Appendix A Calculating principal components

In the very early years of the 20^{th} century, PCA was developed by Pearson [1]. Ever since, PCA has been a popular procedure to reduce the dimensionality of a variable space. For a full coverage on the basics of linear algebra, which is fundamental to PCA but lies outside the scope of this paper, the reader is referred to Jolliffe [2]. In this appendix, we will only introduce the properties of PCA using a graphical representation and give a brief overview of the matrix calculus that is required.

PCA assumes that the true rank of a matrix of observations X is less than the number of observations which are made. Consequently, it conjectures that the observations can be projected onto a new set of coordinate axes, thereby removing redundancy and noise from the system. The PCA decomposition method first calculates the principal components of the observation space, i.e. the directions that will make up the new coordinate axes. These principal components are often also named the latent variables, since they represent the underlying (unobservable) factors really influencing the system dynamics. Historically, the development of the Non Iterative Partial Least Squares (NIPALS) algorithm meant the real breakthrough for PCA. Consider a $(n \times P)$ matrix X that is a collection of n measurements for a P-variate randomly distributed variable $\mathbf{x}$. The first principal component of $\mathbf{x}$ is defined as the vector of coefficients $\mathbf{p}_1$ for which the linear combination $t_1 = \mathbf{x}\mathbf{p}_1$ captures as much as possible of the variance contained in X, subject to $|\mathbf{p}_1| = 1$. The second principal component is then the vector of coefficients $\mathbf{p}_2$ for which the linear combination $t_2 = \mathbf{x}\mathbf{p}_2$ contains as much of the variance from X that is not captured within t_1. Additional principal components up to P are similarly defined. Some of the literature on the NIPALS algorithm is worth a read as it can provide insight into the characteristics of PCA [3, 4]. In addition, Shlens [5] gives an intuitive explanation of how PCA is used to reduce the dimension of the problem.

In more recent work, PCA is always performed using the computationally more efficient singular value decomposition (SVD, [6]) of X. The SVD algorithm decomposes X into three matrices as depicted in equation (A1): a $(n \times P)$ matrix U, a $(P \times P)$ diagonal matrix L and a $(P \times P)$ matrix A.

$$X = \mathrm{ULA}^T \tag{A1}$$

The matrix U, whose columns contain the left singular vectors of X, is multiplied with the diagonal matrix L, containing the singular values of X, to form the matrix of scores T. The matrix A contains the right singular vectors of X and is equal to the matrix of loadings P. The PCA decomposition, illustrated in figure A1, can then be written as

$$X = \mathrm{TP}^T. \tag{A2}$$

The singular values on the diagonal of L are equal to the square roots of the eigenvalues ($\sqrt{l_i^2}, \forall i \in \{1, \dots, P\}$) of $X^T X$. Using these square roots, we can find the standard deviation of the i^{th} principal component $s_{\mathbf{t_i}}$ as follows:

$$s_{\mathbf{t_i}} = \sqrt{\frac{l_i^2}{n-1}} \tag{A3}$$

The score-loading nomenclature is very common in PCA literature and is therefore adopted here. The matrix of loadings $\mathrm{P} = \begin{bmatrix} \mathbf{p}_1 & \mathbf{p}_2 & \dots & \mathbf{p}_P \end{bmatrix}$ can be seen as a $(1 \times P)$ collection of $(P \times 1)$ vectors of coefficients for the linear combination $t_i = \mathbf{x}\mathbf{p}_i$ that defines the i^{th} principal component. A related term "matrix of rotations" expresses the geometrical interpretation of the loadings, as they represent a new coordinate space onto which $\mathbf{x}$ is projected. The $(n \times P)$ matrix of scores T can then be interpreted as the values in the new coordinate space for the collection of n measurements of $\mathbf{x}$. We illustrate this geometrical interpretation of PCA using a trivial example in appendix B. The reader is referred to [4, 5] for further illustrations. Such a geometrical interpretation can be particularly helpful to obtain insights in the relative importance of the different principal components, expressed by their standard deviations. We will now explore this further, while discussing whether all principal components should be retained in the analysis or not.

If all of the P principal components are retained for a further analysis of the data, an observation for the P-variate vector of observations $\mathbf{x}$ can be reconstructed from its $(1 \times P)$ vector of scores $\mathbf{t}$, as demonstrated in equation (A4) and depicted in figure A2.

$$\begin{aligned} \mathbf{x} &= \mathbf{t}\mathrm{P}^T \\ &= \sum_{i=1}^{P} t_i \mathbf{p}_i \end{aligned} \tag{A4}$$

However, since PCA is used to find a solution to the problems associated with the multivariate nature of project data, it will try to reduce the dimensionality of the problem in a structured

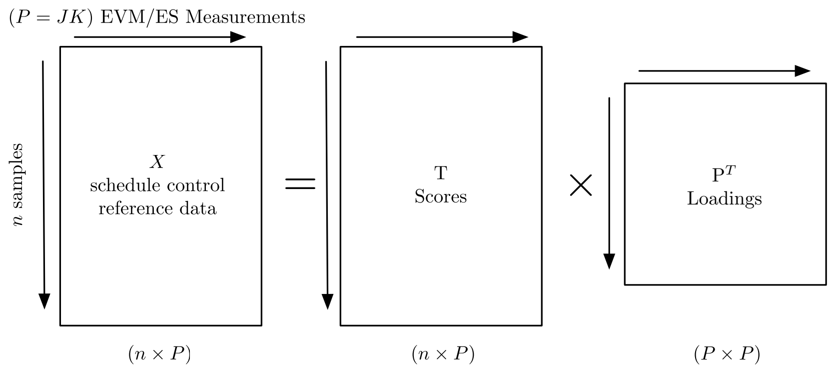

Figure A1: Principal Component Analysis as a data decomposition method ([7]).

