Some classical matrix factorisation techniques

Felix Held[1]

[1]Mathematical Sciences, University of Gothenburg and Chalmers University of Technology, Gothenburg, Sweden, *felix.held@chalmers.se*

Started: 2020-01-10, Last update: 2020–04–02

Contents

1 Introduction

The purpose of this document is to give an overview over classical/traditional matrix factorisation techniques and how they are interpreted. For a start I want to focus on the three classical techniques Principal Component Analysis (PCA), Independent Component Analysis (ICA), and Canonical Correlation Analysis (CCA). As a working tool,

the Rayleigh quotient and the Singular Value Decomposition (SVD) of a matrix will be introduced as well.

The exposition assumes that the reader has a background in multivariate calculus and basic multivariate statistics. The material presented here is somewhat standard and presented in many books and other material across the internet. Some of my sources were Falk et al. (2002), Härdle and Simar (2007) and Murphy (2012).

2 Rayleigh quotient

2.1 The basic Rayleigh quotient

The Rayleigh quotient for a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{p \times p}$ is a useful computational tool. It is defined for vectors $\mathbf{0} \neq \boldsymbol{x} \in \mathbb{R}^p$ as

$$J(\boldsymbol{x}) = \frac{\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}}. \tag{1}$$

Note that it is enough to normalize $\boldsymbol{x}$ and calculate the Rayleigh quotient for the normalized vector since

$$J(\boldsymbol{x}) = \frac{\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} = \frac{\|\boldsymbol{x}\|^2}{\|\boldsymbol{x}\|^2}\frac{\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}} = \frac{\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}^\top \boldsymbol{A}\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}}{\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}^\top \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}} = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}^\top \boldsymbol{A}\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}. \tag{2}$$

One says the Rayleigh quotient is scale-invariant. A common task is to find a $\widehat{\boldsymbol{x}}$ that maximizes the Rayleigh quotient, i.e.

$$\widehat{\boldsymbol{x}} \in \arg\max_{\boldsymbol{x} \neq \mathbf{0}} \frac{\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{x}}. \tag{3}$$

As noted above the Rayleigh quotient is scale-invariant. Therefore, for any solution $\widehat{\boldsymbol{x}}$ of the maximization problem in Eq. (3) the vector $c \cdot \widehat{\boldsymbol{x}}$ for arbitrary $c \in \mathbb{R}\backslash\{0\}$ is a solution as well. To restrict the space of possible solutions and to make the optimization problem identifiable we require that $\|\widehat{\boldsymbol{x}}\| = 1$. This results in the equality-constrained minimization problem

$$\min_{\boldsymbol{x} \neq \mathbf{0}} -\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} \quad \text{subject to} \quad \boldsymbol{x}^\top \boldsymbol{x} = 1. \tag{4}$$

The Lagrangian (Boyd and Vandenberghe, 2004, Section 5.1) of this constrained maximization problem is

$$L(\boldsymbol{x}, \lambda) = -\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} + \lambda(\boldsymbol{x}^\top \boldsymbol{x} - 1) \tag{5}$$

where λ is a Lagrange multiplier. The Lagrange dual function is therefore

$$g(\lambda) = \inf_{\boldsymbol{x} \neq \mathbf{0}} \boldsymbol{x}^\top(-\boldsymbol{A} + \lambda \boldsymbol{I}_p)\boldsymbol{x} - \lambda \tag{6}$$

For the infimum to be finite $-\boldsymbol{A} + \lambda \boldsymbol{I}_p$ has to be positive semi-definite (psd). If $-\boldsymbol{A} + \lambda \boldsymbol{I}_p$ has at least one negative eigenvalue μ with corresponding eigenvector $\boldsymbol{u}$, then

$c^2 \boldsymbol{u}^\top(-\boldsymbol{A} + \lambda \boldsymbol{I}_p)\boldsymbol{u} = c^2 \mu \|\boldsymbol{u}\|_2^2 \to -\infty$ for $c \to \infty$. This implies that the infimum is $-\infty$ in the non-psd case.

Since $\boldsymbol{A}$ is symmetric, let $\boldsymbol{U} \in \mathbb{R}^{p \times p}$ contain an orthonormal basis of eigenvectors of $\boldsymbol{A}$, with the i-th basis vector $\boldsymbol{u}_i$ in the i-th row of $\boldsymbol{U}$, and $\boldsymbol{D}$ the corresponding matrix of eigenvalues of $\boldsymbol{A}$. Then $\boldsymbol{A} = \boldsymbol{U}^\top \boldsymbol{D} \boldsymbol{U}$ and

$$\boldsymbol{x}^\top(-\boldsymbol{A} + \lambda \boldsymbol{I}_p)\boldsymbol{x} = \boldsymbol{x}^\top \boldsymbol{U}^\top(-\boldsymbol{D} + \lambda \boldsymbol{I}_p)\boldsymbol{U}\boldsymbol{x} = \sum_{i=1}^{p} (\lambda - \mu_i)(\boldsymbol{u}_i^\top \boldsymbol{x})^2. \tag{7}$$

This implies especially that the matrix $-\boldsymbol{A} + \lambda \boldsymbol{I}_p$ is psd if and only if $\lambda \geq \mu_{\max}(\boldsymbol{A})$ where $\mu_{\max}(\boldsymbol{A})$ is the largest eigenvalue of $\boldsymbol{A}$. Since every $\boldsymbol{x} = \sum_{i=1}^{p} a_i \boldsymbol{u}_i$ for some a_i it follows that $\sum_{i=1}^{p} (\lambda - \mu_i)(\boldsymbol{u}_i^\top c\boldsymbol{x})^2 = c^2 \sum_{i=1}^{p} (\lambda - \mu_i) a_i^2 \to 0$ for $c \to 0$. This implies that the infimum is $-\lambda$ for $\lambda \geq \mu_{\max}(\boldsymbol{A})$ in the psd case.

The Lagrange dual function for the minimization problem in Eq. (4) is therefore given by

$$g(\lambda) = \begin{cases} -\lambda & \text{if } \lambda \geq \mu_{\max}(\boldsymbol{A}) \\ -\infty & \text{otherwise} \end{cases} \tag{8}$$

which implies the dual problem

$$\sup_{\lambda} -\lambda \quad \text{subject to} \quad \lambda \geq \mu_{\max}(\boldsymbol{A}). \tag{9}$$

Clearly, this optimization problem is solved for $\widehat{\lambda} = \mu_{\max}(\boldsymbol{A})$.

It now follows that, since the optimal function value $d^* = -\mu_{\max}(\boldsymbol{A})$ of the dual problem is a lower bound on the optimal function value p^* of the primal problem (Eq. (4)), we get by choosing $\boldsymbol{u}$ as an unit eigenvector corresponding to $\mu_{\max}(\boldsymbol{A})$

$$-\mu_{\max}(\boldsymbol{A}) \leq p^* \leq -\boldsymbol{u}^\top \boldsymbol{A} \boldsymbol{u} = -\mu_{\max}(\boldsymbol{A}). \tag{10}$$

This shows that $\boldsymbol{u} \in \arg\max_{\boldsymbol{x} \neq \boldsymbol{0}} J(\boldsymbol{x})$. Note that $-\boldsymbol{u}$ is also a solution. In case the eigenvalue $\mu_{\max}(\boldsymbol{A})$ appears repeatedly, any $\boldsymbol{u}$ in its eigenspace with length $\|\boldsymbol{u}\| = 1$ will be a solution of the primal problem in Eq. (4).

