

Principal Component Analysis (PCA)

Learning Representations.

Dimensionality Reduction.

Maria-Florina Balcan

04/11/2018

- Big & High-Dimensional Data.
- Useful to learn lower dimensional representations of the data.

Learning Representations

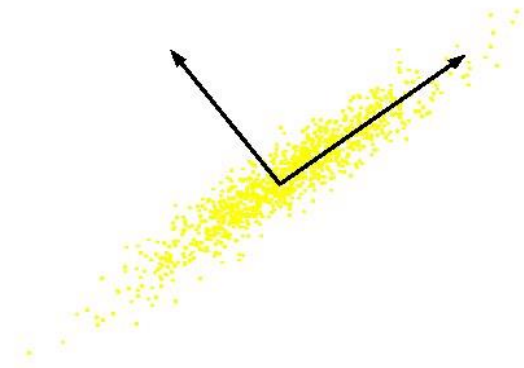
PCA, Kernel PCA, ICA: Powerful unsupervised learning techniques for extracting hidden (potentially lower dimensional) structure from high dimensional datasets.

Useful for:

- Visualization
- More efficient use of resources (e.g., time, memory, communication)
- Statistical: fewer dimensions → better generalization
- Noise removal (improving data quality)
- Further processing by machine learning algorithms

Principal Component Analysis (PCA)

What is PCA: Unsupervised technique for extracting variance structure from high dimensional datasets.

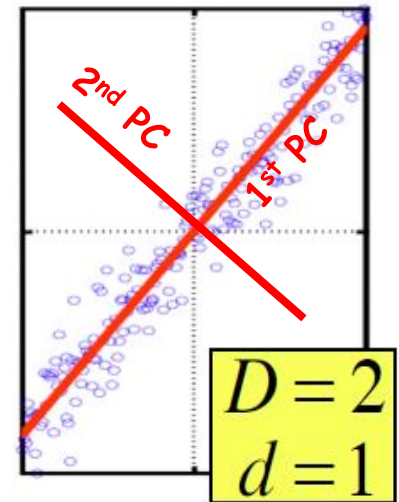
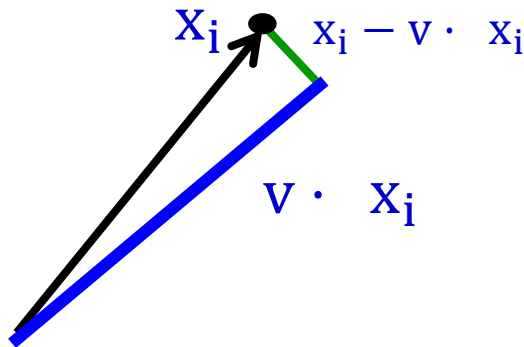


- PCA is an orthogonal projection or transformation of the data into a (possibly lower dimensional) subspace so that the variance of the projected data is maximized.

Principal Component Analysis (PCA)

Principal Components (PC) are orthogonal directions that capture most of the variance in the data.

- 1st PC - direction of greatest variability in data.



- 2nd PC - Next orthogonal (uncorrelated) direction of greatest variability

(remove all variability in first direction, then find next direction of greatest variability)

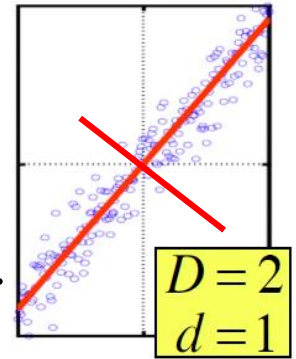
- And so on ...

Principal Component Analysis (PCA)

Let $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_d$ denote the d principal components.

$$\mathbf{v}_i \cdot \mathbf{v}_j = 0, i \neq j \quad \text{and} \quad \mathbf{v}_i \cdot \mathbf{v}_i = 1, i = j$$

Assume data is centered (we extracted the sample mean).



Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]$ (columns are the datapoints)

Find vector that maximizes sample variance of projected data

$$\frac{1}{n} \sum_{i=1}^n (\mathbf{v}^T \mathbf{x}_i)^2 = \mathbf{v}^T \mathbf{X} \mathbf{X}^T \mathbf{v}$$

$$\max_{\mathbf{v}} \mathbf{v}^T \mathbf{X} \mathbf{X}^T \mathbf{v} \quad \text{s.t.} \quad \mathbf{v}^T \mathbf{v} = 1$$

$$\text{Lagrangian: } \max_{\mathbf{v}} \mathbf{v}^T \mathbf{X} \mathbf{X}^T \mathbf{v} - \lambda \mathbf{v}^T \mathbf{v}$$

Wrap constraints into the objective function

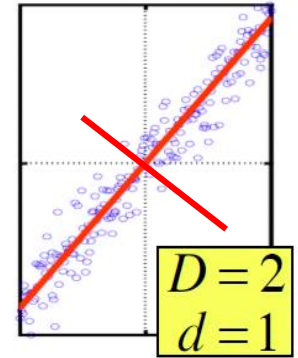
$$\partial / \partial \mathbf{v} = 0 \quad (\mathbf{X} \mathbf{X}^T - \lambda \mathbf{I}) \mathbf{v} = 0 \quad \Rightarrow \quad \boxed{(\mathbf{X} \mathbf{X}^T) \mathbf{v} = \lambda \mathbf{v}}$$

Principal Component Analysis (PCA)

$(X X^T)\mathbf{v} = \lambda\mathbf{v}$, so $\mathbf{v}$ (the first PC) is the eigenvector of sample correlation/covariance matrix $X X^T$

Sample variance of projection $\mathbf{v}^T X X^T \mathbf{v} = \lambda \mathbf{v}^T \mathbf{v} = \lambda$

Thus, the eigenvalue λ denotes the amount of variability captured along that dimension (aka amount of energy along that dimension).

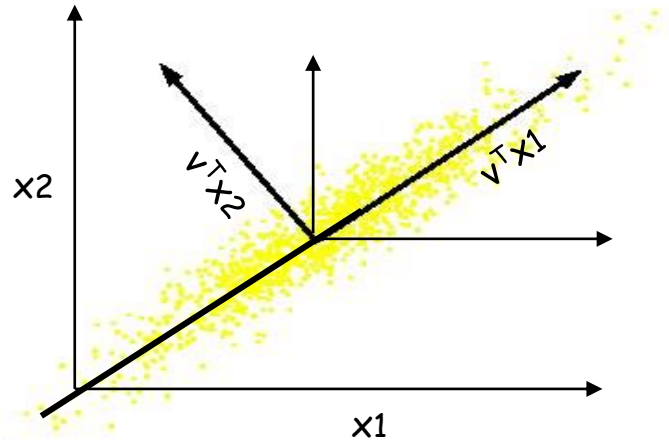


Eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \dots$

- The 1st PC $\mathbf{v}_1$ is the the eigenvector of the sample covariance matrix $X X^T$ associated with the largest eigenvalue
- The 2nd PC $\mathbf{v}_2$ is the the eigenvector of the sample covariance matrix $X X^T$ associated with the second largest eigenvalue
- And so on ...

Principal Component Analysis (PCA)

- So, the new axes are the eigenvectors of the matrix of sample correlations $X X^T$ of the data.
- Transformed features are uncorrelated.



- Geometrically: centering followed by rotation.
 - Linear transformation

Key computation: eigendecomposition of XX^T (closely related to SVD of X).

