Mathematics of Data Science
Chapter I: Examples of Mathematics in Data Science

Prof. Dr. Holger Rauhut
Chair for Mathematics of Information Processing
RWTH Aachen University

WS 2021/22

Role of mathematics in data science

- Acquisition, processing, transmission and storage (compression) of data are usually based on mathematical concepts, methods and algorithms

Role of mathematics in data science

- Acquisition, processing, transmission and storage (compression) of data are usually based on mathematical concepts, methods and algorithms
- Development and understanding of methods of data science often requires significant amount of mathematics

Role of mathematics in data science

- Acquisition, processing, transmission and storage (compression) of data are usually based on mathematical concepts, methods and algorithms
- Development and understanding of methods of data science often requires significant amount of mathematics
- It is often crucial to prove that algorithms in data science work (under suitable conditions), not just to rely on numerical tests

Role of mathematics in data science

- Acquisition, processing, transmission and storage (compression) of data are usually based on mathematical concepts, methods and algorithms
- Development and understanding of methods of data science often requires significant amount of mathematics
- It is often crucial to prove that algorithms in data science work (under suitable conditions), not just to rely on numerical tests
- It is also crucial to understand (prove) limits of methods, i.e., to understand when algorithms fail.

Mathematical fields in data science

Methods of data science require mathematical tools from various fields

- Linear algebra
- Analysis
- Probability theory
- Statistics
- Optimization
- Discrete mathematics (discrete optimization, graph theory)
- Numerical analysis
- Algebra

Mathematical fields in data science

Methods of data science require mathematical tools from various fields

- Linear algebra
- Analysis
- Probability theory
- Statistics
- Optimization
- Discrete mathematics (discrete optimization, graph theory)
- Numerical analysis
- Algebra

We will only cover a small part in this course. More material will be provided in specialized courses.

Example 1 – Regression

Predict weight of a person from sex, age and height!

Person	1	2	$\cdots$	p
x_1: Sex	1 (female)	-1 (male)	$\cdots$	1
x_2: Age (years)	25	37	$\cdots$	45
x_3: Height (cm)	165	182	$\cdots$	175
y: Weight (kg)	52.2	85.3	$\cdots$	55.7

Example 1 – Regression

Predict weight of a person from sex, age and height!

Person	1	2	$\cdots$	p
x_1: Sex	1 (female)	-1 (male)	$\cdots$	1
x_2: Age (years)	25	37	$\cdots$	45
x_3: Height (cm)	165	182	$\cdots$	175
y: Weight (kg)	52.2	85.3	$\cdots$	55.7

Fit linear model to data:

$$y = h_a(x) = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 = \langle a, x \rangle \quad (x_0 = 1)$$

Example 1 – Regression

Predict weight of a person from sex, age and height!

Person	1	2	$\cdots$	p
x_1: Sex	1 (female)	-1 (male)	$\cdots$	1
x_2: Age (years)	25	37	$\cdots$	45
x_3: Height (cm)	165	182	$\cdots$	175
y: Weight (kg)	52.2	85.3	$\cdots$	55.7

Fit linear model to data:

$$y = h_a(x) = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 = \langle a, x \rangle \quad (x_0 = 1)$$

Optimization approach ($x^j \in \mathbb{R}^3, y^j \in \mathbb{R}$ corresponds to person j)

$$\min_{a \in \mathbb{R}^4} \sum_{j=1}^{p} (y^j - h_a(x^j))^2 \quad \text{solution via Linear Algebra}$$

Example 1 – Regression

Predict weight of a person from sex, age and height!

Person	1	2	$\cdots$	p
x_1: Sex	1 (female)	-1 (male)	$\cdots$	1
x_2: Age (years)	25	37	$\cdots$	45
x_3: Height (cm)	165	182	$\cdots$	175
y: Weight (kg)	52.2	85.3	$\cdots$	55.7

Fit linear model to data:

$$y = h_a(x) = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 = \langle a, x \rangle \quad (x_0 = 1)$$

Optimization approach ($x^j \in \mathbb{R}^3, y^j \in \mathbb{R}$ corresponds to person j)

$$\min_{a \in \mathbb{R}^4} \sum_{j=1}^{p} (y^j - h_a(x^j))^2 \quad \text{solution via Linear Algebra}$$

Analysis of accuracy of prediction: Probability Theory and Statistics

Example 2 – Principal Component Analysis

Given data points $x^1, \ldots, x^p \in \mathbb{R}^n$, find a subspace $W \subset \mathbb{R}^n$ of dimension $m < n$ such that the data approximately lie in W!

