Principal components analysis

Victor Kitov

v.v.kitov@yandex.ru

Yandex
School of Data Analysis.

Table of Contents

Scalar product reminer

- 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

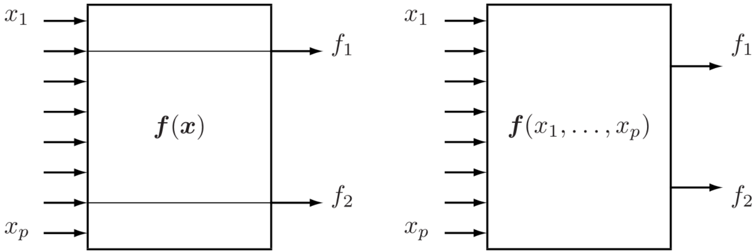
- If for some $A \in \mathbb{R}^{DxD}$ there exist scalar λ and D-dimensional vector v such that $Av = \lambda v$ then
  - v is called eigenvector of A
  - λ is called eigenvalue of A, corresponding to eigenvector v.
- $\exists v \neq 0 : \; Av = \lambda v \Leftrightarrow (A - \lambda I)\, v = 0 \Leftrightarrow det(A - \lambda I) = 0.$ So all eigenvalues satisfy $det(A - \lambda I) = 0$ which
  - is a polynomial equation of order D
  - so has D solutions[1] (accounting for their multiplicity, possibly complex)

[1] According to Fundamental theorem of algebra.

Table of Contents

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
- unsupervied

Mapping to reduced space:

- linear
- non-linear

Principal components analysis - linear unsupervised method of dimensionality reduction.

Table of Contents

3. Principal component analysis
   - Definition
   - Application details
   - Construction of principal components
   - Proof of optimality of principal components

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 $x_1, x_2, ...x_N$ and subspace L find:
  - projections: $p_1, p_2, ...p_N$
  - orthogonal complements: $h_1, h_2, ...h_N$.

Best subspace fit[2]

Definition 1

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

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

Proposition 1

Vectors $v_1, v_2, ...v_k$, solving

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

also define best-fit k-dimensional subspace.

[2] Prove 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 $a_1, a_2, ...a_k$ are vectors, forming orthonormal basis in the k-dimensional subspace of best fit.

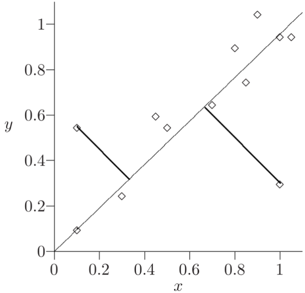
- Properties:
  - Not invariant to translation:
    - center data before PCA:

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

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

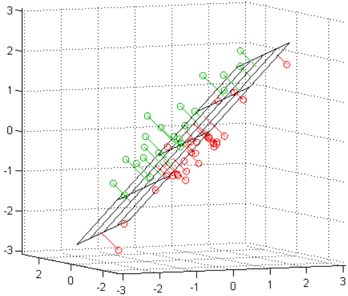
Example: line of best fit

- 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



Quality of approximation

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

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

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

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

Error of this approximation is

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

Contribution of individual component

Contribution of a_k for explaining x is $\langle x, a_k \rangle^2$.
Contribution of a_k for explaining $x_1, x_2, ...x_N$ 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 $x_1, ...x_N$.

Quality of approximation

Using that $a_1, ...a_D$ is an orthonormal set of vectors, we get

$$\|x\|^2 = \langle x, x \rangle = \langle x, a_1 \rangle^2 + ... + \langle x, a_D \rangle^2$$
$$\left\|p^K\right\|^2 = \langle p^K, p^K \rangle = \langle x, a_1 \rangle^2 + ... + \langle x, a_K \rangle^2$$
$$\left\|h^K\right\|^2 = \langle h^K, h^K \rangle = \langle x, a_{K+1} \rangle^2 + ... + \langle x, a_D \rangle^2$$

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

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

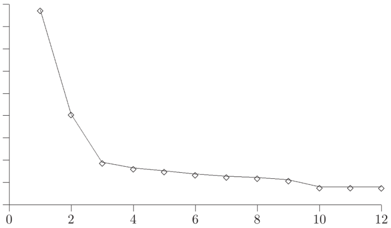
or relative score

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

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

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

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

3 Principal component analysis
- Definition
- Application details
- Construction of principal components
- Proof of optimality of principal components

Constructive definition of PCA

- Principal components $a_1, a_2, ...a_D \in \mathbb{R}^D$ 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:
$$p = A^T x = [\langle a_1, x \rangle, ... \langle a_D, x \rangle]^T$$
where $A = [a_1; a_2; ...a_D] \in \mathbb{R}^{DxD}$.

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 can be proved that:
  - $a_1, ...a_k$ form k-dimensional subspace of best fit.
  - $a_1, a_2, ...$ are first, second,... eigenvectors of $X^T X$ (ordered by decreasing eigenvalue).

Derivation: 1st component

Since

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

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

Comment: If many many eigenvector directions corrsponding to λ_1 exist, select arbitrary eigenvector, satisfying constraint of (**??**).

Derivation: 2nd component

$$\begin{cases} \|Xa_2\|^2 \to \max_{a_2} \\ \|a_2\| = 1 \\ a_2^T a_1 = 0 \end{cases} \tag{3}$$

Lagrangian of optimization problem (3):

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

$$\frac{\partial L}{\partial a_2} = 2X^T X a_2 - 2\mu a_2 - \alpha a_1 = 0 \tag{4}$$

Derivation: 2nd component

By multiplying by a_1^T we obtain:

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

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 (5) simplifies to $\alpha a_1^T a_1 = \alpha = 0$ and (4) 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 corrsponding to λ_2 exist, select arbitrary eigenvector, satisfying constraints of (3).

Componentwise optimization leads to best fit subspace

Theorem 1

Let L_k be the subspace spanned by $a_1, a_2, ...a_k$. 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 $b_1, b_2, ...b_k$ so that

$$\begin{cases} \|b_k\| = 1 \\ b_k \perp a_1, b_k \perp a_2, ...b_k \perp a_{k-1} \end{cases} \tag{6}$$

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

Componentwise optimization leads to best fit subspace

Consider the sum of squared projections:

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

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

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

and

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

since b_k by (6) satisfies constraints of optimization problem (**??**) and a_k is its optimal solution.

Summary

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

$$\|h_1\|^2 + ... + \|h_N\|^2 \to \min_{b_1,...b_k}$$

- 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 $a_1, ...a_k$ are top k eigenvectors of $X^T X$.