Statistica Sinica **19** (2009), 1741-1754

AN EIGENVECTOR VARIABILITY PLOT

I-Ping Tu, Hung Chen and Xin Chen

Academia Sinica, Taiwan, National Taiwan University and UC San Francisco

Abstract: Principal components analysis is perhaps the most widely used method for exploring multivariate data. In this paper, we propose a variability plot composed of measures on the stability of each eigenvector over samples as a data exploration tool. We also show that this variability measure gives a good measure on the intersample variability of eigenvectors through asymptotic analysis. For distinct eigenvalues, the asymptotic behavior of this variability measure is comparable to the size of the asymptotic covariance of the eigenvector in Anderson (1963). Applying this method to a gene expression data set for a gastric cancer study, many hills on the proposed variability plot are observed. We are able to show that each hill groups a set of multiple eigenvalues. When the intersample variability of eigenvectors is considered, the cutoff point on informative eigenvectors should not be on the top of the hill as suggested by the proposed variability plot. We also try the proposed method on functional data analysis through a simulated data set with dimension p greater than sample size n. The proposed variability plot is successful at distinguishing the signal components, noise components and zero eigenvalue components.

Key words and phrases: Dimension reduction, eigenvector, functional data analysis, principal components analysis, resampling, the multiplicity of eigenvalues, the scree plot.

1. Introduction

During the last two decades, analyzing huge data sets has become common to scientists in many fields. Thus one sees microarray data in gene expression experiments, 3-D microscopy reconstruction of macromolecules, and MRI image analysis in brain neural network studies, for example. Among many statistical approaches for huge data sets, principal components analysis (PCA) is often used as the key step for data dimension reduction (Jolliffe (2002)). Often, examination of the chosen components may lead to insight into underlying factors, in the microarray experiments in Raychaudhuri, Stuart, and Altman (2000), for instance.

The *scree* plot (Cattell (1966)) is designed as a visualization tool to search for an elbow or an inflection point on the sorted eigenvalue curve to choose the signal components. When the obvious distinct eigenvalue components do not explain enough variance, a specified percentage of total variance is recruited as an additional criterion to strip off redundant information. However, the specified percentage may not be convincing when the difference between the least eigenvalue of the chosen components and the largest eigenvalue of the unchosen components is not significantly different; for example, see Figure 1 in Section 2 and Mori, Y., Selaru, F. M., Sato, F., Yin, J., Simms, L. A., Xu, Y., Olaru, A., Deacu, E., Wang, S., Taylor, J. M., Young, J., Leggett, B., Jass, J. R., Abraham, J. M., Shibata, D. and Meltzer, S. J. (2003).

In this paper, we propose the stability of eigenvectors as a figure of merit to help make the choice of components. The proposed measure offers a clue as to whether the instability of eigenvectors over different studies can be attributed to sample variability. Thus, in North, Bell, Cahalan and Moeng (1982), one reads

> “That is, a particular sample will lead to one linear combination and another sample may pick out a drastically different linear combination of the nearby eigenvectors. The result is widely differing patterns from one sample to the next.”

To assess stability over sample variability, it is natural to use data resampling, for example the bootstrap (Efron (1979)) as in Beran and Srivastava (1985), to reveal the variability of eigenvectors.

There exists many works on the stability of eigenvectors. Krzanowski (1984) evaluated the robustness of the eigenvector to outliers by adding a user-determined perturbation to the eigenvalues. Sinha and Buchnan (1995) proposed an empirical rule to predict the stability measure as a function of eigenvalues, the dimension p, and the sample size n, after extensive simulation to decide which principal components were stable. Daudin, Dubey and Trecourt (1988) and Besse (1992) used a so-called risk estimate, the number of principal components minus the measure defined in Daudin et al. (1988), and inserted those risk estimates into the scree plots as a graphical device to choose the stable components.

This paper is organized as follows. In Section 2, we describe the proposed stability measure, and we demonstrate it on gastric cancer gene expression data. In Section 3, we present a large sample justification on the proposed measure; proofs are deferred to the Appendix. In Section 4, we bring in simulations to demonstrate the usefulness of the proposed stability measure for both large, $n = 1,000$ and small, $n = 25$, sample size. Furthermore, for $n = 25$, we allow $p = 25$ and 50. The proposed method performs well in all these cases. The paper ends with discussion.

2. Proposed Stability Measure

We employ the bootstrap resampling method to create variations of the data set. A gene expression data set is used to demonstrate the general characteristics

