A Note about Eigenvalues, SVD and PCA

Désiré Sidibé

▶ **To cite this version:**

Désiré Sidibé. A Note about Eigenvalues, SVD and PCA. 2013. hal-00903901

HAL Id: hal-00903901

https://u-bourgogne.hal.science/hal-00903901

Preprint submitted on 13 Nov 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A Note about Eigenvalues, SVD and PCA

Désiré Sidibé*

November 13, 2013

Abstract

Notes on Eigen-decomposition, PCA, SVD and connexions.

1 Matrix preliminaries

A matrix is one way of describing (or representing) a linear transformation between two vector spaces. For instance, a general $m \times n$ matrix A represents a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$.

The matrix acts on vectors $\mathbf{x} \in \mathbb{R}^n$ to produce vectors $\mathbf{y} \in \mathbb{R}^m$ as $\mathbf{y} = A\mathbf{x}$. If we represent the matrix as $A = [\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_n]$, where the $\mathbf{c}_i$'s are the columns of the matrix A, and the vector as $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$, then we have

$$\mathbf{y} = A\mathbf{x} = \sum_{i=1}^{n} x_i \mathbf{c}_i, \tag{1}$$

i.e. $\mathbf{y}$ is a linear combination of the columns of A.

1.1 Column space and nullspace

- The *column space* of A, denoted by $C(A)$ and also called *range* or *span* of A, is the subspace of $\mathbb{R}^m$ such that $y \in C(A)$ if and only if $y = Ax$ for some $x \in \mathbb{R}^n$.
- The *nullspace* of A, denoted by $N(A)$ and also called *kernel*, is the subspace of $\mathbb{R}^n$ such that $x \in N(A)$ if and only if $Ax = 0$.

Note that the column space of a matrix A is exactly the span of all its columns vectors. Therefore, $C(A)$ is just the set of all linear combinations of the columns of A.

*Assistant professor at Université de Bourgogne (dro-desire.sidibe@u-bourgogne.fr)

The nullspace of a matrix A is exactly the set of vectors which are orthogonal to all its row vectors.

For example, consider the following matrix $A = \begin{bmatrix} 1 & 1 & 2 \\ 2 & 1 & 3 \\ 3 & 1 & 4 \\ 4 & 1 & 5 \end{bmatrix}$.

Since the columns of A are in $\mathbb{R}^4$, $C(A)$ is a subspace of $\mathbb{R}^4$. On the other hand, the nullspace $N(A)$ contains all solutions to the equation $A\mathbf{x} = 0$. So $N(A)$ is a subspace of $\mathbb{R}^3$.

1.2 Rank of a matrix

The *rank of a matrix* is the dimension of its column space.

$$rank(A) \doteq dim(C(A)). \tag{2}$$

From this definition, we see that the rank of A is equal to the maximum number of linearly independent columns (or rows) vectors of A.

Properties of rank
For arbitrary $m \times n$ matrix A and $n \times p$ matrix B, rank has the following properties:

1. $0 \leq \text{rank}(A) \leq \min\{m, n\}$
2. $\text{rank}(A) = n - \dim(N(A))$
3. $\text{rank}(A) = \text{rank}(A^T)$
4. $\text{rank}(AB) \leq \min(\text{rank}(A), \text{rank}(B))$
5. $\text{rank}(A^T A) = \text{rank}(AA^T) = \text{rank}(A)$

1.3 Singular and non-singular matrices

A square $n \times n$ matrix A is said to be *non-singular* or *invertible* if there exist a matrix B such that

$$AB = BA = I, \tag{3}$$

where I is the $n \times n$ identity matrix.

B is called inverse of A and denoted by $B = A^{-1}$.

- For a $n \times n$ square matrix A to be invertible, A must be full rank, i.e. $rank(A) = n$.
- If A^{-1} does not exist, we say that the matrix is *singular*.

Assuming A and B are non-singular matrices:

- $(A^{-1})^{-1} = A$
- $(AB)^{-1} = B^{-1}A^{-1}$
- $(A^{-1})^T = (A^T)^{-1}$

2 Eigenvalues/Eigenvectors

Let A be a square $n \times n$ matrix, i.e. a linear map from $\mathbb{R}^n$ to itself.

2.1 Intuitive description

In the following simple example, we set $n = 2$.

Let $A = \begin{bmatrix} 1 & 3 \\ 3 & 1 \end{bmatrix}$, and let $\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\mathbf{x}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ be two vectors in $\mathbb{R}^2$.

