Lecture_notes3

February 17, 2021

1 MATH310 - Lecture_notes3

The Singular value decomposition (SVD) and Principal component analysis (PCA)

- Optimal matrix approximation
- Maximum variance subspace identification

Recommended supplementary videos:

- Gilbert Strang about the SVD
- Gilbert Strang about the Eckart-Young-Mirsky low rank approximation theorem

Recommended supplementary reading:

- Chapter I.8 and I.9 in *Linear Algebra and Learning from Data*
- Chapter 6 in *Matrix Methods in Data Mining and Pattern Recognition*
- Chapter 4.5 in Mathematics for Machine Learning (pdf)

2 The Singular value decomposition

Although the QR decomposition is highly useful for solving least squares problems, and has excellent numerical properties, it has a drawback in only providing an orthonormal basis for the column space of a matrix.

- The **singular value decomposition (SVD)** is a more sophisticated matrix factorization method. It simultaneously deals with both the column- and row spaces of a matrix by providing orthonomal bases for both in a symmetric fashion. Hence, it supplies more information about a matrix than any other matrix factorization method.
- The SVD also sorts the information content of a matrix so that its "dominant parts" becomes easily accessible. This is what makes the SVD so useful in both machine learning and may other areas of applied mathematics.

The following presentation of the SVD and PCA with illustrations are much influenced by the presentation in Lars Eldéns book Matrix Methods in Data Mining and Pattern Recognition.

2.1 Theorem (SVD)

Any $m \times n$ matrix $\mathbf{A}$, with $m \geq n$, can be factorized as

$U^t U = I_m$

$V^t V = I_n$

$$\boxed{A = U\Sigma V^t}$$

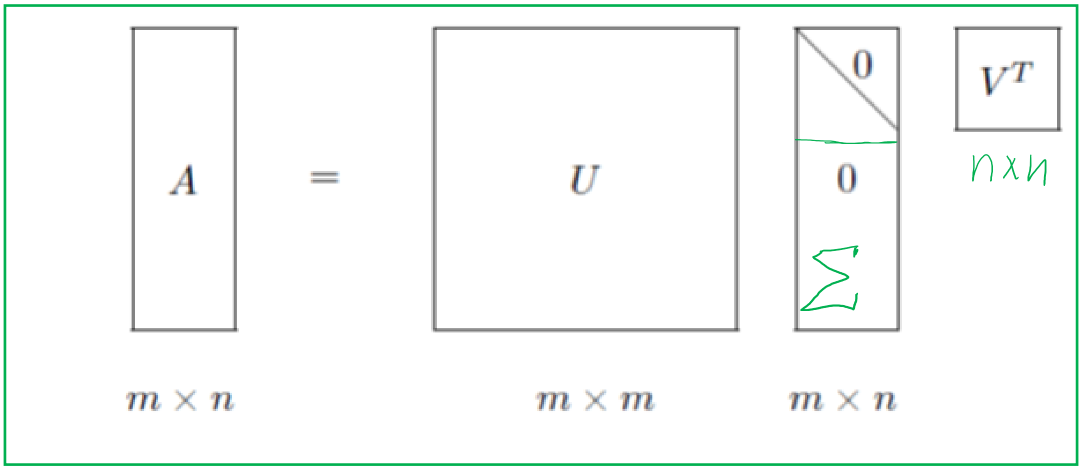
where $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ are orthogonal, and $\Sigma \in \mathbb{R}^{m\times n}$ is diagonal with diagonal elements $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_n \geq 0$ ■

Note that the assumption $m \geq n$ in the SVD-theorem is not really a restriction since for matrices where $n > m$ we can transpose and apply the theorem to A^t by just switching the notation for U and V.

