COMP229: Introduction to Data Science
Lecture 29: Singular Value Decomposition

Olga Anosova, O.Anosova@liverpool.ac.uk
Autumn 2023, Computer Science department
University of Liverpool, United Kingdom

Lecture plan & learning outcomes

On this lecture we should learn

- What is Singular Value Decomposition
- How to perform SVD
- Ways to evaluate PCA
- Assumptions of PCA
- Why infinity is not a number

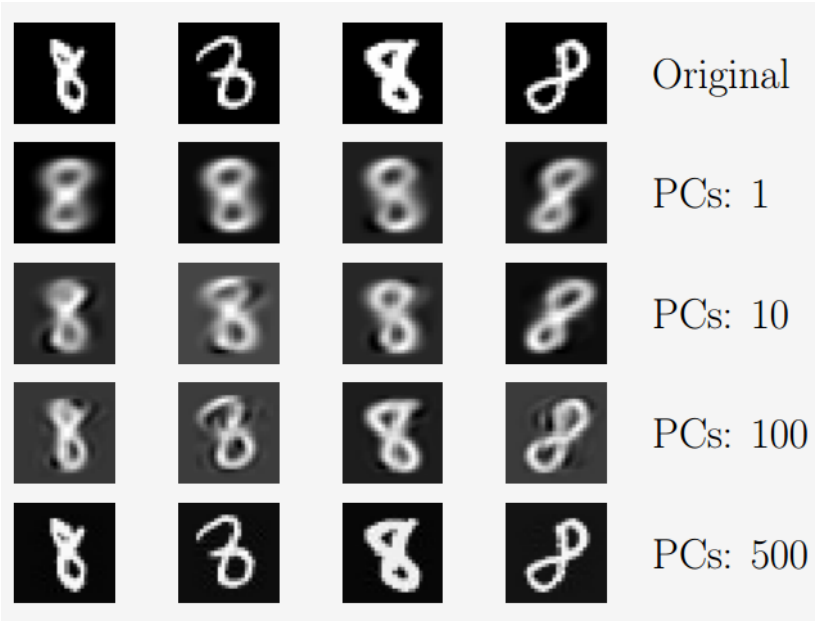
Reminder: PCA via Eigendecomposition (EIG)

- Step 1: subtract the means so that the rows of the sample $k \times n$ matrix S have mean 0.
- Step 2: find the $k \times k$ covariance matrix $M = \frac{SS^T}{n-1}$.
- Step 3: a few eigenvectors of M with largest eigenvalues span an approximating subspace.

Remember to check the assumptions!

PCA and image compression

Each digit 8 from the MNIST database is a grayscale image of size $28 \times 28 = 784$ pixels.



Every image is represented by a vector in $\mathbb{R}^{784}$. The picture shows reconstructions of images from a principal subspace with 1, 10, 100, 500 principle components.

Image from the book 'Mathematics for Machine Learning'.

SVD is related to PCA

We'll find principal components of data by a more general Singular Value Decomposition (SVD), which decomposes any rectangular matrix into a product of 3 simpler matrices (two rotations, one scaling).

Recall that any sample data can be represented by a $k \times n$ matrix S, where s_{ij} is the j-th sample value of the i-th feature, i.e. the rows of S correspond to measurement types, the columns represent trials (several attempts to measure the same feature).

