# Singular Value Decomposition and Principal Component Analysis

Rasmus Elsborg Madsen, Lars Kai Hansen and Ole Winther

February 2004

## Introduction

This note is intended as a brief introduction to singular value decomposition (SVD) and principal component analysis (PCA). These are very useful techniques in data analysis and visualization. Further information can found for example in Numerical Recipes, section 2.6,available free online:

`http://www.nr.com/`
`http://www.library.cornell.edu/nr/bookcpdf.html`

and in C. Bishop, *Neural Networks for Pattern Recognition*, Chapter 8.

The note is organized as follows: first we establish the linear algebra of SVD, then we discuss simple properties of the data matrix and principal component analysis and finally we discuss how to use SVD for PCA and some practical issues in connection with using SVD for PCA in matlab.

### Definitions

- The **Singular Values** of the square matrix $\mathbf{A}$ is defined as the square root of the eigenvalues of $\mathbf{A}^T\mathbf{A}$.
- The **Condition Number** is the ratio of the largest to the smallest singular value.
- A matrix is **Ill Conditioned Matrix** if the condition number is too large. How large the condition number can be, before the matrix is ill conditioned, is determined by the machine precision.
- A matrix is **Singular** if the condition number is infinite. The determinant of a singular matrix is 0.
- The **Rank** of a matrix, is the dimension of the range of the matrix. This corresponds to the number of non-singular values for the matrix, i.e. the number of linear independent rows of the matrix.

## Spectral decomposition of a square matrix

Any real symmetric $m \times m$ matrix $\mathbf{A}$ has a spectral decomposition of the form,

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T \tag{1}$$

where $\mathbf{U}$ is an orthonormal matrix (matrix of orthogonal unit vectors: $\mathbf{U}^T\mathbf{U} = \mathbf{I}$ or $\sum_k U_{ki}U_{kj} = \delta_{ij}$) and $\mathbf{\Lambda}$ is a diagonal matrix. The columns of $\mathbf{U}$ are the eigenvectors of matrix $\mathbf{A}$ and the diagonal elements of $\mathbf{\Lambda}$ are the eigenvalues. If $\mathbf{A}$ is positive-definite, the eigenvalues will all be positive. Multiplying with $U$, equation 1 can be re-written to,

$$\mathbf{A}\mathbf{U} \;=\; \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{\Lambda} \tag{2}$$

This can be written as a normal eigenvalue equation by defining the $i$th column of $\mathbf{U}$ as $\mathbf{u}_i$ and the eigenvalues as $\lambda_i = \Lambda_{ii}$:

$$\mathbf{A}\mathbf{u}_i \;=\; \lambda_i\mathbf{u}_i\ . \tag{3}$$

## Singular Value Decomposition

A real $(n \times m)$ matrix, where $n \geq m$ $\mathbf{B}$ has the decomposition,

$$\mathbf{B} = \mathbf{U}\mathbf{\Gamma}\mathbf{V}^T\ , \tag{4}$$

where $\mathbf{U}$ is a $n \times m$ matrix with orthonormal columns ($\mathbf{U}^T\mathbf{U} = \mathbf{I}$), while $\mathbf{V}$ is a $m \times m$ orthonormal matrix ($\mathbf{V}^T\mathbf{V} = \mathbf{I}$), and $\mathbf{\Gamma}$ is a $m \times m$ diagonal matrix with positive or zero elements, called the singular values.

From $\mathbf{B}$ we can construct two positive-definite symmetric matrices, $\mathbf{B}\mathbf{B}^T$ and $\mathbf{B}^T\mathbf{B}$, each of which we can decompose

$$\mathbf{B}\mathbf{B}^T \;=\; \mathbf{U}\mathbf{\Gamma}\mathbf{V}^T\mathbf{V}\mathbf{\Gamma}\mathbf{U}^T = \mathbf{U}\mathbf{\Gamma}^2\mathbf{U}^T \tag{5}$$

$$\mathbf{B}^T\mathbf{B} \;=\; \mathbf{V}\mathbf{\Gamma}^2\mathbf{V}^T \tag{6}$$

Keep in mind that $n \geq m$. We can now show that $\mathbf{B}\mathbf{B}^T$ which is $n \times n$ and $\mathbf{B}^T\mathbf{B}$ which is $m \times m$ will share $m$ eigenvalues and the remaining $n - m$ eigenvalues of $\mathbf{B}\mathbf{B}^T$ will be zero.

