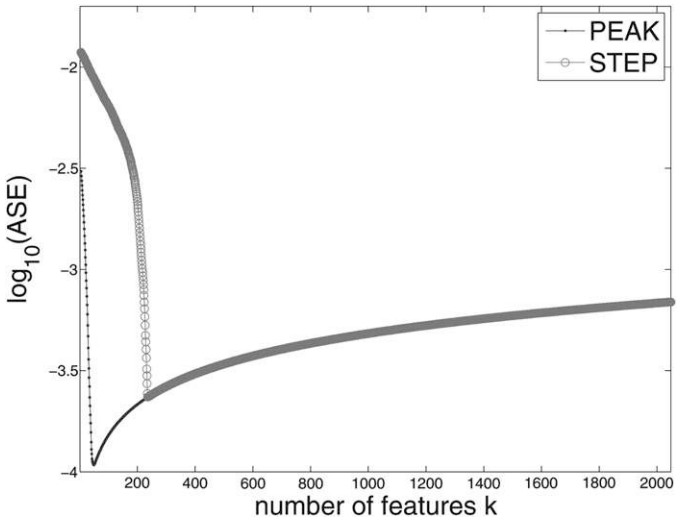

Figure 1. The theoretically optimal reconstruction error versus the number of chosen features, Equation (3).

us the best (large magnitude) k features of the unknown vector ρ. For the three-peak function represented in the `symmlet` wavelet basis, with $n = 1{,}024$ observations, the minimal error 1.1×10^{-4} is obtained with roughly 50 features, whereas for the step function with the `Haar` basis, the optimal error is 2.35×10^{-4} with roughly 240 features. Comparison of these numbers with Table 1 shows that although the thresholding procedure suggested by Johnstone and Lu leads to considerably smaller reconstruction errors in comparison with those of PCA on all variables, there may still be room for improved methods either for variable selection or for covariance regularization. Also note from Figure 1 the high sensitivity of reconstruction errors to mistakes, either by exclusion of important signal features or by incorrect inclusion of noise variables as signals.

3. SPARSITY AND REGULARIZATION

The theoretical analysis of the previous section showed that there is some gap between the performance of the sparse PCA method of Johnstone and Lu and the possibly optimal one. In Table 1 of this discussion, the mean number of indices chosen by the initial thresholding step of the sparse PCA algorithm is shown. We note that, in accordance with the theoretical analysis of the previous section, the initial thresholding step chooses many more variables than necessary for both the three-peak and the step functions. Moreover, in both cases, some of the optimal 50 or 240 features are not always part of this initial set. In other words, quite a few noise variables find their way in and some signal features are regretfully left out. These findings explain the deterioration in performance of the sparse PCA algorithm, the relative success of the post-thresholding step, and suggest that either one of methods (a) or (b) in the article by Johnstone and Lu for initial variable selection may not work well, in particular, at low signal-to-noise ratios.

Can one do better by other methods? First, let us consider a different approach for regularization recently suggested by Bickel and Levina (2008). Their method assumes that the covariance matrix is sparse and computes a "thresholded" covariance matrix $T[\mathbf{S}_n]$, where only entries larger than some

constant s are retained. Then the eigenvalues and eigenvectors of this thresholded matrix are computed. In Bickel and Levina (2008), a specific cross-validation method was suggested for computing this threshold. I have implemented their method and found that for a signal strength of $\|\rho\| = 25$, the threshold is approximately $s \approx 3.5\sqrt{\log p/n}$. To save computational time, this fixed threshold was used for all subsequent experiments. Table 1 also shows the reconstruction errors with this regularization method (denoted BL). As seen from Table 1, this method performs remarkably well. In comparison with sparse PCA it gives slightly lower errors in the three-peak case, but significantly smaller errors for the step function. Yet, even with this approach, there is still a gap from the optimal achievable errors.

At this point I would like to emphasize an important difference between the approaches of Johnstone and Lu and of Bickel and Levina (2008). Although the sparse PCA approach of Johnstone and Lu assumes that all the individual signals are simultaneously sparse, and hence the resulting eigenvector must be sparse as well, the covariance regularization approach of Bickel and Levina (2008), only assumes that the covariance matrix is sparse, but not necessarily its eigenvectors. Simple examples of the latter are the identity matrix and the covariance matrix of an autoregressive process of order one.

Our key observation is that the assumption of Johnstone and Lu—that individual signals are simultaneously sparse in some unknown basis—implies more than just having relatively few features with large variance. It also implies that these features should be highly correlated among themselves. As an example, Figure 2 presents the empirical correlation matrix of data from the three-step function (represented in the `symmlet` wavelet basis). As clearly seen in Figure 2, not only do signal features typically have a larger variance, but they are also highly correlated among themselves. Similar, although much more complicated, structures are typically seen in correlation matrices arising in various applications, including microarray data and text documents.

Under the assumption of uncorrelated Gaussian noise, this observation suggests an alternate, more refined approach to feature selection. Rather than working only with the covariance matrix, we also analyze the structure of the correlation matrix and look for highly correlated variables. Our suggested procedure, denoted CORR, works as follows:

1. Given a data matrix $X_{n,p}$, compute the covariance and correlation matrices $\mathbf{S}_n$, $\mathbf{C}_n$, respectively.
2. Estimate the noise variance as in the sparse PCA algorithm,
$$\hat{\sigma}^2 = \text{median}(\mathbf{S}_n(i,i)).$$
3. Find the sure signal features:
$$I_s = \left\{ i \left| \frac{\mathbf{S}_n(i,i)}{\hat{\sigma}^2} > 1 + \sqrt{\frac{2}{n}} t(\alpha, p) \right. \right\},$$
where
$$t(\alpha,p) = \sqrt{2\ln p} - \frac{\ln(4\pi \ln p)}{2\sqrt{2\ln p}} - \frac{\ln \alpha}{\sqrt{2\ln p}}$$
and α is the confidence level chosen by the user.
3. For each $i = 1, \ldots, p$, and $i \notin I_s$, compute
$$E_i = \frac{1}{|I_s|} \sum_{j \in I_s} \mathbf{C}_n(i,j)^2.$$
4. Denote by I_c the set of variables highly correlated to those in I_s:
$$I_c = \left\{ j \left| j \notin I_s, E_j > \frac{1}{n-1} \left(1 + \sqrt{2}\, t(\alpha,p)\right) \right. \right\}.$$
4. Compute the leading eigenvectors of the covariance matrix $\mathbf{S}_n$, restricted to the set $I = I_s \cup I_c$.

