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

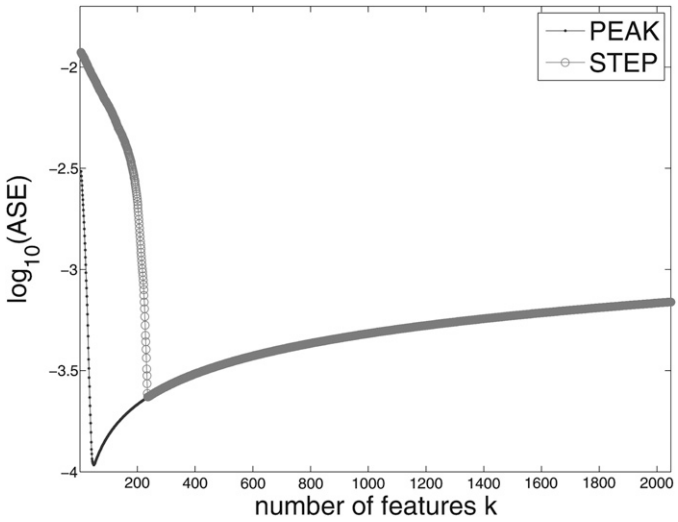

Figure 1. The theoretically optimal reconstruction error versus the number of chosen features, Equation (3).

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.

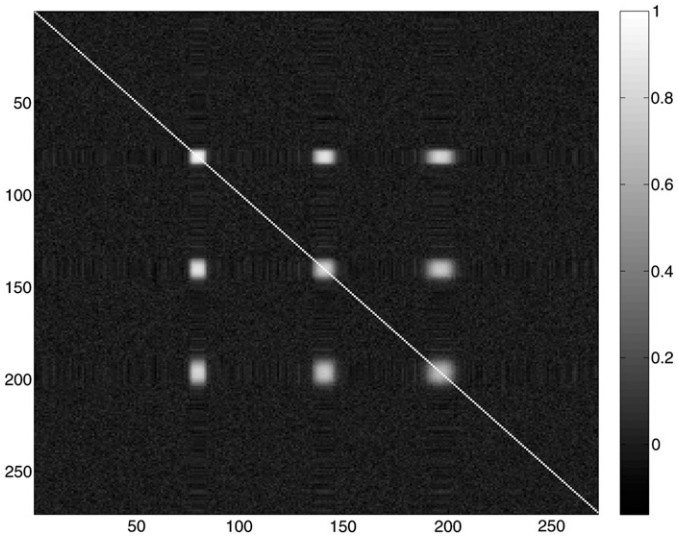

Figure 2. First 250 entries of the empirical correlation matrix of data from the three-peak function in the symmlet wavelet basis.

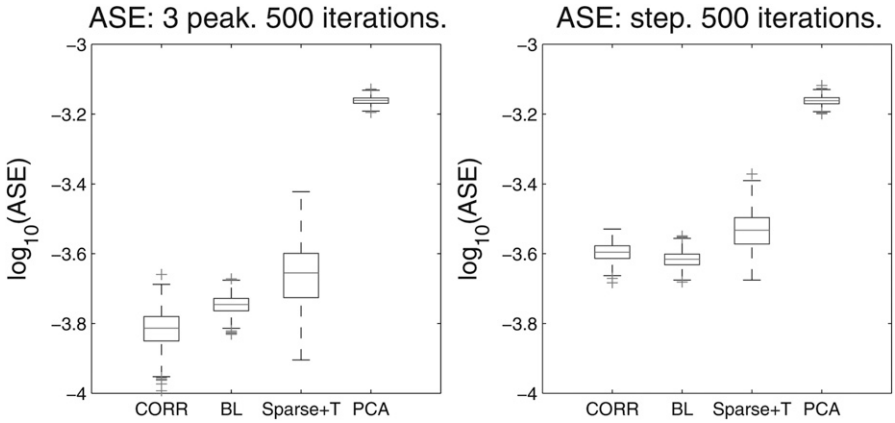

Figure 3. Empirical averaged root squared error of various algorithms.

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.