A more hand-waving derivation is as follows: We want to find a maximum of $L(\boldsymbol{x}, \lambda)$ with respect to $\boldsymbol{x}$ and λ will be determined along the way. The gradient of $L(\boldsymbol{x}, \lambda)$ with respect to $\boldsymbol{x}$ is

$$\frac{\partial L(\boldsymbol{x}, \lambda)}{\partial \boldsymbol{x}} = -\boldsymbol{A}\boldsymbol{x} + \lambda \boldsymbol{x} \tag{11}$$

and setting Eq. (11) equal to $\boldsymbol{0}$ leads to

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}. \tag{12}$$

Since $\boldsymbol{x} \neq \boldsymbol{0}$ this is an eigenvalue equation and λ has to be one of p eigenvalues of $\boldsymbol{A}$. Using Eq. (12) in our original optimization problem Eq. (4) gives

$$\max_{\boldsymbol{x},\, \|\boldsymbol{x}\|=1} \boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x} = \max_{\substack{\boldsymbol{x},\, \|\boldsymbol{x}\|=1 \\ \boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}}} \lambda \boldsymbol{x}^\top \boldsymbol{x} = \max_{\substack{\boldsymbol{x},\, \|\boldsymbol{x}\|=1 \\ \boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}}} \lambda \tag{13}$$

The optimization problem is therefore solved by an eigenvector $\boldsymbol{x}$ of $\boldsymbol{A}$ with $\|\boldsymbol{x}\| = 1$ such that the corresponding eigenvalue λ is maximal among all eigenvalues of $\boldsymbol{A}$. Note that there are always two solutions to this problem. For every $\boldsymbol{x}$ with $\|\boldsymbol{x}\| = 1$ maximizing the Rayleigh quotient, the sign-flipped vector $-\boldsymbol{x}$ is a solution with $\|\boldsymbol{x}\| = 1$ as well.

Note that since $\boldsymbol{A}$ is real and symmetric, a theorem from linear algebra guarantees the existence of p real eigenvalues.

2.2 The generalised Rayleigh quotient

A variant of the Rayleigh quotient assumes that there are two symmetric matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{p\times p}$ where $\boldsymbol{B}$ is positive definite[1]. The Rayleigh quotient is then defined for $\boldsymbol{0} \neq \boldsymbol{x} \in \mathbb{R}^p$ as

$$J(\boldsymbol{x}) = \frac{\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x}}. \tag{14}$$

As above, $J(\boldsymbol{x})$ is invariant to scaling of $\boldsymbol{x}$ and to make the optimization problem uniquely solvable (up to inverting the direction of the solution over the origin) we need to fixate a scaling.

From linear algebra we know that there exists an orthogonal matrix[2] $\boldsymbol{U} \in \mathbb{R}^{p\times p}$ and a diagonal matrix $\boldsymbol{D} \in \mathbb{R}^{p\times p}$ such that $\boldsymbol{B} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top$. Define $\boldsymbol{D}^{1/2} := \mathrm{diag}(\sqrt{\boldsymbol{D}_{ii}}, i = 1, \ldots, p)$ and $\boldsymbol{B}^{1/2} := \boldsymbol{U}\boldsymbol{D}^{1/2}\boldsymbol{U}^\top$, then $\boldsymbol{B}^{1/2}\boldsymbol{B}^{1/2} = \boldsymbol{B}$[3]. By convention, the inverse of $\boldsymbol{B}^{1/2}$ is denoted as $\boldsymbol{B}^{-1/2}$.

To fixate a scaling for the input to the generalised Rayleigh quotient we require

$$\|\boldsymbol{B}^{1/2}\boldsymbol{x}\| = \boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x} = 1, \tag{15}$$

which leads to the optimization problem[4]

$$\max_{\boldsymbol{x}} \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} \quad \text{subject to} \quad \boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x} = 1. \tag{17}$$

Since $\boldsymbol{B}^{1/2}$ is of full-rank[5], we can make the change of variable $\boldsymbol{w} \leftarrow \boldsymbol{B}^{1/2}\boldsymbol{x}$ in Eq. (17), resulting in

$$\max_{\boldsymbol{w}} \boldsymbol{w}^\top \boldsymbol{B}^{-1/2}\boldsymbol{A}\boldsymbol{B}^{-1/2}\boldsymbol{w} \quad \text{subject to} \quad \boldsymbol{w}^\top \boldsymbol{w} = 1. \tag{18}$$

This is the optimization problem for the basic Rayleigh quotient with symmetric matrix $\boldsymbol{B}^{-1/2}\boldsymbol{A}\boldsymbol{B}^{-1/2}$. Any $\hat{\boldsymbol{w}}$ which is a unit eigenvector corresponding to $\mu_{\max}(\boldsymbol{B}^{-1/2}\boldsymbol{A}\boldsymbol{B}^{-1/2})$

[1] i.e. all eigenvalues are positive, $\boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x} > 0$ for all $\boldsymbol{x} \neq \boldsymbol{0}$ and $\boldsymbol{B}$ is invertible. Therefore, the numerator of the general Rayleigh quotient is non-zero and positive as long as $\boldsymbol{x} \neq \boldsymbol{0}$.

[2] $\boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_n$

[3] $\boldsymbol{B}^{1/2}$ is sometimes called the matrix square root of $\boldsymbol{B}$.

[4] Note that this is equivalent to solving

$$\max_{\boldsymbol{x}} \frac{\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^\top \boldsymbol{B}\boldsymbol{x}} \quad \text{subject to} \quad \boldsymbol{x}^\top \boldsymbol{x} = 1 \tag{16}$$

since for every solution $\hat{\boldsymbol{x}}$ to Eq. (16) the vector $\boldsymbol{z} = \hat{\boldsymbol{x}}/\|\boldsymbol{B}^{1/2}\hat{\boldsymbol{x}}\|$ also maximizes the Rayleigh quotient with $\boldsymbol{z}^\top \boldsymbol{B}\boldsymbol{z} = 1$ and is therefore a solution to Eq. (17). On the other hand, every solution $\hat{\boldsymbol{x}}$ to Eq. (17) induces a solution to Eq. (16) by setting $\hat{\boldsymbol{z}} = \hat{\boldsymbol{x}}/\|\hat{\boldsymbol{x}}\|$.

[5] Follows from $\boldsymbol{B}$ being psd.

is therefore a solution to Eq. (18). Corresponding solutions for Eq. (17) can be recovered by setting $\hat{\boldsymbol{x}} = \boldsymbol{B}^{-1/2}\hat{\boldsymbol{w}}$.

3 Principal component analysis (PCA)

When we have quantitative data, a natural choice of coordinate system is one where the axes point in the directions of largest variance and are orthogonal to each other. It is also natural to sort these in descending order, since most information will be gained by observing the most variable direction. Assume we have a data matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ with rows $\boldsymbol{x}_i^\top$. To determine the first principal component we are looking for a direction $\boldsymbol{r}_1$ (a unit vector, i.e. $\|\boldsymbol{r}_1\| = 1$) in which the variance of $\boldsymbol{X}$ is maximal. Define $s_i = \boldsymbol{r}_1^\top \boldsymbol{x}_i$, which is the coefficient of $\boldsymbol{x}_i$ projected onto $\boldsymbol{r}_1$. The variance in the direction of $\boldsymbol{r}_1$ is

$$\sum_{i=1}^{n}(s_i - \overline{s})^2 = \sum_{i=1}^{n}\left(\boldsymbol{r}_i^\top(\boldsymbol{x}_i - \overline{\boldsymbol{x}})\right)^2 \tag{19}$$

where $\overline{\boldsymbol{x}}$ is the mean over all observations. We want to find $\boldsymbol{r}_1$ such that the variance in Eq. (19) becomes maximal. Note that

$$\begin{aligned}\sum_{i=1}^{n}\left(\boldsymbol{r}^\top(\boldsymbol{x}_i - \overline{\boldsymbol{x}})\right)^2 &= \sum_{i=1}^{n}\boldsymbol{r}^\top(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^\top\boldsymbol{r}\\ &= \boldsymbol{r}^\top\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^\top\boldsymbol{r}\\ &= (n-1)\,\boldsymbol{r}^\top\hat{\boldsymbol{\Sigma}}\boldsymbol{r}\end{aligned} \tag{20}$$

where $\hat{\boldsymbol{\Sigma}}$ is the empirical covariance matrix of the data. Since $\hat{\boldsymbol{\Sigma}}$ is a symmetric matrix and it is required that $\|\boldsymbol{r}\| = 1$, maximizing Eq. (20) is equivalent to solving the basic Rayleigh quotient maximisation problem in Section 2. It therefore follows that $\boldsymbol{r}_1$ is an eigenvector of $\hat{\boldsymbol{\Sigma}}$ corresponding to its largest eigenvalue λ_1. Since we required $\boldsymbol{r}_1$ to be of length one, this problem is solved uniquely up to sign (i.e. $-\boldsymbol{r}_1$ is also a solution). Note especially that the variance of the s_i, that we tried to maximize in the original problem (Eq. (19)) is equal to λ_1.