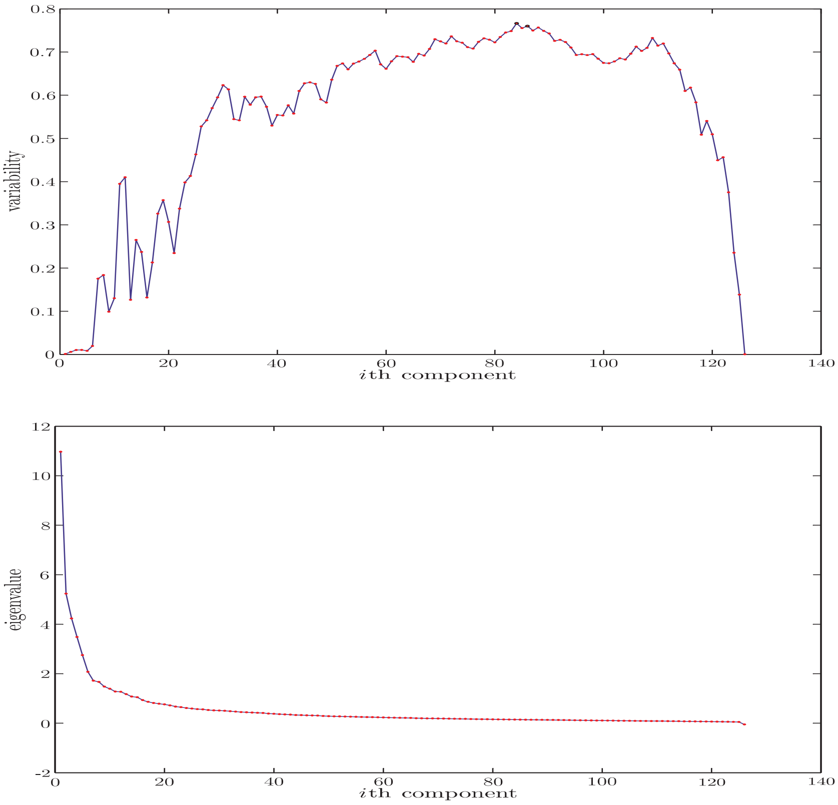

Figure 1. Variability plot and scree plot for gene expression data set: PCA is applied on the covariance matrix. This plot shows the possible multiplicity pairs among the first 20 components: (7, 8), (11, 12), (14, 15), (18, 19) and (19, 20).

of this plot that are further explained by an asymptotic analysis in Section 3. Let $\mathbf{X} = (x_{ij})$ denote the $n \times p$ data matrix of n observations in p variables and let $\mathbf{S}_n$ be the sample covariance matrix. We set

$$\bar{A}_k = 1 - \frac{1}{B} \sum_{b=1}^{B} |\mathbf{e}_k \cdot \mathbf{e}_k^{*b}|,$$

where $\mathbf{e}_k$ is the kth eigenvector of $\mathbf{S}_n$, and $\mathbf{e}_k^{*b}$ is the kth eigenvector of the bth resample covariance matrix. The number B needs to be large enough such that $\bar{A}_k$ is close to the population mean $1 - E|\mathbf{e}_k \cdot \mathbf{e}_k^{*}|$ under bootstrap resampling, $B = 600$ appeared to suffice here. A scatter plot of $(k, \bar{A}_k)$ is generated as a visualization tool for helping choose the components, henceforth, referred to as the variability plot.

Table 1. Comparison of Chosen Number of Components for the Gene Expression Data Set. The proposed variability plot suggests to look into four batches of components since they will remain stable using the sampling variability criteria. They are the first 10, the first thirteen, the first sixteen, and the first twenty one.

Methods	Suggested Significant PC numbers			
scree plot	somewhere between 15 to 30			
Cumulated Percentage Cuts	13(60%)	26(70%)	32(80%)	53(90%)
Variability Plot-cov	10	13	16	21

An application of the variability plot on a gene expression data set is presented in Figure 1. It is common to use array sample as the variable when applying PCA on gene expression data (Alter, Brown and Botstein (2000), Mori et al. (2003) and Raychaudhuri et al. (2000)). Such an analysis creates a set of "principal array components" that indicate the experimental conditions or sample characteristics which best explain the gene behaviors they elicit (Raychaudhuri et al. (2000)). Throughout this paper, array samples are used as the variables in performing PCA.

The gene expression data set is for a study on gastric cancer in Leung, Chen, Chu, Yuen, Mathy, Ji, Chan, Li, Law, Troyanskaya, Tu, Wong, So, Botstein and Brown (2002). The authors profiled a total of $p = 126$ mRNA populations from 90 primary gastric adenocarcinomas, 14 lymph node metastases, and 22 samples of nonneoplastic gastric mucosa; $n = 6,688$ clones were extracted by preprocessing filtration. The first six eigenvalues are obviously distinct from all other eigenvalues on the scree plot which accumulate less than 40% of total variance. One usual approach is to identify an elbow in the scree plot from the right, thus, the choice of the first 15 to 30 components is reasonable.

The variability plots describe the variability level for each eigenvector. The top plot of Figure 1 shows strong stability for the first six eigenvectors where the $\bar{A}_k$ is close to 0. The level of variability increases as the eigenvalues aggregate more seriously. Of most interest on the variability plot are the four hills between the highly stable eigenvectors (the first six) and the highly varied eigenvectors after component 21. Each hill represents a group of non-distinct eigenvalues which are confirmed by Bartlett's test of sphericity (Bartlett (1950)). The p-values are shown in Figure 2.

Thus, we suggest not choosing the eigenvectors up to the point where there is a high point on a hill in the variability plot. Table 1 shows the numbers of chosen components by the percentage of total variance criteria and by the variability plot. We also list the accumulated percentages of total variance accordingly.

3. Asymptotic Results

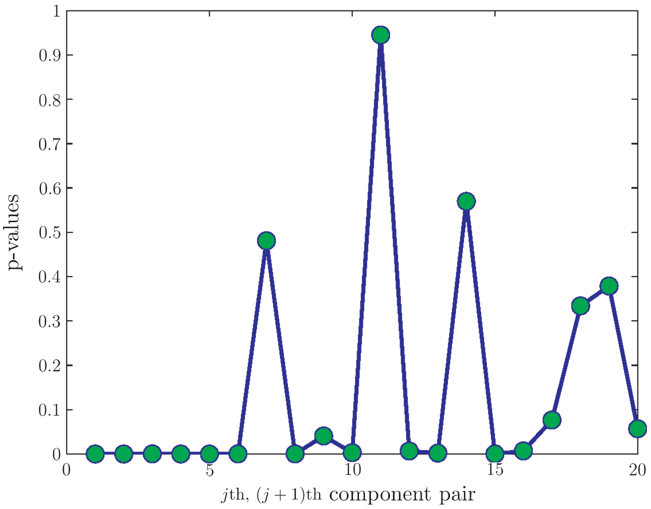

Figure 2. Marginal p-values for Bartlett's test for sphericity applied to test the first 20 (jth, $(j+1)$th) pairs under the null hypothesis $\lambda_j = \lambda_{j+1}$. Low p-values correspond to distinct eigenvalue, while the pairs (7, 8), (11, 12), (14, 15), (18, 19), and (19, 20) are in agreement with the null hypothesis.