When applying the linear transformation A to $\mathbf{x}_1$ and $\mathbf{x}_2$ wet get two new vectors in $\mathbb{R}^2$, respectively $A\mathbf{x}_1 = \begin{bmatrix} 4 \\ 4 \end{bmatrix}$ and $A\mathbf{x}_2 = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$.

We can see that $A\mathbf{x}_1$ and $\mathbf{x}_1$ are "parallel" vectors (i.e. they point in the same direction) whereas $A\mathbf{x}_2$ and $\mathbf{x}_2$ are in different directions. So, they are some vectors in $\mathbb{R}^2$ which are invariant (talking here about their direction) when multiplying by A. Those vectors are called *eigenvectors*, and the scaling factors (in our example $A\mathbf{x}_1 = 4\mathbf{x}_1$, so the factor is $\lambda = 4$) are called *eigenvalues* of A.

Therefore, intuitively, eigenvectors are vectors of $\mathbb{R}^n$ that are "invariant" under the linear transformation represented by A.

2.2 Definition

Given a square $n \times n$ matrix A, we say that $\lambda \in \mathbf{C}$ is an *eigenvalue* of A and $\mathbf{x} \in \mathbf{C}$ in the corresponding *eigenvector* if

$$A\mathbf{x} = \lambda\mathbf{x}, \ \mathbf{x} \neq 0. \tag{4}$$

The set of all eigenvalues of a matrix A is called its *spectrum*, denoted by $\sigma(A)$.

The Matlab command $[V, D] = eig(A)$ produces a diagonal matrix D of eigenvalues and a full-rank matrix V whose columns are the corresponding eigenvectors, so that $AV = VD$.

Properties of eigenvalues

- The sum of the eigenvalues of A is equal to its trace

$$\text{trace}(A) = \sum_{i=1}^{n} A_{ii} = \sum_{i=1}^{n} \lambda_i.$$

- The determinant of A is equal to the product of its eigenvalues

$$\det(A) = |A| = \prod_{i=1}^{n} \lambda_i.$$

- The rank of A is equal to the number of non-zero eigenvalues.

- If A is a non-singular matrix (all of its eigenvalues are non-zero) then $1/\lambda_i$ is an eigenvalue of A^{-1} with associated eigenvector $\mathbf{x}_i$.

Finding eigenvalues and eigenvectors

We can rewrite Eq. 4 as $(A - \lambda I)\mathbf{x} = 0,\ \mathbf{x} \neq 0$, which mean that the non-zero vector $\mathbf{x}$ is in the nullspace of the matrix $(A - \lambda I)$.

From the properties of rank, this also means that the matrix $(A - \lambda I)$ is singular. Therefore, we have

$$\det(A - \lambda I) = |A - \lambda I| = 0. \tag{5}$$

To sum up, in order to find the eigenvalues/eigenvectors of A, we have to:

1. First find the eigenvalues which are the roots of the characteristic equation $det(A - \lambda I) = 0$.
2. Once we know the eigenvalues, λ's, we just have to find the nullspaces of the matrices $(A - \lambda I)$ and any vector in these linear spaces is an eigenvector.

Consider an example, $A = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$.

We first find the eigenvalues of A

$$det(A - \lambda I) = \begin{vmatrix} 3-\lambda & 1 \\ 1 & 3-\lambda \end{vmatrix} = (3-\lambda)^2 - 1 = (\lambda - 4)(\lambda - 2)$$

Thus the eigenvalues are equal to $\begin{cases} \lambda_1 = 4 \\ \lambda_1 = 2 \end{cases}$

We finally find the eigenvectors

- For $\lambda_1 = 4$

$$A - \lambda_1 I = A - 4I = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}$$

So a vector in the nullspace of $(A - 4I)$ is $x_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

- For $\lambda_2 = 2$

$$A - \lambda_2 I = A - 2I = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

So a vector in the nullspace of $(A - 2I)$ is $x_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$

The reader can easily verify that $Ax_1 = \lambda_1 x_1$ and $Ax_2 = \lambda_2 x_2$. We can also check that $\lambda_1 + \lambda_2 = \text{trace(A)} = 6$ and $\lambda_1 \lambda_2 = \det(A) = 8$.