Let us briefly explain the theoretical motivation for this algorithm. First, the empirical variance of a noise variable is distributed as a χ^2_n/n random variable, which for large n can be approximated as $1 + \sqrt{2/n}\,\mathcal{N}(0,1)$. Hence, in Step 2 we choose variables that, with high probability, contain a significant signal contribution. In Step 3, for the remaining variables, we use the well-known fact (see, for example, Anderson (2003, section 4.2)), that under the null assumption that variables i and j are independent Gaussian variables, $\mathbf{C}_n(i, j)$ has density
$$f(r) = \frac{\Gamma(\frac{n-1}{2})}{\Gamma(\frac{n-2}{2})\sqrt{\pi}} (1 - r^2)^{(n-4)/2}$$
or, equivalently, $\mathbf{C}_n(i,j)/\sqrt{1 - \mathbf{C}_n(i,j)^2}$ follows a t-distribution with $n - 2$ degrees of freedom. In particular, $\mathbb{E}[\mathbf{C}_n(i, j)] = 1/(n-1)$, and $\text{var}[\mathbf{C}_n(i,j)] \approx 2/(n-1)^2$. Step 3 detects additional variables that, despite having relatively smaller variance, are significantly correlated to the signal variables already found, and hence are also signal variables with high probability.

In Table 1 we present the reconstruction errors and the number of features chosen by our suggested method with a confidence level $\alpha = 0.02$. Our procedure obtained smaller errors than the sparse PCA method for both signals, although for the step function, which is not so sparse, the covariance thresholding method of Bickel and Levina (2008), performed slightly better. A graphical comparison of the performance of the different algorithms using boxplots is shown in Figure 3.

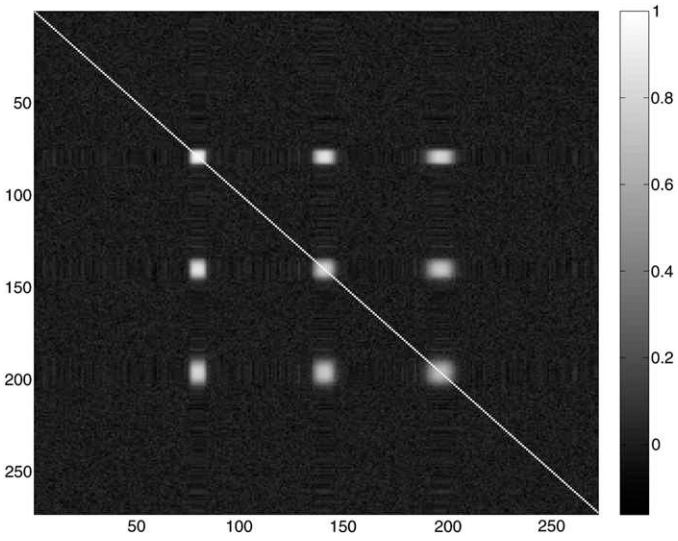

Figure 2. First 250 entries of the empirical correlation matrix of data from the three-peak function in the symmlet wavelet basis.

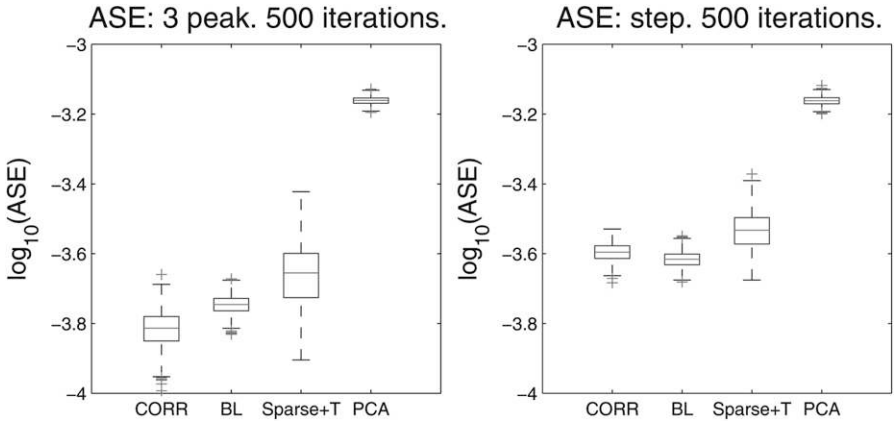

Figure 3. Empirical averaged root squared error of various algorithms.

The specific model suggested by Johnstone and Lu, of a factor model with a relatively small number of components with signals that are all sparse, and our (relatively simple) attempt to detect groups of correlated variables via the correlation matrix structure raises some interesting theoretical questions: For example, what are information limits for detection of sparse structures in a covariance matrix, and what are good algorithms to achieve them? In the presence of multiple signals, this problem relates to our ability to cluster the nodes of an adjacency graph. In this respect, we mention that in both computer science and in statistical physics, a similar problem has received a lot of attention—the so-called ''planted partition problem.'' Perhaps some connections between these results and statistical inference and sparsity should be further explored.