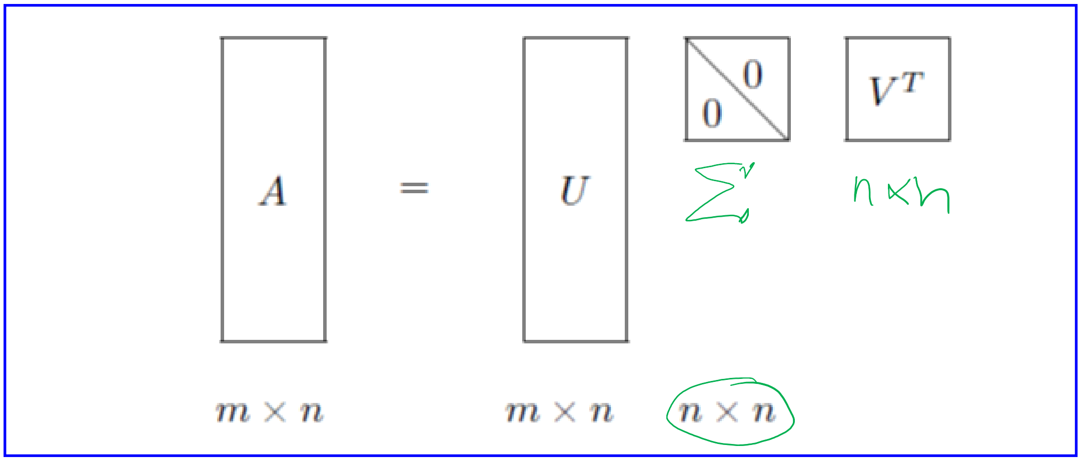
The columns of U and V are called the left- and right singular vectors, respectively, and the diagonal elements σ_i of Σ are called the *singular values*.

Note that the SVD is much more than a fancy theoretical result. There are highly efficient and numerically accurate algorithms for calculating the SVD making it highly powerful for real ML-problems.

Pictorially we can think of the SVD as follows



and by ignoring the last $m - n$ columns of U that are not contained in the column space of the matrix A we obtain the corresponding "thin" version of the SVD:



If we consider the associated matrix equations of the thin SVD including only n columns in U, i.e.

$$AV = U\Sigma \underbrace{V^tV}_{I} = U\Sigma$$

$$\boxed{AV = U\Sigma, \quad A^tU = V\Sigma}$$

column by column of $V = [\mathbf{v}_1 \ \mathbf{v}_2 \ \cdots \ \mathbf{v}_n]$ and $U = [\mathbf{u}_1 \ \mathbf{u}_2 \ \cdots \ \mathbf{u}_n]$, we get the equivalent equations

$$\boxed{A\mathbf{v}_i = \sigma_i \mathbf{u}_i, \quad A^t\mathbf{u}_i = \sigma_i \mathbf{v}_i, \quad i = 1, ..., n.}$$

The SVD can also conveinently be written as an *outer product expansion* of the matrix

$$\boxed{A = \sum_{i=1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^t}$$

rank 1 matrices,

derived by starting from the detailed notation for thin SVD

$$A = U\Sigma V^t = [\mathbf{u}_1 \ \mathbf{u}_2 \ \cdots \ \mathbf{u}_n] \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_n \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^t \\ \mathbf{v}_2^t \\ \vdots \\ \mathbf{v}_n^t \end{bmatrix}$$

$$= \boxed{[\sigma_1\mathbf{u}_1 \ \sigma_2\mathbf{u}_2 \ \cdots \ \sigma_n\mathbf{u}_n] \begin{bmatrix} \mathbf{v}_1^t \\ \mathbf{v}_2^t \\ \vdots \\ \mathbf{v}_n^t \end{bmatrix} = \sum_{i=1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^t.}$$

If A has not full rank, i.e. $rank(A) = r < n$ we must have the signgular values $\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \sigma_{r+2} = \cdots = \sigma_n = 0$, and the outer product expansion obviously requires only inclusion of the first r terms:

$$\boxed{A = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^t.}$$

Think of A and A^t as lin. transf.

2.2 Fundamental Subspaces

The SVD provides orthogonal bases for the four fundamental subspaces of a matrix A.

The range $\mathcal{R}(A)$ of the matrix A, when considered as a linear transformation, is the linear subspace

$$\boxed{\mathcal{R}(A) = \{\mathbf{y} | \mathbf{y} = A\mathbf{x}, \text{ for some } \mathbf{x} \in \mathbb{R}^n\} = Col(A),}$$

where $Col(A)$ denotes the column space of A spanned by all possible linear combinations of the A-columns.

If $rank(A) = r$, $Col(A)$ is spanned by the r first left singular vectors $\mathbf{u}_1, \cdots, \mathbf{u}_r$ corresponding to the r positive singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > \sigma_{r+1} = 0$ of A.