Example 2 – Principal Component Analysis

Given data points $x^1, \ldots, x^p \in \mathbb{R}^n$, find a subspace $W \subset \mathbb{R}^n$ of dimension $m < n$ such that the data approximately lie in W!

Application: Reduce from a large number of variables to a smaller set that characterizes well the data.
Example: Collection of medical data may be reduced to a few essential ones (saves cost and pain for the patient).

Example 2 – Principal Component Analysis

Given data points $x^1, \dots, x^p \in \mathbb{R}^n$, find a subspace $W \subset \mathbb{R}^n$ of dimension $m < n$ such that the data approximately lie in W!

Application: Reduce from a large number of variables to a smaller set that characterizes well the data.
Example: Collection of medical data may be reduced to a few essential ones (saves cost and pain for the patient).

Method: Set $X = (x^1|x^2|\cdots|x^p) \in \mathbb{R}^{n\times p}$ and represent W as the range of a matrix $U \in \mathbb{R}^{n\times m}$.
Find minimizer of optimization problem

$$\min_{U\in\mathbb{R}^{n\times m}} \sum_{j=1}^{p} \|x^j - UU^T x^j\|_2^2$$

Minimizer can be computed using the singular value decomposition (SVD) of X (linear algebra)

Example 3 – Supervised Learning

Given an image, automatically determine whether it contains a cat or not!

Supervised Learning

Mathematical problem: Given pairs (x^j, y^j),$j = 1, \ldots, p$ of (training) input data with $x^j \in \mathbb{R}^p$ and $y^j \in \mathbb{R}$, or $y^j \in \{-1, 1\}$ find a function $h : \mathbb{R}^p \to \mathbb{R}$ that accurately predicts labels y of future data x.

Supervised Learning

Mathematical problem: Given pairs (x^j, y^j), $j = 1, \dots, p$ of (training) input data with $x^j \in \mathbb{R}^p$ and $y^j \in \mathbb{R}$, or $y^j \in \{-1, 1\}$ find a function $h : \mathbb{R}^p \to \mathbb{R}$ that accurately predicts labels y of future data x.

Given a loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ (e.g. $\ell(y, z) = (y - z)^2$) and a set $\mathcal{H}$ of possible hypothesis functions $h : \mathbb{R}^p \to \mathbb{R}$ find minimizer of optimization problem

$$\min_{h \in \mathcal{H}} \sum_{j=1}^{p} \ell(h(x^j), y^j)$$

Supervised Learning

Mathematical problem: Given pairs (x^j, y^j),$j = 1, \dots, p$ of (training) input data with $x^j \in \mathbb{R}^p$ and $y^j \in \mathbb{R}$, or $y^j \in \{-1, 1\}$ find a function $h : \mathbb{R}^p \to \mathbb{R}$ that accurately predicts labels y of future data x.

Given a loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ (e.g. $\ell(y, z) = (y - z)^2$) and a set $\mathcal{H}$ of possible hypothesis functions $h : \mathbb{R}^p \to \mathbb{R}$ find minimizer of optimization problem

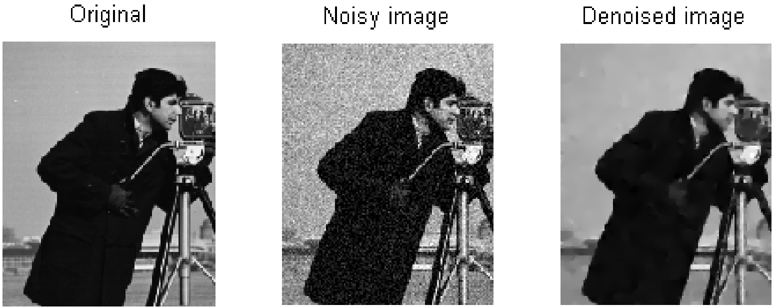
$$\min_{h \in \mathcal{H}} \sum_{j=1}^{p} \ell(h(x^j), y^j)$$

Analysis of prediction error requires techniques from probability theory.

Example for $\mathcal{H}$ (deep learning):
Deep neural networks of a prescribed structure (parametrization).

Details in course:
Mathematical Foundations of Machine Learning

Example 4 - Image Denoising



Reconstruct original image $u \in \mathbb{R}^{m \times n}$ from measured noisy image $\widetilde{u} = u + e$, where $e \in \mathbb{R}^{m \times n}$ represents noise!

Example 4 - Image Denoising

Original

Noisy image

Denoised image

Reconstruct original image $u \in \mathbb{R}^{m \times n}$ from measured noisy image $\widetilde{u} = u + e$, where $e \in \mathbb{R}^{m \times n}$ represents noise!

