DATA.STAT.770

Dimensionality Reduction and Visualization

Spring 2021, 5 ects credits

Jaakko Peltonen

Lecture 1: Introduction, properties of high-dim. data

Practical Information

- Lectures on Tuesdays 14-16 each week in Zoom, from January 12 onward.
- **No exercise sessions**, instead **home exercise packs**.
- Language: English
- **You must sign up for the course** using Sisu. If there is any problem, contact the lecturer at jaakko.peltonen [at] tuni.fi .

Material:

- course slides, additional-reading articles
- Slides originally in part by Kerstin Bunte, Francesco Corona, Manuel Eugster, Amaury Lendasse
- Exercise packs released during the spring. Will contain some mathematical exercises, some implementation & testing of methods, either from scratch or using pre-existing toolboxes.
- **Course Moodle: https://moodle.tuni.fi/course/view.php?id=15590**
- A discussion area is available in Moodle

Practical Information, cont.

Grading (note: preliminary, may change):

- Each exercise graded 0-2 (integer),
  exercise packs total graded 0-5.
- Exam on final lecture, graded 0-5.

- To pass the course, you must pass the **exam** (grade 1 or more) and pass **exercise packs** (grade 1 or more).
- Passing grades are kept fractional between 1 and 5 (e.g. ”3.437”)
- Final course grade
  = round(0.8 * ExamGrade + 0.2 * ExercisesGrade)
  (e.g. 3.499 rounds to 3, 3.501 rounds to 4)

Preliminary Schedule (may change!)

Lecture 1 (January 12, 2021, 14:15-16): Introduction, properties of high-dimensional data.

Lecture 2 (January 19, 2021, 14:15-16): Feature selection.

Lecture 3 (January 26, 2021, 14:15-16): Linear dimensionality reduction.

Lecture 4 (February 2, 2021, 14:15-16): Linear dimensionality reduction continued.

Lecture 5 (February 9, 2021, 14:15-16): Graphical excellence.

Lecture 6 (February 16, 2021, 14:15-16): Human perception.

Lecture 7 (February 23, 2021, 14:15-16): Nonlinear dimensionality reduction, part 1.

Lecture 8 (March 9, 2021, 14:15-16): Nonlinear dimensionality reduction, part 2.

Lecture 9 (March 16, 2021, 14:15-16): Nonlinear dimensionality reduction, part 3.

Lecture 10 (March 23, 2021, 14:15-16): Metric learning.

Lecture 11 (March 30, 2021, 14:15-16): Neighbor embedding, part 1.

Lecture 12 (April 6, 2021, 14:15-16): Neighbor embedding, part 2.

Lecture 13 (April 13, 2021, 14:15-16): Graph visualization.

Lecture 14 (April 20, 2021, 14:15-16): Dimensionality reduction for graph layout.

Lecture 15 (April 27, 2021, 14:15-16): Lectures 13-14 continued and conclusion.

Preliminary date for first exam: May 11, 18 or 25.

A world of high-dimensional measurements

Motivation – high-dimensional data

- In bioinformatics, expressions of tens of thousands of genes can be measured from each tissue sample.
- In social networks, each person may be associated with hundreds or thousands of events (tweets, likes, friendships, interactions etc.)
- In weather and climate prediction, multiple types of information (temperature, sunshine, precipitation etc.) are measured at each moment at thousands of stations across Europe – see http://eca.knmi.nl/
- In finance, stock markets involve changing prices of thousands of stocks at each moment

Our capacity to measure a phenomenon can in some cases exceed our capacity to analyze it (in any complex way)

Motivation

High-dimensional data:

- World is multidimensional: bees, ants, neurons

- In technology (computer networks, sensor arrays, etc.):
    - Combination of many simple units allows complex tasks
    - cheaper than creating a specific device and robust: malfunction of a few units does not impair whole system

Motivation

High-dimensional data:

- World is multidimensional: bees, ants, neurons
- In technology (computer networks, sensor arrays, etc.):
    - Combination of many simple units allows complex tasks
    - cheaper than creating a specific device and robust: malfunction of a few units does not impair whole system
- Efficient management or understanding of all units requires taking redundancy into account.
  ---> summarize as a smaller set with no or less redundancy:
  Dimensionality Reduction (DR)
- Goal:
    - Extract information hidden in the data
    - Detect variables relevant for a specific task and how variables Interact with each other
      ---> Reformulate data with less variables

Demonstration example

Sometimes distance information of higher-dimensional entities can be shown on a display without errors.
3D Probability Density: $x + y + z = 1$

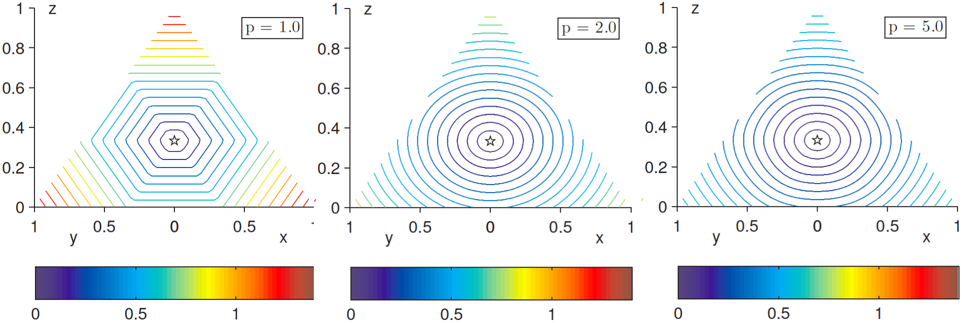
The objects are different probability distributions (different choices x,y,z such that x+y+z=1).

Distances between probability distributions can be computed by various metrics such as Minkowski distances (next slide). It turns out the result can be illustrated on a display.

Demonstration example

Sometimes distance information of higher-dimensional entities can be shown on a display without errors.
3D Probability Density: $x + y + z = 1$

Equidistant lines with the Minkowski metric for 3D probability densities



$$D_{\text{Minkowski}}(\mathbf{u}, \mathbf{v}) = \left(\sum_{i=1}^{d} |u_i - v_i|^p \right)^{\frac{1}{p}}$$

Distances are important for many methods later in the course.

Why reduce dimensionality – different uses

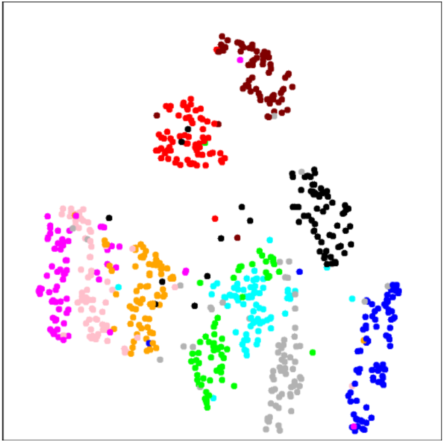
For automated use by computers:

- Saves the cost of observing the features
- Takes less memory, storage, transmission time
- Reduces subsequent computation cost
- Reduces number of parameters
- Simpler models are more robust on small datasets

For use by humans:

- More interpretable; simpler explanations
- Data visualization (structure, groups, outliers, etc) if plotted in 2 or 3 dimensions