The covariance matrix of data

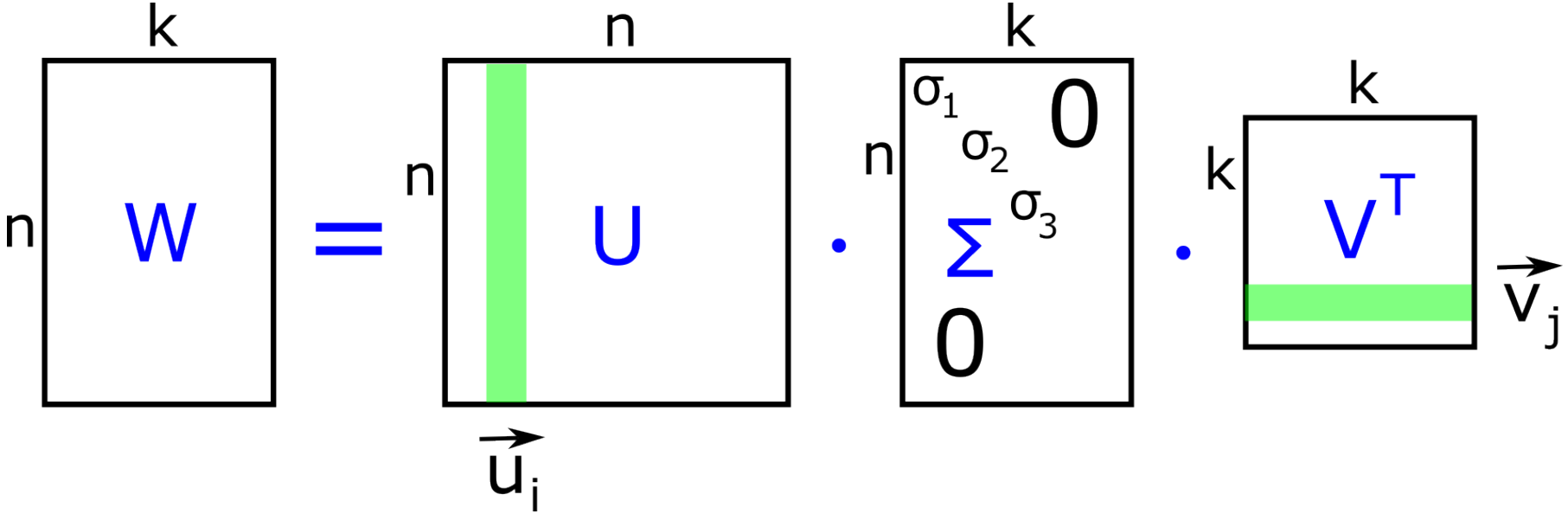
Lecture 28 studied principal components (directions) in a data cloud of n points in $\mathbb{R}^k$ as eigenvectors of the covariance $k \times k$ matrix $M = \dfrac{SS^T}{n-1}$.

If rows of S have zero means, consider the $n \times k$ matrix $W = \dfrac{S^T}{\sqrt{n-1}}$ whose columns have means 0.

Then $W^T W = \dfrac{(S^T)^T S^T}{n-1} = \dfrac{SS^T}{n-1} = M$ is the covariance matrix of the data. The SVD will be applied to W.

