Principal component analysis

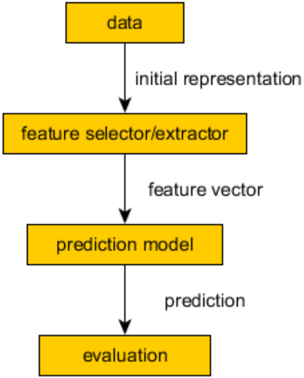
Victor Kitov

v.v.kitov@yandex.ru

Yandex School of Data Analysis

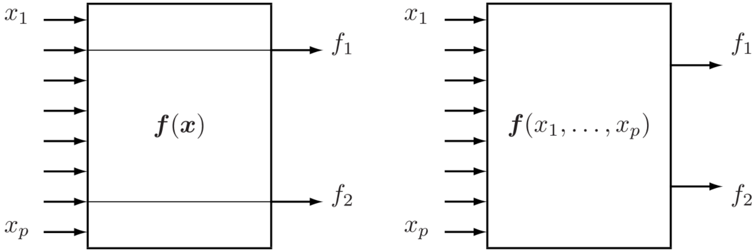
Table of Contents

General modelling pipeline



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.

We will consider unsupervised dimensionality reduction methods, which try to preserve geometrical properties of the data.

Applications of dimensionality reduction

Applications:

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

Categorization

Supervision in dimensionality reduction:

- supervised (such as Fisher's direction)
- unsupervied

Mapping to reduced space:

- linear
- non-linear

Table of Contents

2. Principal component analysis
   - Definition
   - Applications of PCA
   - Application details

Best hyperplane fit

- For point x and subspace L denote:
  - p-the projection of x on L
  - h-orthogonal complement
- $x = p + h$, $\langle p, h \rangle = 0$.

Proposition 1

For x, its projection p and orthogonal complement h

$$\|x\|^2 = \|p\|^2 + \|h\|^2 .$$

- Prove proposition 1.
- For training set $x_1, x_2, ... x_N$ and subspace L we can also find:
  - projections: $p_1, p_2, ... p_N$
  - orthogonal complements: $h_1, h_2, ... h_N$.

Best subspace fit

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 2

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

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

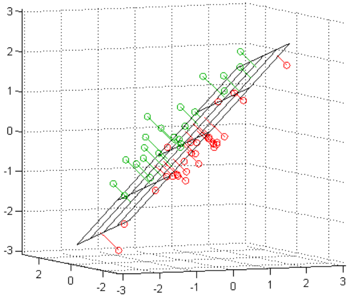
also define best-fit k-dimensional subspace.

- Prove 2 using proposition 1.

Definition 2

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

Best hyperplane fit



Subspace L_k or rank k best fits points $x_1, x_2, ...x_D$.

Properties of PCA

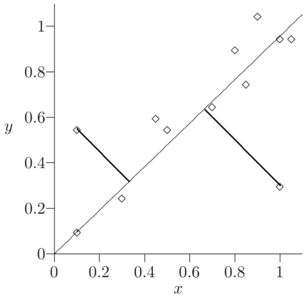
- Properties:
  - Not invariant to translation:
    - Before applying PCA, it is recommended to center objects:

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

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

Example: line of best fit

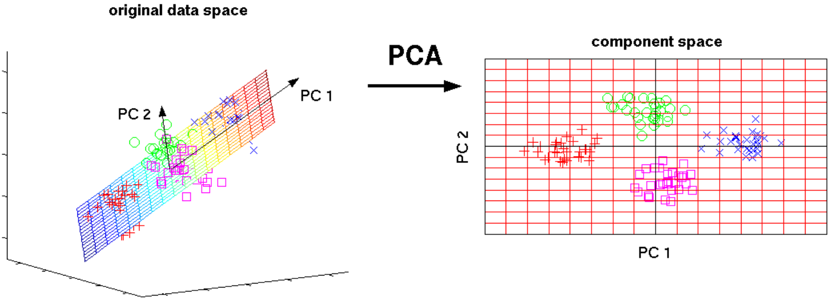
- In PCA the sum of squared perpendicular distances to line is minimized:



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

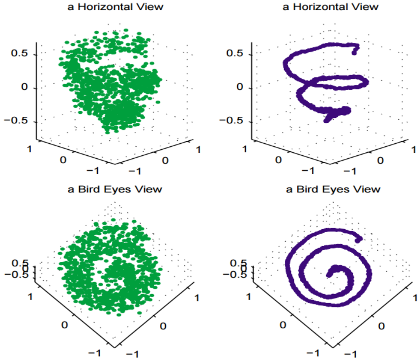
2. Principal component analysis
   - Definition
   - Applications of PCA
   - Application details

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/.