To conclude this section, we note that there are some possible advantages to analyzing the correlation matrix structure. It does not require an estimate of the noise level, and the procedure is more robust to heteroscedastic noise, which is uncorrelated but may have a different strength in different variables. The case of correlated noise requires further research beyond the scope of this discussion. Finally, we remark that in a different although related context, we recently used both the correlation and the covariance matrices to construct a multiscale representation of given data (see Lee, Nadler, and Wasserman 2008).

4. COMPUTATIONAL COMPLEXITY AND NUMBER OF SIGNALS

The second issue raised in the article by Johnstone and Lu is the computational savings of the thresholding procedure. In table 1 in Johnstone and Lu, they show that significant computational savings can be achieved by computing the eigenvalues and eigenvectors of a smaller covariance matrix. A few words regarding this issue are needed. According to the Matlab code supplied by Johnstone and Lu, regardless of the fact that only one eigenvector is of interest, all eigenvalues and eigenvectors of the relevant covariance matrix are computed. This step has computational complexity $O(p^3)$, where p is the size of the relevant covariance matrix. However, under the assumption that the data have an intrinsic low dimensionality—say, bounded above by a small number k—then for dimensionality reduction purposes, only the largest k eigenvalues and eigenvectors of the covariance matrix need to be computed. Furthermore, when these eigenvalues are isolated from the rest, these can be computed (iteratively) much faster than $O(p^3)$. In Matlab, this can be achieved via the function `eigs(A,k)` instead of the function `eig(A)`.

Computation of only a few of the eigenvalues and eigenvectors raises a different theoretical question: How does one determine what is the ''dimensionality'' of the data? That is, how many of the largest eigenvalues indeed correspond to signals and not to noise? In a recent article, Kritchman and Nadler (2008) developed an algorithm to solve this problem (by the way, using another notable result of Iain Johnstone, regarding the convergence of the largest noise eigenvalue to a Tracy-Widom distribution). A careful inspection of that algorithm shows that if one knows in advance that the dimensionality of the data is smaller than k, then only the k largest sample eigenvalues and the trace of the covariance matrix are required to estimate the true dimensionality of the data. Finally, the computational complexity of this algorithm is negligible with respect to the calculation of the eigenvalues themselves. After implementing these changes, all the procedures described in this article have similar running times. In Table 1, I report these CPU running times averaged over 500 iterations, as used by Matlab on an Intel Quad Core CPU Q6600 at 2.40 GHz (without multithreading).

[Received January 2009. Revised January 2009.]

REFERENCES

Anderson, T. W. (2003), *An Introduction to Multivariate Statistical Analysis* In (3rd ed.), New York: Wiley.

Bickel, P. J., and Levina, E. (2008), ''Covariance Regularizing by Thresholding,'' *The Annals of Statistics,* 36, 2577–2604.

Buckheit, J., and Donoho, D. L. (1995), ''Improved Linear Discrimination Using Time Frequency Dictionaries,'' *Proceedings of the Society for Photo-Instrumentation Engineers,* 2569, 540–551.

Kritchman, S., and Nadler, B. (2008), ''Determining the Number of Components in a Factor Model from Limited Noisy Data,'' *Chemometrics and Intelligent Laboratory Systems,* 94, 19–32.

Lee, A. B., Nadler, B., and Wasserman, L. (2008), ''Treelets: An Adaptive Multi-Scale Basis for Sparse Unordered Data,'' *Annals of Applied Statistics,* 2, 435–471.

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

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

Discussion

Daniela M. WITTEN, Trevor HASTIE, and Robert TIBSHIRANI

We congratulate the authors on an intriguing and timely article. There is a great deal of interest in sparsity and its applications to high-dimensional data analysis. In our comment, we would like to relate this work to research by ourselves (and others) on sparse principal components via the lasso.

Given an $n \times p$ data matrix $\mathbf{X}$, with centered columns, perhaps the most natural way to define a sparce principal component is

$$\text{maximize}_{\mathbf{v}}\{\mathbf{v}^T\mathbf{X}^T\mathbf{X}\mathbf{v}\} \text{ subject to } ||\mathbf{v}||_2^2 \le 1, ||\mathbf{v}||_0 \le k. \quad (1)$$

That is, we find the linear combination of variables that has highest variance, among those with at most k nonzero loadings. This is a difficult problem to solve computationally, because the criterion that must be optimized is not convex. There are two reasons that this is the case:

1. It involves *maximizing* the objective function $\mathbf{v}^T\mathbf{X}^T\mathbf{X}\mathbf{v}$, which is convex in $\mathbf{v}$. Convex optimization provides tools for *minimization* of convex functions.
2. The L_0 bound on $\mathbf{v}$ is not convex.

Johnstone and Lu's sparse principal components method can be thought of as an approximation to (1). They define a sparse principal component as

$$\text{maximize}_{\mathbf{v}}\{\mathbf{v}^T\mathbf{X}^T\mathbf{X}\mathbf{v}\} \text{ subject to } ||\mathbf{v}||_2^2 \le 1, 1_{v_j \neq 0} = 1_{\hat{\sigma}_j > \hat{\sigma}_{(p-k)}}, \quad (2)$$

where $\hat{\sigma}_j$ is the variance of column j of $\mathbf{X}$. That is, they perform principal components analysis (PCA) on the k columns of $\mathbf{X}$ with highest variance. Of course, this problem can be solved easily, as it amounts simply to an eigen decomposition of a $k \times k$ submatrix of $\mathbf{X}^T\mathbf{X}$.

An alternative approach to finding sparse principal components is proposed in Jolliffe, Trendafilov, and Uddin (2003); this method also stems from an approximation to the criterion (1). The *SCoTLASS* proposal involves replacing the L_0 bound in (1) with an L_1 bound on the elements of $\mathbf{v}$, as follows:

$$\text{maximize}_{\mathbf{v}}\{\mathbf{v}^T\mathbf{X}^T\mathbf{X}\mathbf{v}\} \text{ subject to } ||\mathbf{v}||_2^2 \le 1, ||\mathbf{v}||_1 \le k. \quad (3)$$

Unfortunately, this problem is still not convex (because it involves maximizing a convex objective function) and computations are difficult.