Variational approach: compute minimizer $\hat{u}$ of optimization program

$$\min_{v \in \mathbb{R}^{m \times n}} \|\widetilde{u} - v\| + \lambda R(v),$$

where $R : \mathbb{R}^{m \times n} \to \mathbb{R}$ is a regularization term, representing prior assumptions on u such as smoothness, sparsity etc.

Example 5 - Analog-to-Digital Conversion

For a digital representation, a continuous time signal (e.g. music) $f : \mathbb{R} \to \mathbb{R}$ is sampled (and quantized) at discrete times, e.g., for $B > 0$,

$$y_j = f\left(\frac{j}{2B}\right), \quad j \in \mathbb{Z}.$$

Task: Reconstruct f accurately from the sequence $y = (y_j)_{j\in\mathbb{Z}}$ (or from a finite, quantized subsequence)!

Example 5 - Analog-to-Digital Conversion

For a digital representation, a continuous time signal (e.g. music) $f : \mathbb{R} \to \mathbb{R}$ is sampled (and quantized) at discrete times, e.g., for $B > 0$,

$$y_j = f\left(\frac{j}{2B}\right), \quad j \in \mathbb{Z}.$$

Task: Reconstruct f accurately from the sequence $y = (y_j)_{j\in\mathbb{Z}}$ (or from a finite, quantized subsequence)!
Fourier transform of f:

$$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(t)e^{-2\pi i\xi t}dt.$$

If f is square-integrable and such that $\hat{f}(\xi) = 0$ for $|\xi| \geq B$ (f belongs to the Paley-Wiener space), then f can be reconstructed exactly via the sampling series

$$f(t) = \sum_{j\in\mathbb{Z}} f\left(\frac{j}{2B}\right) \operatorname{sinc}\left(t - \frac{j}{2B}\right)$$

Example 5 - Analog-to-Digital Conversion

For a digital representation, a continuous time signal (e.g. music) $f : \mathbb{R} \to \mathbb{R}$ is sampled (and quantized) at discrete times, e.g., for $B > 0$,

$$y_j = f\left(\frac{j}{2B}\right), \quad j \in \mathbb{Z}.$$

Task: Reconstruct f accurately from the sequence $y = (y_j)_{j \in \mathbb{Z}}$ (or from a finite, quantized subsequence)!
Fourier transform of f:

$$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(t) e^{-2\pi i \xi t} dt.$$

If f is square-integrable and such that $\hat{f}(\xi) = 0$ for $|\xi| \geq B$ (f belongs to the Paley-Wiener space), then f can be reconstructed exactly via the sampling series

$$f(t) = \sum_{j \in \mathbb{Z}} f\left(\frac{j}{2B}\right) \operatorname{sinc}\left(t - \frac{j}{2B}\right)$$

Mathematical field: Fourier analysis

Example 6 - Dimensionality Reduction

If data points $x^1, \ldots, x^p$ are in a high-dimensional space $\mathbb{R}^n$, then computational effort is often high (e.g. nearest neighbor search in data bases).

Example 6 - Dimensionality Reduction

If data points $x^1, \dots, x^p$ are in a high-dimensional space $\mathbb{R}^n$, then computational effort is often high (e.g. nearest neighbor search in data bases).

Task: Project into a low-dimensional space $\mathbb{R}^m$, $m \ll n$, such that pairwise Euclidean distances are almost preserved,

$$(1-\epsilon)\|x^j - x^k\|_2 \leq \|Px^j - Px^k\|_2 \leq (1+\epsilon)\|x^j - x^k\|_2 \quad \text{for all } j, k$$

Example 6 - Dimensionality Reduction

If data points $x^1, \ldots, x^p$ are in a high-dimensional space $\mathbb{R}^n$, then computational effort is often high (e.g. nearest neighbor search in data bases).

Task: Project into a low-dimensional space $\mathbb{R}^m$, $m \ll n$, such that pairwise Euclidean distances are almost preserved,

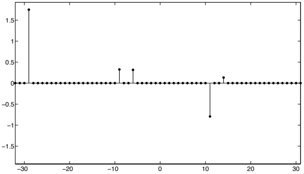
$$(1-\epsilon)\|x^j - x^k\|_2 \leq \|Px^j - Px^k\|_2 \leq (1+\epsilon)\|x^j - x^k\|_2 \quad \text{for all } j, k$$

Johnson-Lindenstrauss Lemma: If $P \in \mathbb{R}^{m \times n}$ is chosen at random (e.g. as Gaussian random matrix), then the inequality holds with probability at least $1 - \delta$ if

$$m \geq C\epsilon^{-2} \log(p/\delta).$$

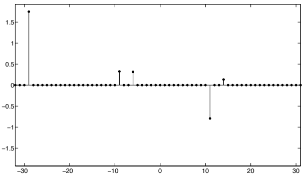
Mathematical tools: Probability Theory in High Dimensions