In this section, we present a theorem and its corollary to describe the consistency and rate of convergence of the proposed variability measure $\bar{A}_k$. Assumption 1, which sets the condition for finite fourth moment of the covariance matrix, is to bound the difference between the resample covariance and sample covariance to within order $1/\sqrt{n}$ (Beran and Srivastava (1985)). Theorem 1(a) states that $\bar{A}_k$ converges to 0 in probability when the corresponding eigenvalue is distinct from all other eigenvalues. Theorem 1(b) states that, for a non-distinct eigenvalue, $\bar{A}_k$ converges to a positive constant. With Assumption 2, which sets the symmetric distribution of the covariance among the multiplicity components, we can derive a lower bound for $\bar{A}_k$ in Theorem 1(c). Corollary 1 extends Theorem 1 to the case of a few distinct eigenvalues and one degenerate eigenvalue, on which the scree plot targets.

Assumption 1. Suppose the observations $\{(x_{i1}, \ldots, x_{ip}); 1 \le i \le n\}$ are i.i.d. $p\times 1$ random vectors with covariance matrix $\Sigma = (\sigma_{j\ell})_{p\times p}$ and finite fourth moment.

Let $\mathbf{S}_n = (s_{j\ell})$ denote the sample covariance matrix. For each $\mathbf{S}_n$, there exists an orthogonal transformation basis $\mathbf{U}$ composed of eigenvectors such that $\mathbf{U}^T\mathbf{S}_n\mathbf{U} = \Lambda$, where Λ is the diagonal matrix with eigenvalues $\lambda_1 \ge \cdots \ge \lambda_p$ along the main diagonal. To simplify the notation, we use the eigenvectors as our coordinate basis. Let $\mathbf{V} = \sqrt{n}(\mathbf{S}^* - \mathbf{S}_n)$, where $\mathbf{S}^*$ is the sample covariance matrix from the resampling.

Assumption 2. $V_{jj}/\sqrt{n}-(\lambda_j-\lambda_1)$, $1 \le j \le p$, are identically and symmetrically distributed with mean 0 such that

$$P\Big(\frac{V_{jj}}{\sqrt{n}} - (\lambda_j-\lambda_1) > \frac{V_{\ell\ell}}{\sqrt{n}} - (\lambda_\ell-\lambda_1)\Big) = P\Big(\frac{V_{\ell\ell}}{\sqrt{n}} - (\lambda_\ell-\lambda_1) > \frac{V_{jj}}{\sqrt{n}} - (\lambda_j-\lambda_1)\Big),$$

and $V_{j\ell}$, $j \ne \ell$, are identically symmetric with mean 0 such that $P(V_{j\ell} > V_{rs}) = P(V_{rs} > V_{j\ell})$.

Theorem 1. *Under Assumption* 1, *for* $1 \le k \le p$,

(a) *if* λ_k *is distinct from all other eigenvalues,* $\bar{A}_k$ *converges to zero in probability and*

$$\bar{A}_k = \frac{1}{2n}\left[\sum_{j\ne k} \frac{EV_{kj}^2}{(\lambda_k-\lambda_j)^2}\right] + O_P\Big(\frac{1}{n^{3/2}}\Big);$$

(b) *when* $\max_j |\lambda_j - \lambda_{j+1}| = O_P(n^{-1/2})$, *all eigenvalues of* $\mathbf{S}^*$ *are around* λ_1 *and the* $\bar{A}_k$*'s converge to nonzero constants;*

(c) *when* $\max_j |\lambda_j - \lambda_{j+1}| = O_P(n^{-1/2})$, *under Assumption* 2, $\bar{A}_k \ge 1 - (1/q) \sum_{\ell=1}^{q}(1/\sqrt{\ell})$ *in probability, where* q *is the size of the multiplicity and* $q = p$ *in this case.*

The proof of Theorem 1 is in the Appendix.

Remark 1. An interesting point is that $\bar{A}_k$ in Theorem 1(a) is comparable with the asymptotic covariance of eigenvector $\mathbf{e}_k$ provided by Anderson (1963) and Mardia, Kent and Bibby (1979). Anderson (1963) derived the asymptotic covariance of eigenvector with distinct eigenvalues under the Gaussian assumption:

$$\lambda_k \sum_{1\le h\ne k\le p} \frac{\lambda_h}{(\lambda_h-\lambda_k)^2}\mathbf{e}_h\mathbf{e}_h^T,$$

comparable to the leading term in Theorem 1(**a**). We suggest a simple version for the kth variability: $\bar{B}_k = \sum_{h\ne k}[\lambda_h\lambda_k/((\lambda_h-\lambda_k)^2)]$. The comparison between $\bar{A}_k$ and $\bar{B}_k$ on the gene expression data set is in Figure 3. For those eigenvalues which are distinct from others, the $(\bar{A}_k, \bar{B}_k)$ fall around a line.

Remark 2. Some $\bar{B}_k$'s become very large for multiple eigenvalues; this is not a surprise because the asymptotic covariance matrix provided by Anderson (1963) and Mardia, Kent and Bibby (1979) does not hold for non-distinct eigenvalues. On the other hand, $\bar{A}_k$ is bounded by 1, by definition, and by the lower bound in Theorem 1(b). Thus, $\bar{A}_k$ presents a more friendly visualization plot.

