Theory related to PCA and SVD

Felix Held

Mathematical Sciences, Chalmers University of Technology and University of Gothenburg

2019-04-08

Contents

1 Rayleigh quotient

1.1 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 $\boldsymbol{0} \neq \boldsymbol{x} \in \mathbb{R}^p$ as

$$J(\boldsymbol{x}) = \frac{\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}}. \tag{1.1}$$

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

$$J(\boldsymbol{x}) = \frac{\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}} = \frac{||\boldsymbol{x}||^2}{||\boldsymbol{x}||^2}\frac{\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}} = \frac{\frac{\boldsymbol{x}}{||\boldsymbol{x}||}^T\boldsymbol{A}\frac{\boldsymbol{x}}{||\boldsymbol{x}||}}{\frac{\boldsymbol{x}}{||\boldsymbol{x}||}^T\frac{\boldsymbol{x}}{||\boldsymbol{x}||}} = \frac{\boldsymbol{x}}{||\boldsymbol{x}||}^T\boldsymbol{A}\frac{\boldsymbol{x}}{||\boldsymbol{x}||}. \tag{1.2}$$

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

$$\widehat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} \frac{\boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{x}}. \tag{1.3}$$

As noted above the Rayleigh quotient is scale-invariant. Therefore, for any solution $\widehat{\boldsymbol{x}}$ of the maximization problem in Eq. (1.3) the vector $c \cdot \widehat{\boldsymbol{x}}$for arbitrary $c \in \mathbb{R}$ is a solution as well. To restrict the space of possible solutions we require that $\|\widehat{\boldsymbol{x}}\| = 1$. This results in the optimization problem

$$\max_{\boldsymbol{x}} \boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x} \quad \text{subject to} \quad \boldsymbol{x}^T\boldsymbol{x} = 1. \tag{1.4}$$

The Lagrangian (optimization course) of this problem is

$$L(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x} - \lambda(\boldsymbol{x}^T\boldsymbol{x} - 1) \tag{1.5}$$

where λ is a Lagrange multiplier. We want to find a maximum of $L(\boldsymbol{x})$ and λ will be determined along the way. The gradient of $L(\boldsymbol{x})$ with respect to $\boldsymbol{x}$ is

$$\frac{\partial L(\boldsymbol{x})}{\partial \boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x} - \lambda\boldsymbol{x} \tag{1.6}$$

and setting Eq. (1.6) equal to $\mathbf{0}$ leads to

$$\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}. \tag{1.7}$$

Since $\boldsymbol{x} \neq \mathbf{0}$ this is an eigenvalue equation and λ has to be one of p eigenvalues of $\boldsymbol{A}$. Using Eq. (1.7) in our original optimization problem Eq. (1.4) gives

$$\max_{\boldsymbol{x},\, \|\boldsymbol{x}\|=1} \boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x} = \max_{\substack{\boldsymbol{x},\, \|\boldsymbol{x}\|=1 \\ \boldsymbol{A}\boldsymbol{x}=\lambda\boldsymbol{x}}} \lambda\boldsymbol{x}^T\boldsymbol{x} = \max_{\substack{\boldsymbol{x},\, \|\boldsymbol{x}\|=1 \\ \boldsymbol{A}\boldsymbol{x}=\lambda\boldsymbol{x}}} \lambda \tag{1.8}$$

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 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.

1.2 A more general 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 $\mathbf{0} \neq \boldsymbol{x} \in \mathbb{R}^p$ as

$$J(\boldsymbol{x}) = \frac{\boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{B}\boldsymbol{x}}. \tag{1.9}$$

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

As before $J(\boldsymbol{x})$ 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[3] $\boldsymbol{D} \in \mathbb{R}^{p\times p}$ such that $\boldsymbol{B} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T$. Define $\boldsymbol{D}^{1/2} = \text{diag}(\sqrt{\boldsymbol{D}_{ii}}, i = 1, \dots, p)$ and $\boldsymbol{B}^{1/2} = \boldsymbol{U}\boldsymbol{D}^{1/2}\boldsymbol{U}^T$, then $\boldsymbol{B}^{1/2}\boldsymbol{B}^{1/2} = \boldsymbol{B}$[4].