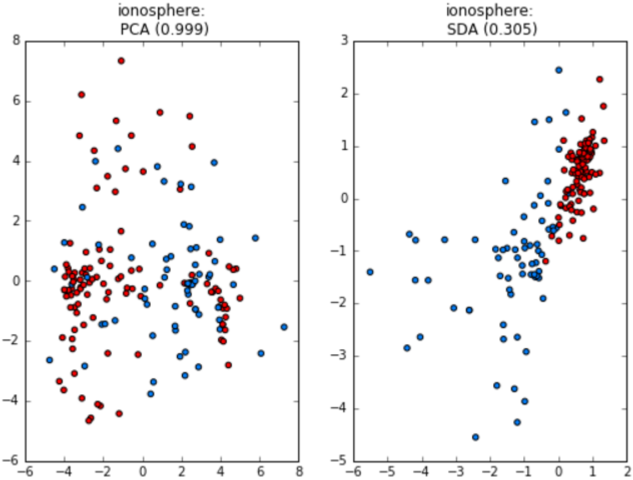
Economic description of data

Faces database:

Eigenfaces

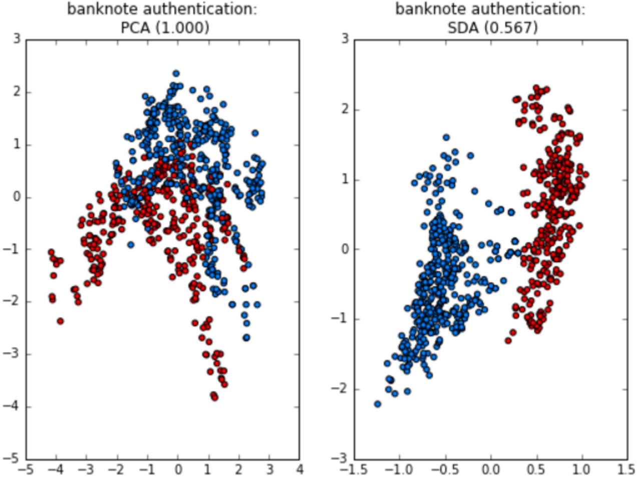
Eigenvectors are called eigenfaces. Projections on first several eigenfaces describe most of face variability.

PCA vs. SDA (not covered here)



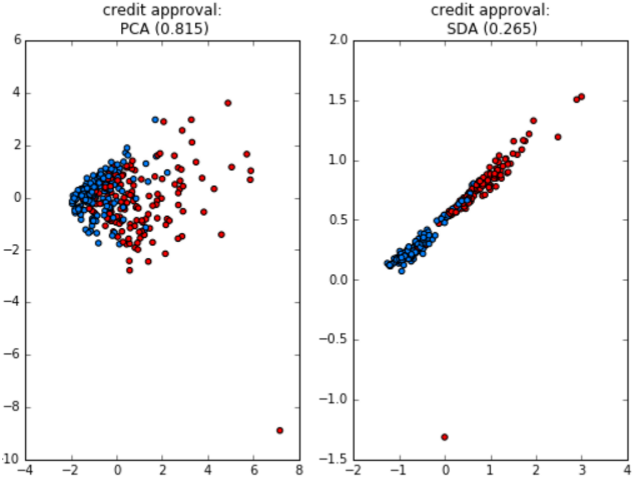
Title format: dataset, method (quality of approximation (2)).

PCA vs. SDA (not covered here)



Title format: dataset, method (quality of approximation (2)).

PCA vs. SDA (not covered here)



Title format: dataset, method (quality of approximation (2)).

2. Principal component analysis
   - Definition
   - Applications of PCA
   - Application details

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:

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

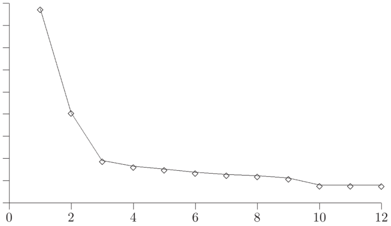
Explained variance ratio:

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

- 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 variance falls sharply down:



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

Conclusion[1]

- For $x \in \mathbb{R}^D$ there exist D principal components.
- Principal component a_i is the i-th eigenvector of $X^T X$, corresponding to i-th largest eigenvalue λ_i.
- Sum of squared projections onto a_i is $\|Xa_i\|^2 = \lambda_i$.
- *Explained variance ratio* by component a_i is equal to

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

[1] Compare dimensionality reduction with PCA and regularization as means of simplification of prediction model.