Corollary 1. *Suppose that Assumption* 1 *holds, that* $\lambda_1, \ldots, \lambda_q$ *are distinct from other eigenvalues, and that* $\lambda_{q+1} - \lambda_p = O_P(1/\sqrt{n})$.

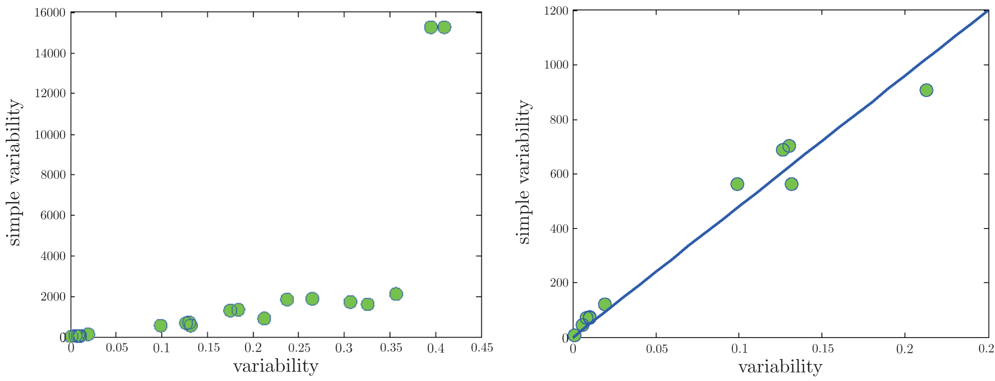

Figure 3. $\bar{A}_k$ vs. $\bar{B}_k$. The top plot is the scatter plot for $(\bar{A}_k, \bar{B}_k)$, where $1 \leq k \leq 20$. The bottom plot is the scatter plot for $(\bar{A}_k, \bar{B}_k)$ excluding non-distinct eigenvalue components.

(a) *For $k \leq q$, $\bar{A}_k$ converges to zero and*

$$\bar{A}_k = \frac{1}{2n} \sum_{j \neq k} \frac{EV_{kj}^2}{(\lambda_k - \lambda_j)^2} + O_p\Big(\frac{1}{n^{3/2}}\Big).$$

(b) *For $k > q$, the $\bar{A}_k$'s converge to nonzero constants.*

(c) *If Assumption 2 holds for those indices $\{q+1, \ldots, p\}$, then $\bar{A}_k \geq 1 - (p - q)^{-1} \sum_{\ell=1}^{p-q} 1/\sqrt{\ell}$ in probability.*

The proof of Corollary 1 is omitted since it follows the arguments found in the proof of Theorem 1.

4. Simulation Studies

We applied the common factor model (Spearman (1904) and Kshirsagar (1961)) to generate data for simulation:

$$\mathbf{x}^T = \sum_{j=1}^{K} \mathbf{g}_j^T f_j + \epsilon^T, \tag{4.1}$$

where $\mathbf{x} = (x_1, \ldots, x_p)$, the $\mathbf{g}_j$'s are constructed as orthogonal vectors with $\mathbf{g}_j = (g_{j1}, \ldots, g_{jp})$, $g_{j\ell} \in \{1, 0, -1\}$, and we let $G_j = \sum_\ell |g_{j\ell}|$. Let $\epsilon = (\epsilon_1, \ldots, \epsilon_p)$ be i.i.d. noise. In this model, the f_j's are unobservable latent variables that influence the surrogate variables x_k's, and the ϵ_k's and f_j's are all uncorrelated, with $Ef_j = 0$, $E(\epsilon_k) = 0$, $\mathrm{Var}\,(f_j) = \sigma_j^2$ and $\mathrm{Var}\,(\epsilon_k) = \sigma_\epsilon^2$. The parameters we used in this comparison were $\sigma_j^2 = j/16$, $\sigma_\epsilon^2 = 1$, p in $\{16, 32, 64, 128\}$, and $G_j = p$. We took $n = 1,000$ and repeated $1,000$ times to calculate the means.