In our recent work on "penalized matrix decompositions" (Witten, Tibshirani, and Hastie 2009), we presented a general class of algorithms that yields new procedures for obtaining sparse matrix decompositions, sparse principal components, and sparse canonical variates. Our procedure reexpresses (3) as

$$\text{maximize}_{\mathbf{u},\mathbf{v}}\{\mathbf{u}^T\mathbf{X}\mathbf{v}\} \text{ subject to } ||\mathbf{v}||_1 \le k, ||\mathbf{u}||_2^2 \le 1, ||\mathbf{v}||_2^2 \le 1, \quad (4)$$

and we show that the solution $\hat{\mathbf{v}}$ is the same for both problems. We call problem (4) "SPC," for sparse principal components. We also show that the sparse principal components criterion of Zou, Hastie, and Tibshirani (2006) is equivalent to this criterion, if a natural symmetric constraint is added to their criterion.

Problem (4) is biconvex, and the following simple alternating algorithm can be used to solve it.

Algorithm: sparse principal components (SPC):

1. Initialize $\mathbf{v}$ to have L_2 norm 1.
2. Iterate until convergence:
   (a) $\mathbf{u} \leftarrow \frac{\mathbf{X}\mathbf{v}}{||\mathbf{X}\mathbf{v}||_2}$.
   (b) $\mathbf{v} \leftarrow \frac{S(\mathbf{X}^T\mathbf{u},\Delta)}{||S(\mathbf{X}^T\mathbf{u},\Delta)||_2}$, where $\Delta = 0$ if this results in $\|\mathbf{v}\|_1 \le k$; otherwise, Δ is chosen to be a positive constant such that $\|\mathbf{v}\|_1 = k$. Here, S is the soft-thresholding operator, defined as $S(a, c) = \text{sgn}(a)(|a| - c)_+$.

Further components are found by taking residuals and applying the algorithm again. That is, if the solutions are $\mathbf{u}_1$ and $\mathbf{v}_1$, and $d_1 = \mathbf{u}_1^T\mathbf{X}\mathbf{v}_1$, then the residuals are $\mathbf{X} - \mathbf{u}_1 d_1 \mathbf{v}_1^T$.

In Johnstone and Lu's procedure, a common set of k features is used for all components. It would seem that a potential advantage of the SPC approach is the fact that different features can be used for different components. We examine this issue in a small simulation example, in which a matrix $\mathbf{X}$ of dimension 100×200 is generated as follows:

$$\mathbf{X} = \mathbf{u}_1\mathbf{v}_1^T + \mathbf{u}_2\mathbf{v}_2^T + \mathbf{Z}. \quad (5)$$

Here, $\mathbf{u}_1$ and $\mathbf{u}_2$ are vectors of normal random variables in $\mathbb{R}^{100}$. The $\mathbf{v}_1$ and $\mathbf{v}_2$ are vectors in $\mathbb{R}^{200}$, each with 50 nonzero (and nonoverlapping) coefficients. $\mathbf{Z}$ is a matrix of normal random variables, and $\mathbf{v}_1$ and $\mathbf{v}_2$ are shown in Figure 1. (Of course, this model is in clear violation of Johnstone and Lu's single component model.) Johnstone and Lu's method and our SPC method both involve tuning parameters that determine the number of nonzero elements of the sparse principal components obtained. We compute the first two sparse principal components for each method. For a given number of nonzero

Daciela M. Witten is Ph.D. student, Department of Statistics, Stanford University, Stanford, CA 94305. Trevor Hastie is Professor, Department of Statistics and Department of Health Research and Policy, Stanford University, Stanford, CA 94305. Robert Tibshirani is Professor, Department of Health Research and Policy and Department of Statistics, Stanford University, Stanford, CA 94305 (E-mail: *tibs@stat.stanford.edu*). Daniela M. Witten was supported by a National Defense Science and Engineering Graduate Fellowship. Trevor Hastie was partially supported by National Science Foundation Grant DMS-0505676 and National Institutes of Health Grant 2R01 CA 72028-07. Robert Tibshirani was partially supported by National Science Foundation Grant DMS-9971405 and National Institutes of Health Contract N01-HV-28183.

Journal of the American Statistical Association
June 2009, Vol. 104, No. 486
DOI 10.1198/jasa.2009.0143

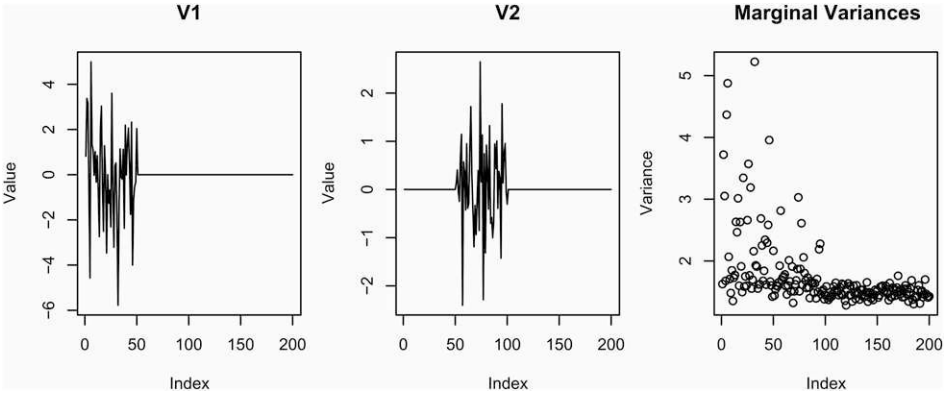

Figure 1. The data were generated using a rank-2 model, plus Gaussian noise. In the model, $\mathbf{v}_1$ and $\mathbf{v}_2$ each have 50 nonzero elements; these elements are nonoverlapping. The left and center panels show $\mathbf{v}_1$ and $\mathbf{v}_2$, and the third shows the observed marginal variances of the $p = 200$ variables.

elements for each sparse principal component, we also determine the number of ''false positives''(i.e., the number of nonzero elements of $\hat{\mathbf{v}}_i$ that are zero in $\mathbf{v}_i$). The number of nonzero elements of $\hat{\mathbf{v}}_i$ is plotted against the number of false positives in Figure 2 for both methods. Our SPC method results in fewer false positives for both the first and second components, because the screening step in Johnstone and Lu's method selects variables based on their marginal variances, which in this case is not necessarily indicative of whether they are nonzero in $\mathbf{v}_i$. The third panel of Figure 1 shows the marginal variances of the variables.

