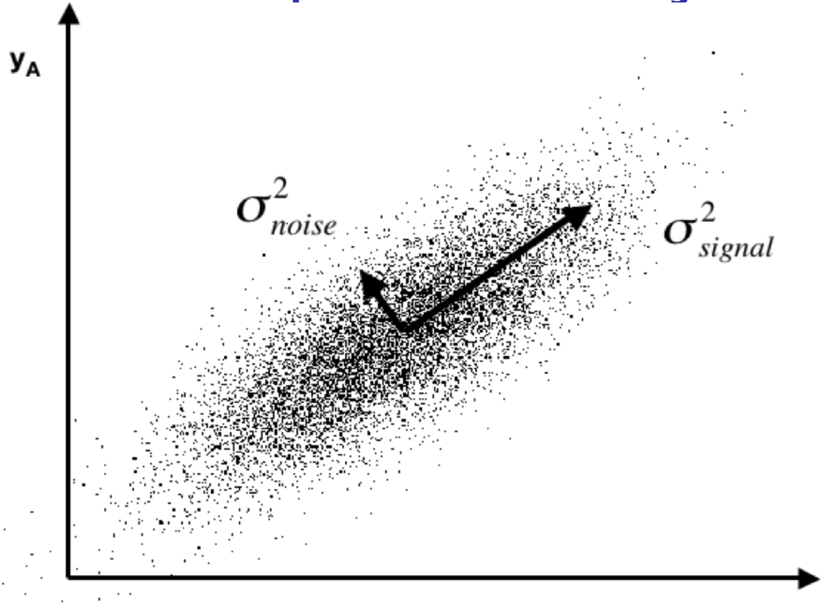
COMP229: Introduction to Data Science
Lecture 23: Principal Component Analysis

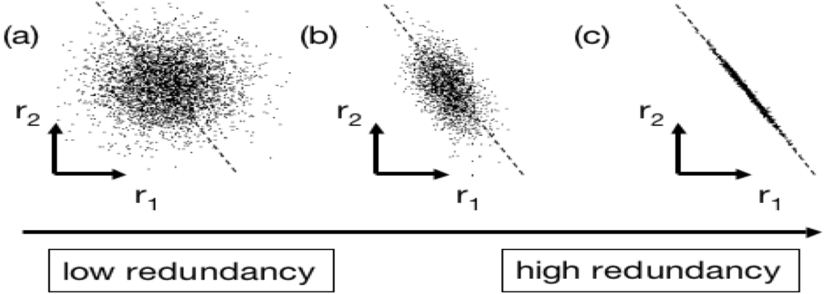
Vitaliy Kurlin, vitaliy.kurlin@liverpool.ac.uk
Autumn 2018, Computer Science department
University of Liverpool, United Kingdom

How to find patterns in noisy data



The signal-to-noise ratio

The *signal-to-noise ratio* $SNR = \frac{\sigma^2_{signal}}{\sigma^2_{noise}}$ is a relative characteristic that may help find signal.



High *SNR* means that data can be *redundant*: in the right picture one can keep only one of r_1, r_2.

A naive approach to finding a signal

The previous pictures show that the signal can be along the direction $\vec{v}_1$ with the highest variance. The next possible direction $\vec{v}_2$ (for a cloud in high dimensions) again has the highest variance, but should be orthogonal to $\vec{v}_1$ to avoid a repetition.

The orthogonality of $\vec{v}_1, \vec{v}_2, \ldots$ will guarantee that the variability along these *principal* directions are not correlated (or have the sample covariance 0).

The key idea is to find the directions that diagonalise the sample covariance matrix.

Assumptions (limitations) of the PCA

Linearity: data are near a linear subspace (a non-linear slower extension is a kernel PCA).

Sufficiency of the mean and variance, e.g. data are normally distributed near a linear subspace.

The signal-to-noise ratio is high, i.e. the signal has a large variance, noise has a low variance.

The principal directions (principal components) of the signal are orthogonal to each other.

The first steps of the PCA

Step 1. For a sample $k \times m$ matrix S, where s_{ij} is the j–th sample value of the i–th feature, subtract the sample means so that each row has mean 0.