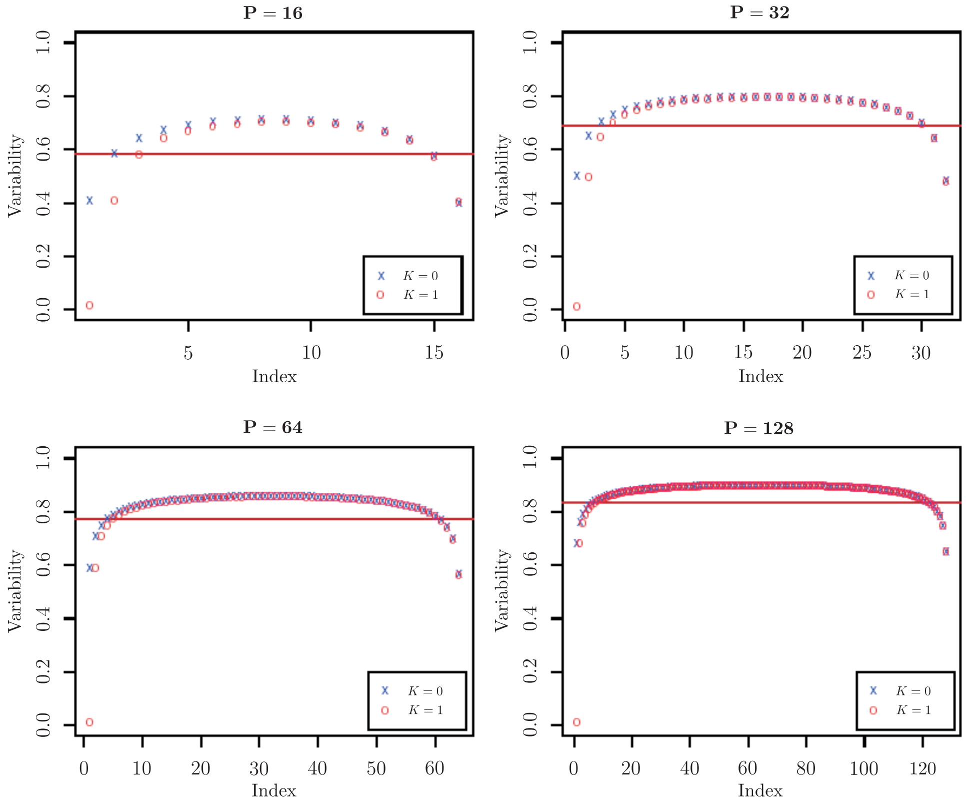

Figure 4. Mean variability plots for $p = 16$, 32, 64, 128, and $K = 1, 0$ are plotted. In all four plots, $\bar{A}_1$ is very close to 0 for $K = 1$. The horizontal lines are the derived lower bounds of the $\bar{A}_k$'s for multiplicity p, under the assumption that the diagonal terms of the resampled covariance matrix have identically symmetric distributions and the off diagonal terms have identically symmetric distributions.

When $K = 1$, one signal component is hidden in p-dimensional i.i.d. noise. When $K = 0$, the data is p-dimensional i.i.d. noise. $\bar{A}_1$ and $\bar{A}_2$ are listed in Table 2. Figure 4 gives the mean curve of those $1,000$ variability plots. The variability measure $\bar{A}_k$ is quite consistent with the asymptotic results. We observe that, as the data dimension p increases, the relevant variability increases also. The derived lower bound in Theorem 1(c), $1 - (1/p)\sum_{\ell=1}^{p} 1/\sqrt{\ell}$, is an increasing function of p. This can be explained by the fact that an increase in multiplicity leads to an increase in the degrees of freedom for the eigenvectors to be reordered.

Figure 5 shows another example with $K = 8$ factors, $n = 1,000$, $p = 64$. The eight common factors were spread over the first $G = G_j = 32$ dimensions for $1 \leq j \leq 8$. PCA was applied on both the correlation matrix and the covariance

Table 2. Mean and SD of mean distances with $\sigma_1^2 = 1/16$, $\sigma_\epsilon^2 = 1$ and $G = p$.

	$K = 0$(No Signal)		$K = 1$(One Signal)	
p	$\bar{A}_1(\rightarrow$ nonzero)	$\bar{A}_2(\rightarrow$ nonzero)	$\bar{A}_1(\rightarrow 0)$	$\bar{A}_2(\rightarrow$ nonzero)
16	0.4098	0.5837	0.0158	0.4040
32	0.4989	0.6510	0.0119	0.5002
64	0.5877	0.7071	0.0100	0.5876
128	0.6797	0.7580	0.0090	0.6805

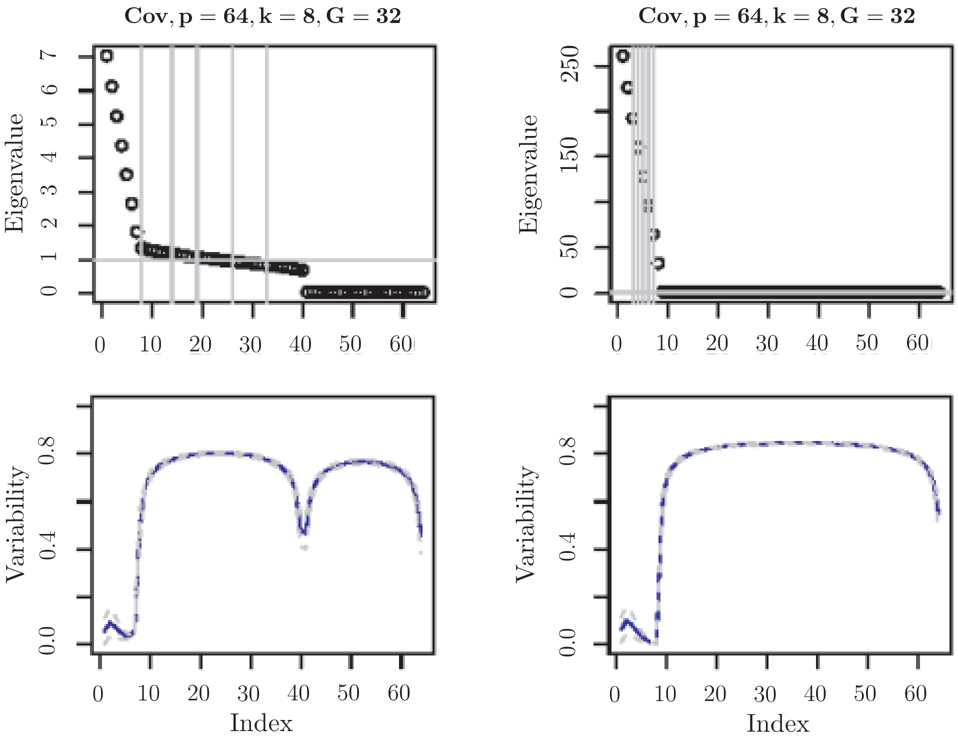

Figure 5. Scree Plot and Variability Plot: PCA was applied to the covariance and the correlation matrix of a data set with $n = 1,000$ and $p = 64$. These plots are based on a mean of 1,000 repeats. The data were generated through the common factor model with $K = 8$ factors and these factors were spread over the first $G = 32$ dimensions, the Gaussian noise was spread out over the $p = 64$ dimensions. Thus, seven distinct eigenvalues and two groups of multiple roots were generated when the correlation matrix was applied.

matrix. The variability plots were successful in grouping the multiplicity eigenvalue components as a hill and the $\bar{A}_k$'s were close to zero for distinct eigenvalues.

