

# Dimensionality reduction

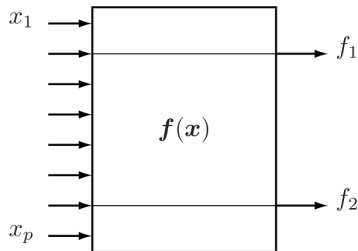
Victor Kitov

# Table of Contents

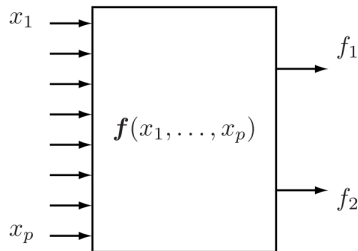
- 1 Dimensionality reduction intro
- 2 Principal component analysis

# 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)
- unsupervised

Mapping to reduced space:

- linear
- non-linear

# Table of Contents

## 1 Dimensionality reduction intro

## 2 Principal component analysis

- Reminder
- Definition
- Applications of PCA
- Application details
- Construction of principal components
- Proof of optimality of principal components

## 2 Principal component analysis

- Reminder
- Definition
- Applications of PCA
- Application details
- Construction of principal components
- 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  $|\langle x, a \rangle| / \|a\|$



## Useful properties

- For any matrix  $X \in \mathbb{R}^{N \times D}$   $X^T X \in \mathbb{R}^{D \times D}$  is symmetric and positive semi-definite:
  - $(X^T X)^T = X^T X$  - symmetric matrix
  - $\forall a \in \mathbb{R}^D : \langle a, X^T X a \rangle = a^T X^T X a = \|Xa\|^2 \geq 0$
- General properties:
  - if all eigenvalues are unique, eigenvectors are also unique (up to scalar multipliers).
  - if  $A \succeq 0$  then all its eigenvalues are non-negative
- Since  $X^T X \succeq 0$  it follows that all its eigenvalues are non-negative.
- We will assume that eigenvalues of  $X^T X$  are  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_D \geq 0$ .

# Useful properties

For any  $x, b \in \mathbb{R}^D$  it holds that<sup>1</sup>:

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

For any  $x \in \mathbb{R}^D$  and symmetric  $B \in \mathbb{R}^{D \times D}$  it holds that<sup>2</sup>:

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

---

<sup>1</sup>prove it

<sup>2</sup>prove it

## 2 Principal component analysis

- Reminder
- **Definition**
- Applications of PCA
- Application details
- Construction of principal components
- Proof of optimality of principal components

# 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  $x_1, x_2, \dots, x_N$  and subspace  $L$  we can also find:
  - projections:  $p_1, p_2, \dots, p_N$
  - orthogonal complements:  $h_1, h_2, \dots, h_N$ .

# Best subspace fit

## 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 2

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.

- Prove 2 using proposition 1.

# 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.

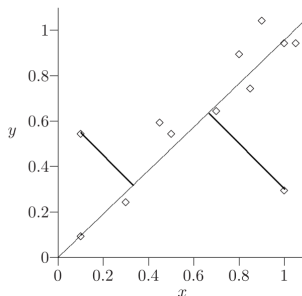
- Properties:
  - Not invariant to translation:
    - Before applying PCA, it is recommended to center objects:

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

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

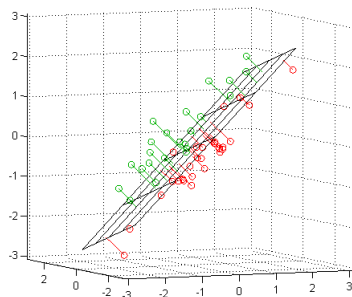
## Example: line of best fit

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



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

# Best hyperplane fit



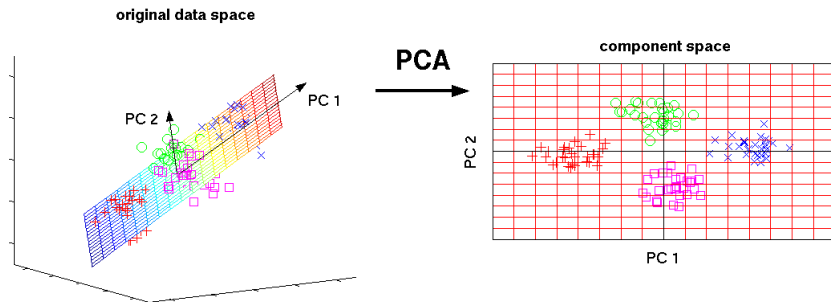
Subspace  $L_k$  or rank  $k$  best fits points  $x_1, x_2, \dots, x_D$ .