Figure 3 shows the first and second sparse principal components estimated using the SPC proposal and Johnstone and Lu's method. (For both methods, tuning parameters were chosen to yield an average of 50 nonzero elements in each sparse principal component.) It turns out that in this example, although Johnstone and Lu's method results in many false positives, these nonzero elements of $\hat{\mathbf{v}}_i$ tend to be extremely small. In fact, all of these false positives would be ''filtered out'' using the thresholding step proposed in the algorithm in Section 4 of Johnstone and Lu's article. These small false positives are perhaps the reason that the extra filtering step is proposed. We note also that this example was simple enough that even hard-thresholding of ordinary principal components works well. It is not difficult to imagine a noisier example in which this might not be the case.

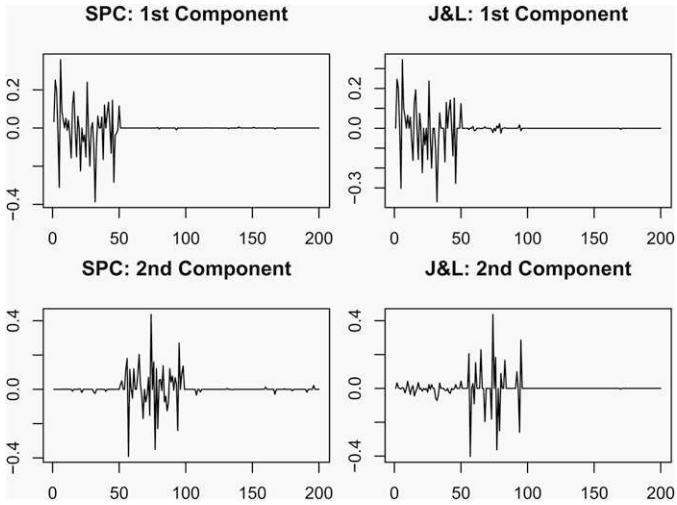

Figure 3. The estimated sparse principal components obtained using the Johnstone and Lu and SPC methods are shown. The false positive elements resulting from the Johnstone and Lu method tend to be quite small in absolute value.

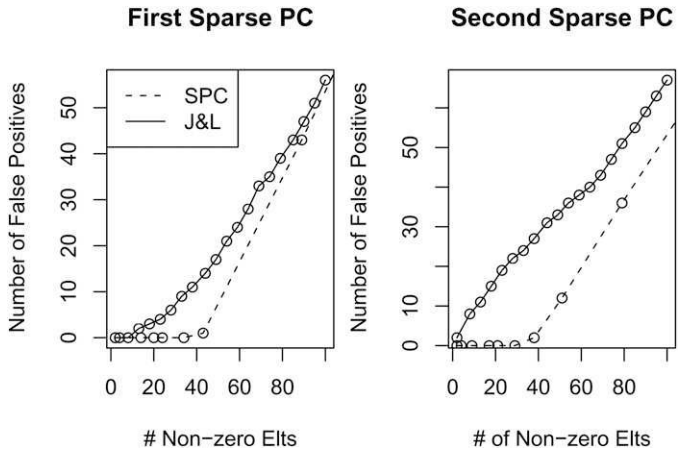

Figure 2. The SPC method results in fewer false positives than the Johnstone and Lu method.

Johnstone and Lu present a theoretical framework in which they prove asymptotic consistency of their sparse principal component approach. This certainly is a very attractive feature of their method. We wonder if the authors' consistency results could be extended to the solutions from the SPC method.

[Received January 2009. Revised January 2009.]

REFERENCES

Jolliffe, I., Trendafilov, N., and Uddin, M. (2003), ''A Modified Principal Component Technique Based on the Lasso,'' *Journal of Computational and Graphical Statistics,* 12, 531–547.

Witten, D., Tibshirani, R., and Hastie, T. (2009), ''A Penalized Matrix Decomposition, with Applications to Sparse Principal Components and Canonical Correlation Analysis,'' *Biostatistics,* to appear.

Zou, H., Hastie, T., and Tibshirani, R. (2006), ''Sparse Principal Component Analysis,'' *Journal of Computational and Graphical Statistics,* 15, 265–286.

Comments

James O. RAMSAY

This splendid paper provides an irresistible opportunity to comment on several issues connected with the analysis of functional data or the use of functional models. The most important of these is, of course, to congratulate the authors warmly on beginning a new chapter on the analysis of data characterized by sharply localized events separated by long, smooth gaps. Wavelets are the obvious choice among bases for this problem, and who is better positioned than they are to put these fascinating little wigglers to work on this, the principal components analysis, problem? Getting along without wavelets is like an orchestra playing music without a percussion section.

In the design stages of our work on functional data (Ramsay and Silverman 2002, 2005), we reflected on giving wavelets more than passing mention as basis systems, but they were still a comparatively exotic topic during the early '90s and we had already committed ourselves to optimizing readability and to a general focus on the estimation of smooth variation. The article by Johnstone and Lu effectively highlights the contrast between smooth and local variation in functional data, and it was gratifying to see that smooth principal components analysis did such a fine job of showing itself to be the wrong tool.