We also applied this algorithm to the case $p \geq n$. We generated a functional

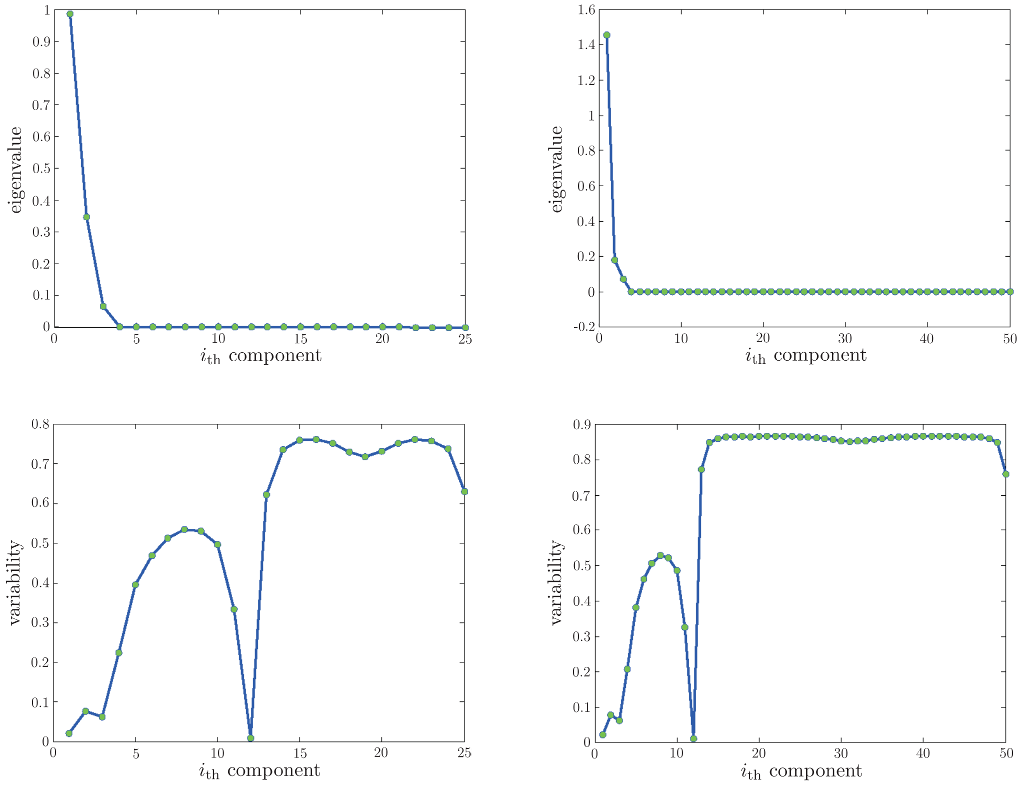

Figure 6. Scree Plot and Variability Plot for $p \geq n$ The data were generated with $n = 25$, 3 signal components and 10 noise components; $p = 25$ for the left column and $p = 50$ for the right column.

data set through the model from Hall and Vial (2007):

$$x(t) = \sum_{j=1}^{K} \eta_j \psi_j(t) + 0.01 n^{-0.25} \sum_{j=2}^{p_0} \zeta_j \chi_j(t),$$

where the η_j's for signal components and the ζ_j's for noise components were uncorrelated Gaussian random variable's with variance j^{-2} and $j^{-1.6}$, $\psi_j(t) = \sqrt{2}\cos(j\pi t)$, and $\chi_j(t) = \sqrt{2}\sin(j\pi t)$. We took $K = 3$, $p_0 = 10$, $n = 25$ and let p be 25 and 50. We let t take values on $2\ell\pi/p$, where $1 \leq \ell \leq p$. Figure 6 shows the scree plots and the variability plots. The number of intrinsic components was $K + p_0 - 1 = 12$, which means that there was multiplicity with size $p - 12$ taking zero value. From the scree plot, the noise components and 0 eigenvalue components were not distinguishable. Variability plots are useful for separating the distinct eigenvalue components, noise components and 0 eigenvalue components.

5. Discussion

As stated in Stewart (1973), a small perturbation of a matrix will only lead to a small change of eigenvalues and also a small change of eigenvectors for distinct eigenvalue, while the behavior of eigenvectors can be quite different for non-distinct eigenvalues as demonstrated in Theorem 1. In this paper, we employ the bootstrap to give a measure of the stability of eigenvectors over different samples. Using the proposed variability plot to choose components, we can avoid the problem, of offering widely differing patterns from one sample to the next, which can be encountered in using a specified percentage of total variance.

We also demonstrated that this algorithm can be applied to the case $p > n$. When the number of intrinsic components is well controlled, then no matter how large p is, the asymptotic results of Theorem 1 still hold. We believe this tool can bring statisticians more insights in exploring high dimensional data. Although the computing time is proportional to $p^2 \times B$, this may not be a major concern. For example, in generating the variability plot for the gene expression data set with $n = 6,688$, $p = 128$ and $B = 1,000$, it took 618 seconds on a PC with Pentium 4, CPU 3.4Ghz and 3GB RAM.