Assume we have found the first $m - 1 < p$ principal components $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_{m-1}$ corresponding to the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_{m-1}$ of $\hat{\boldsymbol{\Sigma}}$. From linear algebra we know that a square matrix $\boldsymbol{P}$ is an orthogonal projection matrix if

$$\boldsymbol{P}^2 = \boldsymbol{P} = \boldsymbol{P}^\top. \tag{21}$$

Projecting a vector onto $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_{m-1}$ is accomplished by the orthogonal projection matrix

$$P = \sum_{i=1}^{m-1}\boldsymbol{r}_i\boldsymbol{r}_i^\top \tag{22}$$

Another property of orthogonal projection matrices is that $\boldsymbol{I} - \boldsymbol{P}$ is an orthogonal projection matrix onto the orthogonal complement of the subspace that $\boldsymbol{P}$ was projecting on, i.e. $\boldsymbol{I} - \boldsymbol{P}_1$ is a projection matrix onto the space of vectors which are orthogonal to $\boldsymbol{r}_1, \dots, \boldsymbol{r}_{m-1}$.

Project the data into this space, after having found the first $m-1$ principal components, i.e. define

$$\boldsymbol{X}_{m-1} = \boldsymbol{X}\left(\boldsymbol{I}_p - \sum_{i=1}^{m-1} \boldsymbol{r}_i \boldsymbol{r}_i^\top\right) \tag{23}$$

and

$$\widehat{\boldsymbol{\Sigma}}_{m-1} = \frac{\boldsymbol{X}_{m-1}^\top \boldsymbol{X}_{m-1}}{n-1} = \left(\boldsymbol{I}_p - \sum_{i=1}^{m-1} \boldsymbol{r}_i \boldsymbol{r}_i^\top\right)^\top \widehat{\boldsymbol{\Sigma}}\left(\boldsymbol{I}_p - \sum_{i=1}^{m-1} \boldsymbol{r}_i \boldsymbol{r}_i^\top\right). \tag{24}$$

The new data matrix is constant (no variance) along the directions of $\boldsymbol{r}_1, \dots, \boldsymbol{r}_{m-1}$. Finding the most variable direction now means to solve

$$\max_{\boldsymbol{r}} \boldsymbol{r}^\top \widehat{\boldsymbol{\Sigma}}_{m-1} \boldsymbol{r} \quad \text{subject to} \quad \boldsymbol{r}^\top \boldsymbol{r} = 1. \tag{25}$$

This is solved by an eigenvector $\boldsymbol{r}_m$ of $\widehat{\boldsymbol{\Sigma}}_{m-1}$ corresponding to its largest eigenvalue λ_m, i.e. $\widehat{\boldsymbol{\Sigma}}_{m-1}\boldsymbol{r}_m = \lambda_m \boldsymbol{r}_m$. It turns out that for $i = 1, \dots, m-1$

$$\begin{aligned} \boldsymbol{r}_m^\top \boldsymbol{r}_i &= \frac{1}{\lambda_m} \boldsymbol{r}_m^\top \widehat{\boldsymbol{\Sigma}}_{m-1} \boldsymbol{r}_i \\ &= \frac{1}{\lambda_m} \boldsymbol{r}_m^\top \left(\boldsymbol{I}_p - \sum_{j=1}^{m-1} \boldsymbol{r}_j \boldsymbol{r}_j^\top\right)^\top \widehat{\boldsymbol{\Sigma}} \underbrace{\left(\boldsymbol{I}_p - \sum_{j=1}^{m-1} \boldsymbol{r}_j \boldsymbol{r}_j^\top\right) \boldsymbol{r}_i}_{=\boldsymbol{0}} = 0 \end{aligned} \tag{26}$$

as well as

$$\begin{aligned} \lambda_m \boldsymbol{r}_m &= \widehat{\boldsymbol{\Sigma}}_{m-1} \boldsymbol{r}_m \\ &= \left(\boldsymbol{I}_p - \sum_{j=1}^{m-1} \boldsymbol{r}_j \boldsymbol{r}_j^\top\right)^\top \widehat{\boldsymbol{\Sigma}} \left(\boldsymbol{I}_p - \sum_{j=1}^{m-1} \boldsymbol{r}_j \boldsymbol{r}_j^\top\right) \boldsymbol{r}_m \\ &= \left(\boldsymbol{I}_p - \sum_{j=1}^{m-1} \boldsymbol{r}_j \boldsymbol{r}_j^\top\right) \widehat{\boldsymbol{\Sigma}} \boldsymbol{r}_m \\ &= \widehat{\boldsymbol{\Sigma}} \boldsymbol{r}_m - \sum_{j=1}^{m-1} \boldsymbol{r}_j \underbrace{\boldsymbol{r}_j^\top \widehat{\boldsymbol{\Sigma}} \boldsymbol{r}_m}_{=\lambda_j \boldsymbol{r}_j^\top \boldsymbol{r}_m = \boldsymbol{0}} \\ &= \widehat{\boldsymbol{\Sigma}} \boldsymbol{r}_m \end{aligned} \tag{27}$$

which shows that $\boldsymbol{r}_m$ is orthogonal to $\boldsymbol{r}_1, \dots, \boldsymbol{r}_{m-1}$ and an eigenvector of $\widehat{\boldsymbol{\Sigma}}$. In addition, since $\lambda_1 \geq \cdots \geq \lambda_{m-1}$ are the $m-1$ largest eigenvalues of $\widehat{\boldsymbol{\Sigma}}$, it must hold that $\lambda_{m-1} \geq \lambda_m$.

As a procedural way of calculating the PCA of a data matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, one has to follow the following steps

1. Centre and standardize the columns of the data matrix $\boldsymbol{X}$ (the variables)
2. Calculate the empirical covariance matrix

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\boldsymbol{X}^\top \boldsymbol{X} \tag{28}$$

3. Determine the eigenvalues λ_i for $i = 1, \dots, p$ of $\widehat{\boldsymbol{\Sigma}}$ and a set of p corresponding orthonormal eigenvectors $\boldsymbol{r}_i$ such that

$$\widehat{\boldsymbol{\Sigma}}\boldsymbol{r}_i = \lambda_i \boldsymbol{r}_i,\ \|\boldsymbol{r}_i\| = 1,\ i = 1, \dots, p \quad \text{and} \quad \boldsymbol{r}_i^\top \boldsymbol{r}_j = 0,\ i \neq j \tag{29}$$

as well as

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \tag{30}$$

4. Set

$$\boldsymbol{R} = \left(\boldsymbol{r}_1, \dots, \boldsymbol{r}_p\right) \in \mathbb{R}^{p\times p} \quad \text{and} \quad \boldsymbol{D} = \operatorname{diag}(\lambda_1, \dots, \lambda_p) \tag{31}$$

so that

$$\widehat{\boldsymbol{\Sigma}} = \boldsymbol{R}\boldsymbol{D}\boldsymbol{R}^\top \tag{32}$$

5. The vectors $\boldsymbol{r}_i$ are the principal component directions, the projections $\boldsymbol{r}_i^\top \boldsymbol{x}$ are called principal components and the corresponding eigenvalues λ_i are the variance of the data in the direction of the principal component.