The null-space $\mathcal{N}(A)$ of the matrix A is the linear subspace

This represents everything that is mapped onto 0 by A.

This represents anything that is mapped onto 0 by A.

$$\boxed{\mathcal{N}(A) = \{\mathbf{x}|A\mathbf{x} = \mathbf{0}\}.}$$

Since matrix-vector product

$$A\mathbf{x} = \boxed{(\sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^t)\mathbf{x}} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i (\underline{\mathbf{v}_i^t \mathbf{x}}), = \sum_{i=1}^{r} \alpha_i \vec{u}_i, \; \alpha_i = \sigma_i(\vec{v}_i^t \vec{x})$$

$A\vec{x}$

Just a scalar that equals 0 if $\vec{v}_i^t \vec{x} = 0$.

it is clear that only the vectors of the form

$$\boxed{\mathbf{x} = \sum_{j=r+1}^{n} c_j \mathbf{v}_j,}$$

i.e. expressed as linear combinations of the right singular vectors corresponding to the singular values $\sigma_{r+1} = \sigma_{r+2} = \cdots \sigma_n = 0$, will satisfy the requirements for membership in the null-space $\mathcal{N}(A)$ due to the orthogonality of the right singular vectors $\mathbf{v}_1, \cdots, \mathbf{v}_n$.

From the corresponding consideration of A^t we can conclude the following theorem:

2.3 Theorem (The four fundamental subspaces of a matrix)

,

1. The left singular vectors $\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_r$ represent an orthonormal basis for $\mathcal{R}(A)$ and

$$rank(A) = \dim(\mathcal{R}(A)) = r.$$

2. The right singular vectors $\mathbf{v}_{r+1}, \mathbf{v}_{r+2}, \cdots, \mathbf{v}_n$ corresponding to the 0-singular values of A is an orthonormal basis for $\mathcal{N}(A)$ and

$$\dim(\mathcal{N}(A)) = n - r.$$

3. The right singular vectors $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_r$ represent an orthonormal basis for $\mathcal{R}(A^t)$ and

$$rank(A^t) = \dim(\mathcal{R}(A)) = r.$$

4. The left singular vectors $\mathbf{u}_{r+1}, \mathbf{u}_{r+2}, \cdots, \mathbf{u}_m$ represent an orthonormal basis for $\mathcal{N}(A^t)$ and