Using the decomposition above, we can identify the eigenvectors and eigenvalues for $\mathbf{B}^T\mathbf{B}$ as the columns of $\mathbf{V}$ and the squared diagonal elements of $\mathbf{\Gamma}$, respectively. (The latter shows that the eigenvalues of $\mathbf{B}^T\mathbf{B}$ must be non-negative). Denoting one such eigenvector by $\mathbf{v}$ and the diagonal element by $\gamma$, we have

$$\mathbf{B}^T\mathbf{B}\mathbf{v} = \gamma^2\mathbf{v} \tag{7}$$

then we can multiply on both sides with $\mathbf{B}$ to get,

$$\mathbf{B}\mathbf{B}^T\mathbf{B}\mathbf{v} = \gamma^2\mathbf{B}\mathbf{v} \tag{8}$$

But this means that we have an eigenvector $\mathbf{u} = \mathbf{B}\mathbf{v}$ and eigenvalue $\gamma^2$ for $\mathbf{B}\mathbf{B}^T$ as well, since

$$(\mathbf{B}\mathbf{B}^T)\mathbf{B}\mathbf{v} = \gamma^2\mathbf{B}\mathbf{v} \tag{9}$$

We have now shown that $\mathbf{B}\mathbf{B}^T$ and $\mathbf{B}^T\mathbf{B}$ share $m$ eigenvalues.

We still need to prove that the remaining $n-m$ eigenvalues of $\mathbf{B}\mathbf{B}^T$ is zero. To do that let us consider an eigenvector for $\mathbf{B}\mathbf{B}^T$, $\mathbf{u}_\perp$: $\mathbf{B}\mathbf{B}^T\mathbf{u}_\perp = \beta_\perp \mathbf{u}_\perp$ which is orthogonal to the $m$ eigenvectors $\mathbf{u}_i$ already determined, i.e. $\mathbf{U}^T\mathbf{u}_\perp = 0$. Using the decomposition $\mathbf{B}\mathbf{B}^T = \mathbf{U}\mathbf{\Gamma}^2\mathbf{U}^T$, we immediately see that the eigenvalues $\beta_\perp$ must all be zero,

$$\mathbf{B}\mathbf{B}^T\mathbf{u}_\perp = \mathbf{U}\mathbf{\Gamma}^2\mathbf{U}^T\mathbf{u}_\perp = 0\mathbf{u}_\perp \ .$$

The Rank $\mathcal{R}$ of $\mathbf{B}\mathbf{B}^T$ is determined by the smallest dimension of $\mathbf{B}$, ($\mathcal{R} \leq m$). This ensures that $\mathbf{B}\mathbf{B}^T$ has at most $m$ eigenvalues larger than zero. Note that the relation for $\mathbf{B}\mathbf{B}^T$ corresponds to the usual spectral decomposition since the "missing" $(n-m)$ eigenvalues are zero. It is then evident that the two square matrices can be interchanged. This is a property we can advantage of when dealing with data matrices where we have many more features than examples.

## Properties of a data matrix – first and second moments

Let $\mathbf{x}$ (with components $x_j$ $j = 1, ..., n$) be a stochastic vector with probability distribution $P(\mathbf{x})$. Let $\{\mathbf{x}^\alpha | \alpha = 1, ..., m\}$ be a sample from $P(\mathbf{x})$. We will choose a convention for the data matrix $\mathbf{X}$, where the rows denote the features $j = 1, ..., n$ and the columns the samples $\alpha = 1, ..., m$: in other words the components are $X_{j,\alpha} = x_j^\alpha$.

Principal component analysis is based on the two first empirical moments of the sample data matrix. The mean vector,

$$\langle \mathbf{x} \rangle \equiv \frac{1}{m} \sum_{\alpha=1}^{m} \mathbf{x}^\alpha \tag{10}$$

and the empirical covariance matrix,

$$\mathbf{C} \equiv \frac{1}{m} \sum_{\alpha=1}^{m} \left(\mathbf{x}^\alpha - \langle \mathbf{x} \rangle\right) \left(\mathbf{x}^\alpha - \langle \mathbf{x} \rangle\right)^T \tag{11}$$

Using the matrix formulation we can write

$$\mathbf{C} \equiv \frac{1}{m} \mathbf{X}\mathbf{X}^T \ , \tag{12}$$

where we have removed the mean of the data: $X_{j,\alpha} := X_{j,\alpha} - \langle x_j \rangle$.

## Principal component analysis (PCA)

In principal component analysis we find the directions in the data with the most variation, i.e. the eigenvectors corresponding to the largest eigenvalues of the covariance matrix, and project the data onto these directions. The motivation for doing this is that the most second order information are in these directions.[1] The choice of the number of directions are often guided by trial and error, but principled methods also exist. If we denote the matrix of eigenvectors sorted according to eigenvalue by $\tilde{\mathbf{U}}$, we can then PCA

[1] This also mean we might discard important non-second order information by PCA.

transformation of the data as $\mathbf{Y} = \tilde{\mathbf{U}}^T\mathbf{X}$. The eigenvectors are called the principal components. By selecting only the first $d$ rows of $\mathbf{Y}$, we have projected the data from $n$ down to $d$ dimensions.

