Consistency of principal component scores in visualizations of high-dimensional data

Kristoffer H. Hellton*
Dept. of Biostatistics, University of Oslo, Oslo, Norway - k.h.hellton@medisin.uio.no

Magne Thoresen
Dept. of Biostatistics, University of Oslo, Oslo, Norway - magne.thoresen@medisin.uio.no

Abstract

Plots of principal component scores are a popular approach to visualize and explore high-dimensional data. However, the inconsistency of high-dimensional eigenvectors prompted the development of sparse principal component analysis (PCA), where sparse eigenvectors can be estimated using different regularization approaches. Still, classical PCA is extensively and successfully used to visualize high-dimensional genetic data. In this work, we try to give an explanation of this paradoxical situation. We show that the visual information in terms of the relative positions of scores will be consistent, if the related signal can be considered to be pervasive, despite of inconsistent eigenvectors. Within the High Dimension Low Sample Size (HDLSS) asymptotic framework, we show that when the spiked population eigenvalues scale linearly with the dimension, the sample principal component scores are scaled and rotated versions of the population principal component scores. Further, we argue that eigenvalues scaling linearly with the dimension can be interpreted as expressions of pervasive signals, specifically given by eigenvectors with an asymptotic non-zero proportion of non-zero coefficients, and we discuss examples of genetic applications where assuming pervasive signals is reasonable.

Keywords: Genomics; HDLSS; PCA; Asymptotic distribution.

1. Introduction

Principal component analysis (PCA) is the workhorse of variable reduction in applied data analysis. It is used to construct a small number of informative scores from the original data, and these scores are then used further in visualization or in conventional classification, clustering or regression methods. This is highly useful in the context of modern high-dimensional data analysis, where the number of measured variables p exceeds the sample size n. Genomics is a typical application where the data exploration is done by visually investigating the first few principal component (PC) scores. The asymptotic behavior of high-dimensional PCA has attracted a substantial amount of attention the last few years. It has been shown, by Paul (2007) and Johnstone and Lu (2009), that the population eigenvalues and -vectors in PCA are not consistently estimated by the sample eigenvalues and -vectors under the finite γ regime, where $p/n = \gamma$ as $p, n \to \infty$. However, in an applied setting, the behavior of the principal component scores is also of interest.

PCA reduces the data dimension by constructing orthogonal linear combinations of the variables, which express maximal variability. The first component is the normalized linear combination of variables with the highest variance, while the second component will be the linear combination, orthogonal to the first, with the highest variance, and so on. The mathematical basis of PCA is the eigendecomposition of the sample covariance matrix. Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ be a $p \times n$ data matrix, where $\mathbf{x}_i = [x_{i1}, \ldots, x_{ip}]^T$ are independent and identically distributed with $\mathrm{E}\,\mathbf{x}_i = \mathbf{0}$ and $\mathrm{var}\,\mathbf{x}_i = \Sigma$.

The eigendecomposition of the covariance matrix is given by

$$\Sigma = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T,$$

where $\mathbf{\Lambda}$ is the diagonal matrix of the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$ and $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_p]$ is the matrix of eigenvectors. The weights of the orthogonal linear combinations are given by the eigenvectors, usually referred to as loadings. We denote the vector of the resulting population component scores by

$$\mathbf{s}_j^T = \mathbf{v}_j^T\mathbf{X} = [\mathbf{v}_j^T\mathbf{x}_1, \ldots, \mathbf{v}_j^T\mathbf{x}_n].$$

The eigenvalues express the variance of the component scores, such that the vector of standardized population component scores is given by

$$\mathbf{z}_j^T = \frac{\mathbf{v}_j^T\mathbf{X}}{\sqrt{\lambda_j}}, \tag{1}$$

where the jth vector of scores is $\mathbf{z}_j^T = [z_{j1}, \ldots, z_{jn}]$ and $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_p]^T$. In applied data analysis, the eigendecomposition is based on the sample covariance matrix with the proper centering, denoted by $\hat{\Sigma} = \frac{1}{n}\mathbf{X}\mathbf{X}^T$:

$$\hat{\Sigma} = \hat{\mathbf{V}}\mathbf{D}\hat{\mathbf{V}}^T.$$

Here $\mathbf{D} = \text{diag}(d_1, \ldots, d_p)$ contains the sample eigenvalues and $\hat{\mathbf{V}} = [\hat{\mathbf{v}}_1, \ldots, \hat{\mathbf{v}}_p]$ the corresponding sample eigenvectors. Following the earlier notation, we construct the sample component scores as

$$\hat{\mathbf{s}}_j^T = \hat{\mathbf{v}}_j^T\mathbf{X},$$

and the sample standardized scores as

$$\hat{\mathbf{z}}_j^T = \frac{\hat{\mathbf{v}}_j^T\mathbf{X}}{\sqrt{d_j}}. \tag{2}$$