This is easier to interpret… … than this



	A	B	C	D	E	F	G	H	I
1	0,143627544	0,76263186	0,512298994	0,798081842	0,250446206	0,198933438	0,373004635	0,179964027	0,786585584
2	0,238272221	0,59648499	0,52698488	0,897050043	0,690755459	0,564704472	0,974296788	0,833605263	0,340292797
3	0,452312219	0,364607266	0,835929876	0,837238703	0,248525629	0,100641438	0,652762643	0,393872306	0,924477495
4	0,577173206	0,492994165	0,333057811	0,385229828	0,037308688	0,469644144	0,314180839	0,243461676	0,184843427
5	0,67891893	0,340443544	0,687615283	0,367366192	0,420620224	0,856951324	0,25335427	0,090625922	0,602530105
6	0,468870715	0,254673163	0,809489606	0,140961417	0,998013817	0,579129822	0,486475348	0,939421771	0,719315503
7	0,958072889	0,615193778	0,805320487	0,732454057	0,263283711	0,369680074	0,77672003	0,587880294	0,500913442
8	0,410643075	0,202900881	0,201630524	0,516237688	0,839719921	0,126377627	0,211997589	0,077192827	0,054175685
9	0,019142448	0,192971776	0,621198974	0,018840329	0,714956274	0,247262027	0,580007352	0,006502912	0,742615423
10	0,992430514	0,569451327	0,847166487	0,211236799	0,875242799	0,900147923	0,57750103	0,986623143	0,357386574
11	0,546098656	0,942951822	0,617923535	0,382720211	0,594655215	0,4078977	0,315143077	0,511397517	0,081731002
12	0,809785753	0,269177106	0,96688787	0,72858903	0,27501773	0,287689517	0,258217208	0,507360818	0,933360928
13	0,762619687	0,10670799	0,444715297	0,839030984	0,50909486	0,596634927	0,594536219	0,656005232	0,013992229
14	0,583634943	0,098460302	0,015527529	0,602532446	0,602813388	0,957901425	0,283072265	0,508866322	0,348537715
15	0,433799997	0,486712324	0,028420831	0,192898353	0,260089411	0,709564135	0,474739488	0,038486801	0,873120818
16	0,467110848	0,662058957	0,903530741	0,316767938	0,906243891	0,247409084	0,448794	0,283702092	0,391870203
17	0,802011862	0,589121008	0,992245409	0,722205046	0,652194193	0,409716809	0,579913608	0,946664964	0,721417716
18	0,96500794	0,838101565	0,577065767	0,389130664	0,897937781	0,579856649	0,959265703	0,693929178	0,383382382
19	0,675932665	0,866526788	0,572390466	0,109605676	0,320051695	0,744457152	0,759240198	0,039603357	0,893583362
20	0,59424673	0,168204394	0,399634255	0,602154047	0,722314151	0,060906403	0,355769501	0,786641984	0,230930249
21	0,305516901	0,390292247	0,945770002	0,441385326	0,521919936	0,229332483	0,253172983	0,653777855	0,401563282
22	0,88403356	0,723829591	0,891987076	0,902963833	0,062350279	0,69443381	0,559043572	0,815094968	0,244848581
23	0,864712216	0,331245804	0,917661169	0,184740538	0,042498818	0,893754464	0,942336603	0,771793981	0,974220777
24	0,630685092	0,826324756	0,000436316	0,018998619	0,78606248	0,468370951	0,049878636	0,969012019	0,311859167
25	0,907125201	0,857508921	0,112309298	0,25492589	0,159733113	0,844925056	0,111669446	0,844417667	0,44014374
26	0,098502477	0,684295712	0,492248837	0,109986083	0,23463731	0,181429059	0,289775145	0,21426066	0,961085635
27	0,456837715	0,822081204	0,68035015	0,873881784	0,028513596	0,802397438	0,190258625	0,032695141	0,859995278
28	0,716668213	0,949602812	0,382940948	0,714633616	0,359653609	0,812883482	0,226834126	0,466200796	0,450994375
29	0,326822215	0,195760563	0,662755339	0,379986736	0,022971772	0,025301282	0,706052747	0,814941845	0,777949337
30	0,358381429	0,045770876	0,158896779	0,882555211	0,534904493	0,204605316	0,297079868	0,606062495	0,08684194
31	0,74568106	0,74293577	0,764655228	0,638629826	0,521018431	0,671149034	0,881393171	0,535097446	0,062761582

Why are advanced methods needed for dimensionality reduction?