2.3 Diagonalization and powers of A

Let A be a $n \times n$ matrix with n independent eigenvectors $X_i, i = 1 \dots n$. If we put all eigenvectors as columns of a matrix S, then we have

$$\begin{aligned} AS &= A[X_1\ X_2\ \cdots X_n] \\ &= [AX_1\ AX_2\ \cdots AX_n] \\ &= [\lambda_1 X_1\ \lambda_2 X_2\ \cdots \lambda_n X_n] \\ &= [X_1\ X_2\ \cdots X_n] \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix} \\ &= S\Lambda \end{aligned}$$

Since the eignevectors of A are independent, the matrix S is invertible. So, we have

$$AS = S\Lambda \Rightarrow A = S\Lambda S^{-1}$$

This is called the *diagonalization* of A.

The reader can verify that the matrix A of the previous example can be diagonalized as:

$$A = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1/2 & 1/2 \\ -1/2 & 1/2 \end{bmatrix}$$

NOTE

- We can diagonalize A only if A has n independent eigenvectors.
- A is guaranteed to have independent eigenvectors if A has n different eigenvalues.
- $\forall\ k \geq 1,\ A^k = S\Lambda^k S^{-1}$

2.4 Symmetric matrices

A square $n \times n$ matrix A is said to be symmetric if $A = A^T$.

Two remarkable properties of symmetric matrices are:

- All the eigenvalues of A are real ($\lambda_i \in \mathbb{R}; \forall i$)
- The eigenvectors of A form an orthonormal basis of $\mathbb{R}^n$.

  So we can diagonalize A as $A = S\Lambda S^T$.

3 SVD

The singular value decomposition (SVD) can be seen as a "generalization" of the concept of eigenvalue-eigenvector pairs to non-square matrices. SVD is also a useful tool to capture essential features of a matrix such as its rank and nullspace. In short, SVD is the most useful factorization of a matrix.

Theorem
Let $A \in \mathbb{R}^{m \times n}$ be a general $m \times n$ matrix with rank equal to r. Furthermore, suppose, without loss of generality, that $m \geq n$. Then

- $\exists\, U \in \mathbb{R}^{m \times r}$ whose columns are orthonormal,
- $\exists\, V \in \mathbb{R}^{n \times r}$ whose columns are orthonormal, and
- $\exists\, \Sigma \in \mathbb{R}^{r \times r},\ \Sigma = \mathit{diag}\{\sigma_1, \sigma_2, \ldots, \sigma_r\}$ diagonal with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$

such that $A = U\Sigma V^T$.

The factorization $A = U\Sigma V^T$ is the SVD of the matrix A and the above theorem shows that such factorization always exists.

It has to be noticed that the rank of A is equal to the number of non-zero singular values.

Relation with eigen-decomposition
Form $A = U\Sigma V^T$, we see that

$$\begin{aligned} A^T A &= (U\Sigma V^T)^T (U\Sigma V^T) \\ &= (V\Sigma U^T)(U\Sigma V^T) \\ &= V\Sigma^2 V^T \end{aligned}$$

This last equation $A^T A = V\Sigma^2 V^T$ is the diagonalization of the symmetric matrix $A^T A$. Similarly, we can show $AA^T = U\Sigma^2 U^T$ which is the eigen-decomposition (or diagonalization) of the symmetric matrix AA^T.

Therefore, we can conclude that:

- the columns of the orthogonal matrix V are the eigenvectors of $A^T A$,
- the columns of the orthogonal matrix U are the eigenvectors of AA^T,
- the singular values of A are the square roots of eigenvalues of $A^T A$ (or AA^T).

Properties of SVD

- The rank of a matrix is equal to the number of non-zero singular values.
- A square $n \times n$ matrix A is non-singular if and only if $\sigma_i \neq 0\ \forall i$.

- If A is a $n \times n$ nonsingular matrix, then A^{-1} is given by

$$A^{-1} = V\Sigma^{-1}U^T,$$

where $\Sigma^{-1} = diag\{\frac{1}{\sigma_1}, \cdots, \frac{1}{\sigma_n}\}$.

3.1 Sum of rank one matrices

The SVD of an $m \times n$ matrix A of rank r is given by $A = U\Sigma V^T$, which can also be written as

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T,$$

where u_i and v_i are the columns of U and V respectively.

Note that each term of this sum, i.e. each $u_i v_i^T$, is a $m \times n$ matrix of rank one. So, the matrix A is a sum of rank one matrices that are orthogonal with respect to the matrix inner product.