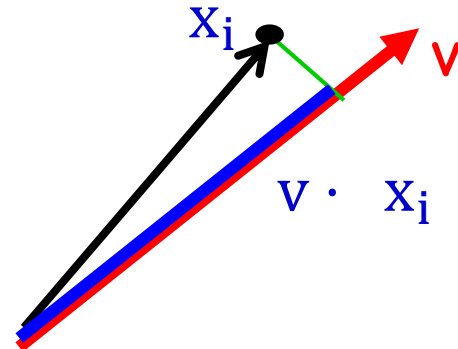
Two Interpretations

So far: **Maximum Variance Subspace**. PCA finds vectors $\mathbf{v}$ such that projections on to the vectors capture maximum variance in the data

$$\frac{1}{n} \sum_{i=1}^n (\mathbf{v}^T \mathbf{x}_i)^2 = \mathbf{v}^T \mathbf{X} \mathbf{X}^T \mathbf{v}$$

Alternative viewpoint: **Minimum Reconstruction Error**. PCA finds vectors $\mathbf{v}$ such that projection on to the vectors yields minimum MSE reconstruction

$$\frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i - (\mathbf{v}^T \mathbf{x}_i) \mathbf{v}\|^2$$



Two Interpretations

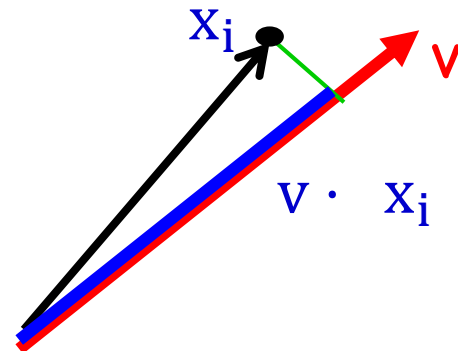
E.g., for the first component.

Maximum Variance Direction: 1st PC a vector v such that projection on to this vector capture maximum variance in the data (out of all possible one dimensional projections)

$$\frac{1}{n} \sum_{i=1}^n (v^T x_i)^2 = v^T X X^T v$$

Minimum Reconstruction Error: 1st PC a vector v such that projection on to this vector yields minimum MSE reconstruction

$$\frac{1}{n} \sum_{i=1}^n \|x_i - (v^T x_i)v\|^2$$



Why? Pythagorean Theorem

E.g., for the first component.

Maximum Variance Direction: 1st PC a vector v such that projection on to this vector capture maximum variance in the data (out of all possible one dimensional projections)

$$\frac{1}{n} \sum_{i=1}^n (v^T x_i)^2 = v^T X X^T v$$

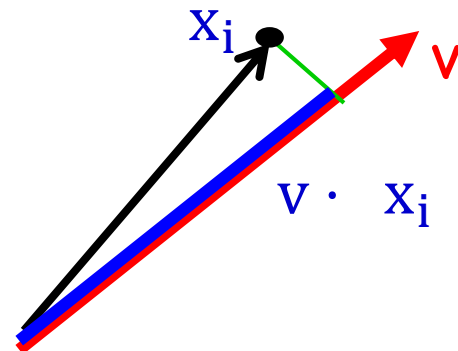
$$\frac{1}{n} \sum_{i=1}^n \|x_i - (v^T x_i)v\|^2$$

Minimum Reconstruction Error: 1st PC a vector v such that projection on to this vector yields minimum MSE reconstruction

$$\text{blue}^2 + \text{green}^2 = \text{black}^2$$

black² is fixed (it's just the data)

So, maximizing blue² is equivalent to minimizing green²

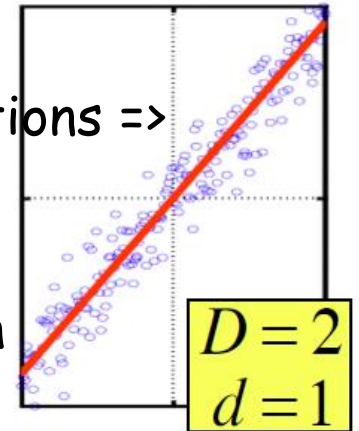


Dimensionality Reduction using PCA

The eigenvalue λ denotes the amount of variability captured along that dimension (aka amount of energy along that dimension).

