

Principal components analysis

Victor Kitov

v.v.kitov@yandex.ru

Table of Contents

- 1 Linear algebra reminder
- 2 Dimensionality reduction intro
- 3 Principal component analysis
- 4 Construction of principal components
- 5 Proof of optimality of principal components

Scalar product reminder

- Here we will assume $\langle a, b \rangle = a^T b$
- $\|a\| = \sqrt{\langle a, a \rangle}$
- Signed projection of x on a is equal to $\langle x, a \rangle / \|a\|$
- Unsigned projection (length) of x onto a is equal to $|\langle x, a \rangle| / \|a\|$

Eigenvectors, eigenvalues

- If for some $A \in \mathbb{R}^{D \times D}$ there exist scalar λ and D -dimensional vector v such that $Av = \lambda v$ then
 - v is called eigenvector of A
 - λ is called eigenvalue of A , corresponding to eigenvector v .
- $\exists v \neq 0 : Av = \lambda v \Leftrightarrow (A - \lambda I)v = 0 \Leftrightarrow \det(A - \lambda I) = 0$. So all eigenvalues satisfy $\det(A - \lambda I) = 0$ which
 - is a polynomial equation of order D
 - so has D solutions¹ (accounting for their multiplicity, possibly complex)

¹According to Fundamental theorem of algebra.

Symmetric matrices

- Matrix $A \in \mathbb{R}^{D \times D}$ is called *symmetric* if $A^T = A$.
- Properties:
 - All eigenvalues of symmetric matrix are real.
 - Eigenvectors, corresponding to different eigenvalues of symmetric matrix B are orthogonal to each other.
 - If $\tilde{\lambda}$ is a repeated root of $\det(A - \lambda I) = 0$ for some symmetric $A \in \mathbb{R}^{D \times D}$ with multiplicity m then there exist m orthonormal eigenvectors of A , corresponding to $\tilde{\lambda}$.
 - For any symmetric matrix $A \in \mathbb{R}^{D \times D}$ there exists orthonormal basis of eigenvectors of this matrix.

Vector derivatives

- Suppose $x = [x^1, \dots, x^D]$ and $f(x) = f(x^1, \dots, x^D)$. Vector derivative

$$\frac{\partial f(x)}{\partial x} := \begin{pmatrix} \frac{\partial f(x)}{\partial x^1} \\ \frac{\partial f(x)}{\partial x^2} \\ \dots \\ \frac{\partial f(x)}{\partial x^D} \end{pmatrix}$$

- For any $x, b \in \mathbb{R}^D$ it holds that²:

$$\frac{\partial [b^T x]}{\partial x} = b$$

- For any $x \in \mathbb{R}^D$ and symmetric $B \in \mathbb{R}^{D \times D}$ it holds that³:

$$\frac{\partial [x^T B x]}{\partial x} = 2Bx$$

²Prove it.

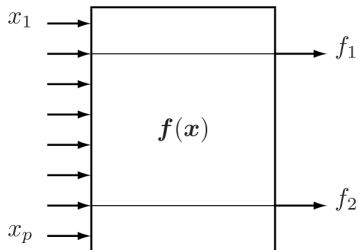
³Prove it. How will the formula change for non-symmetric B ?

Table of Contents

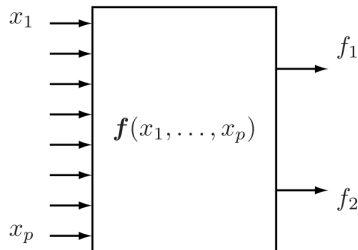
- 1 Linear algebra reminder
- 2 Dimensionality reduction intro
- 3 Principal component analysis
- 4 Construction of principal components
- 5 Proof of optimality of principal components

Dimensionality reduction

Feature selection / Feature extraction



(a) feature selector



(b) feature extractor

Feature extraction: find transformation of original data which extracts most relevant information for machine learning task.

Applications of dimensionality reduction

Applications:

- visualization in 2D or 3D
- reduce operational costs on data storage, transfer and processing
 - memory
 - disk
 - CPU usage
- remove multi-collinearity to improve performance of some machine-learning models

Categorization of dimensionality reduction methods

Supervision:

- supervised
- unsupervised

Mapping to reduced space:

- linear
- non-linear

Principal components analysis - linear unsupervised method of dimensionality reduction.

