Principal components analysis

Victor Kitov

v.v.kitov@yandex.ru

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.

Symmetric matrices

- Matrix $A \in \mathbb{R}^{DxD}$ 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 $A \in \mathbb{R}^{DxD}$ with multiplicity m then there exist m orthonormal eigenvectors of A, corresponding to $\tilde{\lambda}$.
  - For any symmetric matrix $A \in \mathbb{R}^{DxD}$ there exists orthonormal basis of eigenvectors of this matrix.

Spectral decomposition

Theorem 1 (Spectral decomposition.)

Every symmetric $A \in \mathbb{R}^{DxD}$ *can be factorized*

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

where $P \in \mathbb{R}^{DxD}$ *is orthogonal matrix whose columns* $p_1, ...p_D$ *are eigenvectors of* A *and* $\Lambda = \text{diag}\{\lambda_1, ...\lambda_D\}$ *is diagonal matrix with corresponding eigenvalues on the diagonal.*

Intuition: transformation Ax by symmetric matrix is equivalent to

1. rotation of x to ortonormal basis formed by eigenvectors of A
2. scaling coordinates in this basis by eigenvalues $\lambda_1, ...\lambda_D$.
3. reverse rotation to initial basis.

Positivity of matrices

Definition

Symmetric matrix $A \in \mathbb{R}^{DxD}$ is called *positive semi-definite* when

$$\forall x \in \mathbb{R}^D : \langle x, Ax \rangle = x^T Ax \geq 0$$

- Positive semi-definiteness of A is denoted as $A \succcurlyeq 0$.
- Are the following matrix positive semi-definite: $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$?[2]

Theorem

Symmetric matrix A is positive semi-definite <=> all its eigenvalues are non-negative.

[2] Are these matrices $\succcurlyeq 0$ for $D > 2$ dimensional case?

Distribution properties

Let $x_1, ...x_N \in \mathbb{R}^N$ be observations of some vector random variable $x \sim F$. Group these observations into matrix $X = [x_1^T, ...x_N^T]^T \in \mathbb{R}^{NxD}$

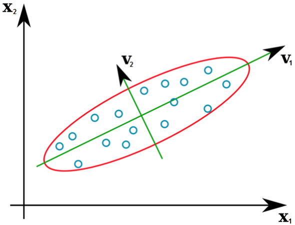
- Expectation: $\mu = \mathbb{E}x$
- Covariance matrix $\Sigma = \mathbb{E}(x - \mu)(x - \mu)^T$.
- Sample mean $\widehat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x_n$
- Sample covariance matrix $\widehat{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)(x_n - \mu)^T = \frac{1}{N}X^T X$

Properties of covariance matrix

- If random vector $x \in \mathbb{R}^D$ has covariance Σ, then random variable $\alpha^T x$ for any $\alpha \in \mathbb{R}^D$ has variance $\alpha^T \Sigma \alpha$.
- Covaraince matrix is symmetric and positive semi-definite.
- For any matrix $X \in \mathbb{R}^{NxD}$ $X^T X \in \mathbb{R}^{DxD}$ is symmetric and positive semi-definite.
    - So all eigenvalues of $X^T X$ are non-negative
- Sample covaraince matrix is symmetric and positive semi-definite.

Estimating scatter by covariance matrix

- For different $\alpha \in \mathbb{R}^D$, $\alpha^T\alpha = 1$ estimate $\text{var}(\alpha^T x) = \alpha^T \Sigma \alpha = \alpha^T P \Lambda P^T \alpha = \left(\Lambda^{1/2} P^T \alpha\right)^T \left(\Lambda^{1/2} P^T \alpha\right)$.
- α gets rotated to new orthonormal basis P, then stretched along axes with factors $\sqrt{\lambda_1}, ...\sqrt{\lambda_D}$.



- We can evaluate scatter by looking at trace $\Sigma = \lambda_1 + ... + \lambda_D$ or det $\Sigma = \lambda_1 \cdot ... \cdot \lambda_D$.
  - This is similar to arithmetic and geometric averaging.

Vector derivatives

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

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

- For any $x,b\in\mathbb{R}^D$ it holds that[3]:

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

- For any $x\in\mathbb{R}^D$ and symmetric $B\in\mathbb{R}^{DxD}$ it holds that[4]:

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

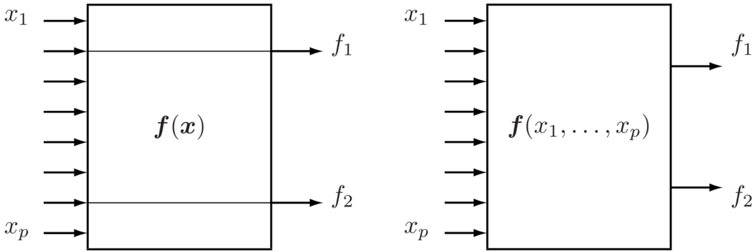
[3] Prove it.

[4] Prove it. How will the formula change for non-symmetric B?

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

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[5]

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.

[5] 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.