Zero eigenvalues indicate no variability along those directions => data lies exactly on a linear subspace

Only keep data projections onto principal components with non-zero eigenvalues, say $v_1, \dots, v_k$, where $k = \text{rank}(X X^T)$



Original representation

Data point

$$x_i = (x_i^1, \dots, x_i^D)$$

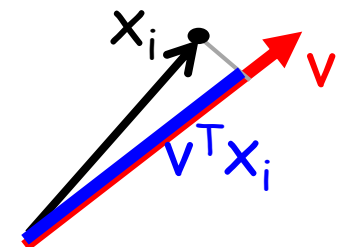
D-dimensional vector

Transformed representation

projection

$$(v_1 \cdot x^i, \dots, v_d \cdot x^i)$$

d-dimensional vector

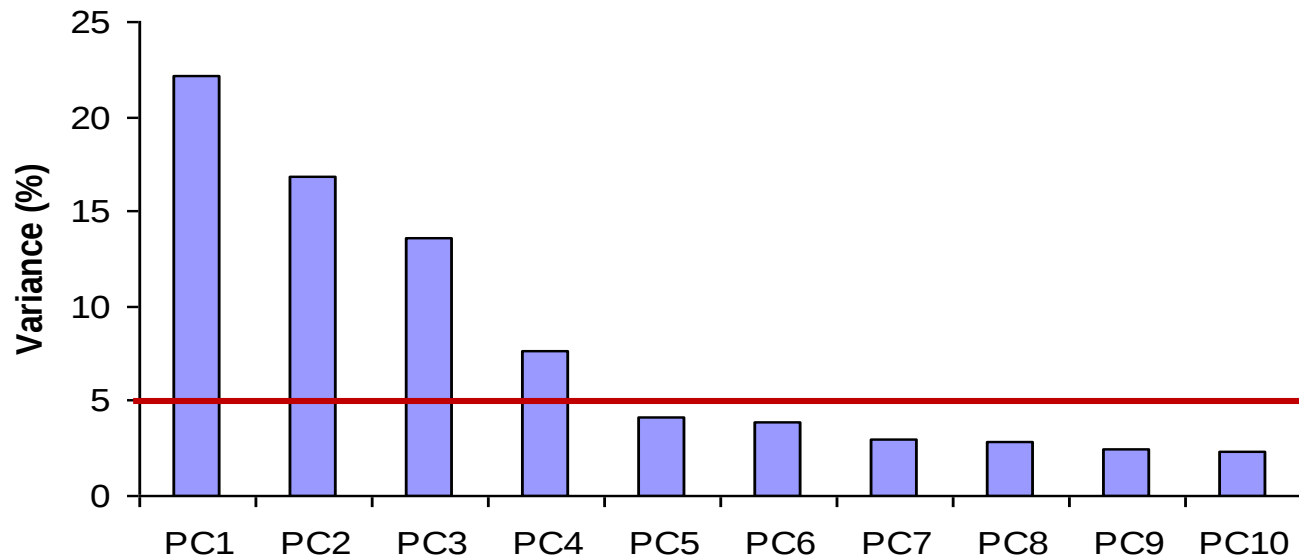


Dimensionality Reduction using PCA

In high-dimensional problems, data sometimes lies near a linear subspace, as noise introduces small variability