Table of Contents

- 1 Linear algebra reminder
- 2 Dimensionality reduction intro
- 3 **Principal component analysis**
 - Definition
 - Applications of PCA
 - Application details
- 4 Construction of principal components
- 5 Proof of optimality of principal components

3 Principal component analysis

- Definition
- Applications of PCA
- Application details

Projections, orthogonal complements

- For point x and subspace L denote:
 - p : the projection of x on L
 - h : orthogonal complement
 - $x = p + h$, $\langle p, h \rangle = 0$.
- For training set x_1, x_2, \dots, x_N and subspace L find:
 - projections: p_1, p_2, \dots, p_N
 - orthogonal complements: h_1, h_2, \dots, h_N .

Best subspace fit⁴

Definition 1

Best-fit k -dimensional subspace for a set of points x_1, x_2, \dots, x_N is a subspace, spanned by k vectors v_1, v_2, \dots, v_k , solving

$$\sum_{n=1}^N \|h_n\|^2 \rightarrow \min_{v_1, v_2, \dots, v_k}$$

Proposition 1

Vectors v_1, v_2, \dots, v_k , solving

$$\sum_{n=1}^N \|p_n\|^2 \rightarrow \max_{v_1, v_2, \dots, v_k}$$

also define best-fit k -dimensional subspace.

⁴Prove 1 using that $\|x\|^2 = \|p\|^2 + \|h\|^2$ for $x = p + h$ and $\langle p, h \rangle = 0$.

Definition of PCA

Definition 2

Principal components a_1, a_2, \dots, a_k are vectors, forming orthonormal basis in the k -dimensional subspace of best fit.

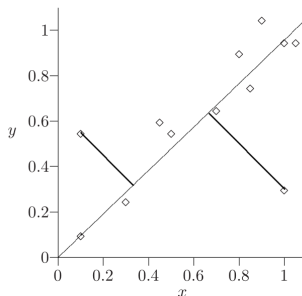
- Properties:
 - Not invariant to translation:
 - center data before PCA:

$$x \leftarrow x - \mu \text{ where } \mu = \frac{1}{N} \sum_{n=1}^N x_n$$

- Not invariant to scaling:
 - scale features to have unit variance before PCA

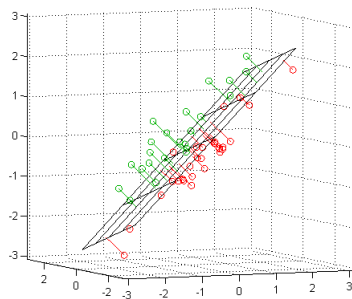
Example: line of best fit

- In PCA the sum of squared perpendicular distances to line is minimized:



- *What is the difference with least squares minimization in regression?*

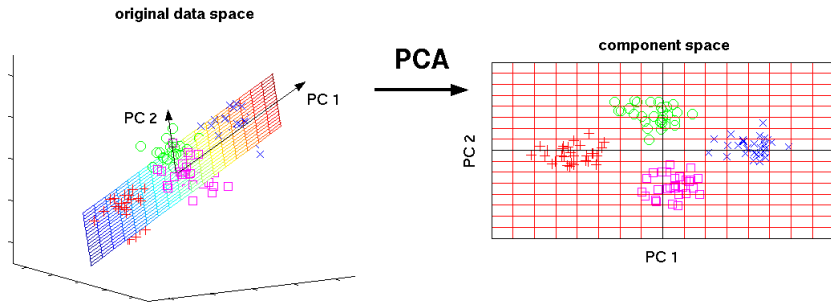
Example: plane of best fit



3 Principal component analysis

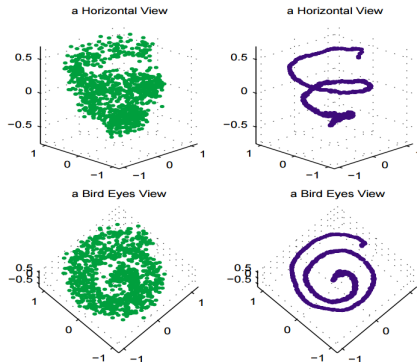
- Definition
- Applications of PCA
- Application details

Visualization



Data filtering

Remove noise to get a cleaner picture of data distribution:



X. Huo and Jihong Chen (2002). Local linear projection (LLP). First IEEE Workshop on Genomic Signal Processing and Statistics (GENSIPS), Raleigh, NC, October.
<http://www.gensips.gatech.edu/proceedings/>.

