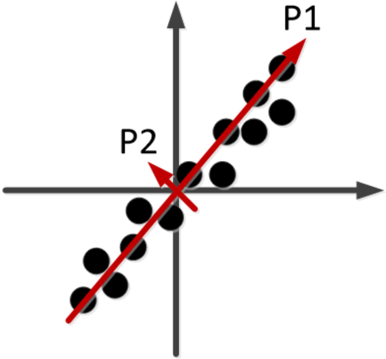
Principal Component Analysis

Department of Electrical & Computer Engineering,
University of Waterloo, ON, Canada

Data and Knowledge Modeling and Analysis (ECE 657A)

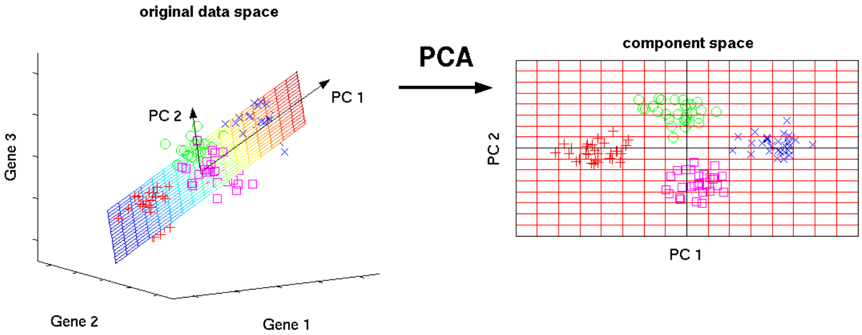
Course Instructor: Prof. Mark Crowley
TA and Presenter of Slides: Benyamin Ghojogh

Goal of PCA: Finding Most Variant Directions



PCA was first proposed in 1901 [1].

Goal of PCA: Finding Most Variant Directions



Credit of image: `http://www.nlpca.org/pca_principal_component_analysis.html`

Lecture Outline

Dataset Notations

$$\text{training dataset: } \{\boldsymbol{x}_i \in \mathbb{R}^d\}_{i=1}^n, \boldsymbol{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_n] \in \mathbb{R}^{d \times n} \tag{1}$$

$$\text{mean of training data: } \mathbb{R}^d \ni \boldsymbol{\mu}_x := \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \tag{2}$$

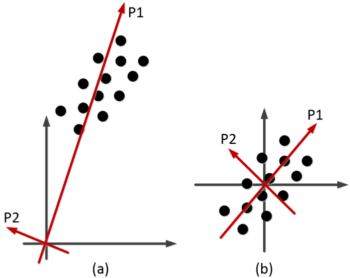
$$\text{test dataset: } \{\boldsymbol{x}_{t,i} \in \mathbb{R}^d\}_{i=1}^{n_t}, \boldsymbol{X}_t = [\boldsymbol{x}_{t,1}, \dots, \boldsymbol{x}_{t,n_t}] \in \mathbb{R}^{d \times n_t} \tag{3}$$

Data Centering?

$$\text{center the data: } \mathbb{R}^{d\times n} \ni \breve{\boldsymbol{X}} := \boldsymbol{X} - \boldsymbol{\mu}_x = \boldsymbol{X}\boldsymbol{H} = \boldsymbol{X} - \boldsymbol{\mu}_x \quad (4)$$

$$\text{centering matrix: } \mathbb{R}^{n\times n} \ni \boldsymbol{H} := \boldsymbol{I} - (1/n)11^\top \quad (5)$$

$$\text{center the test data: } \mathbb{R}^{d\times n_t} \ni \breve{\boldsymbol{X}}_t := \boldsymbol{X}_t - \boldsymbol{\mu}_x. \quad (6)$$