$$\dim(\mathcal{N}(A^t)) = m - r.$$

Not in $R(A) = Col(A)$

2.4 Matrix Approximation - the truncated SVD

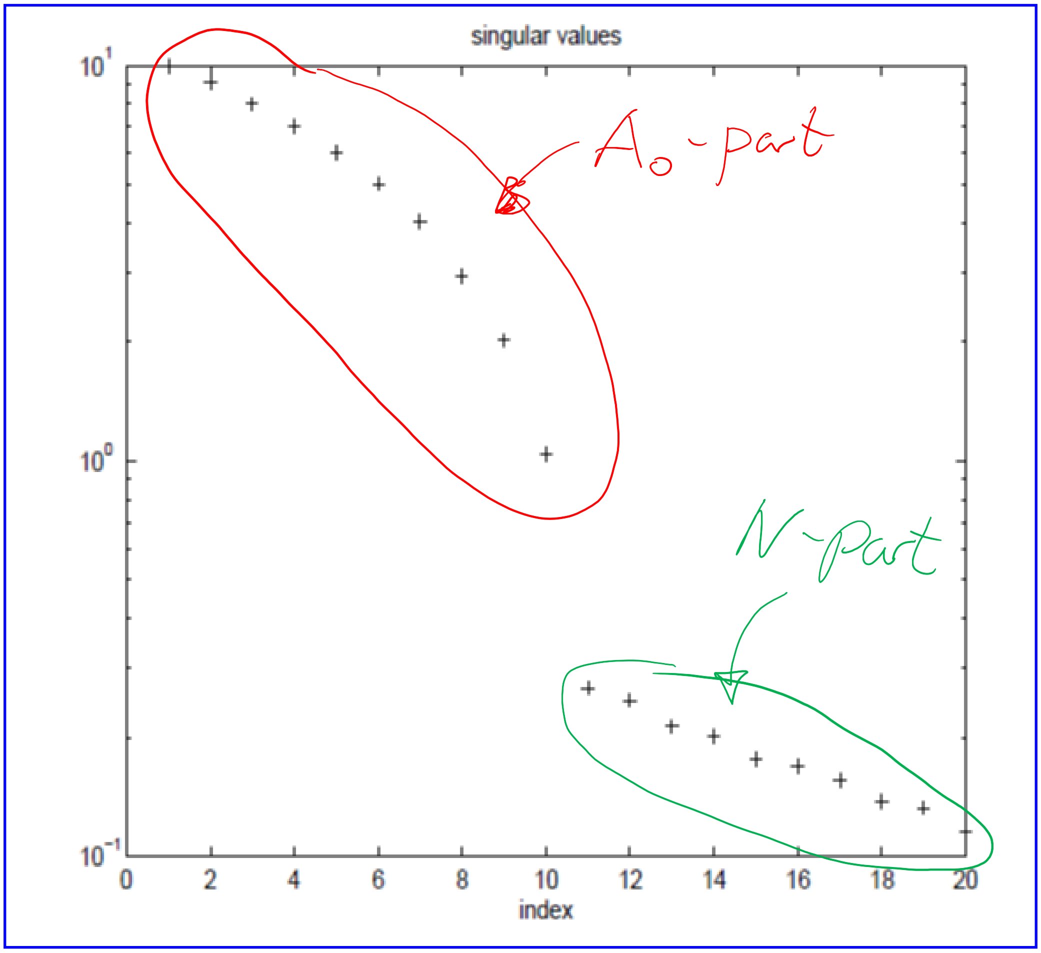
,

Assume that A is the result of a low rank matrix plus noise, i.e.

,

$$\boxed{A} = \boxed{A_0} + \boxed{N},$$

where the matrix N representing the noise is small compared to A_0. In this situation the singular values of A will typically show a pattern as illustrated in the following figure:



In this situation the noisy part of A is typically significantly smaller in magnitude.

The number of large singular values is often referred to as *the numerical rank* of the matrix.

If we know the correct rank of A_0, or are confident in estimating it from inspection of the singular values of A, we can "remove the noisy part" by "truncation" (by setting the smalerl singular values to 0) and approximate the desired A_0 by a matrix A_k of the correct rank.

According to this idea, the approximation of A_0 is obtained by zero-ing out the significantly smaller singular values $\sigma_{k+1}, \sigma_{k+1}, \cdots, \sigma_r$ in the expansion

$$\boxed{A = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^t.}$$

The resulting truncaterd approximation A_k is then given by

$$\boxed{A = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^t \approx \sum_{j=1}^{k} \sigma_j \mathbf{u}_j \mathbf{v}_j^t \stackrel{\text{def}}{=} A_k.}$$

Both A and A_k are $m \times n$ - matrices.

The idea of SVD-truncation is powerful, not only for removing noise i the X-data, but also for compressing data and for stabilizing the solution of problems that are extremely ill-conditioned (situations where the condition number σ_1/σ_r of A is large).

It turns out that the truncated SVD is also the mathematically optimal solution in approximation problems, where one wants the best possible low rank approximation a given matrix A.

The 2-*norm* of a $n \times m$-matrix A is denoted $\|A\|_2$ and is induced by the 2-norm (Eucidean norm) maximization of the resulting vectors $A\mathbf{x}$ restricted to A-multiplications with $\mathbf{x}$-vectors from the unit sphere of $\mathbb{R}^m$:

$$\boxed{\|A\|_2 \overset{\text{def}}{=} \sup_{\|\mathbf{x}\|_2=1} \|A\mathbf{x}\|_2.}$$

The *Frobenius norm* of a $n \times m$-matrix A is denoted $\|A\|_F$ and defined by

$$\|A\|_F \overset{\text{def}}{=} \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij}^2},$$

i.e. the square root of all the squared matrix entries (a_{ij}) added together.