Economic description of data

Faces database:



Eigenvectors (called eigenfaces)

Projections on first several eigenvectors describe most of face variability.



Text analysis

- Objects=text files
- Binary, TF, TF-IDF representations have huge D .
 - math operations with X - inefficient
 - ML methods work longer
- Sparsity induces complications with query matching
 - consider query “automobile”
 - simple cosine-metric matching won’t match documents with “car”, “bus”, etc.

Latent semantic analysis (LSA)

Latent semantic analysis (LSA)

Get economical document representations with coordinates of most important PCA components found without centering.

Comments:

- usually 200-300 components are sufficient.
- Do *not* center X before computing PCA
 - otherwise will lose sparsity of X
 - $\mu \approx 0$ anyway, because most features are 0.
- Technically done with truncated SVD of X .

3 Principal component analysis

- Definition
- Applications of PCA
- Application details

Quality of approximation

Consider vector x . Since all D principal components form a full orthonormal basis, x can be written as

$$x = \langle x, a_1 \rangle a_1 + \langle x, a_2 \rangle a_2 + \dots + \langle x, a_D \rangle a_D$$

Let p^K be the projection of x onto subspace spanned by first K principal components:

$$p^K = \langle x, a_1 \rangle a_1 + \langle x, a_2 \rangle a_2 + \dots + \langle x, a_K \rangle a_K$$

Error of this approximation is

$$h^K = x - p^K = \langle x, a_{K+1} \rangle a_{K+1} + \dots + \langle x, a_D \rangle a_D$$

Quality of approximation

Using that a_1, \dots, a_D is an orthonormal set of vectors, we get

$$\begin{aligned}\|x\|^2 &= \langle x, x \rangle = \langle x, a_1 \rangle^2 + \dots + \langle x, a_D \rangle^2 \\ \|p^K\|^2 &= \langle p^K, p^K \rangle = \langle x, a_1 \rangle^2 + \dots + \langle x, a_K \rangle^2 \\ \|h^K\|^2 &= \langle h^K, h^K \rangle = \langle x, a_{K+1} \rangle^2 + \dots + \langle x, a_D \rangle^2\end{aligned}$$

We can measure how well first K components describe our dataset x_1, x_2, \dots, x_N using relative loss

$$L(K) = \frac{\sum_{n=1}^N \|h_n^K\|^2}{\sum_{n=1}^N \|x_n\|^2} \quad (1)$$

or relative score

$$S(K) = \frac{\sum_{n=1}^N \|p_n^K\|^2}{\sum_{n=1}^N \|x_n\|^2} \quad (2)$$

Evidently $L(K) + S(K) = 1$.

Contribution of individual component

Contribution of a_k for explaining x is $\langle x, a_k \rangle^2$.

Contribution of a_k for explaining x_1, x_2, \dots, x_N is:

$$\sum_{n=1}^N \langle x_n, a_k \rangle^2$$

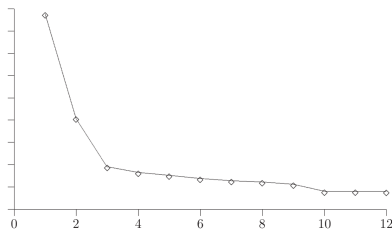
Explained variance ratio:

$$E(a_k) = \frac{\sum_{n=1}^N \langle x_n, a_k \rangle^2}{\sum_{d=1}^D \sum_{n=1}^N \langle x_n, a_d \rangle^2} = \frac{\sum_{n=1}^N \langle x_n, a_k \rangle^2}{\sum_{n=1}^N \|x_n\|^2}$$

- Explained variance ratio measures relative contribution of component a_k to explaining our dataset x_1, \dots, x_N .
- Note that $\sum_{k=1}^K E(a_k) = S(K)$.

How many principal components to select?

- Data visualization: 2 or 3 components.
- Take most significant components until their explained variance ratio falls sharply down:



- Or take minimum K such that $L(K) \leq t$ or $S(K) \geq 1 - t$, where typically $t = 0.95$.