Projection & Reconstruction

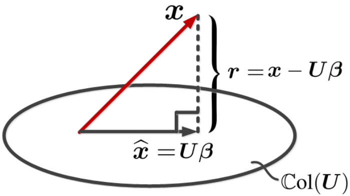
$$\text{projection: } \mathbb{R}^p \ni \widetilde{\boldsymbol{x}} := \boldsymbol{U}^\top \check{\boldsymbol{x}}, \widetilde{\boldsymbol{X}} = [\widetilde{\boldsymbol{x}}_1, \dots, \widetilde{\boldsymbol{x}}_n] \in \mathbb{R}^{p \times n} \tag{7}$$

$$\text{reconstruction: } \mathbb{R}^d \ni \widehat{\boldsymbol{x}} := \boldsymbol{U}\boldsymbol{U}^\top \check{\boldsymbol{x}} + \boldsymbol{\mu}_x = \boldsymbol{U}\widetilde{\boldsymbol{x}} + \boldsymbol{\mu}_x \tag{8}$$

$$\widehat{\boldsymbol{X}} = [\widehat{\boldsymbol{x}}_1, \dots, \widehat{\boldsymbol{x}}_n] \in \mathbb{R}^{d \times n}$$

$$\text{test projection: } \mathbb{R}^{p \times n_t} \ni \widetilde{\boldsymbol{X}}_t = \boldsymbol{U}^\top \check{\boldsymbol{X}}_t, \tag{9}$$

$$\text{test reconst.: } \mathbb{R}^{d \times n_t} \ni \widehat{\boldsymbol{X}}_t = \boldsymbol{U}\boldsymbol{U}^\top \check{\boldsymbol{X}}_t + \boldsymbol{\mu}_x = \boldsymbol{U}\widetilde{\boldsymbol{X}}_t + \boldsymbol{\mu}_x, \tag{10}$$



Formulation

$$\widehat{\boldsymbol{x}} = \boldsymbol{u}\boldsymbol{u}^\top \breve{\boldsymbol{x}} \tag{11}$$

$$\begin{aligned}
||\widehat{\boldsymbol{x}}||_2^2 &= ||\boldsymbol{u}\boldsymbol{u}^\top \breve{\boldsymbol{x}}||_2^2 = (\boldsymbol{u}\boldsymbol{u}^\top \breve{\boldsymbol{x}})^\top (\boldsymbol{u}\boldsymbol{u}^\top \breve{\boldsymbol{x}}) \\
&= \breve{\boldsymbol{x}}^\top \boldsymbol{u} \underbrace{\boldsymbol{u}^\top \boldsymbol{u}}_{1} \boldsymbol{u}^\top \breve{\boldsymbol{x}} = \breve{\boldsymbol{x}}^\top \boldsymbol{u}\, \boldsymbol{u}^\top \breve{\boldsymbol{x}} = \boldsymbol{u}^\top \breve{\boldsymbol{x}}\, \breve{\boldsymbol{x}}^\top \boldsymbol{u}
\end{aligned} \tag{12}$$

$$\sum_{i=1}^{n} ||\widehat{\boldsymbol{x}}_i||_2^2 = \sum_{i=1}^{n} \boldsymbol{u}^\top \breve{\boldsymbol{x}}_i\, \breve{\boldsymbol{x}}_i^\top \boldsymbol{u} = \boldsymbol{u}^\top \Big(\sum_{i=1}^{n} \breve{\boldsymbol{x}}_i\, \breve{\boldsymbol{x}}_i^\top \Big) \boldsymbol{u} \tag{13}$$

$$\begin{aligned}
\text{covariance matrix: } \mathbb{R}^{d\times d} \ni \boldsymbol{S} &:= \sum_{i=1}^{n} \breve{\boldsymbol{x}}_i\, \breve{\boldsymbol{x}}_i^\top = \breve{\boldsymbol{X}}\breve{\boldsymbol{X}}^\top \\
&= \boldsymbol{X}\boldsymbol{H}\boldsymbol{H}^\top \boldsymbol{X}^\top = \boldsymbol{X}\boldsymbol{H}\boldsymbol{H}\boldsymbol{X}^\top = \boldsymbol{X}\boldsymbol{H}\boldsymbol{X}^\top
\end{aligned} \tag{14}$$

Optimization

Hence:

$$\sum_{i=1}^{n} ||\widehat{\boldsymbol{x}}_i||_2^2 = \boldsymbol{u}^\top \boldsymbol{S} \boldsymbol{u} \tag{15}$$

