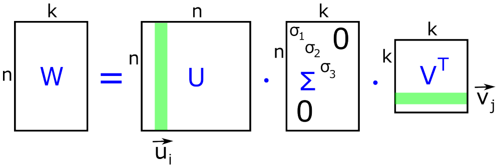
COMP229: Introduction to Data Science
Lecture 30: Revision

Olga Anosova, O.Anosova@liverpool.ac.uk
Autumn 2023, Computer Science department
University of Liverpool, United Kingdom

Recap: SVD

A Singular Value Decomposition of any $n \times k$ matrix W is $U \Sigma V^T$, where U, V are orthogonal matrices (high-dimensional rotations) and Σ is a diagonal (scaling) matrix with ordered square roots of eigenvalues of $W^T W$ (or $W^T W$) on the diagonal.



The columns of V and U are the eigenvectors of $W^T W$ and $W W^T$, respectively.

What key topics did you learn?

Two methods to reduce the dimension of data:

- Linear regression in Lecture 15 covered only dimension 2, a similar approach extends to higher dimensions.
- PCA (EIG and SVD) based on Linear Algebra: Lectures 28-29.

How different are these methods, e.g. do they produce different results for data samples in $\mathbb{R}^2$?

The data example from Lecture 28

Lecture 28 has found the first principal direction $\vec{v}_1 = (\sqrt{5}+1, 2, 0)$ for the 5-point cloud in $\mathbb{R}^3$.

subjects/students	s_{i1}	s_{i2}	s_{i3}	s_{i4}	s_{i5}	mean
Maths x	3	3	2	1	1	$\bar{x} = 2$
English y	2	3	2	2	1	$\bar{y} = 2$

Let's forget about the 3rd coordinate and find the linear regression line $y = ax + b$ for the 5 points projected to $\mathbb{R}^2$: Maths (x) and English (y).

What are the formulae for the coefficients a and b?

Revision of the linear regression

For n points (x_i, y_i), the *least-squares regression line* has an equation $y = ax + b$ that minimises the sum of squared vertical distances (residuals) $f(a, b) = \sum_{i=1}^{n} (ax_i + b - y_i)^2$.

In larger dimensions in matrix form: $\vec{y} = X\vec{\beta} + \vec{\varepsilon}$, where

$$\vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}, \vec{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \vec{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

such that the sum of squared errors $\|\vec{\varepsilon}\|_2^2$ is minimised.

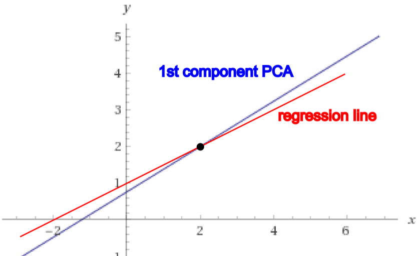
Linear regression vs PCA

The coefficients are $b = \bar{y} - a\bar{x}$ and
$a = r_{xy}\frac{s_y}{s_x} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$, where $\bar{x}$, $\bar{y}$ are the sample means, s_x, s_y are sample deviations.

In this example:

$$a = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{1\cdot 0 + 1\cdot 1 + 0\cdot 0 + (-1)\cdot 0 + (-1)\cdot(-1)}{4} = \frac{1}{2},$$

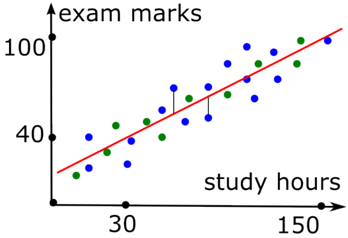
$$b = \bar{y} - a\bar{x} = 2 - 0.5 \cdot 2 = 1.$$

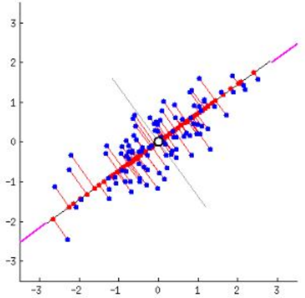


The regression line $y = 0.5x + 1$ differs from the first principal direction parallel to $(\sqrt{5} + 1, 2)$.

Geometry of linear regression vs PCA

The linear regression $y(x)$ minimises the sum of squared *vertical* distances, PCA minimises the variance, i.e. the sum of squared *orthogonal* distances to the final line.