- High-dimensional data has surprising properties
- Hard to intuitively understand them
- We'll discuss many of them on this lecture

- They can also lead to poor modeling performance
- On the other hand, the high-dimensional data are ”real” and we want to preserve their original properties, just in a smaller dimensional setting where it is easier to handle them
- simple reduction would not preserve the high-dimensional properties well

Applications

- Processing of sensor arrays:
  radio telescopes, biomedical (electroencephalograph (EEG), electrocardiogram (ECG)), seismography, weather forecasting
- Image processing:
  digital camera (photosensitive CCD or CMOS captors)
- Multivariate data analysis:
  related measurements coming from different sensors (e.g. cars: rotation-, force-, position-, temperature sensors)

Information discovery and extraction helps to:

- understand existing data: assign class, color and rank
- infer and generalize to new data (“test” or “validation set”)

Theoretical Motivations

- Well-known properties of 2D and 3D Euclidean spaces change with growing dimensions: “curse of dimensionality”
- Visualization regards mainly 2 classes of data:

Spatial data

- spatial: drawing 1 or 2 dimensions straightforward.
  3D already harder

Spatial data

- spatial: drawing 1 or 2 dimensions straightforward. 3D already harder

(perspective still recent discovery: paintings before Renaissance not very different from Egyptian papyri)

Spatial data

- spatial: drawing 1 or 2 dimensions straightforward.
  3D already harder

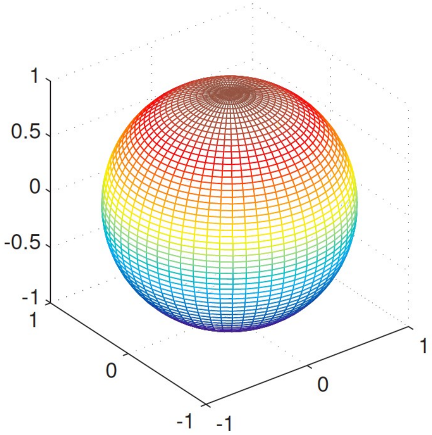
Even today smooth, dynamic and realistic representation of 3D world requires highly specialized chips

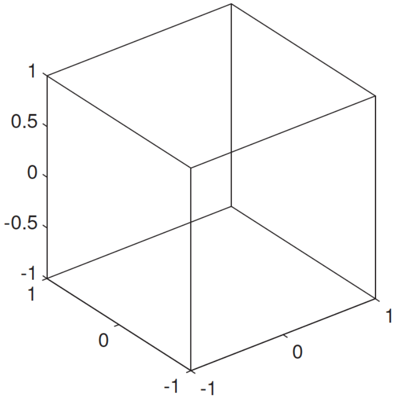
(landscape art from Elder Scolls V: Skyrim, Bethesda Softworks)

Spatial data

Higher dimensions?

- Humans attempt to understand objects same way as in 3D: seeking distances from one point to another, distinguish far from close, follow discontinuities like edges, corners and so on

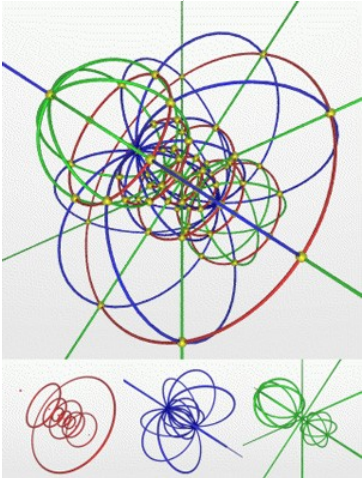




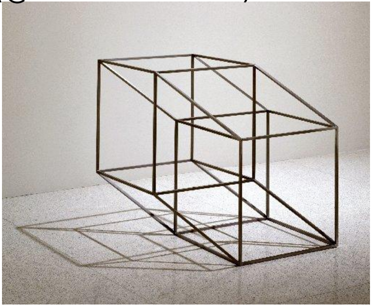
Spatial data

Higher dimensions?