We further assume that the population eigenvalues follow the spiked eigenvalue model, where the first m population eigenvalues are substantially larger than the remaining non-spiked eigenvalues. Anderson (1963) showed that the sample eigenvectors and -values, $\hat{\mathbf{v}}$ and d, will consistently estimate the population eigenvectors and -values, $\mathbf{v}$ and λ, when p is fixed and $n \to \infty$. However, this will not be the case in the high-dimensional setting. Starting with Paul (2007) and Johnstone and Lu (2009), it has been shown that the sample eigenvalues and -vectors are not asymptotically consistent when the population eigenvalues are fixed and $p, n \to \infty$ at a constant ratio $p/n = \gamma > 0$. Jung and Marron (2009) introduced the HDLSS asymptotic framework, where instead n is fixed and the spiked eigenvalues grow with the dimension p, according to

$$\lambda_i = \sigma_i^2 p^\alpha, \quad i = 1, \ldots, m.$$

In this asymptotic setting where only $p \to \infty$, the consistency of PCA depends on α. The eigenvectors are estimated consistently when $\alpha > 1$, while the estimates are strongly inconsistent when $\alpha < 1$. In the boundary case $\alpha = 1$, a situation explored by Jung et al. (2012), the sample eigenvectors are neither consistent nor strongly inconsistent and instead reach a limiting distribution depending on n.

The main focus of the above-mentioned papers has been on the eigenvector inconsistency, and few results are concerned with principal component scores. Shen et al. (2012) investigated the ratio between the individual sample and population scores, and not the inner product between the sample

and population PC score vectors. Following the regime of Jung and Marron (2009), they showed that for $\alpha > 1$, the ratio between the individual scores ratio converges to a random variable common for all the observations within a component. This implies that, asymptotically, a two-dimensional plot of the sample scores will be a scaled version of the population score plot. The visual information displayed by the samples scores will therefore be the same as for the population scores. In this paper, we investigate the same problem as Shen et al. (2012), but in the situation where $\alpha = 1$. We further show that the growth rate of population eigenvalues can be interpreted in terms of the generating mechanism behind the data. Our aim is to translate the assumption about the eigenvalues into an assumption regarding the latent structure and the data-generating mechanism, as this is generally easier to relate to.

2. Pervasive structures

We demonstrate how an assumption on the eigenvector coefficients can lead to a corresponding eigenvalue scaling linearly with the dimension asymptotically, a situation equivalent to the case of $\alpha = 1$ in the HDLSS framework. In the econometrics literature, the concept of pervasiveness is commonly used in latent factor estimation (Fan et al., 2013). A pervasive factor is thought to be an underlying latent variable affecting most or a significant proportion of the observed variables. In a high-dimensional situation where the dimension increases, pervasiveness can be formulated in terms of the asymptotic proportion of non-zero factor loadings, prompting the following definition:

Definition (Pervasiveness)**.** *A sequence of p-dimensional vectors* $\mathbf{v} = [v_1, \ldots, v_p]^T$ *fulfills the pervasiveness assumption, if the proportion of non-zero entries* $r_p = \frac{1}{p}\sum_{i=1}^{p} I_{\{v_i^2 > 0\}}$ *fulfills:*

$$\lim_{p\to\infty} r_p > 0.$$

If the values of the coefficients of $\mathbf{v}$ are constant and the proportion of non-zero coefficients r_p converge to a non-zero value as $p \to \infty$, according to the definition of pervasiveness, there must exist two constant $c_1 \leq r_p \min_j v_j^2$ and $c_2 \geq r_p \max_j v_j^2$. Then the largest population eigenvalue λ_1 of the covariance matrix Σ of $\mathbf{x}_i$ fulfills the bound

$$c_1 p + \sigma^2 \leq \lambda_1 \leq c_2 p + \sigma^2.$$

The remaining population eigenvalues are given by $\lambda_i = \sigma^2$ for $i = 2, \ldots, p$. This argument can also be extended to m components. With the definition of a pervasive factor, it is possible to connect the data-generating mechanism behind the data to the asymptotic behavior of the eigenvalues. There are many examples of applications where pervasive structures are reasonable, suggesting that also the assumption of linearly increasing eigenvalues is reasonable.