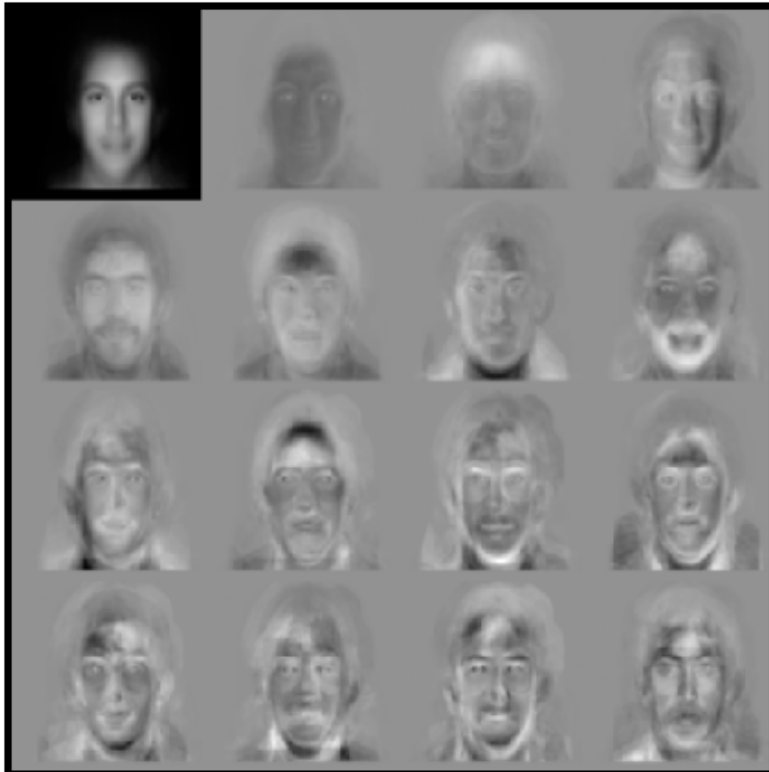
Only keep data projections onto principal components with **large** eigenvalues

Can *ignore* the components of smaller significance.



Might **lose some info**, but if eigenvalues are small, do not lose much

Example: faces



Eigenfaces
from 7562
images:

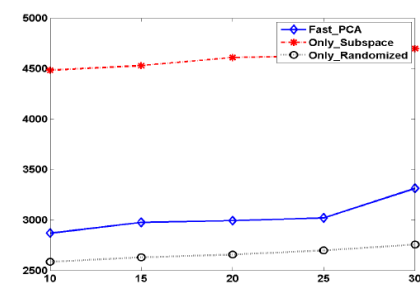
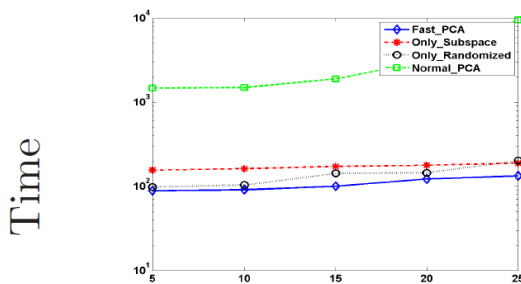
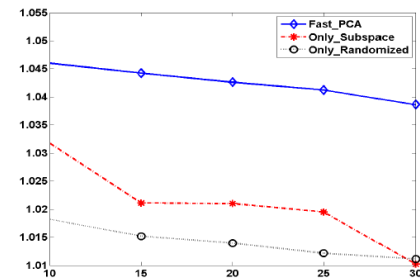
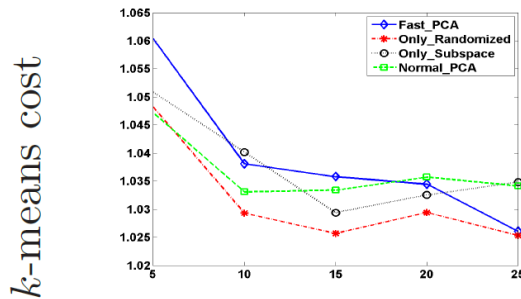
**top left image
is linear
combination
of rest.**

Sirovich & Kirby (1987)
Turk & Pentland (1991)

Can represent a face image using just 15 numbers!

- PCA provably useful before doing k-means clustering and also empirically useful. E.g.,

- ▷ **Performance:** cost increase $< 5\%$; $\times 10$ to $\times 100$ speedup
- ▷ ***k*-Means Clustering:** *k*-means cost/time vs dimension



NewsGroups

BOWpubmed

PCA Discussion

Strengths

Eigenvector method

No tuning of the parameters

No local optima

Weaknesses

Limited to second order statistics

Limited to linear projections

What You Should Know

- Principal Component Analysis (PCA)
 - What PCA is, what is useful for.
 - Both the maximum variance subspace and the minimum reconstruction error viewpoint.