PCA optimization (one projection direction):

$$\begin{aligned} \underset{\boldsymbol{u}}{\text{maximize}} \quad & \boldsymbol{u}^\top \boldsymbol{S} \boldsymbol{u}, \\ \text{subject to} \quad & \boldsymbol{u}^\top \boldsymbol{u} = 1, \end{aligned} \tag{16}$$

PCA optimization (multiple projection directions):

$$\begin{aligned} \underset{\boldsymbol{U}}{\text{maximize}} \quad & \textbf{tr}(\boldsymbol{U}^\top \boldsymbol{S}\, \boldsymbol{U}), \\ \text{subject to} \quad & \boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}, \end{aligned} \tag{17}$$

Optimization Solution

$$
\begin{aligned}
\underset{\boldsymbol{U}}{\text{maximize}} \quad & \mathbf{tr}(\boldsymbol{U}^\top \boldsymbol{S}\, \boldsymbol{U}), \\
\text{subject to} \quad & \boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I},
\end{aligned} \tag{18}
$$

$$
\mathcal{L} = \mathbf{tr}(\boldsymbol{U}^\top \boldsymbol{S}\, \boldsymbol{U}) - \mathbf{tr}\big(\Lambda^\top (\boldsymbol{U}^\top \boldsymbol{U} - \boldsymbol{I})\big) \tag{19}
$$

$$
\mathbb{R}^{d\times p} \ni \frac{\partial \mathcal{L}}{\partial \boldsymbol{U}} = 2\boldsymbol{S}\boldsymbol{U} - 2\boldsymbol{U}\Lambda \overset{\text{set}}{=} 0
$$

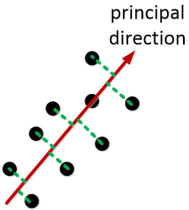
$$
\implies \boxed{\boldsymbol{S}\boldsymbol{U} = \boldsymbol{U}\Lambda} \tag{20}
$$

For maximization, we sort the directions (eigenvectors) from leading to tailing corresponding eigenvalues.
We truncate $\boldsymbol{U} \in \mathbb{R}^{d\times d}$ to $\boldsymbol{U} \in \mathbb{R}^{d\times p}$ (p top directions) where usually $p \ll d$.

Lecture Outline

Reconstruction Error

$$\boldsymbol{R} := \boldsymbol{X} - \widehat{\boldsymbol{X}} = \breve{\boldsymbol{X}} + \boldsymbol{\mu}_x - \boldsymbol{U}\boldsymbol{U}^\top \breve{\boldsymbol{X}} - \boldsymbol{\mu}_x = \breve{\boldsymbol{X}} - \boldsymbol{U}\boldsymbol{U}^\top \breve{\boldsymbol{X}}, \tag{21}$$



Minimization of reconstruction error:

$$\begin{aligned} \underset{\boldsymbol{U}}{\text{minimize}} \quad & ||\breve{\boldsymbol{X}} - \boldsymbol{U}\boldsymbol{U}^\top \breve{\boldsymbol{X}}||_F^2, \\ \text{subject to} \quad & \boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}. \end{aligned} \tag{22}$$

Optimization Solution

Minimization of reconstruction error:

$$\begin{aligned} \underset{\boldsymbol{U}}{\text{minimize}} \quad & ||\breve{\boldsymbol{X}} - \boldsymbol{U}\boldsymbol{U}^\top \breve{\boldsymbol{X}}||_F^2, \\ \text{subject to} \quad & \boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}. \end{aligned} \tag{23}$$

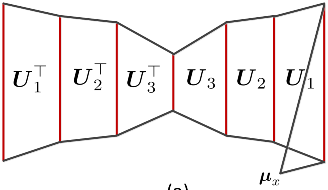
$$||\breve{\boldsymbol{X}} - \boldsymbol{U}\boldsymbol{U}^\top \breve{\boldsymbol{X}}||_F^2 = \textbf{tr}(\breve{\boldsymbol{X}}^\top \breve{\boldsymbol{X}}) - \textbf{tr}(\breve{\boldsymbol{X}}\breve{\boldsymbol{X}}^\top \boldsymbol{U}\boldsymbol{U}^\top). \tag{24}$$

$$\mathcal{L} = \textbf{tr}(\breve{\boldsymbol{X}}^\top \breve{\boldsymbol{X}}) - \textbf{tr}(\breve{\boldsymbol{X}}\breve{\boldsymbol{X}}^\top \boldsymbol{U}\boldsymbol{U}^\top) - \textbf{tr}\big(\Lambda^\top(\boldsymbol{U}^\top \boldsymbol{U} - \boldsymbol{I})\big), \tag{25}$$