One example from genomics is the genetic markers such as single nucleotide polymorphisms (SNPs). SNPs are genetic loci having at least two alleles with an associate allelic frequency in a population. The neutral theory of molecular evolution states that allele frequencies at most genetic loci change due to two stochastic processes; mutation and random drift. If the main variation in the data sample stems from differences between ethnic populations, random allelic drift is the main driver behind changes in the genetic markers. This will give many and randomly distributed differences and when new markers are included, we expect a certain proportion to be informative with respect to ethnicity. This corresponds to our notion of a pervasive association between ethnicity and the SNPs, and one could then assume that the eigenvalue corresponding to the signal of ethnicity would scale linearly with number of included variables.

3. Asymptotic results

We present an asymptotic result regarding the behavior of the sample principal component scores in the case of eigenvalues scaling linearly with the dimension. Theorem 1 is derived within the HDLSS regime (Jung et al., 2012), and we state the same general conditions for the distribution of the standardized population component scores and for the structure of the population eigenvalues:

Conditions.

1. *The standardized principal component scores* $\mathbf{z}_i$ *have finite fourth moments and are uncorrelated, but possibly dependent, fulfilling the* ρ*-mixing condition.*

2. *For the non-spiked eigenvalues* $\lambda_{m+1}, \ldots, \lambda_p$*, it must hold that*

$$\frac{\sum_{i=m+1}^{p} \lambda_i^2}{\left(\sum_{i=m+1}^{p} \lambda_i\right)^2} \to 0, \quad \frac{1}{p} \sum_{i=m+1}^{p} \lambda_i \to \tau^2, \qquad \text{as } p \to \infty.$$

Remark. *The* ρ*-mixing condition is satisfied if the maximal correlation coefficient approach zero,* $\rho(m) \to 0$*, as* $m \to \infty$*, where*

$$\rho(m) = \sup_{j,f,g} |\operatorname{cor}(f, g)|, \quad f \in L_2(\mathcal{F}_{-\infty}^{j}), g \in L_2(\mathcal{F}_{j+m}^{\infty}),$$

and $\mathcal{F}_K^L$ *is the* σ*-field of events generated by the variables* $\mathbf{z}_i, K \leq i \leq L$*. Condition 2 ensures that the non-spiked eigenvalues do not decrease too fast and that the mean converges to* τ^2*. The constant non-spiked eigenvalues* $\lambda_{m+1} = \cdots = \lambda_p = \tau^2$ *in the spiked covariance model is the simplest situation which fulfills condition 2. The scaling constants of the eigenvalues* $\sigma_1^2 \geq \cdots \geq \sigma_m^2 > 0$ *represent the signal strength.*

Let $\mathbf{Z}$ be the population standardized principal component scores. For a m spiked model, we denote the scaled population scores by

$$\tilde{\mathbf{Z}}_{1:m} = [\sigma_1 \mathbf{z}_1, \ldots, \sigma_m \mathbf{z}_m]$$

giving rise to the $m \times m$ matrix $\mathbf{W} = \tilde{\mathbf{Z}}_{1:m}^T \tilde{\mathbf{Z}}_{1:m}$. The eigenvalues and the eigenvectors of the matrix $\mathbf{W}$ are denoted by $\phi_j(\mathbf{W})$ and $\mathbf{v}_j(\mathbf{W})$.