2.5 The Eckart–Young–Mirsky theorem (low-rank approximation of matrices) in two versions

,

Assume that the matrix $A \in \mathbb{R}^{m\times n}$ has rank $r > k$. The two rank k matrix approximation problems

$$\boxed{\min_{rank(Z)=k} \|A - Z\|_2} \text{ and } \boxed{\min_{rank(Z)=k} \|A - Z\|_F}$$

both has the same solution

$$\boxed{Z = A_k \overset{\text{def}}{=} U_k \Sigma_k V_k^t,}$$

where $U_k = [\mathbf{u}_1\ \mathbf{u}_2\ \cdots\ \mathbf{u}_k]$, $V_k = [\mathbf{v}_1\ \mathbf{v}_2\ \cdots\ \mathbf{v}_k]$, and $\Sigma_k = diag(\sigma_1, \cdots, \sigma_k)$.

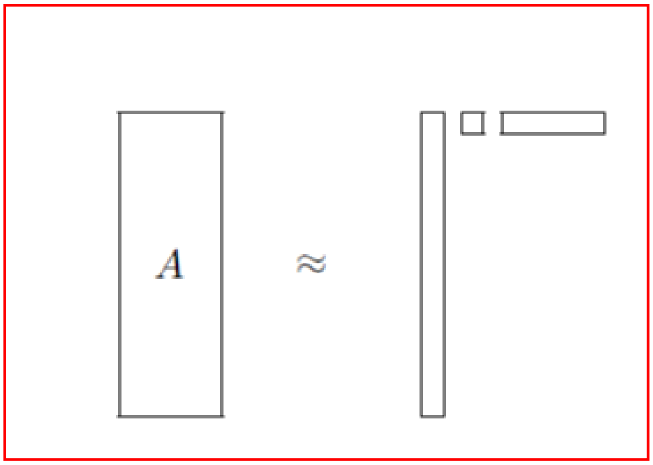
The corresponding minimum values are

$$\boxed{\|A - A_k\|_2 = \sigma_{k+1}} \text{ and } \boxed{\|A - A_k\|_F = \sqrt{\sum_{i=k+1}^{p} \sigma_i^2},}$$

respectively, where $p = \min\{m, n\}$.

A proof of both versions of the Eckart–Young–Mirsky theorem can be found here.

A pictorial illustration of the optimal k rank matrix approximation $A_k = U_k \Sigma_k V_k^t$ of A is



2.6 Principal Component Analysis (PCA)

,

- PCA is much used for both unsupervised learning problems, exploratory data analysis and for making predictive models.

- It is highly useful for dimensionality reduction by projecting the data points onto a set of relatively few principal components to obtain a lower-dimensional approximation of the data that preserves as much of the original data variation as possible.

- The optimal approximation properties of the SVD are appropriate for explaining the fundamental aspects of principal component analysis (PCA).

In the following we assume that the data matrix $X \in \mathbb{R}^{m\times n}$ has rank r, and that each X-column is arranged according to a common ordering of observations of corresponding real-valued random vectors with mean zero. The matrix is assumed to be centered, i.e. the mean of each X-column is equal to zero.

The SVD of the data matrix $X = U_r\Sigma_r V_r^t$, and the right singular vectors $\mathbf{v}_i$ $(i = 1, \cdots, r)$ are called *the principal components directions* (also known as *the loadings*) of X.

The *first principal component* vector

$$\mathbf{t}_1 = X\mathbf{v}_1 = \boxed{\sigma_1\mathbf{u}_1}$$

has the largest possible sample variance among all the possible normalized linear combinations of the X-columns with:

$$Var(\mathbf{t}_1) = Var(X\mathbf{v}_1) = \frac{\sigma_1^2}{m}.$$

In linear algebra terminology finding the vector of maximal variance is equivalent to maximizing the Rayleigh quotient