PCA can be used to reduce the dimension of the data. Since the principal components account for less variance in every step, it is possible that there is little information in the last principal components. An important result from linear algebra is that for a matrix $\boldsymbol{A} \in \mathbb{R}^{p\times p}$ with eigenvalues $\mu_1, \dots, \mu_p$ it holds that

$$\operatorname{tr}(\boldsymbol{A}) = \sum_{i=1}^{p} \mu_i. \tag{33}$$

The empirical covariance matrix has the variance of each variable on its diagonal and therefore

$$\operatorname{tr}\left(\widehat{\boldsymbol{\Sigma}}\right) = \sum_{i=1}^{p} s^2(\boldsymbol{X}_{\cdot i}) = \sum_{i=1}^{p} \lambda_i, \tag{34}$$

where $\boldsymbol{X}_{\cdot i}$ is the i-th column of the data matrix and s^2 is the empirical variance. If the variables are standardized then $s^2(\boldsymbol{X}_{\cdot i}) = 1$ for all i and therefore $\operatorname{tr}\left(\widehat{\boldsymbol{\Sigma}}\right) = p$. This means in particular that the mean of the eigenvalues will be 1[6]. A typical criterion for considering a principal component as important is that the corresponding eigenvalue is larger than the mean of all eigenvalues (in case of standardised data: larger than one). A

[6] $\sum_{i=1}^{p} \lambda_i / p = 1$

tool for the analysis of the information behind the principal components is a *scree plot*. In a scree plot, the eigenvalues (variances) are plotted as a function of the index of the principal components. It is a way to quickly see how many principal components are of interest.

4 Singular value decomposition (SVD)

The singular value decomposition (SVD) of a matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, $n \geq p$, splits the data matrix into a product of three matrices

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top \tag{35}$$

where $\boldsymbol{U} \in \mathbb{R}^{n \times p}$ has orthonormal columns, $\boldsymbol{D} \in \mathbb{R}^{p \times p}$ is a diagonal matrix, and $\boldsymbol{V} \in \mathbb{R}^{p \times p}$ is an orthogonal matrix. Note that

$$\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}_p \quad \text{and} \quad \boldsymbol{V}\boldsymbol{V}^\top = \boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}_p \tag{36}$$

Note that if $n > p$ it cannot hold that $\boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{I}_n$ [7]. The matrix $\boldsymbol{D}$ is diagonal and contains the *singular values* d_i. These are typically sorted such that $d_{i+1} \leq d_i$.

The orthogonality properties in Eq. (36) can now be used to derive the following equations

$$\begin{aligned} \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{U} &= \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top \boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{U}\boldsymbol{D}^2 \\ \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{V} &= \boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\top \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{V}\boldsymbol{D}^2 \end{aligned} \tag{37}$$

Since $\boldsymbol{D}$ is a diagonal matrix, this reduces the problem of determining $\boldsymbol{U}$ and $\boldsymbol{V}$ to solving a series of eigenvalue problems

$$\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{u}_i = d_i^2 \boldsymbol{u}_i \quad \text{and} \quad \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{v}_i = d_i^2 \boldsymbol{v}_i, \quad i = 1, \ldots, p \tag{38}$$

Since $n \geq p$, the matrix $\boldsymbol{X}^\top \boldsymbol{X} \in \mathbb{R}^{p \times p}$ is as large as or smaller than $\boldsymbol{X}\boldsymbol{X}^\top \in \mathbb{R}^{n \times n}$. It is therefore more computationally effective to calculate $\boldsymbol{V}$ by solving the p eigenvalue problems in Eq. (38) and then arrive at $\boldsymbol{U}$ by projecting the observations in $\boldsymbol{X}$ on the space spanned by the columns of $\boldsymbol{V}$ and scaling them by the inverse of the singular values, i.e.

$$\boldsymbol{U} = \boldsymbol{X}\boldsymbol{V}\boldsymbol{D}^{-1}. \tag{39}$$

Note that this approach requires that there are no singular values equal to zero (which is allowed and possible in general).

For $n < p$ the SVD can still be calculated. Note that the SVD can be calculated as above for $\boldsymbol{X}^\top \in \mathbb{R}^{p \times n}$. We get matrices $\boldsymbol{V} \in \mathbb{R}^{p \times n}$ (orthonormal columns), a diagonal matrix $\boldsymbol{D} \in \mathbb{R}^{n \times n}$, and $\boldsymbol{U} \in \mathbb{R}^{n \times n}$ (orthogonal matrix) such that

$$\boldsymbol{X}^\top = \boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\top. \tag{40}$$

By transposing again we get

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top. \tag{41}$$

Note that this time $\boldsymbol{U}$ is square and $\boldsymbol{V}$ is rectangular.

[7] A set of maximally p vectors can be linearly independent in $\mathbb{R}^p$ and the equation above would imply that $n > p$ vectors in $\mathbb{R}^p$ are orthogonal and thus linearly independent.

4.1 SVD and PCA

The SVD of a data matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is usually used to calculate the principal components in the data. Assume that $n \geq p$ for now. Assume that the variables in $\boldsymbol{X}$ have been centred and scaled. Note that the empirical covariance matrix of $\boldsymbol{X}$ is

$$\widehat{\boldsymbol{\Sigma}} = \frac{\boldsymbol{X}^\top \boldsymbol{X}}{n-1}. \tag{42}$$

Using the SVD of $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$ this leads to

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top = \boldsymbol{V}\left(\frac{1}{n-1}\boldsymbol{D}^2\right)\boldsymbol{V}^\top. \tag{43}$$

Comparing this with an orthogonal decomposition of a symmetric real matrix, we get that the eigenvalues of $\widehat{\boldsymbol{\Sigma}}$ are on the diagonal of the matrix $\boldsymbol{D}^2/(n-1)$, and the columns of $\boldsymbol{V}$ are the corresponding eigenvectors. This determines the principal components and the corresponding variance explained by each component.

For $n < p$ the calculations above work out the same, however, $\boldsymbol{V}$ is not a square matrix any longer. The interpretation of eigenvalues and principal components still holds up since

$$\widehat{\boldsymbol{\Sigma}}\boldsymbol{V} = \boldsymbol{V}\left(\frac{1}{n-1}\boldsymbol{D}^2\right) \tag{44}$$

but note that the eigenvalues $\lambda_{n+1}, \ldots, \lambda_p$ are zero.

4.2 SVD and dimension reduction

When using SVD for dimension reduction, we want to reduce the number of variables. Therefore we want to find a subspace of $\mathbb{R}^p$ of dimension $q \leq \min(n, p)$ such that the projections of the observations into this subspace are as similar to the original observations in the Euclidean norm as possible, i.e. the objective is to find a subspace $\widehat{\mathcal{S}} \subset \mathbb{R}^p$, such that

$$\widehat{\mathcal{S}} = \arg\min_{\mathcal{S}} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{P}_{\mathcal{S}}\boldsymbol{x}_i\|_2^2 \tag{45}$$

where $\boldsymbol{P}_{\mathcal{S}}$ is the orthogonal projection of $\mathbb{R}^p$ onto the subspace $\mathcal{S}$. Note that

$$\begin{aligned}
\sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{P}_{\mathcal{S}}\boldsymbol{x}_i\|_2^2 &= \sum_{i=1}^{n} \left(\boldsymbol{x}_i - \boldsymbol{P}_{\mathcal{S}}\boldsymbol{x}_i\right)^\top \left(\boldsymbol{x}_i - \boldsymbol{P}_{\mathcal{S}}\boldsymbol{x}_i\right) \\
&= \sum_{i=1}^{n} \left[\boldsymbol{x}_i^\top\boldsymbol{x}_i - 2\boldsymbol{x}_i^\top\boldsymbol{P}_{\mathcal{S}}\boldsymbol{x}_i + \boldsymbol{x}_i^\top\boldsymbol{P}_{\mathcal{S}}^\top\boldsymbol{P}_{\mathcal{S}}\boldsymbol{x}_i\right] \\
&= \sum_{i=1}^{n} \left[\boldsymbol{x}_i^\top\boldsymbol{x}_i - \boldsymbol{x}_i^\top\boldsymbol{P}_{\mathcal{S}}^\top\boldsymbol{P}_{\mathcal{S}}\boldsymbol{x}_i\right] \\
&= \sum_{i=1}^{n} \left[\|\boldsymbol{x}_i\|_2^2 - \|\boldsymbol{P}_{\mathcal{S}}\boldsymbol{x}_i\|_2^2\right]
\end{aligned} \tag{46}$$