Theorem 1. *Under the Conditions 1 and 2 and the assumption that the* m *spiked eigenvalues scale linearly with* p*,*

$$\lambda_1 = \sigma_1^2 p, \quad \lambda_2 = \sigma_2^2 p \quad \cdots \quad \lambda_m = \sigma_m^2 p,$$

the sample principal component scores converges to the following limiting distribution jointly for all j *as* $p \to \infty$*:*

$$\hat{z}_{ij} \xrightarrow{\mathrm{d}} \sqrt{\frac{n}{\phi_j(\mathbf{W})}} \sum_{l=1}^{m} \sigma_l z_{il} v_{jl}(\mathbf{W}),$$

where $\phi_j(\mathbf{W})$ *and* $\mathbf{v}_j(\mathbf{W})$ *are the* j*th eigenvalue and eigenvector of the stochastic matrix* $\mathbf{W}$*. The joint asymptotic distribution of sample PC scores can be written in matrix notation as*

$$\begin{bmatrix} \hat{z}_{i1} \\ \vdots \\ \hat{z}_{im} \end{bmatrix} \xrightarrow{\mathrm{d}} \begin{bmatrix} \sqrt{n/\phi_1(\mathbf{W})} & & 0 \\ & \ddots & \\ 0 & & \sqrt{n/\phi_m(\mathbf{W})} \end{bmatrix} \begin{bmatrix} \mathbf{v}_1(\mathbf{W}) & \cdots & \mathbf{v}_m(\mathbf{W}) \end{bmatrix} \begin{bmatrix} \sigma_1 z_{i1} \\ \vdots \\ \sigma_m z_{im} \end{bmatrix}, \quad i = 1, \ldots, n.$$

The result is proven by decomposing the distribution of the sample scores and utilizing results found by Jung et al. (2012). The first matrix, given by the sample eigenvalues, is a diagonal matrix and will act as a *scaling matrix*. The second matrix, given by the m sample eigenvectors, is an orthogonal matrix and will therefore act as an m-dimensional *rotation matrix*. Thus the m-dimensional sample score vector will be a scaled and rotated version of the population scores in m-dimensional space. However, for visualizations one are mainly concerned with pairs of scores displayed in two-dimensional plots, as explored in the next section.

4. Implications for visualization

When principal component scores are used to visualize high-dimensional data, it is common to display the m first important sample scores in two-dimensional plots. These score plots are used to explore data structures, compare observations and detect subgroups. However, even though the eigenvectors are inconsistently estimated, Theorem 1 can contribute to the understanding of when a plot of the sample scores will give valid information about the population scores. For $m > 2$, the distribution of the pair of the jth and kth sample principal component score can be decomposed into two parts:

$$\begin{bmatrix} \hat{z}_{ij} \\ \hat{z}_{ik} \end{bmatrix} \xrightarrow{\text{d}} \underbrace{\begin{bmatrix} \sqrt{n/\phi_j(\mathbf{W})} & 0 \\ 0 & \sqrt{n/\phi_k(\mathbf{W})} \end{bmatrix}}_{\text{Scaling}} \underbrace{\begin{bmatrix} v_{jj}(\mathbf{W}) & v_{kj}(\mathbf{W}) \\ v_{jk}(\mathbf{W}) & v_{kk}(\mathbf{W}) \end{bmatrix}}_{\text{Approximate rotation}} \begin{bmatrix} \sigma_j z_{ij} \\ \sigma_k z_{ik} \end{bmatrix} + \begin{bmatrix} \varepsilon_{ij} \\ \varepsilon_{ik} \end{bmatrix} \quad i = 1, \dots, n,$$

where

$$\varepsilon_{ij} = \sqrt{\frac{n}{\phi_j(\mathbf{W})}} \sum_{l \neq j,k}^{m} \sigma_l z_{il} v_{jl}(\mathbf{W}), \quad \varepsilon_{ik} = \sqrt{\frac{n}{\phi_k(\mathbf{W})}} \sum_{l \neq j,k}^{m} \sigma_l z_{il} v_{jl}(\mathbf{W}).$$

Simulations show that the terms ε_{ij} and ε_{ik} can be regarded as noise compared to the scaling and rotation matrix, for moderate sample sizes. The simulations are based on normally distributed standardized population scores. We further investigate the impact of the sample size n, the number of signals m and the signal strengths $\sigma_1^2, \dots, \sigma_n^2$, on the noise using simulations and theoretical results. When the noise term is negligible, the two-dimensional plot of the estimated jth and kth sample scores will be a scaled and approximately rotated version of the two-dimensional plot of the population scores. Thus the relative positions of the population scores, and thereby the visual information, will be preserved by the sample scores. The specific behavior of the scaling and the approximate rotation matrix will depend on the distribution of the random eigenvalues and -vectors, $\phi_j(\mathbf{W})$ and $\mathbf{v}_{ij}(\mathbf{W})$.