In the previous section, we required to fix the scaling

$$\|\boldsymbol{x}\| = \boldsymbol{x}^T\boldsymbol{x} = 1. \tag{1.10}$$

This is very convenient for the applications below and it led for the numerator of the Rayleigh quotient to be one which made the proof above simpler. However, this is not equally convenient here since we have $\boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x}$ in the numerator. We therefore require

$$\|\boldsymbol{B}^{1/2}\boldsymbol{x}\| = \boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x} = 1, \tag{1.11}$$

which leads to the optimization problem

$$\max_{\boldsymbol{x}} \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} \quad \text{subject to} \quad \boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x} = 1. \tag{1.12}$$

Note that this is equivalent to solving

$$\max_{\boldsymbol{x}} \frac{\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x}} \quad \text{subject to} \quad \boldsymbol{x}^T\boldsymbol{x} = 1 \tag{1.13}$$

since for every solution $\hat{\boldsymbol{x}}$ to Eq. (1.13) the vector $\boldsymbol{z} = \hat{\boldsymbol{x}}/\|\boldsymbol{B}^{1/2}\hat{\boldsymbol{x}}\|$ also maximizes the Rayleigh quotient with $\boldsymbol{z}^T\boldsymbol{B}\boldsymbol{z} = 1$ and is therefore a solution to Eq. (1.12). On the other hand, every solution $\hat{\boldsymbol{x}}$ to Eq. (1.12) is a solution to Eq. (1.13) by setting $\boldsymbol{z} = \hat{\boldsymbol{x}}/\|\hat{\boldsymbol{x}}\|$.

Using a Lagrangian, calculating its gradient and setting it to zero (as above for the basic Rayleigh quotient) leads to

$$\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{B}\boldsymbol{x} \tag{1.14}$$

This is called a generalized eigenvalue problem for the symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. Using $\boldsymbol{B}^{1/2}$ as above we get

$$\left(\boldsymbol{B}^{-1/2}\boldsymbol{A}\boldsymbol{B}^{-1/2}\right)\left(\boldsymbol{B}^{1/2}\boldsymbol{x}\right) = \lambda\left(\boldsymbol{B}^{1/2}\boldsymbol{x}\right) \tag{1.15}$$

Defining $\boldsymbol{w} = \boldsymbol{B}^{1/2}\boldsymbol{x}$ results in

$$\left(\boldsymbol{B}^{-1/2}\boldsymbol{A}\boldsymbol{B}^{-1/2}\right)\boldsymbol{w} = \lambda\boldsymbol{w} \tag{1.16}$$

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

[3] The eigenvalues of $\boldsymbol{U}$ are on the diagonal.

[4] Note that there was a similar construct in the lecture. A inverse covariance matrix $\boldsymbol{\Sigma}^{-1}$ was written as $\boldsymbol{\Sigma}^{-1} = \boldsymbol{R}\boldsymbol{D}^{-1}\boldsymbol{R}^T$. With the definition $\boldsymbol{\Sigma}^{-1/2} := \boldsymbol{D}^{-1/2}\boldsymbol{R}^T$ it holds that $\left(\boldsymbol{\Sigma}^{-1/2}\right)^T\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{\Sigma}^{-1}$. Note that there is a transpose and above we are simply multiplying the "square root" matrices.

which is an eigenvalue problem for the symmetric matrix $\boldsymbol{B}^{-1/2}\boldsymbol{A}\boldsymbol{B}^{-1/2}$. This matrix is guaranteed to have p real-valued eigenvalues and corresponding eigenvectors $\boldsymbol{w}_i$ for $i = 1, \ldots, p$. Note that the rank of this matrix is determined by the rank of $\boldsymbol{A}$ since $\boldsymbol{B}^{-1/2}$ is invertible. The original generalized eigenvectors $\boldsymbol{x}$ can be recovered through

$$\boldsymbol{x} = \boldsymbol{B}^{-1/2}\boldsymbol{w}. \tag{1.17}$$

By using Eq. (1.14) in the original optimization problem in Eq. (1.12), show that this more general variant of the Rayleigh quotient is maximized for the vector $\boldsymbol{x}$ such that $\boldsymbol{B}^{-1/2}\boldsymbol{x}$ is an eigenvector of the matrix $\boldsymbol{B}^{-1/2}\boldsymbol{A}\boldsymbol{B}^{-1/2}$ corresponding to its largest eigenvalue λ.