## 2 Principal component analysis

- Reminder
- Definition
- Applications of PCA
- Application details
- Construction of principal components
- Proof of optimality of principal components

# Visualization



## Economic description of data

Faces database:



# Eigenfaces

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



## 2 Principal component analysis

- Reminder
- Definition
- Applications of PCA
- **Application details**
- Construction of principal components
- Proof of optimality of principal components

## 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:

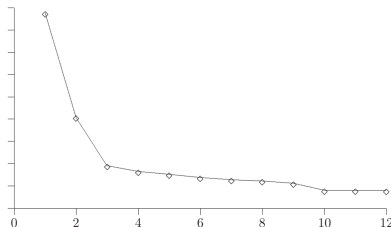
$$E(a_k) = \frac{\sum_{n=1}^N \langle x_n, a_k \rangle^2}{\sum_{d=1}^D \sum_{n=1}^N \langle x_n, a_d \rangle^2} = \frac{\sum_{n=1}^N \langle x_n, a_k \rangle^2}{\sum_{n=1}^N \|x_n\|^2}$$

- 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 variance falls sharply down:



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

## 2 Principal component analysis

- Reminder
- Definition
- Applications of PCA
- Application details
- Construction of principal components
- Proof of optimality of principal components

# Constructive definition of PCA

- Principal components  $a_1, a_2, \dots, a_D \in \mathbb{R}^D$  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.

## 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  $\lambda_1$ .

Comment: If many many eigenvector directions corresponding to  $\lambda_1$  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  $\lambda_2$ .

Comment: If many many eigenvector directions corresponding to  $\lambda_2$  exist, select arbitrary eigenvector, satisfying constraints of (4).

## Derivation: k-th component

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

Lagrangian of optimization problem (7):

$$L(a_k, \mu) = a_k^T X^T X a_k - \mu(a_k^T a_k - 1) - \sum_{j=1}^{k-1} \alpha_j a_k^T a_j \rightarrow \text{extr}_{a_k, \mu, \alpha_1, \dots, \alpha_{k-1}}$$

$$\frac{\partial L}{\partial a_k} = 2X^T X a_k - 2\mu a_k - \sum_{j=1}^{k-1} \alpha_j a_j = 0 \quad (8)$$

## Derivation: k-th component

By multiplying by  $a_i^T$  for any  $i = 1, 2, \dots, k-1$  we obtain:

$$a_i^T \frac{\partial L}{\partial a_1} = 2a_i^T X^T X a_k - 2\mu a_i^T a_k - \alpha_1 a_i^T a_1 - \dots - \alpha_{k-1} a_i^T a_{k-1} = 0 \quad (9)$$

Since  $a_i$  and  $a_j$  are selected to be orthogonal for  $i \neq j$ , we have:

$$2\mu a_i^T a_k = 0, \quad \alpha_j a_i^T a_j = 0 \quad \forall i \neq j$$

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

$$a_i^T X^T X a_2 = \left( a_i^T X^T X a_k \right)^T = a_k^T X^T X a_i = \lambda_i a_k^T a_i = 0$$

It follows that (9) simplifies to  $\alpha_i a_i^T a_i = \alpha_i = 0$ . Since  $i$  was selected arbitrary from  $i = 1, 2, \dots, k-1$ ,  $\alpha_1 = \alpha_2 = \dots = \alpha_{k-1} = 0$  and (8) becomes

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

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

## Derivation: k-th component

Since

$$\|Xa_k\|^2 = (Xa_k)^T Xa_k = a_k^T X^T Xa_k = \lambda a_k^T a_k = \lambda$$

$a_k$  should be the eigenvector, corresponding to the k-th largest eigenvalue  $\lambda_k$ .

Comment: If many many eigenvector directions corresponding to  $\lambda_k$  exist, select arbitrary eigenvector, satisfying constraints of (7).

## 2 Principal component analysis

- Reminder
- Definition
- Applications of PCA
- Application details
- Construction of principal components
- **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

$$\begin{cases} \|b_k\| = 1 \\ b_k \perp a_1, b_k \perp a_2, \dots, b_k \perp a_{k-1} \end{cases} \quad (10)$$

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 (10) satisfies constraints of optimization problem (7) and  $a_k$  is its optimal solution.

## Conclusion

- 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  $\lambda_i$ .
- Sum of squared projections onto  $a_i$  is  $\|X a_i\|^2 = \lambda_i$ .
- *Explained variance ratio* by component  $a_i$  is equal to

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