Appendix B Geometrical interpretation of a principal component analysis.

Suppose we have a vector of two observations $\boldsymbol{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}$, for which $n = 100$ samples are taken. Now let us assume a linear relation between x_1 and x_2, with a small noise component added: $x_2 = x_1 + e$, where $e \sim \mathrm{N}(0, \sigma_e)$ is normally distributed with zero mean and standard deviation σ_e. This situation is depicted on the left pane of figure B1, with values for x_1 uniformly sampled between 0 and 10 and $\sigma_e = 0.5$. The matrix of loadings P is produced by the SVD algorithm, applied to the matrix X (equation (B1)). X is the matrix composed out of 100 rows, which are the independent observations for the vector $\begin{bmatrix} x_1 & x_2 \end{bmatrix}$. The diagonal matrix L is presented in equation (B2). The matrix U is left out due to its size.

$$\mathrm{P} = \begin{bmatrix} .7071 & -.7071 \\ .7071 & .7071 \end{bmatrix} \tag{B1}$$

$$\mathrm{L} = \begin{bmatrix} 83.51 & 0 \\ 0 & 3.60 \end{bmatrix} \tag{B2}$$

A closer study of P reveals that it is very similar to the rotation matrix of a counter-clockwise rotation over an angle $\theta = \pi/4$ in the cartesian x_1x_2-plane, as illustrated in equation (B3).

$$\mathrm{P} \approx \begin{bmatrix} \sqrt{2}/2 & -\sqrt{2}/2 \\ \sqrt{2}/2 & \sqrt{2}/2 \end{bmatrix} = \begin{bmatrix} \cos\pi/4 & -\sin\pi/4 \\ \sin\pi/4 & \cos\pi/4 \end{bmatrix} \tag{B3}$$

Effectively, the PCA decomposition finds the direction of maximal variance ($x_2 = x_1$ in this case, "Varimax rotation" [1]) and looks for an additional direction perpendicular to this. These directions are depicted as arrows on the left pane of figure B1. The matrix of scores T can now be interpreted as the two-dimensional coordinates of the points in this new rotated coordinate space, as displayed on the right pane of figure B1. It should be noted that the arrows representing the principal components are scaled in length to match the standard deviation explained by each, found from applying equation (A3) in Appendix A to the matrix L ($s_{\mathbf{t}_1} = 8.40$, $s_{\mathbf{t}_2} = 0.36$).

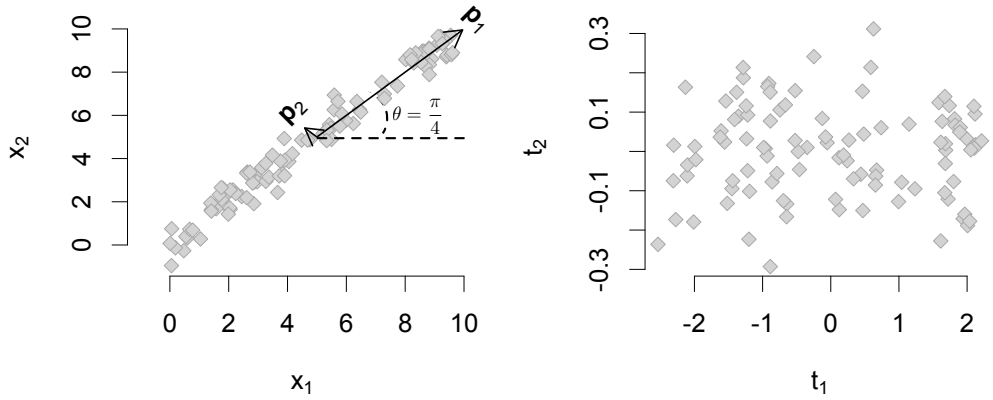

Figure B1: Geometrical interpretation of PCA

References

[1] H. F. Kaiser, The varimax criterion for analytic rotation in factor analysis, Psychometrika 23 (3) (1958) 187–200.