This shows that minimizing the squared distance between the observations and their projections on the subspace is equivalent to maximizing the squared length of the projected vectors[8]. Let's start by looking for the best one-dimensional subspace $\mathcal{S}_1$. Every one-dimensional vector space is spanned by a vector $\boldsymbol{r}_1$ and we can assume that $\|\boldsymbol{r}_1\|_2 = 1$. The projection matrix onto a one-dimensional vector space is $\boldsymbol{P} = \boldsymbol{r}_1\boldsymbol{r}_1^\top$ and thus we are trying to maximize

$$\begin{aligned}\sum_{i=1}^{n} \|\boldsymbol{r}_1\boldsymbol{r}_1^\top \boldsymbol{x}_i\|_2^2 &= \sum_{i=1}^{n} (\boldsymbol{r}_1^\top \boldsymbol{x}_i)^2 \\ &= \boldsymbol{r}_1^\top \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top \right) \boldsymbol{r}_1 \\ &= (n-1)\boldsymbol{r}_1^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{r}_1\end{aligned} \tag{47}$$

The term we are maximizing is the Rayleigh quotient and therefore the optimal subspace is spanned by $\boldsymbol{r}_1$, the eigenvector of $\hat{\boldsymbol{\Sigma}}$ with the largest corresponding eigenvalue. So the optimal one-dimensional sub-space that is closest to the data is spanned by the first principal component, which is also the first column of the matrix $\boldsymbol{V}$ in the SVD of $\boldsymbol{X}$.

Similar arguments as for PCA lead to the conclusion that the best q-dimensional subspace to approximate the data in is the space spanned by the first q principal component directions, which are also the first q columns in the matrix $\boldsymbol{V}$ in the SVD of $\boldsymbol{X}$. Note that for $n < p$ the data is maximally n dimensional and any approximating subspace must therefore have dimension $< n$.

4.3 SVD and orthogonal components

Another interpretation of SVD is that it describes a method to describe the data as a structure of orthogonal components. Let $\boldsymbol{u}_i \in \mathbb{R}^n$ be the columns of $\boldsymbol{U}$ and $\boldsymbol{v}_i \in \mathbb{R}^{\min(n,p)}$ the columns of $\boldsymbol{V}$. Then

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top = \sum_{i=1}^{\min(n,p)} d_i \boldsymbol{u}_i \boldsymbol{v}_i^\top. \tag{48}$$

Each matrix $\boldsymbol{u}_i\boldsymbol{v}_i^\top$ is of rank 1 and these matrices are scaled by the singular values d_i. As we have seen above in Section 4.2, the optimal $q \leq \min(n, p)$ subspace to approximate $\boldsymbol{X}$ is spanned by the first q columns of $\boldsymbol{V}$ and the projection on that subspace is

$$\boldsymbol{P}_q = \sum_{i=1}^{q} \boldsymbol{v}_i \boldsymbol{v}_i^\top. \tag{49}$$

[8] Since $\sum_{i=1}^{n} \|\boldsymbol{x}_i\|_2^2$ is a constant given a dataset. The only object we can control is the subspace we project into and therefore we can only change $\sum_{i=1}^{n} \|\boldsymbol{P}_{\mathcal{S}}\boldsymbol{x}_i\|_2^2$.

Projecting $\boldsymbol{X}$ on this optimal q-dimensional subspace using the projection in Eq. (49) and the representation in Eq. (48) leads to

$$\boldsymbol{X}_q = \boldsymbol{X}\boldsymbol{P}_q = \left(\sum_{i=1}^{\min(n,p)} d_i \boldsymbol{u}_i \boldsymbol{v}_i^\top\right)\left(\sum_{j=1}^{q} \boldsymbol{v}_j \boldsymbol{v}_j^\top\right) = \sum_{i=1}^{q} d_i \boldsymbol{u}_i \boldsymbol{v}_i^\top. \tag{50}$$

So projecting the data into this optimal q-dimensional subspace simply means to only keep components 1 to q. The approximation error in Frobenius norm[9] is

$$\begin{aligned}
\|\boldsymbol{X} - \boldsymbol{X}_q\|_F^2 &= \left\|\sum_{i=q+1}^{\min(n,p)} d_i \boldsymbol{u}_i \boldsymbol{v}_i^\top\right\|_F^2 \\
&= \operatorname{tr}\left[\left(\sum_{i=q+1}^{\min(n,p)} d_i \boldsymbol{u}_i \boldsymbol{v}_i^\top\right)^\top \left(\sum_{j=q+1}^{\min(n,p)} d_j \boldsymbol{u}_j \boldsymbol{v}_j^\top\right)\right] \\
&= \operatorname{tr}\left[\sum_{i,j=q+1}^{\min(n,p)} d_i d_j \boldsymbol{v}_i \boldsymbol{u}_i^\top \boldsymbol{u}_i \boldsymbol{v}_j^\top\right] \\
&= \operatorname{tr}\left[\sum_{i,j=q+1}^{\min(n,p)} d_i d_j \boldsymbol{v}_i \boldsymbol{v}_j^\top\right] \\
&= \sum_{k=1}^{\min(n,p)} \sum_{i,j=q+1}^{\min(n,p)} d_i d_j v_{ik} v_{jk} \\
&= \sum_{i,j=q+1}^{\min(n,p)} d_i d_j \sum_{k=1}^{\min(n,p)} v_{ik} v_{jk} \\
&= \sum_{i,j=q+1}^{\min(n,p)} d_i d_j \, 1(i=j) = \sum_{i=q+1}^{\min(n,p)} d_i^2
\end{aligned} \tag{51}$$

4.4 SVD and regression

In addition to dimension reduction and easy determination of the optimal q-dimensional approximation to the data, SVD can also be a useful tool in regression.

Recall the linear regression problem with response vector $\boldsymbol{y} \in \mathbb{R}^n$ and design matrix $\boldsymbol{X} \in \mathbb{R}^{n\times(p+1)}$, where we adopt the convention that the first column of $\boldsymbol{X}$ is a vector of 1's to encode the intercept. The variable y is modelled as

$$y = \boldsymbol{x}^\top \boldsymbol{\beta} + \varepsilon \tag{52}$$

where $\boldsymbol{x} \in \mathbb{R}^{p+1}$ is a vector of predictors with $x_1 = 1$ and $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ are the regression coefficients. The variable ε is the error and typically one assumes that

$$\varepsilon \sim \text{Normal}(0, \sigma^2) \tag{53}$$

[9] This is a matrix norm defined by $\|\boldsymbol{X}\|_F^2 = \sum_{i=1}^n \sum_{j=1}^n x_{ij}^2$

for some (possibly unknown) variance σ^2. A solution to the regression model can be found with least squares, i.e. solving

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \tag{54}$$

The solution to the least squares problem is given by

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}, \tag{55}$$

which requires $n \geq p$ for the inversion to be possible. The SVD of $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$ can be plugged into this equation to arrive at

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \left(\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top\right)^{-1} \boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\top\boldsymbol{y} \\
&= \left(\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^\top\right)^{-1} \boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\top\boldsymbol{y} \\
&= \boldsymbol{V}\boldsymbol{D}^{-2}\boldsymbol{V}^\top\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\top\boldsymbol{y} \\
&= \boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{U}^\top\boldsymbol{y}
\end{aligned} \tag{56}$$

The expression

$$\boldsymbol{X}^+ = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top = \boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{U}^\top \tag{57}$$

is called the Moore-Penrose pseudo-inverse of $\boldsymbol{X}$. While it is therefore possible to obtain a least squares solution through SVD, there are other simpler algorithms (e.g. QR decomposition) which are preferable.

The fitted values for $\boldsymbol{y}$ are then

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{U}^\top\boldsymbol{y} = \boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{y}. \tag{58}$$

Note that $\boldsymbol{U}\boldsymbol{U}^\top$ is an orthogonal projection matrix, projecting $\boldsymbol{y}$ onto the column space of $\boldsymbol{U}$.