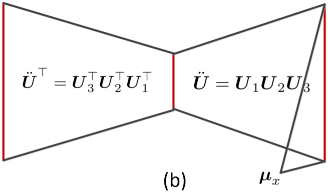
$$\begin{aligned} &\mathbb{R}^{d\times p} \ni \frac{\partial \mathcal{L}}{\partial \boldsymbol{U}} = 2\breve{\boldsymbol{X}}\breve{\boldsymbol{X}}^\top \boldsymbol{U} - 2\boldsymbol{U}\Lambda \overset{\text{set}}{=} 0 \implies \breve{\boldsymbol{X}}\breve{\boldsymbol{X}}^\top \boldsymbol{U} = \boldsymbol{U}\Lambda, \\ &\implies \boldsymbol{S}\boldsymbol{U} = \boldsymbol{U}\Lambda \end{aligned} \tag{26}$$

It is PCA solution!

PCA: Optimal Linear Reconstruction



(a)

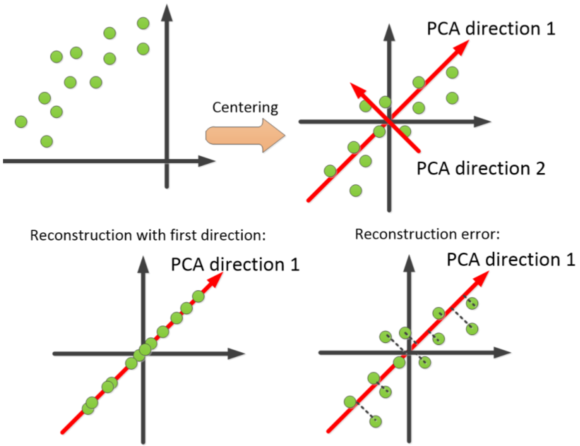


(b)

$$\widehat{\boldsymbol{X}} = \ddot{\boldsymbol{U}} \ddot{\boldsymbol{U}}^\top \boldsymbol{X} + \boldsymbol{\mu}_x \tag{27}$$

PCA: Rotation and Reconstruction

- If using all PCA directions, it acts as **rotation** of coordinates.
- If using some PCA directions (if we use PCA subspace), we will have a **reconstruction error**.



Lecture Outline

PCA: Optimal Linear Reconstruction

$$\text{SVD on centered data: } \mathbb{R}^{d\times n} \ni \breve{\boldsymbol{X}} = \boldsymbol{U}\Sigma\boldsymbol{V}^\top \tag{28}$$

$$\breve{\boldsymbol{X}}\breve{\boldsymbol{X}}^\top = \boldsymbol{U}\Sigma^2\boldsymbol{U}^\top \tag{29}$$

$$\breve{\boldsymbol{X}}^\top\breve{\boldsymbol{X}} = \boldsymbol{V}\Sigma^2\boldsymbol{V}^\top \tag{30}$$

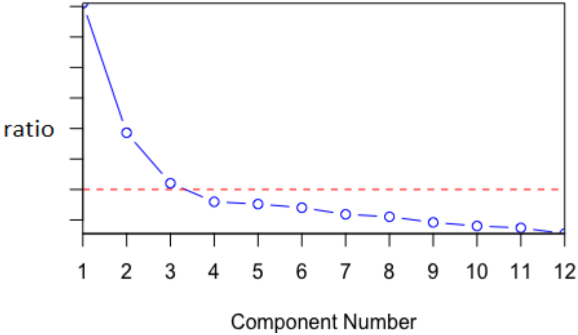
$$\boldsymbol{S} = \breve{\boldsymbol{X}}\breve{\boldsymbol{X}}^\top \implies \boldsymbol{U}: \text{ eigenvectors of covariance matrix} \tag{31}$$

So, we can apply SVD on the centered data and take the left singular vectors ($\boldsymbol{U}$) as the PCA projection matrix.

