 DOI: 10.1086/681083

Appendix A from J. G. Kingsolver et al., "Genetic Variation, Simplicity, and Evolutionary Constraints for Function-Valued Traits" (Am. Nat., vol. 185, no. 6, p. 000)

Simple Basis Analysis (SBA) and Principal Components Analysis (PCA)

The formation of the principal components (PCs) basis and the simple basis (SB) in k dimensions requires a $k \times k$ matrix that is nonnegative definite—that is, a matrix $\mathbf{\Lambda}$ with $\mathbf{v}'\mathbf{\Lambda}\mathbf{v}$ greater than or equal to 0 for any k-vector $\mathbf{v}$. In PCA, the matrix, denoted $\mathbf{G}$ instead of $\mathbf{\Lambda}$, is a covariance matrix that typically corresponds to some data set consisting of N vectors, $\mathbf{g}_1, \dots, \mathbf{g}_N$, each of length k. If the vector $\mathbf{v}$ is chosen to yield a large value of $\mathbf{v}'\mathbf{G}\mathbf{v}$, then the transformed/projected data values, $\mathbf{v}'\mathbf{g}_1, \dots, \mathbf{v}'\mathbf{g}_N$, will have a large variance. In SBA, the required $\mathbf{\Lambda}$ matrix is a simplicity matrix. A large value of the simplicity score $\mathbf{v}'\mathbf{\Lambda}\mathbf{v}$ indicates that $\mathbf{v}$ is simple.

Before explaining the general method for constructing the basis vectors in PCA and SBA, we will show how to write the simplicity measure in the article in terms of a matrix $\mathbf{\Lambda}$. In the article, we define the simplicity of a vector $\mathbf{v} = (v_1, \dots, v_k)'$ in terms of D on the basis of first divided differences:

$$D = \sum_{j=2}^{k} \frac{(v_j - v_{j-1})^2}{t_j - t_{j-1}}.$$

To write D in terms of a matrix, first define the $(k-1) \times k$ difference matrix $\mathcal{D}$

$$\mathcal{D} = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 \\ 0 & -1 & 1 & \cdots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -1 & 1 \end{bmatrix}$$

and the $(k-1) \times (k-1)$ diagonal matrix $\mathcal{W}$

$$\mathcal{W} = \mathrm{diag}\{(t_2 - t_1)^{-1}, (t_3 - t_2)^{-1}, \dots, (t_k - t_{k-1})^{-1}\}.$$

Then one easily verifies that $\mathcal{D}\mathbf{v} = (v_2 - v_1, v_3 - v_2, \dots, v_k - v_{k-1})'$ and

$$D = (\mathcal{D}\mathbf{v})'\mathcal{W}\mathcal{D}\mathbf{v} = \mathbf{v}'\mathcal{D}'\mathcal{W}\mathcal{D}\mathbf{v}.$$

We see that a large value of D means that $\mathbf{v}$ is complex, so we can consider D a complexity measure. We could proceed with a complexity measure, simply minimizing complexity instead of maximizing simplicity. However, we prefer using a simplicity measure, where large values of the measure mean that $\mathbf{v}$ is simple. To achieve this, we use a simplicity measure of the form $a - bD$, with positive a and b chosen to make $a\mathbf{v}'\mathbf{v} - bD$ "nice" in some way. In the article, we define our simplicity measure with $a = 4$ and $b = \min_j\{t_j - t_{j-1}\}$:

$$S = 4\mathbf{v}'\mathbf{v} - \min_j\{t_j - t_{j-1}\}D.$$

We can write S as $\mathbf{v}'\mathbf{\Lambda}\mathbf{v}$ using the $k \times k$ identity matrix $\mathbf{I}$:

$$\begin{aligned} S &= 4\mathbf{v}'\mathbf{I}\mathbf{v} - \min_j\{t_j - t_{j-1}\}\mathbf{v}'\mathcal{D}'\mathcal{W}\mathcal{D}\mathbf{v} \\ &= \mathbf{v}'[4\mathbf{I} - \min_j\{t_j - t_{j-1}\}\mathcal{D}'\mathcal{W}\mathcal{D}]\mathbf{v}. \end{aligned}$$

How do we choose a and b? The choice is just one of interpretability and can be left to the user. Certainly, we would want to choose a and b so that S cannot be negative. Here, we have chosen just such an a and b using a theorem of Schatzman (2002) that states that $\sum_{j=2}^{k}(v_j - v_{j-1})^2 \leq 4\mathbf{v}'\mathbf{v}$ for any vector $\mathbf{v}$. For discussion of other simplicity measures and a general way of choosing a and b, see Zhang et al. (2014).