For ridge regression the problem to be solved is

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{p}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2 \tag{59}$$

for some $\lambda \geq 0$. Assume that the response $\boldsymbol{y}$ and the columns of $\boldsymbol{X}$ are centred. The solution is given by

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_p\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}. \tag{60}$$

Using the SVD of $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$ leads to

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{V}\left(\boldsymbol{D}^2 + \lambda \boldsymbol{I}_p\right)^{-1} \boldsymbol{D}\boldsymbol{U}^\top\boldsymbol{y} = \sum_{i=1}^{p} \frac{d_i}{d_i^2 + \lambda}\boldsymbol{v}_i\boldsymbol{u}_i^\top\boldsymbol{y}. \tag{61}$$

It can be seen that λ can lead to stability in the calculation of the fractions $d_i/(d_i^2 + \lambda)$. If d_i is small, the fraction would become big if $\lambda = 0$. Increasing lambda decreases the magnitude of the fractions and increases numerical stability. Also, an increase in λ decreases the influence of each term in the same and therefore shrinks coefficients towards 0.

5 Factor Analysis (FA)

Factor Analysis (FA) is a simple model that helps in the exploratory data analysis of measurements with many variables. The idea is that there are underlying unobserved factors that can explain the other variables. A major driving force behind the development of FA was to understand and make measurable abstract properties such as intelligence in psychology.

The traditional model description of FA is a linear model

$$\boldsymbol{x}_i = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{z}_i + \boldsymbol{e}_i, \quad i = 1, \dots, n, \tag{62}$$

where

- $\boldsymbol{x}_i \in \mathbb{R}^p$ is the observed data,
- $\boldsymbol{\mu} \in \mathbb{R}^p$ are the mean values of each variable,
- $\boldsymbol{L} \in \mathbb{R}^{p \times k}$ is called the factor loading matrix describing how much each of the k factors contributes to the description of the p variables,
- $\boldsymbol{z}_i \in \mathbb{R}^k$ are the latent scores, describing the extent to which each p-dimensional factor describes observation $\boldsymbol{x}_i$,
- and $\boldsymbol{e}_i$ are residual errors.

Note that the mean values and the factor loading matrix are constant across observations. Denote $\boldsymbol{Z} = (\boldsymbol{z}_1 \dots \boldsymbol{z}_n)$ and $\boldsymbol{E} = (\boldsymbol{e}_1 \dots \boldsymbol{e}_n)$. Typical assumptions are

- $\boldsymbol{Z}$ and $\boldsymbol{E}$ are independent,
- $\mathbb{E}[\boldsymbol{e}_i] = \boldsymbol{0}$ for all i and $\mathrm{Cov}[\boldsymbol{E}] = \boldsymbol{\Psi} \in \mathbb{R}^{p \times p}$ is diagonal,
- and $\mathbb{E}[\boldsymbol{z}_i] = \boldsymbol{0}$ for all i and $\mathrm{Cov}[\boldsymbol{Z}] = \boldsymbol{I}_k$.

The first assumption ensures that all structural information is extracted and that $\boldsymbol{E}$ represents the residual error, unexplained by the model. The second assumption, enforces the interpretation of $\boldsymbol{E}$ as the residual error, especially by ensuring that the covariance is diagonal. This means that any covariance information is supposed to be explained by the factors. The third assumption is mostly for notational convenience. If the mean of the latent scores were not zero, $\boldsymbol{\mu}_0$ say, then $\boldsymbol{L}\boldsymbol{\mu}_0$ could be absorbed into $\boldsymbol{\mu}$. Similarly for the covariance matrix of $\boldsymbol{Z}$. If $\mathrm{Cov}(\boldsymbol{Z}) = \boldsymbol{L}_0$ then setting $\widetilde{\boldsymbol{L}} = \boldsymbol{L}\boldsymbol{L}_0^{-1/2}$ would lead to an equivalent model with unit covariance matrix (see also Murphy (2012, Section 12.1.1)).

Denoting $\boldsymbol{X} = (\boldsymbol{x}_1 \dots \boldsymbol{x}_n)$, the assumptions above ensure that

$$\mathbb{E}(\boldsymbol{x}_i) = \boldsymbol{\mu} \tag{63}$$

$$\mathrm{Cov}(\boldsymbol{X}) = \mathrm{Cov}(\boldsymbol{L}\boldsymbol{Z} + \boldsymbol{E}) = \boldsymbol{L}\boldsymbol{L}^\top + \boldsymbol{\Psi} \tag{64}$$

where in the last equation we used the independence of $\boldsymbol{Z}$ and $\boldsymbol{E}$ as well as the assumptions on the covariances. This shows that factor analysis provides a low-rank parametrisation[10] of the covariance of the variables in $\boldsymbol{X}$. Instead of p^2 parameters only $pk + p$ are needed. Depending on the relative sizes of p and k this can lead to a substantial reduction in the number of necessary parameters.

[10] The rank of $\boldsymbol{L}$ and therefore $\boldsymbol{L}\boldsymbol{L}^\top$ is at most k, if $k < p$.

Note that from a probabilistic perspective, the model and assumptions above are equivalent to the hierarchical model

$$\begin{aligned} \boldsymbol{z}_i &\sim \text{Normal}(\boldsymbol{0}, \boldsymbol{I}_k) \\ \boldsymbol{x}_i|\boldsymbol{z}_i &\sim \text{Normal}(\boldsymbol{L}\boldsymbol{z}_i + \boldsymbol{\mu}, \boldsymbol{\Psi}) \end{aligned} \tag{65}$$

Computing the marginal distribution of $\boldsymbol{x}_i$ reveals the low-rank covariance parametrisation above, i.e.

$$\boldsymbol{x}_i \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{L}\boldsymbol{L}^\top + \boldsymbol{\Psi}). \tag{66}$$

As mentioned in the beginning, a major driving force in the advent of FA was to find interpretable latent factors that can explain the observed variables. The posterior of $\boldsymbol{z}_i|\boldsymbol{x}_i$ is therefore of interest. Using Bayes rule together with the model in Eq. (65) we can compute[11]

$$\boldsymbol{z}_i|\boldsymbol{x}_i \sim \text{Normal}(\boldsymbol{m}_i, \boldsymbol{\Sigma}) \tag{67}$$

where $\boldsymbol{\Sigma} = (\boldsymbol{I}_k + \boldsymbol{L}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{L})^{-1}$ and $\boldsymbol{m}_i = \boldsymbol{\Sigma}\boldsymbol{L}^\top\boldsymbol{\Psi}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})$. Note that a priori, neither $\boldsymbol{L}$, $\boldsymbol{\Psi}$, nor $\boldsymbol{\mu}$ are known.

6 Probabilistic Principal Component Analysis (PPCA)

Probabilistic Principal Component Analysis (PPCA) is a variant of FA and is related to standard PCA described above. Setting $\boldsymbol{\Psi} = \sigma^2\boldsymbol{I}_p$ and choosing k as the number of principal components to retain, the model becomes very similar to PCA with added homogeneous independent Gaussian noise. Note that the predictive distribution for the latent factors can be written as

$$\boldsymbol{z}_i|\boldsymbol{x}_i \sim \text{Normal}(\boldsymbol{M}^{-1}\boldsymbol{L}^\top(\boldsymbol{x}_i - \boldsymbol{\mu}), \sigma^2\boldsymbol{M}^{-1}) \tag{68}$$

where $\boldsymbol{M} = \sigma^2\boldsymbol{I}_k + \boldsymbol{L}^\top\boldsymbol{L}$. It is easily seen that when $\sigma \to 0$ then the PPCA model becomes the standard PCA model, since the covariance vanishes [12] and the mean becomes $(\boldsymbol{L}^\top\boldsymbol{L})^{-1}\boldsymbol{L}^\top(\boldsymbol{x}_i - \boldsymbol{\mu})$ which is the orthogonal projection onto the column space of $\boldsymbol{L}$.