Step 2. The sample covariance matrix $M = \dfrac{SS^T}{n-1}$.

We aim to find a new orthonormal basis (of the principal directions) that makes the covariance matrix diagonal. Let P be the $k \times k$ matrix whose rows are the principal directions that we need.

Then the new sample $k \times n$ matrix is $A = PS$ (the old basis vectors map to the principal directions).

An orthogonal transition matrix

We'll make the new covariance matrix diagonal:
$\frac{AA^T}{n-1} = \frac{(PS)(PS)^T}{n-1} = \frac{P(SS^T)P^T}{n-1} = PMP^T$, where
M is the original sample covariance matrix.

Claim 23.1. Any $k \times k$ matrix P whose rows are orthonormal to each other is an orthogonal matrix in the sense of Definition 19.6: $P^{-1} = P^T$.

Proof. The (i, j) element of PP^T is the scalar (dot) product of the i-th and j-th rows of P (equal to 1 for $i = j$, 0 otherwise), hence $PP^T = I$. $\square$

Eigenvectors of the covariance matrix

By Claim 22.6 the original covariance matrix M is symmetric positive-definite, hence diagonalisable by Claim 21.9, i.e. the transition matrix C whose columns are orthonormal eigenvectors gives the diagonal matrix $D = C^{-1}MC$ (with eigenvalues of M on the diagonal). If we put the eigenvectors (columns of C) into rows of P, then $P = C^T$.

Step 3. Find the eigenvectors of M that form a new basis such that $M = CDC^{-1}$ and the new covariance is $PMP^T = (PP^T)D(PP^T) = D$.

An explicit example for the PCA

The last identity is by $PP^T = I$ in Claim 23.1.

We'll find the principal components for the data.

Subjects	student 1	student 2	student 3	4	5
Maths	3	3	2	1	1
English	2	3	2	2	1
Art	3	1	2	3	1

Find the eigenvalues and eigenvectors of the covariance matrix $M = \begin{pmatrix} 1 & 0.5 & 0 \\ 0.5 & 0.5 & 0 \\ 0 & 0 & 1 \end{pmatrix}$.

Find eigenvalues and eigenvectors

$$\det(M-\lambda I) = \det\begin{pmatrix} 1-\lambda & 0.5 & 0 \\ 0.5 & 0.5-\lambda & 0 \\ 0 & 0 & 1-\lambda \end{pmatrix} =$$

$(1-\lambda)((1-\lambda)(0.5-\lambda)-0.25) =$

$(1-\lambda)(\lambda^2-1.5\lambda+0.25)=0.$ The eigenvalues are ordered: $\lambda_1 = \dfrac{3+\sqrt{5}}{4} > \lambda_2 = 1 > \lambda_3 = \dfrac{3-\sqrt{5}}{4}$.

The eigenvectors (the principal components):

$$\begin{pmatrix} \frac{3+\sqrt{5}}{4} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{3-\sqrt{5}}{4} \end{pmatrix}$$

$$\begin{aligned} \vec{v}_1 &= (\sqrt{5}+1, 2, 0) \text{ for } \lambda_1, \\ \vec{v}_2 &= (0,0,1) \text{ for } \lambda_2 = 1, \\ \vec{v}_3 &= (-2, \sqrt{5}+1, 0) \text{ for } \lambda_3. \end{aligned}$$

Your questions and the quiz

To benefit from the lecture, now you could

- ask or submit your anonymous questions to the COMP229 folder after the lecture;
- write down your summary in 2–3 phrases, e.g. list key concepts you have learned;
- talk to your classmates to revise the lecture.

Question. What are the first two principal components of the given data?

Answer to the quiz and summary

Answer. The first two principal components are along the eigenvectors $\vec{v}_1 = (\sqrt{5}+1, 2, 0)$ and $\vec{v}_2 = (0, 0, 1)$ with the highest eigenvalues.

- Subtract the means so that the rows of the sample $k \times n$ matrix S have zero means.
- Compute the covariance matrix $M = \dfrac{SS^T}{n-1}$.
- Project the data (with all means 0) to the smaller space of a few first eigenvectors of M.