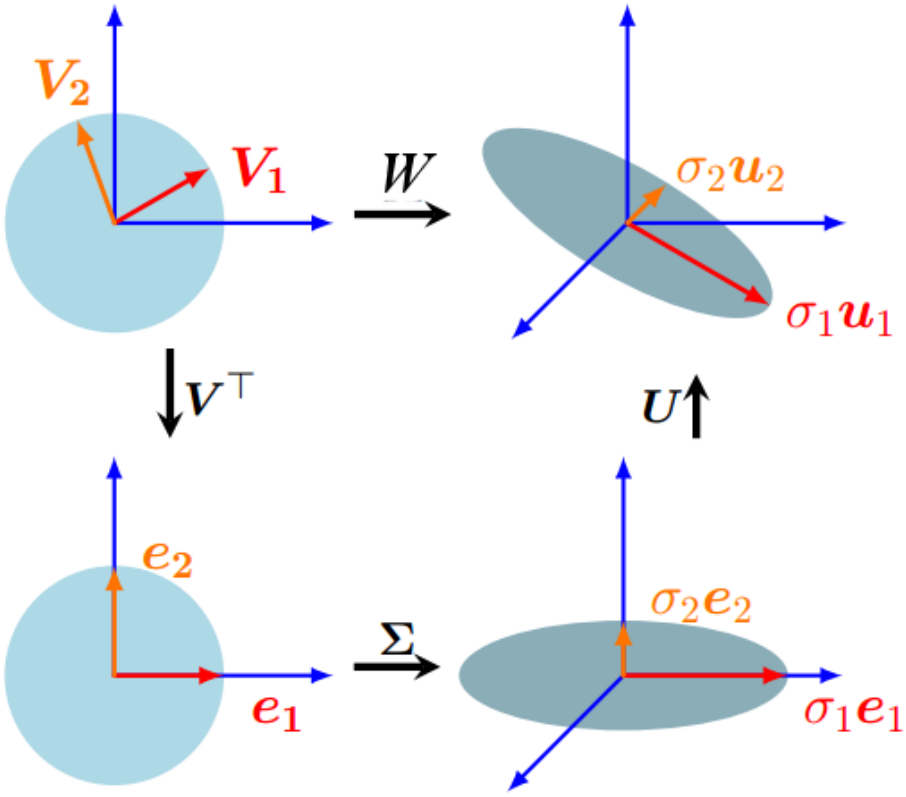
Singular Value Decomposition

Definition 29.1. A **Singular Value Decomposition** of any $n \times k$ matrix W is $U\Sigma V^T$, where U, V are orthogonal matrices (high-dimensional rotations) and Σ is a diagonal (scaling) matrix with ordered **singular** values $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant 0$ on the diagonal.



Vectors $\vec{u}_i$ and $\vec{v}_j$ are **left-singular** and **right-singular** vectors.

Two rotation matrices in the SVD



Informal interpretation of SVD: any linear map $\mathbb{R}^k \to \mathbb{R}^n$ represented by a matrix W can be written as a rotation $\times$ scaling $\times$ rotation.

In $W = U \cdot \Sigma \cdot V^T$, the first $n \times n$ matrix U is a rotation in the space $\mathbb{R}^n$ of trials, the last $k \times k$ matrix V is a rotation in the space $\mathbb{R}^k$ of features.

Orthogonality

By Claim 21.8 any orthogonal matrix satisfies $V^{-1} = V^T$. Let $\vec{u}_1, \ldots, \vec{u}_n$ be the columns of U (*left-singular* vectors of W), and $\vec{v}_1, \ldots, \vec{v}_k$ be the columns of V (*right-singular* vectors of W) .

By Claim 21.9 orthogonality of U, V means that

$\vec{u}_1, \ldots, \vec{u}_n$ is an orthonormal basis in $\mathbb{R}^n$: $\vec{u}_i \cdot \vec{u}_j = \delta_{ij}$,

$\vec{v}_1, \ldots, \vec{v}_k$ is an orthonormal basis in $\mathbb{R}^k$: $\vec{v}_i \cdot \vec{v}_j = \delta_{ij}$,

with the **Kronecker delta** symbol $\delta_{ij} = \begin{cases} 0, & \text{if } i \neq j, \\ 1, & \text{if } i = j \end{cases}$.

The basis maps to the scaled basis

Claim 29.2. $W\vec{v}_i = \sigma_i \vec{u}_i$ for $i \leqslant \min\{k, n\}$.

Proof. The SVD formula $W = U \cdot \Sigma \cdot V^T$ implies that $W \cdot V = (U\Sigma)(V^T V) = U \cdot \Sigma$. Split the last matrix identity into the identities for columns:

the i-th column of $W \cdot V$ is the vector $W\vec{v}_i$,

the i-th column of $U \cdot \Sigma$ is $\sigma_i \vec{u}_i$.

So W maps the vectors $\vec{v}_i$ to the scaled vectors $\sigma_i \vec{u}_i$ for $i \leqslant \min\{k, n\}$, otherwise to $\vec{0}$. □

Properties of the transpose

Claim 29.3. (a) $(AB)^T = B^T A^T$, (b) $(A^T)^{-1} = (A^{-1})^T$, (c) $(A^T)^T = A$, (d) $(A+B)^T = A^T + B^T$.