It can be shown that the maximum likelihood estimators of $\boldsymbol{L}$ and σ^2 recover the usual interpretation of PCA, in the sense that if $\boldsymbol{L} \in \mathbb{R}^{n\times k}$ then $\boldsymbol{L}$ will be estimated to a

[11] Using Bayes rule and calculating on the log-scale

$$\begin{aligned} \log p(\boldsymbol{z}_i|\boldsymbol{x}_i) &= \log p(\boldsymbol{x}_i|\boldsymbol{z}_i) + \log p(\boldsymbol{z}_i) + \text{const} \\ &= \log \text{Normal}(\boldsymbol{x}_i; \boldsymbol{L}\boldsymbol{z}_i + \boldsymbol{\mu}, \boldsymbol{\Psi}) + \log \text{Normal}(\boldsymbol{z}_i; \boldsymbol{0}, \boldsymbol{I}_k) + \text{const} \\ &= -\boldsymbol{z}_i\boldsymbol{L}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{L}\boldsymbol{z}_i + 2(\boldsymbol{x}_i - \boldsymbol{\mu})^\top\boldsymbol{\Psi}^{-1}\boldsymbol{L}\boldsymbol{z}_i - \boldsymbol{z}_i^\top\boldsymbol{z}_i + \text{const} \\ &= -\boldsymbol{z}_i\left(\boldsymbol{I}_k + \boldsymbol{L}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{L}\right)\boldsymbol{z}_i \\ &\quad + 2\left(\left(\boldsymbol{I}_k + \boldsymbol{L}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{L}\right)^{-1}\boldsymbol{L}^\top\boldsymbol{\Psi}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})\right)^\top\left(\boldsymbol{I}_k + \boldsymbol{L}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{L}\right)\boldsymbol{z}_i + \text{const} \end{aligned}$$

where the symmetry of $\boldsymbol{\Psi}$ was used.

[12] i.e. the model becomes degenerate/singular

slightly penalised version of the first k principal components and σ^2 is the average of the remaining singular values.

There are multiple advantages to the probabilistic formulation of PCA. One advantage of the probabilistic formulation of PCA is that it allows for data driven selection of the number of latent components to retain. One possible solution for this is Automatic Relevance Determination (ARD). Another advantage is that it is easier to adapt PCA to other data types than continuous data as well as the incorporation of missing data (Roweis, 1998; Tipping and Bishop, 1999).

7 Canonical Correlation Analysis (CCA)

The previous methods focused only on one dataset and projections into lower dimensions that emphasize certain aspects (e.g. variance or correlations) of the data. When two datasets are analysed jointly, one question that can be asked is how to project each dataset into a lower-dimensional space such that the most similar directions in each dataset are uncovered.

The traditional view/interpretation of CCA is as follows: Given two samples of (possibly different) features, each sample vector of dimension p and q, respectively, find all pairs of directions in the respective feature space (i.e. linear combinations of the feature vectors) that are maximally correlated and orthogonal to all previous found pairs.

This can be formalised as follows. Assume that there are data matrices $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ and $\boldsymbol{Y} \in \mathbb{R}^{n\times q}$. The task of CCA in its first step is to find $\boldsymbol{a} \in \mathbb{R}^p$ and $\boldsymbol{b} \in \mathbb{R}^q$ such that the sample correlation between vectors $\boldsymbol{Xa}$ and $\boldsymbol{Yb}$ is maximal. Let $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{XY}}$ be the cross-covariance matrix between $\boldsymbol{X}$ and $\boldsymbol{Y}$, and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{XX}}$ as well as $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{YY}}$ the respective covariance matrices. A necessary assumption is that the covariance matrices of $\boldsymbol{X}$ and $\boldsymbol{Y}$ are of full rank, i.e. are postive definite. The objective function is then

$$r(\boldsymbol{a}, \boldsymbol{b}) = s_{\boldsymbol{Xa},\boldsymbol{Yb}} = \frac{\boldsymbol{a}^\top \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{XY}} \boldsymbol{b}}{\left(\boldsymbol{a}^\top \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{XX}} \boldsymbol{a}\right)^{1/2} \left(\boldsymbol{b}^\top \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{YY}} \boldsymbol{b}\right)^{1/2}}. \tag{69}$$

Note that $r(\boldsymbol{a}, \boldsymbol{b})$ is invariant under scaling of either argument and therefore we can impose the restrictions

$$\boldsymbol{a}^\top \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{XX}} \boldsymbol{a} = 1 \quad \text{and} \quad \boldsymbol{b}^\top \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{YY}} \boldsymbol{b} = 1. \tag{70}$$

Using the variable substitutions $\boldsymbol{c} \leftarrow \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{XX}}^{1/2} \boldsymbol{a}$ and $\boldsymbol{d} \leftarrow \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{YY}}^{1/2} \boldsymbol{b}$ the objective function is transformed to

$$r(\boldsymbol{a}, \boldsymbol{b}) = \boldsymbol{c}^\top \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{XX}}^{-1/2} \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{XY}} \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{YY}}^{-1/2} \boldsymbol{d}. \tag{71}$$

Denote $\boldsymbol{K} = \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{XX}}^{-1/2} \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{XY}} \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{YY}}^{-1/2}$ and the optimization problem to be solved, formulated as a minimization problem, becomes

$$\inf_{\boldsymbol{c},\boldsymbol{d}} -\boldsymbol{c}^\top \boldsymbol{K} \boldsymbol{d} \quad \text{such that} \quad \boldsymbol{c}^\top \boldsymbol{c} = 1,\ \boldsymbol{d}^\top \boldsymbol{d} = 1. \tag{72}$$

where $\boldsymbol{a}$ and $\boldsymbol{b}$ can be recovered by re-substitution. The Lagrangian for the corresponding minimization problem is

$$L(\boldsymbol{c}, \boldsymbol{d}, \lambda) = -\boldsymbol{c}^\top \boldsymbol{K}\boldsymbol{d} + \lambda_1(\boldsymbol{c}^\top \boldsymbol{c} - 1) + \lambda_2(\boldsymbol{d}^\top \boldsymbol{d} - 1). \tag{73}$$

The dual function of the Lagrangian is

$$g(\lambda) = \inf_{\boldsymbol{c},\boldsymbol{d}} -\boldsymbol{c}^\top \boldsymbol{K}\boldsymbol{d} + \lambda_1(\boldsymbol{c}^\top \boldsymbol{c} - 1) + \lambda_2(\boldsymbol{d}^\top \boldsymbol{d} - 1) \tag{74}$$

and can be rewritten as either

$$g(\lambda) = \inf_{\boldsymbol{c},\boldsymbol{d}} \frac{1}{4\lambda_1} \left\{ \left\| 2\lambda_1 \boldsymbol{c} - \boldsymbol{K}\boldsymbol{d} \right\|_2^2 + \boldsymbol{d}^\top \left(-\boldsymbol{K}^\top \boldsymbol{K} + 4\lambda_1\lambda_2 \boldsymbol{I}_q \right) \boldsymbol{d} \right\} - \lambda_1 - \lambda_2, \tag{75}$$

or

$$g(\lambda) = \inf_{\boldsymbol{c},\boldsymbol{d}} \frac{1}{4\lambda_2} \left\{ \left\| 2\lambda_2 \boldsymbol{d} - \boldsymbol{K}^\top \boldsymbol{c} \right\|_2^2 + \boldsymbol{c}^\top \left(-\boldsymbol{K}\boldsymbol{K}^\top + 4\lambda_1\lambda_2 \boldsymbol{I}_p \right) \boldsymbol{c} \right\} - \lambda_1 - \lambda_2. \tag{76}$$

The reformulation in Eq. (75) shows directly that if $\lambda_1 < 0$ then the infimum is $-\infty$ by choosing $\boldsymbol{d} = \boldsymbol{0}$ and letting $\|\boldsymbol{c}\|_2 \to \infty$. An analogous result for $\lambda_2 < 0$ follows from Eq. (76). It follows from Eq. (74) that the infimum is $-\infty$ if $\lambda_1 = 0$ by choosing $\boldsymbol{c} = \tau \boldsymbol{K}\boldsymbol{d}$ and letting $\tau \to \infty$ for any $\boldsymbol{d}$ such that $\boldsymbol{K}\boldsymbol{d} \neq \boldsymbol{0}$ (exists as long as $\boldsymbol{K} \neq \boldsymbol{0}$). An analogous result follows for $\lambda_2 = 0$.