2 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 $\boldsymbol{x}_i^T$. 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 $s_i = \boldsymbol{r}_1^T\boldsymbol{x}_i$, which is the coefficient of $\boldsymbol{x}_i$ projected onto $\boldsymbol{r}_1$. The variance in the direction of $\boldsymbol{r}_1$ is

$$\sum_{i=1}^{n}(s_i - \overline{s})^2 = \sum_{i=1}^{n}\left(\boldsymbol{r}_i^T(\boldsymbol{x}_i - \overline{\boldsymbol{x}})\right)^2 \tag{2.1}$$

where $\overline{\boldsymbol{x}}$ is the mean over all observations. We want to find $\boldsymbol{r}_1$ such that the variance in Eq. (2.1) becomes maximal. Note that

$$\begin{aligned}\sum_{i=1}^{n}\left(\boldsymbol{r}^T(\boldsymbol{x}_i - \overline{\boldsymbol{x}})\right)^2 &= \sum_{i=1}^{n}\boldsymbol{r}^T(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T\boldsymbol{r} \\ &= \boldsymbol{r}^T\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T\boldsymbol{r} \\ &= (n-1)\,\boldsymbol{r}^T\widehat{\boldsymbol{\Sigma}}\boldsymbol{r}\end{aligned} \tag{2.2}$$

where $\widehat{\boldsymbol{\Sigma}}$ is the empirical covariance matrix of the data. Since $\widehat{\boldsymbol{\Sigma}}$ is a symmetric matrix and it is required that $\|\boldsymbol{r}\| = 1$, maximizing Eq. (2.2) is equivalent to solving the basic Rayleigh quotient maximisation problem in Section 1. It therefore follows that $\boldsymbol{r}_1$ is an eigenvector of $\widehat{\boldsymbol{\Sigma}}$ 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. (2.1)) 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 $\lambda_1 \geq \ldots \geq \lambda_{m-1}$ of $\widehat{\boldsymbol{\Sigma}}$. From linear algebra we know that

a square matrix $\boldsymbol{P}$ is an orthogonal projection matrix if

$$\boldsymbol{P}^2 = \boldsymbol{P} = \boldsymbol{P}^T. \tag{2.3}$$

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

$$P = \sum_{i=1}^{m-1} \boldsymbol{r}_i \boldsymbol{r}_i^T \tag{2.4}$$

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

$$\boldsymbol{X}_{m-1} = \boldsymbol{X}\left(\boldsymbol{I}_p - \sum_{i=1}^{m-1} \boldsymbol{r}_i \boldsymbol{r}_i^T\right) \tag{2.5}$$

and

$$\widehat{\boldsymbol{\Sigma}}_{m-1} = \frac{\boldsymbol{X}_{m-1}^T \boldsymbol{X}_{m-1}}{n-1} = \left(\boldsymbol{I}_p - \sum_{i=1}^{m-1} \boldsymbol{r}_i \boldsymbol{r}_i^T\right)^T \widehat{\boldsymbol{\Sigma}}\left(\boldsymbol{I}_p - \sum_{i=1}^{m-1} \boldsymbol{r}_i \boldsymbol{r}_i^T\right). \tag{2.6}$$

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

$$\max_{\boldsymbol{r}} \boldsymbol{r}^T \widehat{\boldsymbol{\Sigma}}_{m-1} \boldsymbol{r} \quad \text{subject to} \quad \boldsymbol{r}^T \boldsymbol{r} = 1. \tag{2.7}$$