But are smoothness and sparsity really such different concepts? Functions can be "small" in two quite orthogonal ways—most obviously in terms of their amplitude variation, but perhaps less clearly to those less familiar with higher analysis, in terms of the dimension of a function space required to represent the signal adequately. The latter concept, function complexity, is often expressed in terms of the rate of decay of an infinite series of smooth basis functions, but wavelets put the matter up front where nobody can miss the point.

The high-dimensional modeling context, referred to often as $p >> n$, is receiving a lot of attention these days, but in the end our strategy does not seem to change much. We seek a model structure that is usefully identified by the available data. What has changed is the range of options for model structures, with the possibility of defining simplicity in new ways, such as the number and complexity of a set of nonlinear differential equations with solutions that exhibit catastrophic or chaotic variation, as deviations from the quantile function for the uniform distribution, or as the number lines of computer code required to display data structure. Combining wavelets with principal components analysis is the ideal strategy for the exploration of sparse variation, and both concepts are deeply rooted in the traditions of applied mathematics and statistics.

Where do we go from here? The focus of the article by Johnstone and Lu is rather exclusively on variation defined by a single principal component, and maybe this problem has rather more to do with data smoothing than it has to do with classic principal components analysis. For example, would it be better to treat the scores on this single component as multiplicative random effects, so that their variances have parametric status rather than being just handy indices for defining hard thresholds?

Substantial contributions from even a few additional principal components makes the problem much tougher by deemphasizing eigenvectors or eigenfunctions, and calling for coordinate-free methods for assessing subspace identification. I hope to see a flurry of new papers extending this work in this direction, as well as in the direction of functional linear modeling and dynamic systems identification.

The bipolar spikes in the right panels of the displays of ECG variation are, in my experience, the signature of phase variation, and seem most likely to be the result of slight misalignments of the R wave, perhaps related to the use of dyadic sampling point sequences. It seems intuitive that sparse signals of this nature will routinely show horizontal as well as vertical displacements, and that registration methodology will wind up being as important for the analysis of sparsity as it has been for smoothness.

[Received January 2009. Revised January 2009.]

REFERENCES

Ramsay, J. O., and Silverman, B. W. (2002), *Applied Functional Data Analysis,* New York: Springer.

Ramsay, J. O., and Silverman, B. W. (2005), *Functional Data Analysis* (2nd ed.), New York: Springer.

James O. Ramsay is Professor, Department of Psychology, McGill University, 2748 Howe St., Ottawa, Ontario, K2B 6W9, Canada. (E-mail: *ramsay@psych.mcgill.ca*)

Journal of the American Statistical Association
June 2009, Vol. 104, No. 486
DOI 10.1198/jasa.2009.0139

Rejoinder

Iain M. JOHNSTONE and Arthur Yu LU

We thank the editors for inviting this discussion and especially the discussants for their kind remarks and thoughtful contributions, which in some cases are so thorough as to verge on articles in their own right!

Before turning to specific comments, we set the stage by noting that this article is a reflection of the Ph.D. dissertation of Lu, awarded in 2002. Its chief, original purpose was to provide the first rigorous proof of the inconsistency phenomenon, reviewed in section 2 of our article, and to propose a simple algorithm that demonstrates that exploitation of sparsity allows one to recover consistency. There have been many subsequent developments during the years that this article has been in publication review. In particular, as the various references and the current discussants show, there are now several algorithms and theoretical results that may improve in various respects on the original proposal.

For example, already in his Ph.D. dissertation, Paul (2005) recognized that selection of variables based on sample variances alone was inefficient and could only detect components of size $(\log(n \vee p)/n)^{1/4}$. He developed a two-stage estimator, using information about the correlation between the variables to augment the selected variable set at the second stage, and showed that this successfully selected components of size down to $O(\log(n \vee p)/n)^{1/2}$. Boaz Nadler's discussion independently makes a somewhat similar remark about exploiting correlation information, and develops an alternate estimator with properties that appear attractive and worthy of further study. In a technical tour de force, Paul (2005) also developed lower bounds and studied the asymptotic properties of his two-stage estimator to show that it achieves (within logarithmic factors) optimal rates of convergence.

Nadler also points to the good performance, on our two examples, of the covariance matrix thresholding estimator of Bickel and Levina (2008). Bickel and Levina (2008) were careful to establish convergence properties of their estimator in operator norm, which implies convergence of eigenvectors, and so the attractive performance of the derived eigenvector estimator is, in retrospect (!), not so surprising. One should caution that further investigation is needed, because our two examples are admittedly at relatively high signal-to-noise ratios. In a recent preprint in the related setting of banded covariance estimation, Cai, Zhang, and Zhou (2008) have shown that threshold choice based on a function of $(\log p)/n$ need not be optimal, at least over specific function classes.

Nadler provides an interesting analysis of reconstruction error, using what one might call a "projection oracle"—namely, an estimator that knows the best subset of k variables to retain. The discussion and his figure 1 suggest that (i) reconstruction error is highly sensitive to incorrect inclusion of noise variables and that (ii) there may be considerable room for reconstruction error improvement. We comment on these in turn.

Avoidance of the first point was the reason for inclusion of thresholding as a final stage in our algorithm. No theoretical analysis of thresholding was included in the article, so we attempt some heuristics here. To continue with the notation used by Nadler, suppose that $\boldsymbol{\rho}_k$ is the response vector $\boldsymbol{\rho}$ restricted to the optimal k variables and $\hat{\boldsymbol{\rho}}_k$ is the eigenvector of sample principal component analysis (PCA) computed only on these k chosen variables. In this relatively high signal-to-noise setting, we may heuristically imagine $\hat{\boldsymbol{\rho}}_k$ as approximately derived from a Gaussian signal plus noise model $\hat{\boldsymbol{\rho}}_k = \boldsymbol{\rho}_k + \epsilon \boldsymbol{z}_k$ with $z_k \sim N_k(0, \boldsymbol{I})$ and $\epsilon^2 = 1/n$ (some support for this heuristic can be found in Paul's (2005) dissertation).