Lecture Outline

Scree Plot

$$\text{ratio} = \frac{\lambda_j}{\sum_{k=1}^{d} \lambda_k} \tag{32}$$



Lecture Outline

Training Projection & Reconstruction in Dual PCA

Projection in dual PCA:

$$\breve{\boldsymbol{X}} = \boldsymbol{U}\Sigma\boldsymbol{V}^\top \implies \widetilde{\boldsymbol{X}} = \boldsymbol{U}^\top\breve{\boldsymbol{X}} = \underbrace{\boldsymbol{U}^\top\boldsymbol{U}}_{\boldsymbol{I}}\Sigma\boldsymbol{V}^\top = \boxed{\Sigma\boldsymbol{V}^\top} \tag{33}$$

Reconstruction in dual PCA:

$$\breve{\boldsymbol{X}} = \boldsymbol{U}\Sigma\boldsymbol{V}^\top \implies \breve{\boldsymbol{X}}\boldsymbol{V} = \boldsymbol{U}\Sigma\underbrace{\boldsymbol{V}^\top\boldsymbol{V}}_{\boldsymbol{I}} = \boldsymbol{U}\Sigma \implies \boxed{\boldsymbol{U} = \breve{\boldsymbol{X}}\boldsymbol{V}\Sigma^{-1}} \tag{34}$$

$$\widehat{\boldsymbol{X}} = \boldsymbol{U}\widetilde{\boldsymbol{X}} + \boldsymbol{\mu}_x = \breve{\boldsymbol{X}}\boldsymbol{V}\Sigma^{-1}\widetilde{\boldsymbol{X}} + \boldsymbol{\mu}_x = \breve{\boldsymbol{X}}\boldsymbol{V}\underbrace{\Sigma^{-1}\Sigma}_{\boldsymbol{I}}\boldsymbol{V}^\top + \boldsymbol{\mu}_x$$

$$\implies \boxed{\widehat{\boldsymbol{X}} = \breve{\boldsymbol{X}}\boldsymbol{V}\boldsymbol{V}^\top + \boldsymbol{\mu}_x} \tag{35}$$

Test Projection & Reconstruction in Dual PCA

Test projection in dual PCA:

$$\begin{aligned} \boldsymbol{U} = \breve{\boldsymbol{X}}\boldsymbol{V}\Sigma^{-1} &\implies \boldsymbol{U}^\top = \Sigma^{-\top}\boldsymbol{V}^\top\breve{\boldsymbol{X}}^\top = \Sigma^{-1}\boldsymbol{V}^\top\breve{\boldsymbol{X}}^\top \\ &\implies \widetilde{\boldsymbol{x}}_t = \boldsymbol{U}^\top\breve{\boldsymbol{x}}_t = \boxed{\Sigma^{-1}\boldsymbol{V}^\top\breve{\boldsymbol{X}}^\top\breve{\boldsymbol{x}}_t} \end{aligned} \tag{36}$$

Test reconstruction in dual PCA:

$$\begin{aligned} \boldsymbol{U} = \breve{\boldsymbol{X}}\boldsymbol{V}\Sigma^{-1} &\implies \boldsymbol{U}\boldsymbol{U}^\top = \breve{\boldsymbol{X}}\boldsymbol{V}\Sigma^{-1}\Sigma^{-1}\boldsymbol{V}^\top\breve{\boldsymbol{X}}^\top \\ &\implies \boxed{\widehat{\boldsymbol{x}}_t = \breve{\boldsymbol{X}}\boldsymbol{V}\Sigma^{-2}\boldsymbol{V}^\top\breve{\boldsymbol{X}}^\top\breve{\boldsymbol{x}}_t + \boldsymbol{\mu}_x} \end{aligned} \tag{37}$$

Why Dual PCA?

