Why principal component scores are a good exploratory tool for high-dimensional data

Kristoffer Hellton, Magne Thoresen

Department of Biostatistics, Institute of Basic Medical Sciences, University of Oslo k.h.hellton@medisin.uio.no

Aim

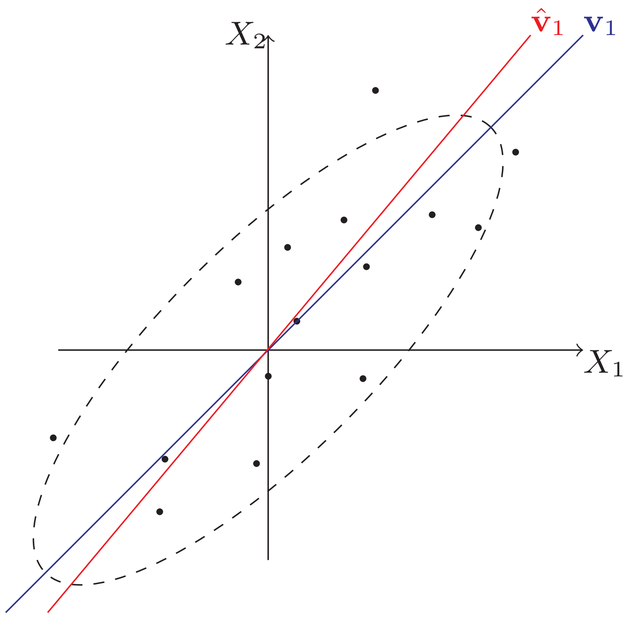
Principal component analysis (PCA) is a widely used method for reducing the dimension of high-dimensional data, even though the estimated eigenvectors are asymptotically inconsistent. We instead investigate the asymptotic behavior of the principal component scores to explain why PCA can still work as an exploratory tool.

Problem

PCA reduces a p-dimensional vector $\mathbf{x} = [x_1, \ldots, x_p]^T$ to a set of scores $\mathbf{z}_j$ by constructing linear combinations

$$\mathbf{z}_j = \mathbf{v}_j^T \mathbf{x} = v_{i1}x_1 + v_{i2}x_2 + \cdots + v_{ip}x_p,$$

where $\mathbf{v}_j$ are the eigenvectors of the covariance matrix Σ of $\mathbf{x}$. In practice, one uses the eigenvector $\hat{\mathbf{v}}_j$ of the empirical covariance matrix $\mathbf{S}$, usually referred to as the sample eigenvector. In the classical case ($p \ll n$), $\hat{\mathbf{v}}_j$ is a consistent estimator for $\mathbf{v}_j$.



However, in the high-dimensional setting, $p \gg n$, the sample eigenvectors $\hat{\mathbf{v}}_j$ do not in general converge to the population vectors $\mathbf{v}_j$, (Johnstone and Lu, 2009).

Assumptions

We investigate the asymptotic behavior of PC scores as n is fixed and $p \to \infty$, when the m leading eigenvalues scale linearly with the dimension:

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

a special case of the HDLSS setting. We also assume $\mathbf{Z}_j \sim \mathcal{N}(0, I)$ and $\lambda_j = \tau^2$ for $j = m+1, \ldots, p$, which can be generalized Jung *et al.* (2012). The situation was investigated by Shen *et al.* (2012) for $\lambda \sim p^\alpha, \alpha > 1$.

Data model and pervasiveness

The coefficients of the eigenvector $\mathbf{v} = [v_1, v_2, \ldots, v_p]^T$ can be interpreted as the effects of the latent factor z_i on the observed variable $\mathbf{x}$. To construct a data model which fulfills our asymptotic assumptions, we introduce the concept of pervasiveness: An eigenvector is pervasive, if the number of non-zero coefficients is asymptotically a non-zero *proportion* of the dimension. If an eigenvector is pervasive with fixed coefficients, the corresponding eigenvalue must scale linearly with the dimension:

$$\lambda_i \sim p.$$

Biological interpretation

From the area of genomics and genetic data, we have at least two situations where pervasive effects are reasonable from a biological perspective:

- Genetic markers, such as SNPs, from different populations where ethnicity is a latent factor.
- Microarray expression data from cancer cases and controls where disease status is a latent factor.

Asymptotic ratio

Under the specified assumptions, the ratio between the sample and population principal component scores is shown to converge to the following limit, as $p \to \infty$:

$$\frac{\hat{Z}_{ij}}{Z_{ij}} \to R_j + \varepsilon_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m,$$