- Humans attempt to understand objects same way as in 3D: seeking distances from one point to another, distinguish far from close, follow discontinuities like edges, corners and so on



4D Hypersphere and Hypercube projected onto 3D (parallels, meridians, hypermeridians) (@ClaudioRocchini)



Temporal data

- Because of time-information geometrical representation no longer unique
- draw evolution of each variable as function of time:
- temporal representation easily generalizes to more than 3 dimensions (for example EEG)

  → harder to perceive similarities and dissimilarities

Temporal data

- Because of time-information geometrical representation no longer unique
- draw evolution of each variable as function of time:
- temporal representation easily generalizes to more than 3 dimensions (for example EEG)
  → harder to perceive similarities and dissimilarities

Properties of High-dimensional Data

Curse of dimensionality

- Term first coined by Bellman 1961:
  Considering a cartesian grid of spacing 1/10 on the unit cube in 10D equals 10^{10} number of points.
  For 20D cube number of points increases to 10^{20}
- Bellman's interpretation:
  optimizing a function over a continuous domain of a few dozen variables by exhaustive searching a discrete space defined by crude discretization can easily face tens of trillions evaluations of the function
- amount of available data generally restricted to few observations→ high-D inherently sparse
- unexpected properties

Hypervolume of Cubes and Spheres

Volume of a Hypersphere:

$$V_{\mathrm{sphere}}(r) = \frac{\pi^{\frac{d}{2}} r^d}{\Gamma(1 + \frac{d}{2})}$$

corresp. circumscripted Hypercube (edges=sphere diameter)

$$V_{\mathrm{cube}}(r) = (2r)^d$$

Hypervolume of Cubes and Spheres

Volume of a Hypersphere:

$$V_{\mathrm{sphere}}(r) = \frac{\pi^{\frac{d}{2}} r^d}{\Gamma(1 + \frac{d}{2})}$$

corresp. circumscripted Hypercube (edges=sphere diameter)

$$V_{\mathrm{cube}}(r) = (2r)^d$$

Ratio $\lim_{d\to\infty} \frac{V_{\mathrm{sphere}}}{V_{\mathrm{cube}}} = 0 \to$ Cube becomes more and more spiky like a sea urchin, while the spherical body gets smaller and smaller

For $r = 0.5 \to V_{\mathrm{cube}} = 1 \Rightarrow \lim_{d\to\infty} V_{\mathrm{sphere}}(r) = 0$

→ nearly all high-D space is far away from the center

Hypervolume of a Thin Shell

$$\frac{V_{\mathrm{sphere}}(r) - V_{\mathrm{sphere}}(r(1-\epsilon))}{V_{\mathrm{sphere}}(r)} \qquad (\epsilon << 1)$$

Hypervolume of a Thin Shell

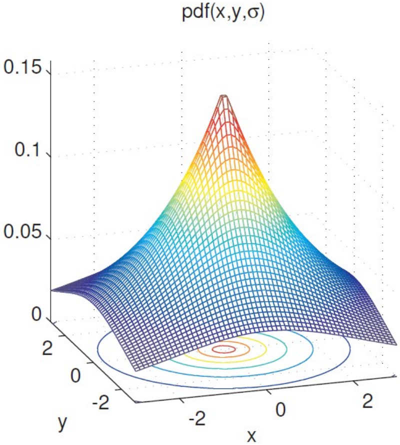
$$\frac{V_{\mathrm{sphere}}(r) - V_{\mathrm{sphere}}(r(1-\epsilon))}{V_{\mathrm{sphere}}(r)} \sim \frac{1^d - (1-\epsilon)^d}{1^d} \quad (\epsilon << 1)$$

For increasing dimensionality the ratio tends to 1
→ the shell contains almost all the volume (Wegman 1990)

Tail Probability of Isotropic Gaussian Distributions

Probability density function (pdf) of isotropic Gaussian distribution

$$p(\mathbf{v}) = \frac{1}{\sqrt{(2\pi)^d}} \exp\left(-\frac{1}{2}\frac{\|\mathbf{v}-\mu_{\mathbf{v}}\|^2}{\sigma^2}\right) \qquad \begin{cases} \mathbf{v} \in \mathbb{R}^d \\ \mu_{\mathbf{v}} \ (d\text{-dim. mean}) \\ \sigma^2 \ (\text{scalar variance}) \end{cases}$$