A part of this behavior is demonstrated in Figure 1, displaying two different realizations of the sample standardized principal component scores, Eq. (2), compared to the population standardized principal component scores, Eq. (1). The sample standardized scores are shown as black dots, while the population standardized scores are shown as circle. For a specific pair of PC scores, if the scaling in both components are significantly larger or smaller than 1 and there is no significant rotation, the sample scores will appear as a radial shift compared to the population scores, as seen in Figure 1a). If instead the approximate rotation is significant, but there is no significant scaling, the sample scores will appear as a rotation of the population scores around the origin, as seen in Figure 1b).

5. Conclusions

The use of PCA in high-dimensional data suffers from the somewhat paradoxical situation where, theoretically, the eigenvectors and -values are not correctly estimated, but the method can be highly

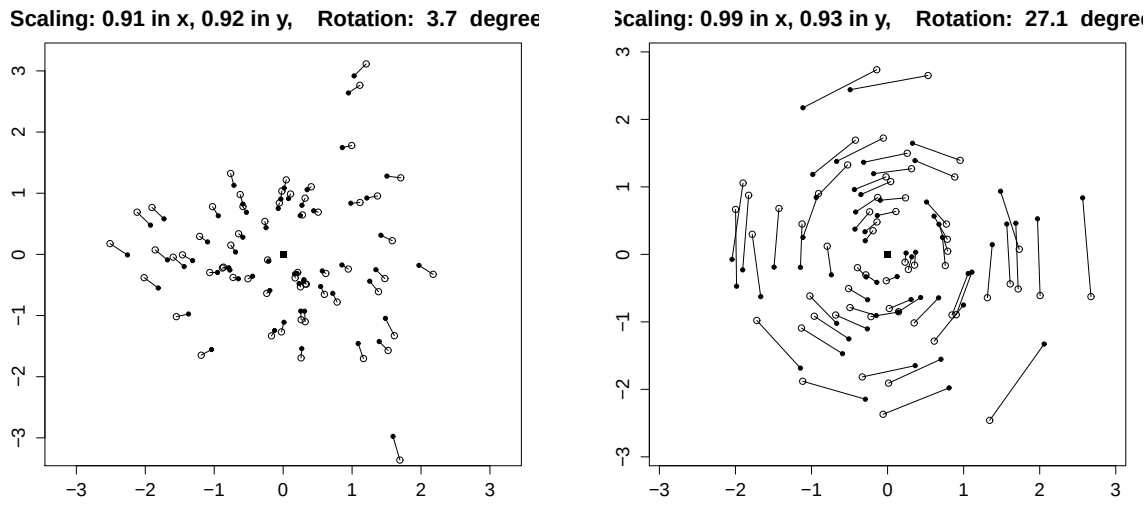

Figure 1: Two different realizations of the sample PC scores (black dots) compared to the population PC scores (circles). Panel a) displays a situation with mainly scaling and b) displays mainly rotation.

successful in practice. In this paper, we attempt to bridge this gap by showing that the relative positions, and thereby the visual content, of pairs of principal component scores are preserved if the corresponding eigenvectors are pervasive. For data situations where pervasive underlying variables are the main drivers of the variability, classical PCA will still give valid visual information regarding the population, despite of the eigenvectors being inconsistent. We also discuss data examples in genomics, such as genetic SNP markers and ethnicity, where pervasive signals can be considered as a reasonable assumption. In future work we will consider the implication of these results, when using principal component scores in further analyzes, such as regression and clustering.

References

Fan, J., Y. Liao, and M. Mincheva (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75*(4), 603–680.

Johnstone, I. M. and A. Y. Lu (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association 104*(486), 682–693.

Jung, S. and J. S. Marron (2009). PCA consistency in high dimension, low sample size context. *Annals of Statistics 37*(6B), 4104–4130.

Jung, S., A. Sen, and J. S. Marron (2012). Boundary behavior in high dimension, low sample size asymptotics of PCA. *Journal of Multivariate Analysis 109*(3), 190–203.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica 17*(4), 1617–1642.

Shen, D., H. Shen, H. Zhu, and J. S. Marron (2012). High dimensional principal component scores and data visualization. *arXiv preprint arXiv:1211.2679*.