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 $i = 1, \dots, m-1$

$$\begin{aligned} \boldsymbol{r}_m^T \boldsymbol{r}_i &= \frac{1}{\lambda_m} \boldsymbol{r}_m^T \widehat{\boldsymbol{\Sigma}}_{m-1} \boldsymbol{r}_i \\ &= \frac{1}{\lambda_m} \boldsymbol{r}_m^T \left(\boldsymbol{I}_p - \sum_{j=1}^{m-1} \boldsymbol{r}_j \boldsymbol{r}_j^T\right)^T \widehat{\boldsymbol{\Sigma}} \underbrace{\left(\boldsymbol{I}_p - \sum_{j=1}^{m-1} \boldsymbol{r}_j \boldsymbol{r}_j^T\right) \boldsymbol{r}_i}_{=\boldsymbol{0}} = 0 \end{aligned} \tag{2.8}$$

as well as

$$\begin{aligned}
\lambda_m \boldsymbol{r}_m &= \widehat{\boldsymbol{\Sigma}}_{m-1} \boldsymbol{r}_m \\
&= \left(\boldsymbol{I}_p - \sum_{j=1}^{m-1} \boldsymbol{r}_j \boldsymbol{r}_j^T \right)^T \widehat{\boldsymbol{\Sigma}} \left(\boldsymbol{I}_p - \sum_{j=1}^{m-1} \boldsymbol{r}_j \boldsymbol{r}_j^T \right) \boldsymbol{r}_m \\
&= \left(\boldsymbol{I}_p - \sum_{j=1}^{m-1} \boldsymbol{r}_j \boldsymbol{r}_j^T \right) \widehat{\boldsymbol{\Sigma}} \boldsymbol{r}_m \\
&= \widehat{\boldsymbol{\Sigma}} \boldsymbol{r}_m - \sum_{j=1}^{m-1} \boldsymbol{r}_j \underbrace{\boldsymbol{r}_j^T \widehat{\boldsymbol{\Sigma}} \boldsymbol{r}_m}_{= \lambda_j \boldsymbol{r}_j^T \boldsymbol{r}_m = \boldsymbol{0}} \\
&= \widehat{\boldsymbol{\Sigma}} \boldsymbol{r}_2
\end{aligned} \tag{2.9}$$

which shows that $\boldsymbol{r}_m$ is orthogonal to $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_{m-1}$ and an eigenvector of $\widehat{\boldsymbol{\Sigma}}$. In addition, since $\lambda_1 \geq \ldots \geq \lambda_{m-1}$ 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

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n-1} \boldsymbol{X}^T \boldsymbol{X} \tag{2.10}$$

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

$$\widehat{\boldsymbol{\Sigma}} \boldsymbol{r}_i = \lambda_i \boldsymbol{r}_i, \ \|\boldsymbol{r}_i\| = 1, \ i = 1, \ldots, p \quad \text{and} \quad \boldsymbol{r}_i^T \boldsymbol{r}_j = 0, \ i \neq j \tag{2.11}$$

as well as

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \tag{2.12}$$

4. Set

$$\boldsymbol{R} = \left(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_p \right) \in \mathbb{R}^{p \times p} \quad \text{and} \quad \boldsymbol{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_p) \tag{2.13}$$

so that

$$\widehat{\boldsymbol{\Sigma}} = \boldsymbol{R} \boldsymbol{D} \boldsymbol{R}^T \tag{2.14}$$

5. The vectors $\boldsymbol{r}_i$ are the principal component directions, the projections $\boldsymbol{r}_i^T \boldsymbol{x}$ 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 $\mu_1, \ldots, \mu_p$ it holds that

$$\operatorname{tr}(\boldsymbol{A}) = \sum_{i=1}^{p} \mu_i. \tag{2.15}$$

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

$$\operatorname{tr}\left(\hat{\boldsymbol{\Sigma}}\right) = \sum_{i=1}^{p} s^2(\boldsymbol{X}_{\cdot i}) = \sum_{i=1}^{p} \lambda_i, \tag{2.16}$$

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 $\operatorname{tr}\left(\hat{\boldsymbol{\Sigma}}\right) = p$. This means in particular that the mean of the eigenvalues will be 1[5]. 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 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.

3 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

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T \tag{3.1}$$

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

$$\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}_p \quad \text{and} \quad \boldsymbol{V}\boldsymbol{V}^T = \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}_p \tag{3.2}$$

Note that if $n > p$ it cannot hold that $\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{I}_n$ [6]. 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. (3.2) can now be used to derive the following equations

$$\begin{aligned} \boldsymbol{X}\boldsymbol{X}^T\boldsymbol{U} &= \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{U}\boldsymbol{D}^2 \\ \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{V} &= \boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{V}\boldsymbol{D}^2 \end{aligned} \tag{3.3}$$

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

[6] 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.