1. The formulae for dual PCA only include $\boldsymbol{V}$ and not $\boldsymbol{U}$.
   - The columns of $\boldsymbol{V}$ are the eigenvectors of $\breve{\boldsymbol{X}}^\top \breve{\boldsymbol{X}} \in \mathbb{R}^{n \times n}$
   - The columns of $\boldsymbol{U}$ are the eigenvectors of $\breve{\boldsymbol{X}} \breve{\boldsymbol{X}}^\top \in \mathbb{R}^{d \times d}$
   - In case $n \ll d$, computation of eigenvectors of $\breve{\boldsymbol{X}}^\top \breve{\boldsymbol{X}}$ is easier and faster than $\breve{\boldsymbol{X}} \breve{\boldsymbol{X}}^\top$ and also requires less storage.
2. Some inner product forms, such as $\breve{\boldsymbol{X}}^\top \breve{\boldsymbol{x}}_t$, have appeared in the formulae of dual PCA.
   - It provides the opportunity for kernelizing PCA to have kernel PCA using the so-called kernel trick.

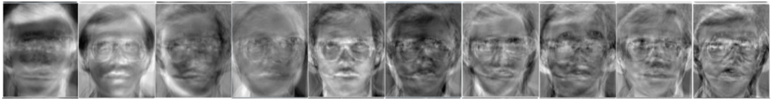
Lecture Outline

1. PCA Using Eigen-decomposition
2. Reconstruction Error in PCA
3. PCA Using Singular Value Decomposition
4. Finding the Number of Principal Directions
5. Dual PCA
6. Eigenfaces

Eigenfaces on ORL (or AT&T) Face Dataset

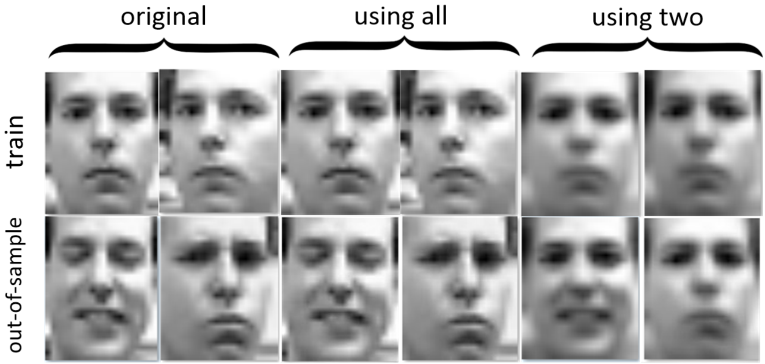
Some samples from the ORL face dataset:

Some eigenfaces [2, 3] on ORL face dataset:

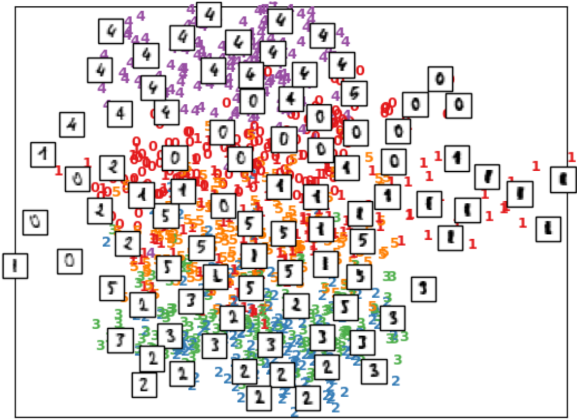


Reconstruction on Frey Face Dataset

Some reconstructions on the Frey face dataset:



Example: PCA on MNIST Data



The credit of this image is for: `https://scikit-learn.org/stable/auto_examples/manifold/plot_lle_digits.html`

Useful Resources To Read

- Tutorial paper: "Unsupervised and supervised principal component analysis: Tutorial" [4]
- Tutorial YouTube videos by Prof. Ali Ghodsi at University of Waterloo: [Click here] and [Click here]
- A book on PCA: [5]

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*, vol. 2, no. 11, pp. 559–572, 1901.

[2] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of cognitive neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[3] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on*, pp. 586–591, IEEE, 1991.

[4] B. Ghojogh and M. Crowley, "Unsupervised and supervised principal component analysis: Tutorial," *arXiv preprint arXiv:1906.03148*, 2019.

[5] I. Jolliffe, *Principal component analysis*.
Springer, 2011.