PCA solution

- Center x_1, \dots, x_N to have zero mean.
- Scale x_1, \dots, x_N to have equal variance.
- Form $X = [x_1^T; \dots, x_N^T]^T \in \mathbb{R}^{N \times D}$
- Estimate sample covariance matrix of x : $\hat{\Sigma} = \frac{1}{N} X^T X$
- Find eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_D \geq 0$ and corresponding eigenvectors a_1, a_2, \dots, a_D .
- a_1, a_2, \dots, a_k are first k principal components, $k = 1, 2, \dots, D$.
- Sum of squared projections onto a_i is $\|X a_i\|^2 = \lambda_i$.
- *Explained variance ratio* by component a_i is equal to

$$\frac{\lambda_i}{\sum_{d=1}^D \lambda_d}$$

Table of Contents

- 1 Linear algebra reminder
- 2 Dimensionality reduction intro
- 3 Principal component analysis
- 4 Construction of principal components**
- 5 Proof of optimality of principal components

Constructive definition of PCA

- Principal components $a_1, a_2, \dots, a_D \in \mathbb{R}^D$ are found such that
$$\langle a_i, a_j \rangle = \begin{cases} 1, & i = j \\ 0 & i \neq j \end{cases}$$
- Xa_i is a vector of projections of all objects onto the i -th principal component.
- For any object x its projections onto principal components are equal to:

$$p = A^T x = [\langle a_1, x \rangle, \dots, \langle a_D, x \rangle]^T$$

where $A = [a_1; a_2; \dots, a_D] \in \mathbb{R}^{D \times D}$.

Constructive definition of PCA

- ① a_1 is selected to maximize $\|Xa_1\|$ subject to $\langle a_1, a_1 \rangle = 1$
- ② a_2 is selected to maximize $\|Xa_2\|$ subject to $\langle a_2, a_2 \rangle = 1$,
 $\langle a_2, a_1 \rangle = 0$
- ③ a_3 is selected to maximize $\|Xa_3\|$ subject to $\langle a_3, a_3 \rangle = 1$,
 $\langle a_3, a_1 \rangle = \langle a_3, a_2 \rangle = 0$
etc.
- It turns out that:
 - a_1, \dots, a_k form k -dimensional subspace of best fit.
 - a_1, a_2, \dots are first, second, ... eigenvectors of $X^T X$ (ordered by decreasing eigenvalue).

Derivation: 1st component

$$\begin{cases} \|Xa_1\|^2 \rightarrow \max_{a_k} \\ \|a_1\| = 1 \end{cases} \quad (3)$$

Lagrangian of optimization problem (3):

$$L(a_1, \mu) = a_1^T X^T X a_1 - \mu(a_1^T a_1 - 1) \rightarrow \text{extr}_{a_1, \mu}$$

$$\frac{\partial L}{\partial a_1} = 2X^T X a_1 - 2\mu a_1 = 0$$

so a_1 is selected from a set of eigenvectors of $X^T X$.

Derivation: 1st component

Since

$$\|Xa_1\|^2 = (Xa_1)^T Xa_1 = a_1^T X^T Xa_1 = \lambda a_1^T a_1 = \lambda$$

a_1 should be the eigenvector, corresponding to the largest eigenvalue λ_1 .

Comment: If many many eigenvector directions corresponding to λ_1 exist, select arbitrary eigenvector, satisfying constraint of (3).

Derivation: 2nd component

$$\begin{cases} \|Xa_2\|^2 \rightarrow \max_{a_2} \\ \|a_2\| = 1 \\ a_2^T a_1 = 0 \end{cases} \quad (4)$$

Lagrangian of optimization problem (4):

$$L(a_2, \mu) = a_2^T X^T X a_2 - \mu(a_2^T a_2 - 1) - \alpha a_1^T a_2 \rightarrow \text{extr}_{a_2, \mu, \alpha}$$

$$\frac{\partial L}{\partial a_2} = 2X^T X a_2 - 2\mu a_2 - \alpha a_1 = 0 \quad (5)$$

Derivation: 2nd component

By multiplying by a_1^T we obtain:

$$a_1^T \frac{\partial L}{\partial a_1} = 2a_1^T X^T X a_2 - 2\mu a_1^T a_2 - \alpha a_1^T a_1 = 0 \quad (6)$$

Since a_2 is selected to be orthogonal to a_1 :

$$2\mu a_1^T a_2 = 0$$

Since $a_1^T X^T X a_2$ is scalar and a_1 is eigenvector of $X^T X$:

$$a_1^T X^T X a_2 = \left(a_1^T X^T X a_2 \right)^T = a_2^T X^T X a_1 = \lambda_1 a_2^T a_1 = 0$$

It follows that (6) simplifies to $\alpha a_1^T a_1 = \alpha = 0$ and (5) becomes

$$X^T X a_2 - \mu a_2 = 0$$

So a_2 is selected from a set of eigenvectors of $X^T X$.