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

$$\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{u}_i = d_i^2\boldsymbol{u}_i \quad \text{and} \quad \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{v}_i = d_i^2\boldsymbol{v}_i, \quad i = 1, \dots, p \tag{3.4}$$

Since $n \geq p$, the matrix $\boldsymbol{X}^T\boldsymbol{X} \in \mathbb{R}^{p\times p}$ is as large as or smaller than $\boldsymbol{X}\boldsymbol{X}^T \in \mathbb{R}^{n\times n}$. It is therefore more computationally effective to calculate $\boldsymbol{V}$ by solving the p eigenvalue problems in Eq. (3.4) 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.

$$\boldsymbol{U} = \boldsymbol{X}\boldsymbol{V}\boldsymbol{D}^{-1}. \tag{3.5}$$

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 $\boldsymbol{X}^T \in \mathbb{R}^{p\times n}$. 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

$$\boldsymbol{X}^T = \boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^T. \tag{3.6}$$

By transposing again we get

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T. \tag{3.7}$$

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

3.1 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

$$\widehat{\boldsymbol{\Sigma}} = \frac{\boldsymbol{X}^T\boldsymbol{X}}{n-1}. \tag{3.8}$$

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

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T = \boldsymbol{V}\left(\frac{1}{n-1}\boldsymbol{D}^2\right)\boldsymbol{V}^T. \tag{3.9}$$

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

$$\widehat{\boldsymbol{\Sigma}}\boldsymbol{V} = \boldsymbol{V}\left(\frac{1}{n-1}\boldsymbol{D}^2\right) \tag{3.10}$$

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

3.2 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

$$\widehat{\mathcal{S}} = \arg\min_{\mathcal{S}} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{P}_{\mathcal{S}}\boldsymbol{x}_i\|_2^2 \tag{3.11}$$

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

$$\begin{aligned}
\sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{P}_{\mathcal{S}}\boldsymbol{x}_i\|_2^2 &= \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{P}_{\mathcal{S}}\boldsymbol{x}_i)^T (\boldsymbol{x}_i - \boldsymbol{P}_{\mathcal{S}}\boldsymbol{x}_i) \\
&= \sum_{i=1}^{n} \left[\boldsymbol{x}_i^T\boldsymbol{x}_i - 2\boldsymbol{x}_i^T\boldsymbol{P}_{\mathcal{S}}\boldsymbol{x}_i + \boldsymbol{x}_i^T\boldsymbol{P}_{\mathcal{S}}^T\boldsymbol{P}_{\mathcal{S}}\boldsymbol{x}_i\right] \\
&= \sum_{i=1}^{n} \left[\boldsymbol{x}_i^T\boldsymbol{x}_i - \boldsymbol{x}_i^T\boldsymbol{P}_{\mathcal{S}}^T\boldsymbol{P}_{\mathcal{S}}\boldsymbol{x}_i\right] \\
&= \sum_{i=1}^{n} \left[\|\boldsymbol{x}_i\|_2^2 - \|\boldsymbol{P}_{\mathcal{S}}\boldsymbol{x}_i\|_2^2\right]
\end{aligned} \tag{3.12}$$

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[7]. 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 $\boldsymbol{P} = \boldsymbol{r}_1\boldsymbol{r}_1^T$ and thus we are trying to maximize

$$\begin{aligned}
\sum_{i=1}^{n} \|\boldsymbol{r}_1\boldsymbol{r}_1^T\boldsymbol{x}_i\|_2^2 &= \sum_{i=1}^{n} (\boldsymbol{r}_1^T\boldsymbol{x}_i)^2 \\
&= \boldsymbol{r}_1^T \left(\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^T\right) \boldsymbol{r}_1 \\
&= (n-1)\boldsymbol{r}_1^T\widehat{\boldsymbol{\Sigma}}\boldsymbol{r}_1
\end{aligned} \tag{3.13}$$

The term we are maximizing is the Rayleigh quotient and therefore the optimal subspace is spanned by $\boldsymbol{r}_1$, the eigenvector of $\widehat{\boldsymbol{\Sigma}}$ 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 $\boldsymbol{X}$.

[7] 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 $\sum_{i=1}^{n} \|\boldsymbol{P}_{\mathcal{S}}\boldsymbol{x}_i\|_2^2$.