Given a $\mathbf{G}$ matrix and a $\mathbf{\Lambda}$ matrix, both the PC basis vectors and the SB vectors are defined sequentially and can be computed by an appropriate eigenanalysis. The first vector $\mathbf{v}_1$ in the PC basis is defined as the vector of length 1 that maximizes $\mathbf{v}'\mathbf{G}\mathbf{v}$. We say that $\mathbf{v}_1$ points in the direction of maximum variability in the data vectors. The first vector $\mathbf{w}_1$ in the SB is defined as the vector of length 1 that maximizes $\mathbf{w}'\mathbf{\Lambda}\mathbf{w}$. We call $\mathbf{w}_1$ the simplest vector. The second PC basis vector, $\mathbf{v}_2$, is the vector that maximizes $\mathbf{v}'\mathbf{G}\mathbf{v}$ over all $\mathbf{v}$'s of length 1 that are perpendicular to $\mathbf{v}_1$. The third PC basis vector, $\mathbf{v}_3$, is the vector that maximizes $\mathbf{v}'\mathbf{G}\mathbf{v}$ over all $\mathbf{v}$'s of length 1 that are perpendicular to both $\mathbf{v}_1$ and $\mathbf{v}_2$. Similarly, the second SB vector $\mathbf{w}_2$ maximizes $\mathbf{w}'\mathbf{\Lambda}\mathbf{w}$ over all $\mathbf{w}$'s of length 1 that are perpendicular to $\mathbf{w}_1$. We say that $\mathbf{w}_2$ is the simplest vector perpendicular to $\mathbf{w}_1$. The construction of the set of basis vectors continues in this way.

The two resulting sets of k basis vectors are simply eigenvectors of the corresponding matrix ($\mathbf{G}$ or $\mathbf{\Lambda}$) and thus are easily computed. Furthermore, the eigenvalues of $\mathbf{G}$ are equal to the variances of the transformed data values, and the eigenvalues of $\mathbf{\Lambda}$ are equal to the simplicity measures of the SB vectors.

In addition to the PC basis and the SB, we consider a mixed basis consisting of the first m PC basis vectors along with the SB for the $(k-m)$–dimensional nearly null space—the space that is perpendicular to the first m PC basis vectors. We define the first SB vector as the simplest vector of length 1 in the nearly null space, that is, $\mathbf{w}_1$ maximizes $\mathbf{w}'\mathbf{\Lambda}\mathbf{w}$ over all vectors of length 1 that are perpendicular to the first m PCA basis vectors. The second SB vector is the simplest vector of length 1 that is perpendicular to the first m PCA basis vectors and to $\mathbf{w}_1$. The remaining SB vectors are defined similarly. The variance associated with the vector $\mathbf{w}$ is equal to $\mathbf{w}'\mathbf{G}\mathbf{w}$.

Again, the resulting set of $k-m$ SB vectors can be computed using an eigenanalysis, as follows. Let P be the $k \times (k-m)$ matrix with columns containing the last $k-m$ PC basis vectors. Then the SB of the nearly null space is $\mathbf{w}_1 = P\mathbf{u}_1, \dots, \mathbf{w}_{k-m} = P\mathbf{u}_{k-m}$, where $\mathbf{u}_1, \dots, \mathbf{u}_{k-m}$ are the eigenvectors of $P'\mathbf{\Lambda}P$. The simplicity scores of $P\mathbf{u}_1, \dots, P\mathbf{u}_{k-m}$ are the eigenvalues of $P\mathbf{\Lambda}P$.

Literature Cited Only in Appendix A

Schatzman, M. 2002. Numerical analysis: a mathematical introduction. Clarendon, Oxford.