Consistent with the proposal in section 4.1 of our original article, let $\tilde{\boldsymbol{\rho}}_k$ denote the result of hard thresholding at $\epsilon\lambda_k$, where $\lambda_k = \sqrt{2 \log k}$. Bounds for the risk of hard thresholding

$$\mathbb{E} \parallel \tilde{\boldsymbol{\rho}}_k - \boldsymbol{\rho}_k \parallel^2 \ \le R(\tilde{\boldsymbol{\rho}}_k, \boldsymbol{\rho}_k),$$

may be obtained, for example, from Johnstone (2002, lemma 11.5), and yield

$$R(\tilde{\boldsymbol{\rho}}_k, \boldsymbol{\rho}_k) \le \epsilon^2 \sum_j \bar{r}_H(\rho_{k,j}/\epsilon, \lambda_k),$$

where

$$\bar{r}_H(\mu, \lambda) = \begin{cases} (6/5)[2\lambda\phi(\lambda) + 2\tilde{\Phi}(\lambda) + \mu^2] & \text{if} \quad \mu \le \lambda, \\ 1 + \mu^2\tilde{\Phi}(\mu - \lambda) & \text{if} \quad \mu > \lambda, \end{cases}$$

and $\Phi = 1 - \tilde{\Phi}$ is the standard Gaussian cumulative distribution function. Figure 1 shows plots of

$$\widetilde{ASE}_k = \frac{1}{p}\left(\parallel \boldsymbol{\rho}_k^{\perp} \parallel^2 + R(\tilde{\boldsymbol{\rho}}_k, \boldsymbol{\rho}_k)\right)^{1/2}$$

for the step and three-peak functions, superimposed on Nadler's graphs. From these plots, it is apparent that thresholding essentially removes the high sensitivity of reconstruction errors to incorrect inclusion of noise variables.

A general remark is that one should be cautious using oracles as indicators of the amount of improvement that may be possible. There is a price to be paid for selecting variables based on data. This is partially visible in the plot of $\widetilde{ASE}_k$ for the three-peak function, which has a minimum of 1.3×10^{-4}, already 20% larger than for the minimum projection oracle, and this still assumes the optimal set of k variables is known. It may be that it is not possible in these two cases to improve much on CORR (for three-peak) and Bickel and Levina (2008) (for step).

Nadler remarks on the use of special-purpose routines, such as MATLAB's `eigs` to compute only the largest eigenvalues

Iain M. Johnstone is Professor of Statistics and Biostatistics, Stanford University, Stanford, CA 94305. Arthur Yu Lu is Principal at Renaissance Technologies Corp., East Setauket, NY 11733. Work on this rejoinder supported in part by NIH EB R01 EB001988. We are grateful to Zongming Ma for his continuing work on the ASPCALab package, for his remarks on this rejoinder, and for allowing us to include details of his forthcoming work.

Journal of the American Statistical Association
June 2009, Vol. 104, No. 486
DOI 10.1198/jasa.2009.0158

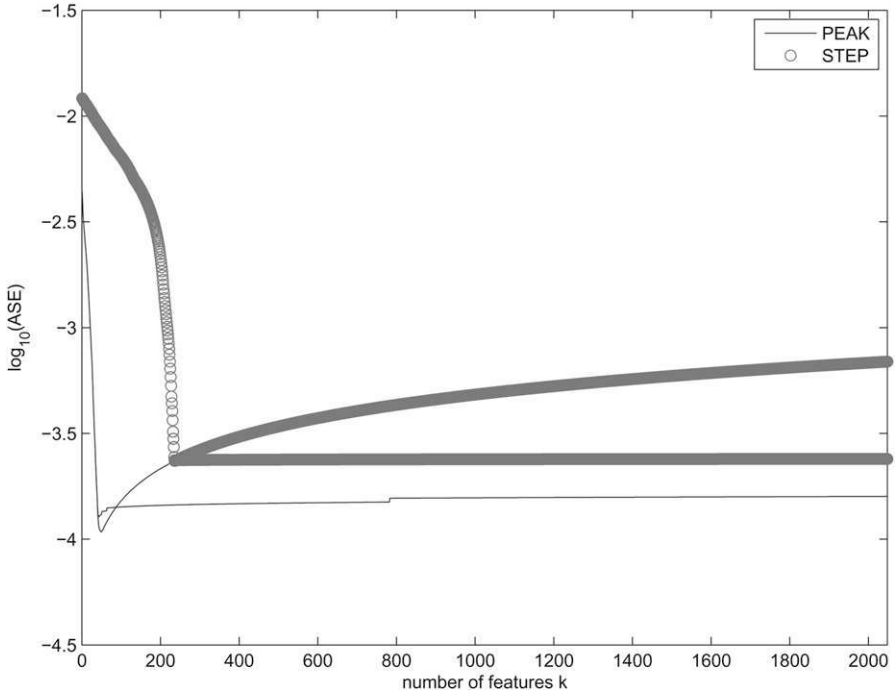

Figure 1. Reconstruction error based on thresholding bound.

and eigenvectors of the covariance matrix. The help page (at *www.mathworks.com*) for `eigs` says "`eigs` is not a substitute for [running `eig` and selecting largest eigenvectors] but is most appropriate for large sparse matrices," which is not necessarily the case for a noisy covariance matrix, especially in the lower signal-to-noise settings. Indeed, the `eigs` algorithm is based on a Lanczos method, with a convergence rate that depends on a function of the ratio of the largest to second largest eigenvalue (more precisely, it is based on an implicitly restarted Arnoldi method, (Lehoucq and Sorensen 1996, Morgan 1996)). In the three-peak and step examples of our original article, this ratio is deliberately chosen to be large, but there is no reason for this to be the case in general. We have included both methods as options in the ASPCALab package, and users can let their circumstances determine the choice.