Assume random vector **v** has zero mean and unit variance, radius of equiprobable contours are spherical:

$$p(\mathbf{v}) = K(r) = \frac{1}{\sqrt{(2\pi)^d}} \exp\left(-\frac{r^2}{2}\right)$$

Tail Probability of Isotropic Gaussian Distributions

Surface of d-dimensional Hypersphere:

$$S_{\mathrm{sphere}}(r) = \frac{2\pi^{\frac{d}{2}} r^{d-1}}{\Gamma\left(\frac{d}{2}\right)}$$

Assume $r_{0.95}$ being the radius of a hypersphere that contains 95% of the distribution:

$$\frac{\int_0^{r_{0.95}} S_{\mathrm{sphere}}(r) K(r) dr}{\int_0^{\infty} S_{\mathrm{sphere}}(r) K(r) dr} = 0.95$$

→ $r_{0.95}$ grows with increasing dimensionality, larger and larger radius is needed to capture 95%

solutions of $r_{0.95}$ by numerical integration:

d	1	2	3	4	5	6
$r_{0.95}$	1.96	2.45	2.80	3.08	3.33	3.55

Concentration of Norms and Distances

- With growing dimensionality the contrast provided by usual metrics decreases
- The distribution of norms in a given distribution of points tends to concentrate→ *concentration phenomenon*
- Euclidean norm of iid (independent identical distributed) random vectors behaves unexpectedly

$$\|\mathbf{u}-\mathbf{v}\|_2 = \sqrt{\sum_{k=1}^{d}(u_k - v_k)^2}$$

$$\|\mathbf{a}\|_2 = \sqrt{<\mathbf{a},\mathbf{a}>} \qquad <\mathbf{u},\mathbf{v}> = \mathbf{u}^\top\mathbf{v} = \sum_{k=1}^{d} a_k b_k$$

iid random vectors distribute close to the surface of a hypersphere

→ Euclidean distance between any two vectors is approximately constant: $\lim_{d\to\infty} \frac{\mathrm{dist}_{\max}-\mathrm{dist}_{\min}}{\mathrm{dist}_{\min}} \to 0$

Diagonal of a Hypercube

Hypercube $[-1,1]^d$ and diagonal vectors **v** from center to a corner (2^d vectors of the form $[\pm 1, \pm 1, \dots, \pm 1]^T$)

- the angle between a diagonal **v** and an Euclidean coordinate axis $\mathbf{e}_j$=[0,...,1,...,0] is:

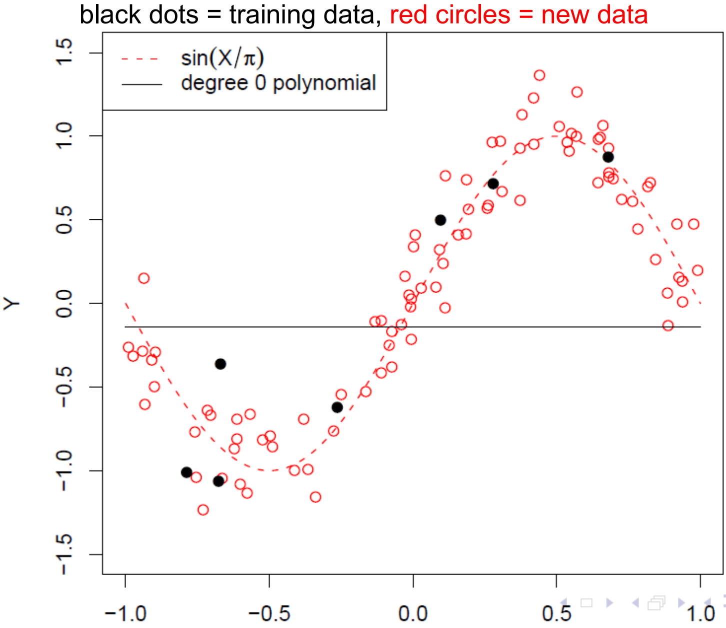
$$\cos\theta_d = \frac{<\mathbf{v}, \mathbf{e}_j>}{\sqrt{<\mathbf{v}, \mathbf{v}><\mathbf{e}_j, \mathbf{e}_j>}} = \frac{\pm 1}{\sqrt{d}} \xrightarrow[d \to \infty]{} 0$$