Proof. (a) AB has the element $\sum_{k=1}^{n} a_{ik} b_{kj}$ in row i, column j.

Then $((AB)^T)_{ij} = \sum_{k=1}^{n} a_{jk} b_{ki} = \sum_{k=1}^{n} (B^T)_{ik} (A^T)_{kj} = (B^T A^T)_{ij}$.

(b) The transpose of the inverse is found as follows:

$I = I^T = (AA^{-1})^T = (A^{-1})^T A^T$.

(c),(d) are obvious.

(c) means that $(A^{-1})^T$ coincides with $(A^T)^{-1}$. □

The covariance matrix M of S

The original sample $k \times n$ matrix S has the following covariance matrix (use Claim 29.3):

$M = W^T W = (U \cdot \Sigma \cdot V^T)^T (U \cdot \Sigma \cdot V^T) =$
$(V \Sigma^T U^T) U \Sigma V^T = V(\Sigma^T \Sigma) V^{-1}$, where $\Sigma^T \Sigma$ is the $k \times k$ matrix with the diagonal elements σ_i^2.

Compare to PCA in Lecture 28: $M = BDB^{-1}$, where $B = V$ is an orthogonal matrix with columns equal to the eigenvectors of M, $D = \Sigma^T \Sigma$ is the diagonal matrix consisting of the eigenvalues of M.

Columns of the matrices U, V

Claim 29.4. In $W = U \cdot \Sigma \cdot V^T$, the columns of V are the eigenvectors of $W^T W$. The columns of U are the eigenvectors of WW^T. The squares σ_i^2 are the eigenvalues of WW^T (also of $W^T W$).

Proof of the 2nd part is similar to the 1st part on the last slide: $WW^T = U \cdot \Sigma \cdot V^T (U \cdot \Sigma \cdot V^T)^T = U \cdot \Sigma \cdot V^T (V \Sigma^T U^T) = U(\Sigma \Sigma^T) U^{-1}$.

The final expression has the diagonal $n \times n$ matrix $\Sigma \Sigma^T$ consisting of σ_i^2 on the diagonal. $\square$

Comparison of two decompositions

The eigendecomposition (EIG) is $M = BDB^T$.
The SVD decomposition is $W = U\Sigma V^T$.

The SVD is a more general method, because SVD works for any matrix $W_{n\times k}$, the eigendecomposition is defined for a subclass of square diagonalizable matrices $M_{k\times k}$.

The EIG algorithm is usually faster than SVD.
But the EIG is less numerically stable than SVD, because the product $M = W^T W$ can have very small entries even if W is good, which leads to loss of accuracy.

An example of the SVD

Data S: $k = 2$ marks, $n = 5$ students, means =0.

subjects/students	s_{i1}	s_{i2}	s_{i3}	s_{i4}	s_{i5}
Maths	1	0	0	-1	0
English	0	1	0	0	-1

Then $W = \dfrac{S^T}{\sqrt{n-1}} = \dfrac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ -1 & 0 \\ 0 & -1 \end{pmatrix}$, $W^T = \dfrac{1}{2}\begin{pmatrix} 1 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & -1 \end{pmatrix}$

Find the covariance matrix $M = W^T W$.

Eigenvalues and singular values

The covariance $M = W^T W = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}$ has the eigenvalues $\lambda_1 = \lambda_2 = 0.5$ and singular values $\sigma_1 = \sigma_2 = \frac{1}{\sqrt{2}}$ with the orthonormal eigenvectors $\vec{v}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $\vec{v}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$.

Then $V = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and $\Sigma = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$.

Now write the 5×5 matrix U.