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 $< n$.

3.3 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

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T = \sum_{i=1}^{\min(n,p)} d_i \boldsymbol{u}_i \boldsymbol{v}_i^T. \tag{3.14}$$

Each matrix $\boldsymbol{u}_i \boldsymbol{v}_i^T$ is of rank 1 and these matrices are scaled by the singular values d_i. As we have seen above in Section 3.2, 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

$$\boldsymbol{P}_q = \sum_{i=1}^{q} \boldsymbol{v}_i \boldsymbol{v}_i^T. \tag{3.15}$$

Projecting $\boldsymbol{X}$ on this optimal q-dimensional subspace using the projection in Eq. (3.15) and the representation in Eq. (3.14) leads to

$$\boldsymbol{X}_q = \boldsymbol{X}\boldsymbol{P}_q = \left(\sum_{i=1}^{\min(n,p)} d_i \boldsymbol{u}_i \boldsymbol{v}_i^T \right) \left(\sum_{j=1}^{q} \boldsymbol{v}_j \boldsymbol{v}_j^T \right) = \sum_{i=1}^{q} d_i \boldsymbol{u}_i \boldsymbol{v}_i^T. \tag{3.16}$$

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[8] is

$$\begin{aligned}
\|\boldsymbol{X} - \boldsymbol{X}_q\|_F^2 &= \left\| \sum_{i=q+1}^{\min(n,p)} d_i \boldsymbol{u}_i \boldsymbol{v}_i^T \right\|_F^2 \\
&= \operatorname{tr}\left[\left(\sum_{i=q+1}^{\min(n,p)} d_i \boldsymbol{u}_i \boldsymbol{v}_i^T \right)^T \left(\sum_{j=q+1}^{\min(n,p)} d_j \boldsymbol{u}_j \boldsymbol{v}_j^T \right) \right] \\
&= \operatorname{tr}\left[\sum_{i,j=q+1}^{\min(n,p)} d_i d_j \boldsymbol{v}_i \boldsymbol{u}_i^T \boldsymbol{u}_i \boldsymbol{v}_j^T \right] \\
&= \operatorname{tr}\left[\sum_{i,j=q+1}^{\min(n,p)} d_i d_j \boldsymbol{v}_i \boldsymbol{v}_j^T \right] \\
&= \sum_{k=1}^{\min(n,p)} \sum_{i,j=q+1}^{\min(n,p)} d_i d_j v_{ik} v_{jk} \\
&= \sum_{i,j=q+1}^{\min(n,p)} d_i d_j \sum_{k=1}^{\min(n,p)} v_{ik} v_{jk} \\
&= \sum_{i,j=q+1}^{\min(n,p)} d_i d_j \, \mathbb{1}(i=j) = \sum_{i=q+1}^{\min(n,p)} d_i^2
\end{aligned} \tag{3.17}$$

3.4 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

$$y = \boldsymbol{x}^T \boldsymbol{\beta} + \varepsilon \tag{3.18}$$

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

$$\varepsilon \sim \text{Normal}(0, \sigma^2) \tag{3.19}$$

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

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \tag{3.20}$$

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

The solution to the least squares problem is given by

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}, \tag{3.21}$$

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

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= \left(\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T\right)^{-1}\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{y}\\ &= \left(\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^T\right)^{-1}\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{y}\\ &= \boldsymbol{V}\boldsymbol{D}^{-2}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{y}\\ &= \boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{U}^T\boldsymbol{y}\end{aligned} \tag{3.22}$$

The expression

$$\boldsymbol{X}^+ = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T = \boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{U}^T \tag{3.23}$$

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

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{U}^T\boldsymbol{y} = \boldsymbol{U}\boldsymbol{U}^T\boldsymbol{y}. \tag{3.24}$$

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

For ridge regression the problem to be solved is

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2 \tag{3.25}$$

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

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}. \tag{3.26}$$

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

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{V}\left(\boldsymbol{D}^2 + \lambda\boldsymbol{I}_p\right)^{-1}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{y} = \sum_{i=1}^{p}\frac{d_i}{d_i^2+\lambda}\boldsymbol{v}_i\boldsymbol{u}_i^T\boldsymbol{y}. \tag{3.27}$$

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.