Example 7 - Signal Reconstruction (Compressed Sensing)

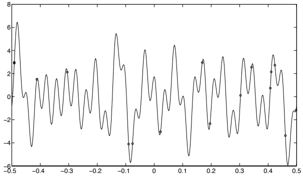


Fourier-Coefficients

Example 7 - Signal Reconstruction (Compressed Sensing)

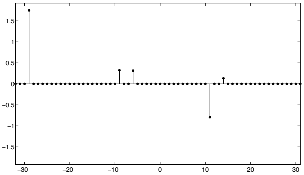


Fourier-Coefficients

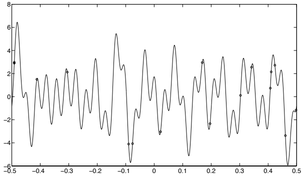


Time-Domain Signal with 16 Samples

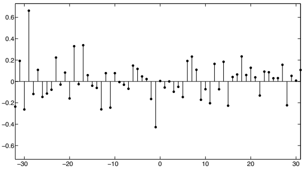
Example 7 - Signal Reconstruction (Compressed Sensing)



Fourier-Coefficients

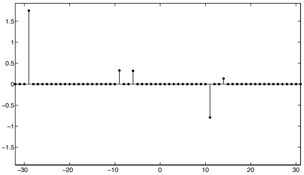


Time-Domain Signal with 16 Samples

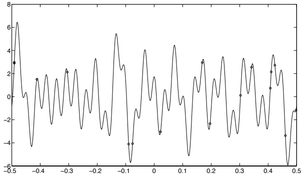


Traditional Reconstruction

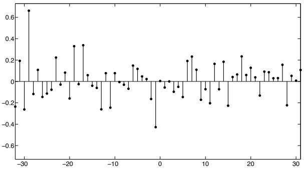
Example 7 - Signal Reconstruction (Compressed Sensing)



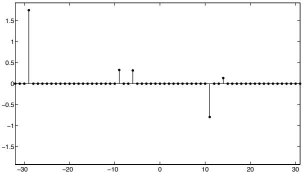
Fourier-Coefficients



Time-Domain Signal with 16 Samples



Traditional Reconstruction



Reconstruction via ℓ_1-minimization

Compressed Sensing

Solve underdetermined system

$$y = Ax, \quad \text{where } A \in \mathbb{R}^{m \times n} \quad \text{with } m \ll n.$$

Without further assumptions impossible!

Compressed Sensing

Solve underdetermined system

$$y = Ax, \quad \text{where } A \in \mathbb{R}^{m \times n} \quad \text{with } m \ll n.$$

Without further assumptions impossible!

Assume x is (approximately) s-sparse: $\|x\|_0 := \#\{\ell : x_\ell \neq 0\} \leq s$

Compressed Sensing

Solve underdetermined system

$$y = Ax, \quad \text{where } A \in \mathbb{R}^{m \times n} \quad \text{with } m \ll n.$$

Without further assumptions impossible!

Assume x is (approximately) s-sparse: $\|x\|_0 := \#\{\ell : x_\ell \neq 0\} \leq s$

Convex optimization approach:

$$\min_{z \in \mathbb{R}^n} \|z\|_1 \quad \text{subject to } Az = y.$$

Compressed Sensing

Solve underdetermined system

$$y = Ax, \quad \text{where } A \in \mathbb{R}^{m \times n} \quad \text{with } m \ll n.$$

Without further assumptions impossible!

Assume x is (approximately) s-sparse: $\|x\|_0 := \#\{\ell : x_\ell \neq 0\} \leq s$

Convex optimization approach:

$$\min_{z \in \mathbb{R}^n} \|z\|_1 \quad \text{subject to } Az = y.$$

If A is random draw of a Gaussian matrix, then one can show that x is reconstructed exactly (!) with high probability provided that

$$m \gtrsim 2s \ln(en/s).$$

Compressed Sensing

Solve underdetermined system

$$y = Ax, \quad \text{where } A \in \mathbb{R}^{m \times n} \quad \text{with } m \ll n.$$

Without further assumptions impossible!

Assume x is (approximately) s-sparse: $\|x\|_0 := \#\{\ell : x_\ell \neq 0\} \leq s$

Convex optimization approach:

$$\min_{z \in \mathbb{R}^n} \|z\|_1 \quad \text{subject to } Az = y.$$

If A is random draw of a Gaussian matrix, then one can show that x is reconstructed exactly (!) with high probability provided that

$$m \gtrsim 2s \ln(en/s).$$

Details: Course on Compressive Sensing.