6. Appendix

Proof of Theorem 1(a). Write the bth bootstrap resampling covariance matrix as

$$\mathbf{S}_b^* = \mathbf{S}_n + \frac{1}{\sqrt{n}}\mathbf{V}.$$

Let $\mathbf{e}_k$ and λ_k, $1 \leq k \leq p$, be the eigenvectors and the eigenvalues of $\mathbf{S}_n$. According to Beran and Srivastava (1985), $\mathbf{V}_{j\ell} = O_P(1)$ for any $1 \leq j \neq \ell \leq p$. For the bth resampling, denote the kth eigenvalue and the kth eigenvector by λ_k^{*b} and $\mathbf{e}_k^{*b}$, respectively. When λ_k is the simple root of the characteristic polynomial of $\mathbf{S}_n$, it follows from Theorems 8.1.4, 8.1.5, and 8.1.12 of Golub and Van Loan (1996) that

$$\lambda_k^{*b} = \lambda_k + \frac{\delta}{\sqrt{n}} \quad \text{and} \quad \mathbf{e}_k^{*b} = \mathbf{e}_k + \frac{1}{\sqrt{n}}\mathbf{f}_k,$$

where the Frobenius norm of $\mathbf{f}_k$ is bounded.

By comparing the equations

$$\mathbf{S}_n\mathbf{e}_k = \lambda_k\mathbf{e}_k,$$
$$\Big(\mathbf{S}_n + \frac{1}{\sqrt{n}}\mathbf{V}\Big)\Big(\mathbf{e}_k + \frac{1}{\sqrt{n}}\mathbf{f}_k\Big) = \Big(\lambda_k + \frac{\delta}{\sqrt{n}}\Big)\Big(\mathbf{e}_k + \frac{1}{\sqrt{n}}\mathbf{f}_k\Big),$$

we get

$$(\mathbf{S}_n - \lambda_k\mathbf{I})\mathbf{f}_k = (\delta\mathbf{I} - \mathbf{V})\mathbf{e}_k + O_P\Big(\frac{1}{\sqrt{n}}\Big).$$

Take the inner product of the above with $\mathbf{e}_k$ to get $\delta = \mathbf{V}_{kk} + O_P(1/\sqrt{n})$, and take the inner product with $\mathbf{e}_j$ to get

$$(\lambda_j - \lambda_k) f_{kj} = -\mathbf{V}_{kj} + O_P\Big(\frac{1}{\sqrt{n}}\Big),$$

where $f_{kj} = \mathbf{f}_k \cdot \mathbf{e}_j$. With the constraint $|\mathbf{e}_k^{*b}| = 1$, we have

$$\left(1 + \frac{f_{kk}}{\sqrt{n}}\right)^2 + \sum_{j \neq k} \left(\frac{f_{kj}}{\sqrt{n}}\right)^2 = 1.$$

Thus

$$f_{kk} = -\frac{1}{2\sqrt{n}} \sum_{j \neq k} \left(\frac{\mathbf{V}_{kj}}{\lambda_j - \lambda_k}\right)^2 + O_P\Big(\frac{1}{n}\Big),$$

which leads to

$$\bar{A}_k = 1 - \frac{1}{B} \sum_{b=1}^{B} |\mathbf{e}_k \cdot \mathbf{e}_k^{*b}| \to 1 - E|\mathbf{e}_k \cdot \mathbf{e}_k^{*b}| = \frac{1}{2n} \sum_{j \neq k} \frac{E\mathbf{V}_{jk}^2}{(\lambda_j - \lambda_k)^2} + O_P\Big(\frac{1}{n^{3/2}}\Big)$$

as $B \to \infty$.

Proof of Theorem 1(b)

For non-distinct eigenvalues $\lambda_1, \ldots \lambda_p$, we give a proof by arguing that $E|\mathbf{e}_k \cdot \mathbf{e}_k^*| < 1$. We first determine the eigenvalues of $\mathbf{S}^*$ that are roots of $\det|\Lambda + \mathbf{V}/\sqrt{n} - \lambda^* \mathbf{I}| = 0$, where Λ is the eigenvalue diagonal matrix, using

$$(\lambda^* - \lambda_k)^p - \left[\sum_{j=1}^{p} \left(\frac{V_{jj}}{\sqrt{n}} - \tau_j\right)\right] (\lambda^* - \lambda_1)^{p-1}$$
$$+ \left[\sum_{j<\ell} \left(\frac{V_{jj}}{\sqrt{n}} - \tau_j\right)\left(\frac{V_{\ell\ell}}{\sqrt{n}} - \tau_\ell\right) - \sum_{j<\ell} \frac{V_{j\ell}^2}{n}\right] (\lambda^* - \lambda_1)^{p-2} + O_P\Big(\frac{1}{n^{3/2}}\Big) = 0, \quad (6.1)$$

where $\tau_j = \lambda_1 - \lambda_j$ is of the order $1/\sqrt{n}$.

The p roots of (6.1) are $\lambda_1 + o(1)$. The kth eigenvector $\mathbf{e}_k^*$ is determined by the linear system

$$\left(\lambda_1 + \frac{V_{jj}}{\sqrt{n}} - \tau_j\right) e_{kj}^* + \sum_{\ell \neq j} \frac{V_{j\ell}}{\sqrt{n}} e_{k\ell}^* = \lambda_k^* e_{kj}^*, \quad 1 \leq j \leq p, \quad (6.2)$$

where $e_{kj}^* = \mathbf{e}_k^* \cdot \mathbf{e}_j$. Recall that

$$\bar{A}_k = 1 - \frac{1}{B} \sum_{b=1}^{B} |\mathbf{e}_k \cdot \mathbf{e}_k^{*b}| \to 1 - E|e_{kk}^*| = 1 - E\frac{|e_{kk}^*|}{\sqrt{\sum_{j=1}^{p} (e_{kj}^*)^2}}$$
$$= 1 - E\frac{1}{\sqrt{1 + \sum_{j \neq k} (|e_{kj}^* / e_{kk}^*|^2)}}. \quad (6.3)$$

Now $\bar{A}_k \to 0$ if and only if $e^*_{kj} = 0$ with probability 1 for all $j \neq k$, which cannot be true in view of (6.1) and (6.2). This concludes the proof.