Witten, Hastie, and Tibshirani give a simulation example to show that their attractive penalized matrix decomposition (PMD) method selects fewer false positives than our approach, at least prior to thresholding. As they note, the thresholding step in our method is indeed designed to "clean up" the final estimate, and their figure 3 suggests that the final results may not be so different. Indeed, a comparison of reconstruction errors would be a natural next step.

Witten, Hastie, and Tibshirani also ask if there may be some theoretical analysis possible for penalized multivariate methods. We can report on some related work in progress by Ma (2009). Consider the single component model and let $\boldsymbol{S}$ denote the sample covariance matrix. As in our original article, we suppose that the data have been converted to an appropriate transform domain. Instead of the optimization problem (0.3) proposed in the discussion by Witten, Hastie, and Tibshirani, one can consider the Lagrangian form

$$\max_{v} \boldsymbol{v}^T \boldsymbol{S}\boldsymbol{v} - \lambda \parallel \boldsymbol{v} \parallel_1, \text{subject to } \parallel \boldsymbol{v} \parallel_2 \ \leq 1.$$

This problem may be solved by the following algorithm, which is apparently simpler than their Sparse Principal Components (SPC).

Algorithm: Iterative Soft-Thresholding PCA (IS-PCA).

1. Initialize $\boldsymbol{v}$ with a unit 2-norm vector $\boldsymbol{v}_0$.
2. Iterate until convergence:
   (a) $\boldsymbol{v} \leftarrow \eta_S(\boldsymbol{S}\boldsymbol{v}, \lambda/2)$.
   (b) $\boldsymbol{v} \leftarrow \boldsymbol{v}/ \parallel \boldsymbol{v} \parallel_2$.

Here, $\eta_S(x, c) = \text{sgn}(x)(|x| - c)_+$ is the soft-thresholding function.

The two tuning parameters in the previous algorithm (also present in SPC) are the penalty parameter λ and the initial value $\boldsymbol{v}_0$. Both are important to the properties of the final estimator. We propose initializing $\boldsymbol{v}_0$ with the Johnstone and Lu estimator, and picking λ as $O(\log(n \vee p)/n)^{1/2}$. With such choices, the output of the IS-PCA algorithm can be shown to be not only consistent, but also rate optimal (within a logarithmic factor).

Methods for Choice of k. Section 4.2 in our original article proposed two methods for data-based selection of k using the sample variances, which one might loosely call the "alpha" and "percent" criteria respectively. They are available as `ASPCAalp.m` and `ASPCA.m` respectively in the MATLAB package described at the end of section 4.1. The "percent" method was used for the computations and figures, whereas the "alpha" method was used for the proofs.

In response to thoughtful correspondence from Boaz Nadler and to the remarks of Witten, Hastie, and Tibshirani, we did a comparison of the two approaches on our test functions. To save space, we give a summary of the conclusions here. Further details are available on request.

For both methods, the larger the feature set chosen (larger percent or smaller alpha), the better the average estimation

error. The computation time increases, but not greatly, and certainly sublinearly in the number of features chosen. The thresholding step, of course, has a progressively more significant effect as the number of features grows.

The percent method is considerably more variable (over replications) in the number of features selected. Because of thresholding, this does not have bad consequences for the final result. However, if this is a concern, then the alpha method is much more stable, with a coefficient of variation for k that is generally less than 4% in our examples.

We agree with Ramsay that one should not oppose smoothness and sparsity, and that it is often more helpful to regard sparsity as an extension of smoothness (which may be thought of as sparsity when concentrated at low frequencies).

The exposition of our original article focused on a single component model largely for simplicity in the proofs and displays of figures. As noted in sections 2 and 3.2, perhaps with insufficient emphasis, the method, results, and proofs do extend to additional principal components, at least with distinct eigenvalues. We do agree with Ramsay that some additional issues, presumably not insurmountable, arise when several eigenvalues coalesce and the issue becomes one of subspace identification.

In connection with the ECG example, Ramsay raises some stimulating questions regarding registration. We attempted to deal with the main additive effect of horizontal displacement by anchoring the maximum of the R wave at the 150th position in each cycle. Stimulated by Ramsay, one might then ask whether the faster rise/slower fall nature of the first mode of variation (especially for sample 1) would be expressed differently if a curve-specific warping were introduced, along the lines of Chapters 7 in the two Ramsay and Silverman books. This is indeed an interesting question for further analysis, along with the evident identifiability issues it raises.

[Received January 2009. Revised January 2009.]

REFERENCES

Bickel, P. J., and Levina, E. (2008), ''Covariance Regularization by Thresholding,'' *The Annals of Statistics*, 36, 2577–2604.

Cai, T. T., Zhang, C.- H., and Zhou, H. H. (2008), ''Optimal Rates of Convergence for Covariance Matrix Estimation,'' Technical Report, University of Pennsylvania.

Johnstone, I. M. (2002), *Function Estimation and Gaussian Sequence Models*, book manuscript available at *www-stat.stanford.edu/~imj*.

Lehoucq, R. B., and Sorensen, D. C. (1996), ''Deflation Techniques for an Implicitly Restarted Arnoldi Iteration,'' *SIAM Journal on Matrix Analysis and Applications*, 17, 789–821.

Ma, Z. (2009), ''IS-PCA: An Iterative Soft-Thresholding Approach to Sparse Principal Component Analysis.'' Manuscript in preparation.

Morgan, R. B. (1996), ''On Restarting the Arnoldi Method for Large Nonsymmetric Eigenvalue Problems,'' *Mathmatics of Computation,* 65, 1213–1230.

Paul, D. (2005), ''Nonparametric Estimation of Principal Components,'' Ph.D. dissertation, Stanford University, Dept. of Statistics.