1.Introductory statistics

L2: descriptive statistics
sample mean, sample standard deviation

L3: box plot summaries
median, mode, quartiles, outliers, plots

2.Probability distributions

L4: probabilities, exclusivity, independence
axioms, sum rule, product rule

L5: conditional probability, Bayes formula

L6: probabilistic paradoxes
conditional, mean, aggregation, finite approximation

L7: dicsrete distributions
random variable, expectation, cdf, pmf, uniform, Bernoully, Binomial

L8: continuous distributions
pdf vs cdf, uniform, exponential, curse of dimensionality

3. About Normal

L9: normal distribution
CLT, CIs

L10: hypotheses and significance
hypothesis testing, P-value

L11: moving from normality
Cauchy distribution, limitations of common approaches

- **4. Equivalences and vectors**
  - **L12: equivalence relations, vector spaces**
    fixed and free vectors, angles
  - **L13: vector operations**
    scalar product, Cauchy inequality
- **5.Correlation and regression**
  - **L14: correlation & scatterplots**
    normalisation, covariance
  - **L15: simple linear regression**
    formulae, Anscombe's quartet
- **6. Clustering**
  - **L16: metric axioms**
    types of clustering, L_p metrics, k-means objective
  - **L17: intro to clustering**
    types, hierarchical, DTW
  - **L18: Lloyd's k-means**
    Voronoi cells

7. Linear maps and isometries

L19: matrices of linear maps
linear vs affine, standard basis, matrix multiplication

L20: isometries
the kernel trick, translations, rotations, reflections

L21: orthogonal map
connection between linear, affine maps, bijections and isometries

L22: isometry invariants
invariants, complete invariants, pairwise distances

8. Invariants of linear maps

L23: determinant of a matrix
linear independence of vectors, properties of determinant

L24: areas of planar polygons
shoelace formula, convex hull, geometry of a determinant

L25: change of linear basis
formula, vector space models in NLP

L26: eigenthings
conjugation, trace, PageRank

9. Dimensionality reduction

L27: covariance matrix

L28: Principal Component Analysis (PCA) via eigendecomposition

L29: Singular Value Decomposition (SVD)

What to expect at the exam

Your mark for the module = 70% exam + 30% mid-term test

Time: 2 hours, 5 questions, all questions will be marked.

Exam consists of 5 written questions (no MCQ part) that sum up to 100 points.

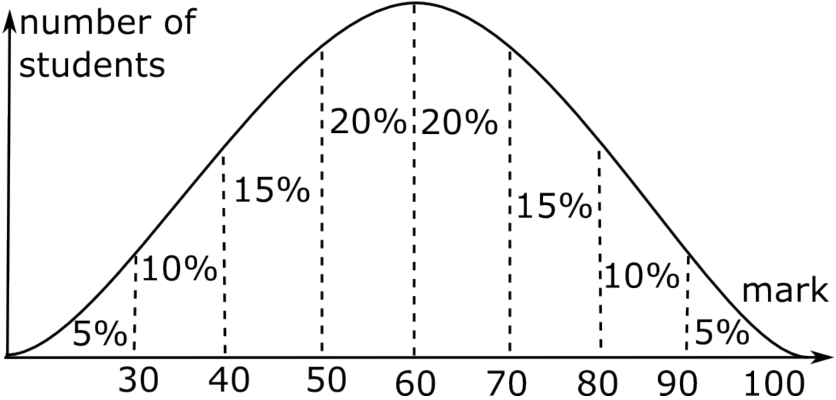
Each question is worth 20 points, subparts of questions will have points written next to them.

All questions requre understanding of definitions and formulas.

Calculators and tables of normal distributions are neither needed nor allowed.

An example distribution of marks

The pass mark is 40, first class marks are 70+.



An expected average mark should be in [60,65].

Advice on revisions before the exam

- Short regular revisions are better (1-2 hours per module each working day) than a non-stop rush only few days before the actual exam.
- During revisions, focus on simpler concepts rather than wasting time on harder topics.
- The pass mark is 40%. Your exam marks will likely be forgotten even by you in a few weeks.
- Rest or sleep between revisions! A good sleep strengthens neural connections, otherwise your learning is wasted.

Tutorials are continuing next week, I will be here at the usual lecture time on Monday the 4th December and will cover DBSCAN clustering and answer your revision questions.

From theory to coding

A short implementation of SVD in Python is available on Canvas (file svd.html)