In the case $\lambda_1 > 0$ and $\lambda_2 > 0$, it is enough to focus on Eq. (75) if $p > q$ or Eq. (76) if $q > p$. Assume $p > q$ here[13]. The arguments are similar to the analysis of the dual function of the basic Rayleigh coefficient.

If the matrix $(-\boldsymbol{K}^\top \boldsymbol{K} + 4\lambda_1\lambda_2 \boldsymbol{I}_q)$ is not psd there exists a negative eigenvalue μ with corresponding eigenvector $\boldsymbol{u}$. Setting $\boldsymbol{d} = \tau \boldsymbol{u}$ and $\boldsymbol{c} = \frac{1}{2\lambda_1} \boldsymbol{K}\boldsymbol{d}$ leads to

$$\left\| 2\lambda_1 \boldsymbol{c} - \boldsymbol{K}\boldsymbol{d} \right\|_2^2 + \boldsymbol{d}^\top \left(-\boldsymbol{K}^\top \boldsymbol{K} + 4\lambda_1\lambda_2 \boldsymbol{I}_q \right) \boldsymbol{d} = \tau^2 \mu \|\boldsymbol{u}\|_2^2 \to -\infty \quad \text{for} \quad \tau \to \infty. \tag{77}$$

For the infimum to be greater than $-\infty$ it is therefore necessary that $(-\boldsymbol{K}^\top \boldsymbol{K} + 4\lambda_1\lambda_2 \boldsymbol{I}_q)$ is psd. Let the SVD of $\boldsymbol{K}$ be given by $\boldsymbol{K} = \boldsymbol{U}\boldsymbol{\Delta}\boldsymbol{V}^\top$ assuming that the singular values $\delta_1 \geq \delta_2 \geq \cdots \geq \delta_q \geq 0$ in $\boldsymbol{\Delta}$ are sorted from largest to smallest. Denote the columns of $\boldsymbol{V}$ by $\boldsymbol{v}_i$ and the columns of $\boldsymbol{U}$ by $\boldsymbol{u}_i$. Since $p > q$ it holds that $\boldsymbol{V}\boldsymbol{V}^\top = \boldsymbol{I}_q$ which implies

$$\begin{aligned} \boldsymbol{d}^\top \left(-\boldsymbol{K}^\top \boldsymbol{K} + 4\lambda_1\lambda_2 \boldsymbol{I}_q \right) \boldsymbol{d} &= \boldsymbol{d}^\top \boldsymbol{V} \left(-\boldsymbol{\Delta}^2 + 4\lambda_1\lambda_2 \boldsymbol{I}_q \right) \boldsymbol{V}^\top \boldsymbol{d} \\ &= \sum_{i=1}^{q} (4\lambda_1\lambda_2 - \delta_i^2)(\boldsymbol{v}_i^\top \boldsymbol{d})^2 \end{aligned} \tag{78}$$

for all $\boldsymbol{d} \in \mathbb{R}^q$. The matrix $(-\boldsymbol{K}^\top \boldsymbol{K} + 4\lambda_1\lambda_2 \boldsymbol{I}_q)$ is therefore psd if $4\lambda_1\lambda_2 \geq \delta_1^2$. The norm and quadratic terms in Eq. (75) are therefore greater or equal zero and both become zero for $\boldsymbol{c} = \boldsymbol{0} = \boldsymbol{d}$. The dual function is therefore

$$g(\lambda) = \begin{cases} -\lambda_1 - \lambda_2 & \text{if } 4\lambda_1\lambda_2 \geq \delta_1^2,\ \lambda_1 > 0,\ \lambda_2 > 0, \\ -\infty & \text{otherwise} \end{cases} \tag{79}$$

[13] The arguments for $q > p$ are symmetric using Eq. (76) instead.

and implies the dual problem

$$\sup_{\lambda} -\lambda_1 - \lambda_2 \quad \text{such that} \quad 4\lambda_1\lambda_2 \geq \delta_1^2,\ \lambda_1 > 0,\ \lambda_2 > 0. \tag{80}$$

It can be easily seen[14] that the supremum is achieved for $\widehat{\lambda_1} = \delta_1/2 = \widehat{\lambda_2}$ which leads to a dual objective function value of $d^* = -\delta_1$.

Let $\boldsymbol{c}_1 = \boldsymbol{u}_1$ and $\boldsymbol{d}_1 = \boldsymbol{v}_1$. Since the primal objective value p^* in Eq. (72) is bounded from below by d^*, it holds that

$$d^* = -\delta_1 \leq p^* \leq -\boldsymbol{u}_1^\top \boldsymbol{K} \boldsymbol{v}_1 = -\delta_1. \tag{81}$$

This shows that $p^* = -\delta_1$ and that the optimisation problem in Eq. (72) is solved by $\boldsymbol{c}_1$ and $\boldsymbol{d}_1$. By resubstitution it follows that $\boldsymbol{a}_1 = \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{XX}}^{-1/2} \boldsymbol{u}_1$ and $\boldsymbol{b}_1 = \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{YY}}^{-1/2} \boldsymbol{v}_1$ solve the original optimisation problem of maximizing $r(\boldsymbol{a}_1, \boldsymbol{b}_1)$ in Eq. (69).

By restricting the search space for $\boldsymbol{c}$ to vectors that are orthogonal to $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_{i-1}$ and for $\boldsymbol{d}$ to vectors that are orthogonal to $\boldsymbol{d}_1, \ldots, \boldsymbol{d}_{i-1}$ it follows easily—by the arguments above—that the restriction in the dual problem becomes $4\lambda_1\lambda_2 \geq \delta_i^2$ which ultimately leads to the solution $p^* = -\delta_i$ which is achieved for $\boldsymbol{c}_i = \boldsymbol{u}_i$ and $\boldsymbol{d}_i = \boldsymbol{v}_i$. By defining $\boldsymbol{\eta}_i = \boldsymbol{X}\boldsymbol{a}_i$ and $\boldsymbol{\rho}_i = \boldsymbol{Y}\boldsymbol{b}_i$, q pairs[15] of canonical directions $(\boldsymbol{\eta}_1, \boldsymbol{\rho}_1), \ldots, (\boldsymbol{\eta}_q, \boldsymbol{\rho}_q)$ can be found such that

$$\widehat{\boldsymbol{\Sigma}}_{(\boldsymbol{\eta},\boldsymbol{\rho})(\boldsymbol{\eta},\boldsymbol{\rho})} = \begin{pmatrix} \boldsymbol{I}_p & \boldsymbol{\Delta} \\ \boldsymbol{\Delta} & \boldsymbol{I}_q \end{pmatrix}. \tag{82}$$

CCA can also be interpreted as a probabilistic model as shown by Bach and Jordan (2006).

8 Independent Component Analysis (ICA)

References

F. R. Bach and M. I. Jordan. A Probabilistic Interpretation of Canonical Correlation Analysis. Technical Report 688, Department of Statistics, University of California, Berkeley, 2006. (page 17)

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. (page 2)

M. Falk, F. Marohn, and B. Tewes. *Foundations of Statistical Analyses and Applications with SAS*. Birkhäuser, 2002. (page 2)

W. Härdle and L. Simar. *Applied Multivariate Statistical Analysis*. Springer-Verlag Berlin Heidelberg, 2 edition, 2007. (page 2)

[14] TODO Create graphic

[15] If $q > p$, then only p pairs can be found.

K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. Massachusetts Institute of Technology, 2012. (pages 2, 13)

S. T. Roweis. EM Algorithms for PCA and SPCA. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, *Advances in Neural Information Processing Systems 10*, pages 626–632. MIT Press, 1998. (page 15)

M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *J. R. Statist. Soc. B*, 61(Part 3):611–622, 1999. (page 15)