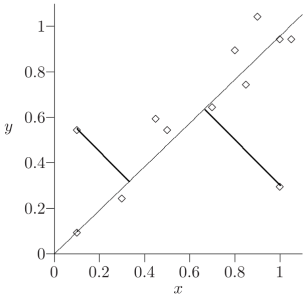
- Properties:
  - Not invariant to translation:
    - center data before PCA:

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

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

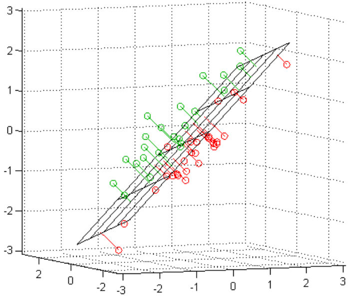
Example: line of best fit

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

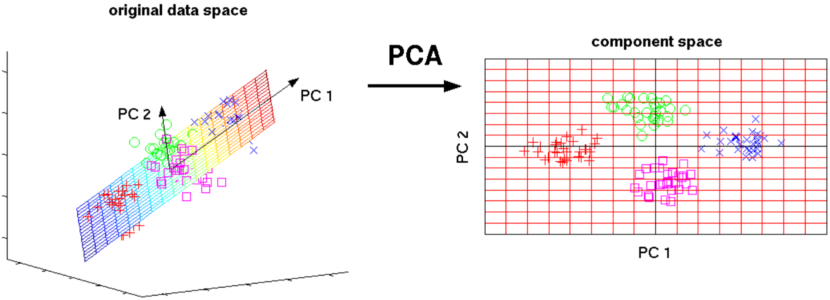


- *What is the difference with least squares minimization in regression?*

Example: plane of best fit

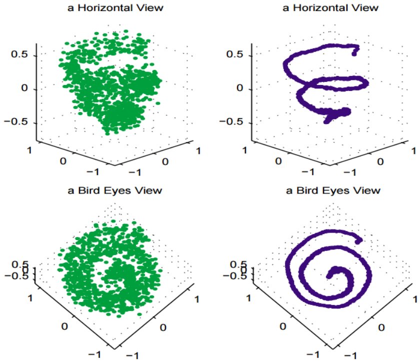


Visualization



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.
- Techically done with truncated SVD of X.

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

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

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

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:

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

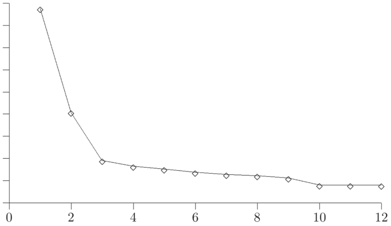
Explained variance ratio:

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

- Explained variance ratio measures relative contribution of component a_k to explaining our dataset $x_1, ...x_N$.
- 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 $t = 0.95$.

Transformation $\xi \rightleftarrows x$

Dependence between original and transformed features:

$$\xi = A^T(x-\mu),\; x = A\xi + \mu,$$

where $\mu = \frac{1}{N}\sum_{n=1}^{N} x_n$.
Taking first r components - $A_r = [a_1|a_2|...|a_r]$, we get the image of the reduced transformation:

$$\xi_r = A_r^T(x-\mu)$$

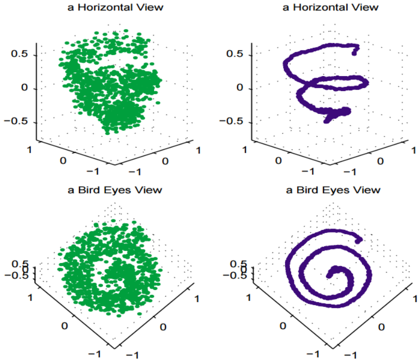
ξ_r will correspond to

$$x_r = A\begin{pmatrix}\xi_r\\ 0\end{pmatrix} + \mu = A_r\xi_r + \mu$$

$$x_r = A_rA_r^T(x-\mu) + \mu$$

$A_rA_r^T$ is projection matrix with rank r
(follows from the property $rank\left[A_rA_r^T\right] = rank\left[A_r^TA_r\right]$ for arbitrary A_r).

Local linear projection



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/.

Local linear projection

Local linear projection method makes denoised version of original data by locally projecting it onto hyperplane of small rank.

INPUT:
  p-local dimensionality of data
  K-number of nearest neighbours

for each x_i in X:
  1) find K nearest neighbours of x_i: $x_{j(i,1)}, ... x_{j(i,K)}$
  2) find linear hyperplane L_p of dimensionality p,
    describing $x_{j(i,1)}, ... x_{j(i,K)}$ # hyperplane-subspace with offset
  3) let $\hat{x}_i$ be the projection of x_i onto this hyperplane

OUTPUT:
  denoised version of objects $\hat{x}_1, \hat{x}_2, ... \hat{x}_K$.

- Projection is made on hyperplane, not subspace!

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
    - Applications of PCA
    - 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 turns out 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).

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

Componentwise optimization leads to best fit subspace

Theorem 2

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

$$\begin{cases} \|b_k\| = 1 \\ b_k \perp a_1, b_k \perp a_2, ...b_k \perp a_{k-1} \end{cases} \tag{3}$$

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:

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

and

$$\|Xb_k\|^2 \le \|Xa_k\|^2$$

since b_k by (3) satisfies constraints of optimization problem (**??**) 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 $\frac{1}{N}X^TX$ is symmetric and $\succcurlyeq 0$.
  - so it has non-negative eigenvalues $\lambda_1 \geq ... \geq \lambda_D \geq 0$ with corresponding eigenvectors $a_1, ...a_D$.
  - 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 $x_1, ...x_N$ 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 $X^T X$.