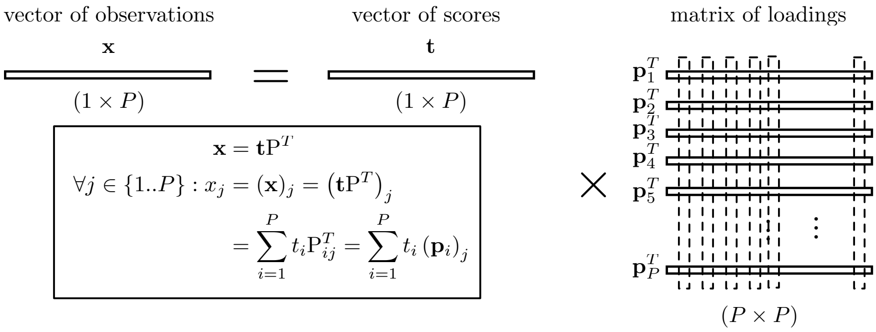

Figure A2: Principal Component Analysis interpretation for a vector of observations

manner, without losing valuable information. We assume that an integer k ($0 < k \leq P$) exists such that the last $P - k$ principal components do not represent valuable information for our system. By definition, each principal component will only explain a very small part of the original variation contained in X. Later in this paper we will test different choices for k principal components to retain in the analysis. For now we state that if only k principal components are retained, the original observation for the P-variate vector of observations $\mathbf{x}$ can be estimated as $\hat{\mathbf{x}}$ from its ($1 \times k$) vector of scores $\mathbf{t}$, as depicted in equation (A5) and figure A3.

$$\begin{aligned} \hat{\mathbf{x}} &= \mathbf{t}\mathrm{P}^T \\ &= \sum_{i=1}^{k} t_i \mathbf{p}_i \end{aligned} \tag{A5}$$

The matrix form for the collection of n observations for $\mathbf{x}$ in X is presented in equation (A6), where T_k is the ($n \times k$) matrix of scores, P_k is the ($P \times k$) matrix of loadings when only k principal components are retained and E_k is the ($n \times P$) error matrix.

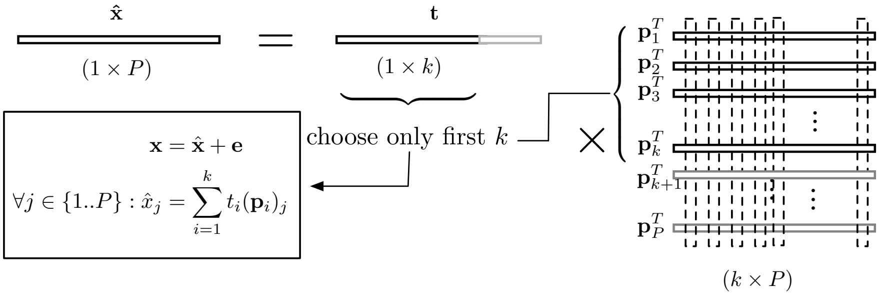

Figure A3: Interpretation for the PCA estimate of a vector of observations

$$X = \sum_{i=1}^{k} \mathrm{T}_k \mathrm{P}_k^T + \mathrm{E}_k \tag{A6}$$

E_k can be seen as the collection of error vectors $\mathbf{e}$, each corresponding to the vector $\mathbf{x}$ when only k principal components are retained:

$$\begin{aligned} \mathbf{e} &= \mathbf{x} - \hat{\mathbf{x}} \\ &= \sum_{i=k+1}^{P} t_i \mathbf{p}_i \end{aligned} \tag{A7}$$

References

[1] K. Pearson, Liii. on lines and planes of closest fit to systems of points in space, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2 (11) (1901) 559–572.

[2] I. Jolliffe, Principal component analysis, Wiley Online Library, 2005.

[3] T. Kourti, Application of latent variable methods to process control and multivariate statistical process control in industry, International Journal of adaptive control and signal processing 19 (4) (2005) 213–246.

[4] H. Abdi, Partial least squares regression and projection on latent structure regression (pls regression), Wiley Interdisciplinary Reviews: Computational Statistics 2 (1) (2010) 97–106.

[5] J. Shlens, A tutorial on principal component analysis, Systems Neurobiology Laboratory, University of California at San Diego.

[6] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammerling, A. McKenney, et al., LAPACK Users' guide, Vol. 9, Siam, 1999.

[7] J. F. MacGregor, Using on-line process data to improve quality: Challenges for statisticians*, International Statistical Review 65 (3) (1997) 309–323.