Derivation: 2nd component

Since

$$\|Xa_2\|^2 = (Xa_2)^T Xa_2 = a_2^T X^T Xa_2 = \lambda a_2^T a_2 = \lambda$$

a_2 should be the eigenvector, corresponding to second largest eigenvalue λ_2 .

Comment: If many many eigenvector directions corresponding to λ_2 exist, select arbitrary eigenvector, satisfying constraints of (4).

Derivation: k-th component

$$\begin{cases} \|Xa_k\|^2 \rightarrow \max_{a_k} \\ \|a_k\| = 1 \\ a_k^T a_1 = \dots = a_k^T a_{k-1} = 0 \end{cases} \quad (7)$$

Table of Contents

- 1 Linear algebra reminder
- 2 Dimensionality reduction intro
- 3 Principal component analysis
- 4 Construction of principal components
- 5 Proof of optimality of principal components

Componentwise optimization leads to best fit subspace

Theorem 1

Let L_k be the subspace spanned by a_1, a_2, \dots, a_k . Then for each k L_k is the best-fit k -dimensional subspace for X .

Proof: use induction. For $k = 1$ the statement is true by definition since projection maximization is equivalent to distance minimization.

Suppose theorem holds for $k - 1$. Let L_k be the plane of best-fit of dimension with $\dim L = k$. We can always choose an orthonormal basis of L_k b_1, b_2, \dots, b_k so that

$$\begin{cases} \|b_k\| = 1 \\ b_k \perp a_1, b_k \perp a_2, \dots, b_k \perp a_{k-1} \end{cases} \quad (8)$$

by setting b_k perpendicular to projections of a_1, a_2, \dots, a_{k-1} on L_k .

Componentwise optimization leads to best fit subspace

Consider the sum of squared projections:

$$\|Xb_1\|^2 + \|Xb_2\|^2 + \dots + \|Xb_{k-1}\|^2 + \|Xb_k\|^2$$

By induction proposition $L[a_1, a_2, \dots, a_{k-1}]$ is space of best fit of rank $k-1$ and $L[b_1, \dots, b_{k-1}]$ is some space of same rank, so sum of squared projections on it is smaller:

$$\|Xb_1\|^2 + \|Xb_2\|^2 + \dots + \|Xb_{k-1}\|^2 \leq \|Xa_1\|^2 + \|Xa_2\|^2 + \dots + \|Xa_{k-1}\|^2$$

and

$$\|Xb_k\|^2 \leq \|Xa_k\|^2$$

since b_k by (8) satisfies constraints of optimization problem (7) and a_k is its optimal solution.

Summary

- Every symmetric matrix A can be decomposed into rotation, scaling and backward rotation:

$$A = P\Lambda P^T$$

- Sample covariance matrix $\frac{1}{N}X^T X$ is symmetric and $\succcurlyeq 0$.
 - so it has non-negative eigenvalues $\lambda_1 \geq \dots \geq \lambda_D \geq 0$ with corresponding eigenvectors a_1, \dots, a_D .
 - spread of distribution is characterized by eigenvalues.

Summary

- Dimensionality reduction - common preprocessing step for efficiency and numerical stability.
- Subspace of best fit of rank k for training set x_1, \dots, x_N is k -dimensional subspace $\mathcal{L}(b_1, \dots, b_k)$, minimizing:

$$\|h_1\|^2 + \dots + \|h_N\|^2 \rightarrow \min_{b_1, \dots, b_k}$$

- Solution vectors are called top k principal components.
- Principal component analysis - expression of x in terms of first k principal components.
- It is unsupervised linear dimensionality reduction.
- Solution: principal components a_1, \dots, a_k are top k eigenvectors of $X^T X$.