Proof of Theorem 1(c) We give the lower bound of $\bar{A}_k$ under the symmetric assumption by showing that the projections of the resample eigenvector on the multiple root eigenvectors have the same distribution. Therefore, the probability for each projection to be of an order from 1 to p is equally likely. Note that the solutions for e^*_{kj} are functions of $V_{j\ell}$ and τ_j, where $1 \leq j,\ \ell \leq p$. Due to the symmetric expression of $e^*_{k1}, \ldots, e^*_{kp}$ at (4), both e^*_{k1} and $e^*_{k\ell}$ can be written as a function of V_{rs} and τ_r, $1 \leq r,\ s \leq p$, in which the indices for V_{rs} and τ_r are adjusted by permuting index 1 with index ℓ, where $1 \leq \ell \leq p$.

Under Assumption 2: symmetric distributions of $V_{jj}/\sqrt{n} - \tau_j$, $1 \leq j \leq p$, and symmetric distribution of $V_{j\ell}$, $1 \leq j \neq \ell \leq p$, the rank of $|e^*_{kj}|$ among $\{|e^*_{k1}|, \ldots, |e^*_{kp}|\}$ is equally likely from 1 to p. Thus we can give a crude lower bound of $\bar{A}_k$: each term $|e^*_{kj}/e^*_{kk}|$ in the denominator is replaced by 0 if it is less than 1, and by 1 otherwise. Then the denominator is replaced by $\sqrt{j}$ when $|e^{*b}_{kk}|$ is the jth order statistic of $\{|e^*_{k\ell}|, 1 \leq \ell \leq p\}$. Accordingly, the probability of the above event is $1/p$. Thus we have

$$\bar{A}_k \geq 1 - \frac{1}{p}\sum_{\ell=1}^{p}\frac{1}{\sqrt{\ell}} \ \text{ in probability.}$$

Acknowledgement

This research is partially supported by Taiwan's NSC grant (93-2118-M-001-008-MY2). The authors would like to thank the precious comments from the referees. We are also very grateful to Hsin-Ping Wu and Hsin-Hsiung Huang for providing programming help.

References

Alter, O., Brown, P. O. and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA* **97**, 10101-6.

Anderson, T. W. (1963). Asymptotic theory for principal component analysis. *Ann. Math. Statist.* **34**, 122-148.

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Psychology* **3**, 77-85.

Beran, R. and Srivastava, M. S. (1985). Bootstrap tests and confidence regions for functions of a covariance matrix. *Ann. Statist.* **13**, 95-115.

Besse, P. (1992). PCA stability and choice of dimensionality. *Statist. Probab. Lett.*. **13**, 405-10.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research* **1**, 245-76.

Daudin, J. J., Dubey, C. and Trecourt, P. (1988). Stability of principal component analysis studied by the bootstrap method. *Statistics* **19**, 241-58.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7**, 1-26.

Golub, G. and Van Loan, C. F. (1996). *Matrix Computations.* Third Edition. The Johns Hopkins University Press, Baltimore.

Hall, P. and Vial, C. (2007). Assessing the finite dimensionality of functional data. *J. Roy. Statist. Soc. Ser. B.* **68**, 689-705.

Jolliffe, I. T. (2002). *Principal Component Analysis*, Wiley, New York.

Krzanowski, W. J. (1984). Sensitivity of principal components. *J. Roy. Statist. Soc. Ser. B.* **46**, 558-63.

Kshirsagar, A. M. (1961). The goodness of fit of a single (non-isotropic) hypothetical principal component. *Biometrika* **48**, 397-407.

Leung, S. Y., Chen, X., Chu, K. M., Yuen, S. T., Mathy, J., Ji, J., Chan, A. S., Li, R., Law, S., Troyanskaya, O. G., Tu, I. P., Wong, J., So, S., Botstein, D. and Brown, P. O. (2002). Phospholipase A2 group IIA expression in gastric adenocarcinoma is associated with prolonged survival and less frequent metastasis. *Proc. Natl. Acad. Sci. USA* **99**, 16203-8.

Mardia, K. V., Kent, T. J. and Bibby, J. M. (1979). *Multivariate Analysis.* Academic Press, New York.

Mori, Y., Selaru, F. M., Sato, F., Yin, J., Simms, L. A., Xu, Y., Olaru, A., Deacu, E., Wang, S., Taylor, J. M., Young, J., Leggett, B., Jass, J. R., Abraham, J. M., Shibata, D. and Meltzer, S. J., (2003). The impact of microsatellite instability on the molecular phenotype of colorectal tumors. *Cancer Research* **63**, 4577-82.

North, G. R., Bell, T. L., Cahalan, R. F. and Moeng, F. J. (1982). Sampling errors in the estimation of empirical orthogonal functions. *Monthly Weather Rev.* **110**, 699-706.

Raychaudhuri, S., Stuart, J. M., and Altman, R. B. (2000). Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pacific Symposium on Biocomputing*, 455-66.

Sinha, A. R. and Buchnan, B. S. (1995). Assessing the stability of principal components using regression. *Psychometrika* **60**, 355-69.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology* **15**, 72-101.

Stewart, G. W. (1973). Error and perturbation bounds for subspaces associated with certain eigenvalue problems. *SIAM Rev.* **15**, 727-64.

Institute of Statistical Science, Academia Sinica, Taipei, 11529 Taiwan.

E-mail: iping@stat.sinica.edu.tw

Department of Mathematics and Institute of Epidemiology, National Taiwan University, Taipei, 10617 Taiwan.

E-mail: hchen@math.ntu.edu.tw

Department of Biopharmaceutical Sciences, UC San Francisco, CA 94143, USA.

E-mail: xinchen@itsa.ucsf.edu

(Received April 2007; accepted August 2008)