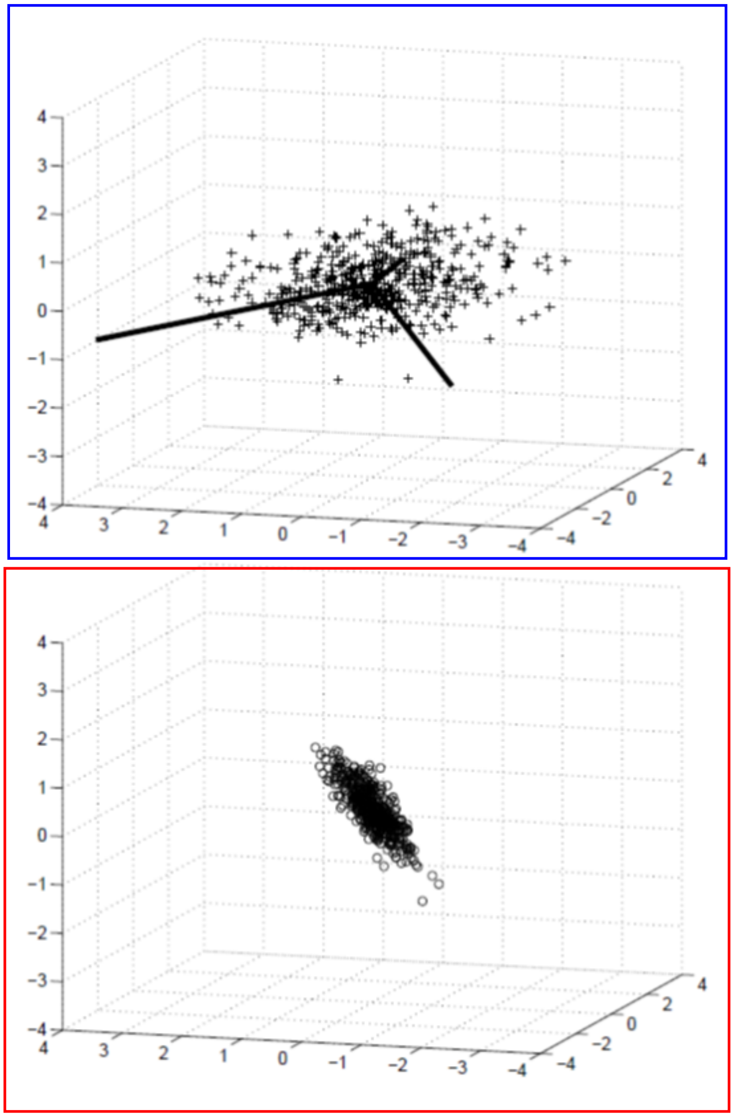
$\sigma_1^2 = \max_{\mathbf{v}\neq\mathbf{0}} \frac{\mathbf{v}^t X^t X\mathbf{v}}{\mathbf{v}^t\mathbf{v}}$ where the corresponding optimal solution vector $\mathbf{v}_1 = \arg\max_{\mathbf{v}\neq\mathbf{0}} \frac{\mathbf{v}^t X^t X\mathbf{v}}{\mathbf{v}^t\mathbf{v}}$.

- The normalized variable $\mathbf{u}_1$ is called the *normalized first principal component* of X.

- The second principal component $\boxed{\mathbf{t}_2 = X\mathbf{v}_2 = \sigma_2\mathbf{u}_{\overset{2}{1}}}$ is the vector of the largest possible sample variance that can be obtained from the deflated data matrix $X_{(1)} = X - \sigma_1\mathbf{u}_1\mathbf{v}_1^t$, and so on.
- Equivalently, any subsequent principal component is defined as the vector of maximal variance subject to the constraint that it must be orthogonal to the previous ones.

The following figure illustrates PCA of a data matrix $X \in \mathbb{R}^{500\times 3}$ containing 500 data points generated artificially from a correlated normal distribution in $\mathbb{R}^3$.

The original data points and the principal components are illustrated in the upper plot. After a deflation of X with respect to $\sigma_1\mathbf{u}_1\mathbf{v}_1^t$, the corresponding data points of the deflated $X_{(1)} = X - \sigma_1\mathbf{u}_1\mathbf{v}_1^t$ are shown in the bottom plot:



It is straightforward to show that the principal components are eigenvectors of the empirical covariance matrix

$$C = \frac{1}{m}X^tX.$$

Thus, the principal components can also be found from the eigendecomposition of the $n \times n$-matrix C.

Since the scalar $\frac{1}{m}$ does not affect the resulting the principal components, it can be omitted from the calculations without loss of information.

- The amount of (empirical) X-variance *explained* by the first k principal compoenets is given by

$$\boxed{\frac{\sum_{i=1}^{k} \sigma_i^2}{m}.}$$

- The corresponding amount in percent (%) of X-variance *explained* by the first k principal compoenets is

$$\boxed{100 \cdot \frac{\sum_{i=1}^{k} \sigma_i^2}{\sum_{i=1}^{p} \sigma_i^2},}$$

where $p = \min\{m, n\}$.