Truncating the sum at p terms defines a rank p matrix $A_p = \sum_{i=1}^{p} \sigma_i u_i v_i^T$. If we approximate A with A_p, then we make an error equals to $E_p = A - A_p = \sum_{i=p+1}^{k} \sigma_i u_i v_i^T$. It can be shown that A_p is the best rank p approximation to the matrix A.

The low rank approximation of a matrix can, for example, be used for image compression.

4 PCA

Principal component analysis (PCA) is one of the most widely used technique for data analysis.

The main goal of PCA is to reduce a complex data set to a lower dimension to reveal the, sometimes hidden, simplified structure that underlines it. So, we can see PCA essentially as a change of basis; and we would like to compute the most meaningful basis to re-express our data with the hope that the new basis will reveal hidden structure of the data and remove the redundancy.

4.1 Data representation

Say we have N samples, each of which is a point in $\mathbb{R}^L$. We represent our data as a data matrix $\mathbf{X} = [X_1, \cdots, X_N]$, where each column X_i represents a sample point of dimension L. Thus, $\mathbf{X}$ is a matrix of size $L \times N$.

The covariance matrix of the data is then given by the equation:

$$\mathbf{C}_X = \widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T, \tag{6}$$

where $\widetilde{\mathbf{X}} = [X_1 - \bar{X}, \cdots, X_N - \bar{X}]$ is the zero-mean data matrix ($\bar{X}$ being the mean vector).

The covariance matrix $\mathbf{C}_X$ is a square $L \times L$ symmetric matrix that encodes the correlation between the different features; the diagonal element of $\mathbf{C}_X$ contain the variances of each feature, and the off-diagonal elements correspond to the covariances between different features.

4.2 PCA derivation

First, note that if all our features were uncorrelated then the covariance matrix would be diagonal. Since the goal is to remove redundancy in the data (which corresponds to removing correlation between the variables), we have to find a transformation (a change of basis) that makes $\mathbf{C}_X$ is diagonal matrix.

In other words, we want to find a matrix P, such that if $\mathbf{Y} = P\mathbf{X}$, then the covariance matrix of $\mathbf{Y}$ is diagonal.

Since $\mathbf{C}_X$ is a symmetric matrix, we can diagonalize it as $\mathbf{C}_X = V\Lambda V^T$. Which gives $\Lambda = V^T\mathbf{C}_X V$.

Let choose $P = V^T$, i.e. $\mathbf{Y} = P\mathbf{X} = V^T\mathbf{X}$. Then, the covariance matrix of transformed data $\mathbf{Y}$ is

$$\begin{aligned} \mathbf{C}_Y &= \mathbf{Y}\mathbf{Y}^T \\ &= (V^T\mathbf{X})(V^T\mathbf{X})^T \\ &= V^T(\mathbf{X}\mathbf{X}^T)V \\ &= V^T\mathbf{C}_X V \\ &= \Lambda \end{aligned}$$

So, setting $P = V^T$ makes the covariance matrix of transformed data to be diagonal, what we wanted to achieve.

The rows of the matrix P are the vectors of the new basis on which to re-express the data. This vectors are called *principal components*.

We can observe that

- The principal components of $\mathbf{X}$ are the eigenvectors of the covariance matrix $\mathbf{C}_X$.
- The corresponding eigenvalues give the amount of information carried by each principal component.

4.3 Dimension reduction

Usually, we want to reduce the dimensionality of the problem, i.e. we want to represent each of our data point in a lower dimensional space $\mathbb{R}^K$, with $K \ll L$.

Dimensionality reduction is achieved by projecting the data points onto the K principal components corresponding to the K largest eigenvalues of the covariance matrix $\mathbf{C}_X$.

One question is how to choose the number of principal components (or how to fix the value of K)? To choose K, we use the following criterion which takes into account the amount of information carried by each eigenvector:

How many components to keep?
Choose K such that

$$(\sum_{i=1}^{K} \lambda_i)/(\sum_{i=1}^{N} \lambda_i) > Threshold.$$

Typical threshold values are 0.9 or 0.95.

It can be shown that the reconstruction error $e = \left\|X - \widehat{X}\right\|$ is minimized using the principal components. This error is equal to

$$e = \frac{1}{2} \sum_{i=K+1}^{N} \lambda_i.$$

Data normalization
Finally, note that the principal components depend on the *units* and *range* of the original data. Therefore, we should always normalize the data prior to using PCA. A common normalization method is to transform all the data to have zero mean and unit standard deviation:

$$X_i' = \frac{X_i - \mu}{\sigma},$$

μ and σ being, respectively, the mean and standard deviation of the X_i's.