The new 5×5 matrix WW^T

The columns of U in the SVD formula $W = U\Sigma V^T$ are orthonormal eigenvectors of WW^T.

$$W = \frac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ -1 & 0 \\ 0 & -1 \end{pmatrix}, \text{ then } WW^T = \frac{1}{4}\begin{pmatrix} 1 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 & 1 \end{pmatrix}$$

has eigenvectors equal to the columns of U (the left singular vectors of W).

Eigenvalues λ_i of W^TW and WW^T

The non-zero eigenvalues are the same as for W^TW:

$\lambda_1 = \lambda_2 = 0.5 > \lambda_3 = \lambda_4 = \lambda_5 = 0.$

Check: divide the eigenvalues of $4WW^T$ by 4

$\det(4WW^T - \lambda I) = -\lambda^3(\lambda - 2)^2 = 0.$

Instead of directly finding eigenvectors of WW^T, use Claim 29.2: $W\vec{v}_i = \sigma_i \vec{u}_i$ for $i \leqslant \min\{k, n\}$,

or $\vec{u}_i = \dfrac{1}{\sigma_i} W\vec{v}_i$ for $i = 1, 2.$

Then $\vec{u}_1 = \frac{1}{\sqrt{2}}(1, 0, 0, -1, 0)^T$, $\vec{u}_2 = \frac{1}{\sqrt{2}}(0, 1, 0, 0, -1)^T$.

If $\lambda_3 = \lambda_4 = \lambda_5 = 0$, then $\vec{u}_3 = (0, 0, 1, 0, 0)^T$,

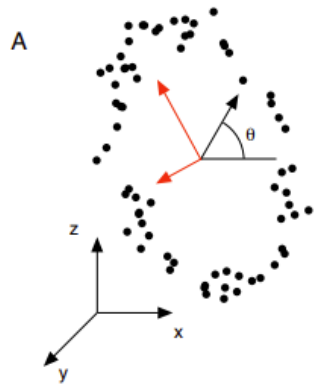
$\vec{u}_4 = \frac{1}{\sqrt{2}}(1, 0, 0, 1, 0)^T$, $\vec{u}_5 = \frac{1}{\sqrt{2}}(0, 1, 0, 0, 1)^T$.

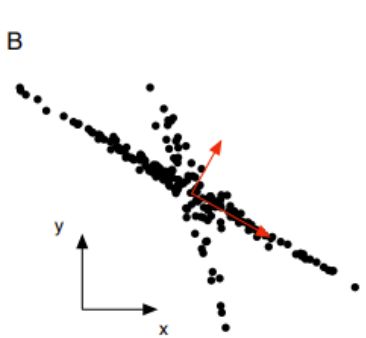
SVD formula: $W = U\Sigma V^T$

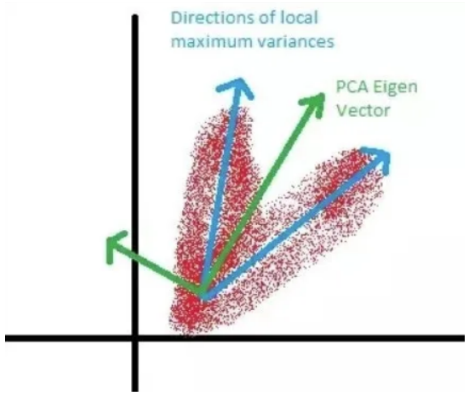
$$W = \frac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ -1 & 0 \\ 0 & -1 \end{pmatrix},\ U = \begin{pmatrix} \frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} & 0 \\ 0 & \frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} \\ 0 & 0 & 1 & 0 & 0 \\ -\frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} & 0 \\ 0 & -\frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{\sqrt{2}} \end{pmatrix},$$

$$\Sigma = \begin{pmatrix} \frac{1}{\sqrt{2}} & 0 \\ 0 & \frac{1}{\sqrt{2}} \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix},\ V = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$ Check now that $W = U\Sigma V^T$.