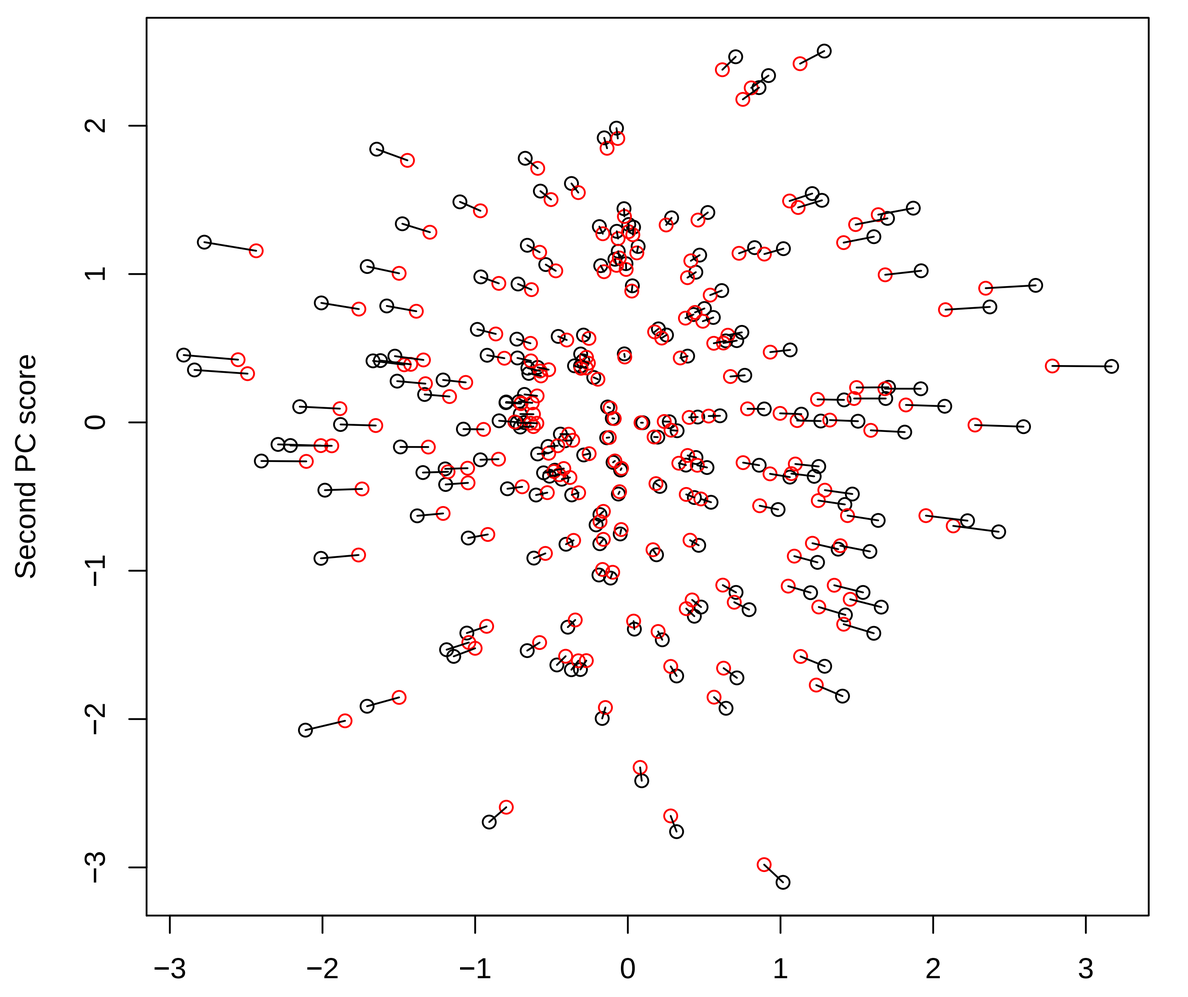
where the ratio R_j and ε_{ij} are distributed as

$$R_j \sim \sqrt{\frac{n}{\phi_j(\mathbf{W})}}\, \sigma_j v_{jj}(\mathbf{W}), \quad \varepsilon_{ij} \sim \sqrt{\frac{n}{\phi_j(\mathbf{W})}} \sum_{k=1, k\neq j}^{m} \sigma_k \frac{z_{ik}}{z_{ij}}\, v_{jk}(\mathbf{W}),$$

where $\mathbf{W} \sim Wishart(\mathrm{diag}(\sigma_1, \ldots, \sigma_m), n)$ with the stochastic eigenvectors and -values $\mathbf{v}_j(\mathbf{W})$ and $\lambda_j(\mathbf{W})$.

Simulations

Simulations show that for realistic parameters the last terms $\delta_{ij} + \varepsilon_{ij}$ will be small compared to R_j. As R_j is independent of observation index i, the ratio between population and sample scores will for each component be approximately equal for *all observations*. Visually, we see the difference only as a common scaling, where the relative positions of the scores will remain the same. In the following simulation, black circles are the true scores and red circles the estimated scores:



References

[1] Johnstone, Lu (2009) *On consistency and sparsity for principal component analysis*

[2] Jung, Sen, Marron (2012) *Boundary behavoir in high dimension, low sample size asymptotics for PCA*

[3] Shen, Shen, Zhu, Marron (2012) *High dimensional principal components scores and data visualization*

[4] K. Hellton, M. Thoresen (2013) *Asymptotic distribution of principal component scores connected to pervasive, high-dimensional eigenvectors*

Conclusion

We offer an explanation for the practical success of PCA with certain high-dimensional data: In situations with pervasive signals, the asymptotic inconsistency in eigenvectors is limited to a common scaling for the scores. This will conserve the relative positions and thereby the visual information of the population scores in the estimated scores.