4.4 PCA algorithm

Here, we summarize the methodology to perform PCA. Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ be vectors in $\mathbb{R}^L$.

- Step 1: compute the mean vector $\overline{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i$.
- Step 2: subtract the mean $\Phi_i = \mathbf{x}_i - \overline{\mathbf{x}}$ for $i = 1, \ldots, N$.
- Step 3: form the $L \times N$ matrix $\mathbf{A} = [\Phi_1 \; \Phi_2 \; \cdots \; \Phi_N]$ and compute

$$\Sigma = \frac{1}{L} \sum_{i=1}^{N} \Phi_i \Phi_i^T = \frac{1}{L} \mathbf{A}\mathbf{A}^T.$$

- Step 4: compute the eigenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_L$ and the corresponding eigenvectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_L$ of Σ.
- Step 5: since Σ is symmetric, its eigenvectors form a basis. So any vector $\mathbf{x} - \overline{\mathbf{x}}$ can be expressed as $\mathbf{x} - \overline{\mathbf{x}} = \sum_{i=1}^{L} b_i \mathbf{u}_i$.

  To perform dimensionality reduction, keep only the vectors corresponding to the K largest eigenvalues:

$$\widehat{\mathbf{x}} - \overline{\mathbf{x}} = \sum_{i=1}^{K} b_i \mathbf{u}_i \text{ where } K << N.$$

4.5 Size trick

In many applications, it happens that we have few data but many variables, i.e. $N \ll L$. For example, if we have 100 images each of size 400x600, then our data matrix $\mathbf{X}$ has dimensions $N = 100$ and $L = 240,000$.

In such a case, computing the eigenvalues/eigenvectors of $\mathbf{C}_X$ as above may be difficult if L is too large.

The **size trick** here is to work with the other covariance matrix

$$\mathbf{C}'_X = \widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}.$$

This new matrix is of size $N \times N$ instead of $L \times L$ as $\mathbf{C}_X$.

Let $\mathbf{C}'_X = V'\Lambda'V'^T$, be the eigen-decomposition of $\mathbf{C}'_X$.

The fact is that both matrices $\mathbf{C}_X$ and $\mathbf{C}'_X$ have the same non-zero eigenvalues (they have same rank). Hence, we have $\Lambda = \Lambda'$.

What about the eigenvectors?

Let $\mathbf{x}$ be an eigenvector of $\mathbf{C}'_X$. Then $\mathbf{C}'_X\mathbf{x} = \lambda\mathbf{x}$.

That is $\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}\mathbf{x} = \lambda\mathbf{x} \Rightarrow (\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^T)\widetilde{\mathbf{X}}\mathbf{x} = \lambda\widetilde{\mathbf{X}}\mathbf{x} \Rightarrow \Sigma(\widetilde{\mathbf{X}}\mathbf{x}) = \lambda\widetilde{\mathbf{X}}\mathbf{x}$.

In other words, we have $\mathbf{C}_X(\widetilde{\mathbf{X}}\mathbf{x}) = \lambda(\widetilde{\mathbf{X}}\mathbf{x})$, and we can conclude that the eigenvectors of $\mathbf{C}_X$ and those of C'_X are related by the equation

$$V = \widetilde{\mathbf{X}}V'.$$

4.6 PCA & SVD

From Section 3, we know that the SVD of a matrix A is given by $A = U\Lambda V^T$, where U and V are orthogonal matrices.

Moreover, we have also seen that the columns of V are the eigenvectors of AA^T, while the columns of U are eigenvectors of A^TA.

So, we can perform PCA without computing the covariance matrix, but from SVD directly (of course, after transforming the data to have zero-mean). The advantage of this option is that we directly apply the size-trick. In short, if $\mathbf{X}$ is an $L \times N$ matrix with $N < L$, then we know the rank of $\mathbf{X}$ is at most equal to N (see properties of rank in Section 1.2). So we don't need to compute all L singular values (and vectors) of $\mathbf{X}$ as many of them will be zero. We can simply compute N singular values (and the corresponding vectors), and this is exactly the size-trick explained above.

References

1. Gilbert strang, "Introduction to Linear Algebra", 4th edition, Wellesley Cambridge Press, 2009.

2. Désiré Sidibé, "Applied Mathematics", Lecture Notes, Université de Bourgogne, 2012.