Does PCA always work?



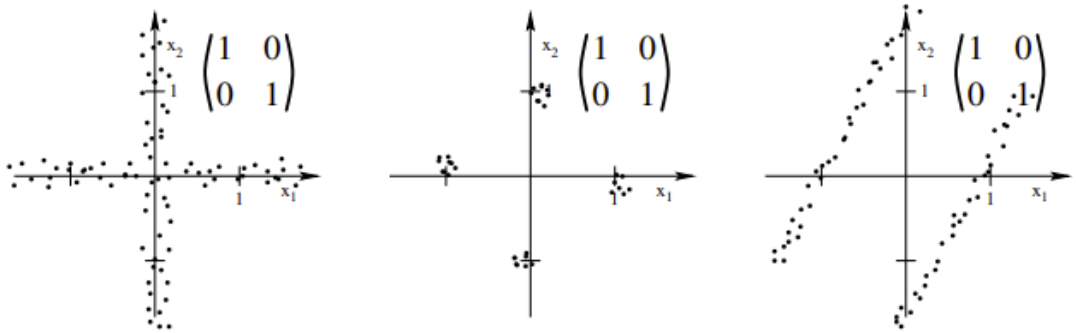




On A the dynamics can be described by the non-linear elliptic "rotation" with just one paramenter.

On the other two the data is non-Gaussian with non-orthogonal axes.

Covariance matrix does not distinguish non-Normal distributions:



Images from Interview questions, PCA tutorial and Stanford module.

Optimality of PCA

To address these problems, we must define what we consider **optimal results**.

In dimentionaluty reduction, it's reconstruction from the reduced data, the quality of which is measured via an error (or loss) function.

The usual loss function is the *mean squared error*, and it has been proven that PCA provides the optimal reduced representation of the data. This seems to contradict the pictures above.

What's the goal? Does our goal coinside with the one set in the method? The goal of the method is defined by (is hidden in) the assumptions: PCA goal is to decorrelate the data.

Assumptions of PCA

Linearity: given data are near a linear subspace. *A non-linear slower extension exits and is called a kernel PCA (KPCA).*

Sufficiency of the mean and variance: the data is *normally distributed* near the linear subspace above.

The signal-to-noise ratio is high, i.e. the data points are distributed along a linear signal with a large variance, while the noise has a low variance. *Which is sensitive to the choice of measurement units, ok for unitless values.*

The principal directions are orthogonal to each other. *Independent Component Analysis (ICA) is an extension for statistically independent components, it's again slower.*

PCA is **not a feature selection** as principal directions are not the subset of the initial features.

Advantage of PCA

PCA is completely **nonparametric**:
any data can be plugged in and an answer comes out, it is unique and independent of the user.

Which is both a blessing and a curse.

But, as discovered, lots of things can be linearly approximated.

Even a derivative. Though a derivative of a simplest possible real-life function, i.e. a step function does not really exist (or is not a function).
Because many mathematical tools are hiding infinity, which is not a number, hence **require understanding**.

Time to revise and ask questions

- Singular value decomposition: $W = U\Sigma V^T$, where U, V are orthogonal, Σ is diagonal with square roots of eigenvalues of $W^T W$.
- This method will always produce a result, hence relies on our understanding!

$$0 = 0 + 0 + 0 + 0 + \dots$$
$$0 = (1-1) + (1-1) + (1-1) + \dots$$
$$0 = 1 - 1 + 1 - 1 + 1 - 1 + \dots$$
$$0 = 1 + (-1) + 1 + (-1) + 1 + (-1) + \dots$$
$$0 = 1 + (-1 + 1) + (-1 + 1) + (-1 + 1) + \dots$$
$$0 = 1 + 0 + 0 + 0 + \dots$$
$$0 = 1$$

Additional references

- Classic PCA tutorial
- PCA lecture notes from Stanford Computer graphics lab
- Job interview questions on PCA