## PCA by SVD

We can use SVD to perform PCA. We decompose $\mathbf{X}$ using SVD, i.e.

$$\mathbf{X} = \mathbf{U}\mathbf{\Gamma}\mathbf{V}^T$$

and find that we can write the covariance matrix as

$$\mathbf{C} = \frac{1}{n}\mathbf{X}\mathbf{X}^T = \frac{1}{n}\mathbf{U}\mathbf{\Gamma}^2\mathbf{U}^T \ .$$

In this case $\mathbf{U}$ is a $n \times m$ matrix. Following from the fact that SVD routine order the singular values in descending order we know that, if $n < m$, the first $n$ columns in $\mathbf{U}$ corresponds to the sorted eigenvalues of $\mathbf{C}$ and if $m \geq n$, the first $m$ corresponds to the sorted non-zero eigenvalues of $\mathbf{C}$. The transformed data can thus be written as

$$\mathbf{Y} = \tilde{\mathbf{U}}^T\mathbf{X} = \tilde{\mathbf{U}}^T\mathbf{U}\mathbf{\Gamma}\mathbf{V}^T \ ,$$

where $\tilde{\mathbf{U}}^T\mathbf{U}$ is a simple $n \times m$ matrix which is one on the diagonal and zero everywhere else. To conclude, we can write the transformed data in terms of the SVD decomposition of $\mathbf{X}$.

## PCA by SVD in Matlab

It is common in image processing, sound processing, text processing etc. that we have many more features than samples, $n \ll m$. The covariance matrix itself is therefore very unpleasant to work with because it is very large and as we have proved above singular. However, using the relations eqs. (7) and (9), we find that is suffices to decompose the smaller $m \times m$ matrix

$$\mathbf{D} \equiv \frac{1}{m}\mathbf{X}^T\mathbf{X} \tag{13}$$

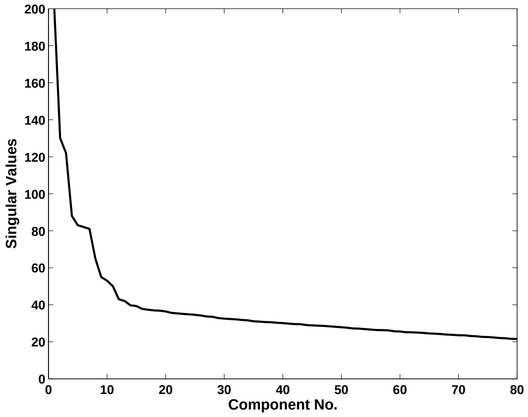
Given a decomposition of $\mathbf{D}$ we can find the interesting non-zero principal directions and components for $\mathbf{C}$, $\mathbf{U} = \mathbf{X}\mathbf{V}\mathbf{S}^{-1}$. You can instruct matlab to always use the smallest matrix by using the command '[u s v] = svd(X,0)', see also 'help svd' in matlab. However, in that case we have to be careful about which matrices to use for the transformation.

## More samples than variables

In some cases, the number of variables is smaller than the number of examples ($n < m$). In these cases, decomposition and dimension reduction might still be desirable for the $n \times m$ matrix $\mathbf{X}$. Dimension change on $\mathbf{X}$ however also results in dimension change on $\mathbf{U}$,$\mathbf{\Gamma}$ and $\mathbf{V}$, who respectively get the sizes $(n \times n)$, $(n \times m)$ and $(m \times m)$. The dimension changes the svd routine in matlab slow and adds unnecessary rows to the $\mathbf{V}$ matrix. The problem can be avoided using "$[V, S, U] = svd(X', 0); U = U'; V = V';$", in the cases where $(n < m)$.

## Number of Principal Directions

The no of principal components to use $d$, is not always easy to determine. The energy fraction could be used to argue for the usage of a given number of principal components. The number of components could also be determined from the characteristics of the singular values. When the singular values stabilize, the remaining components is usually contaminated with much noise and therefore not useful. In the figure below, an example of singular values is shown. From component number 13 and up, the singular values are almost constant, indicating that $d$ should be 12.



## Similar Methods for Dimensionality Reduction

There exists multiple methods that can be used for dimensionality reduction. Some of them are given in the list below.

- Singular Value Decomposition (SVD)
- Independent Component Analysis (ICA)
- Non-negative Matrix Factorization (NMF)
- Eigen Decomposition
- Random Projection
- Factor Analysis (FA)