- The diagonals are nearly orthogonal to all coordinate axes for large d!
- Plotting a subset of 2 coordinates on a plane can be misleading: cluster of points lying near a diagonal will be plotted near the origin, whereas a cluster lying near a coordinate axis should be visible in some plot

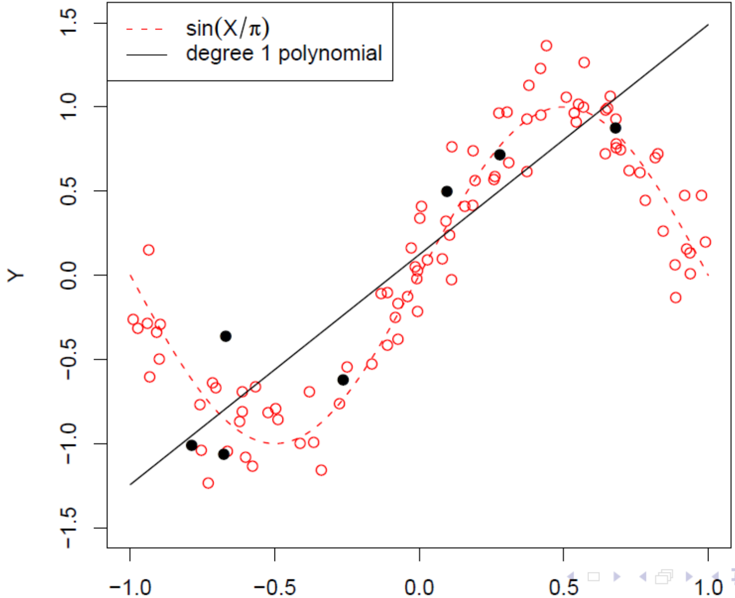
Curse of dimensionality and overfitting

- Many statistical models need ever more parameters when applied in higher dimensional spaces. E.g. Gaussian: needs d*d parameters in covariance matrix.
- Few data, many parameters –---> **overfitting**
- In overfitting, the model mistakes measurement noise for real effects. Parameters are adjusted to explain the noise.
- Result: the model fits the set of training data apparently well, but predicts poorly for new data.

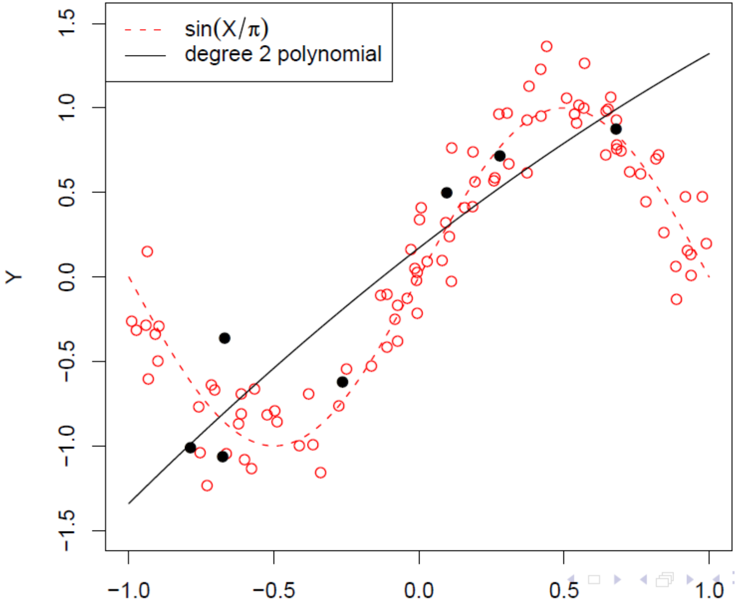
Example: Least square polynomial regression (1D)



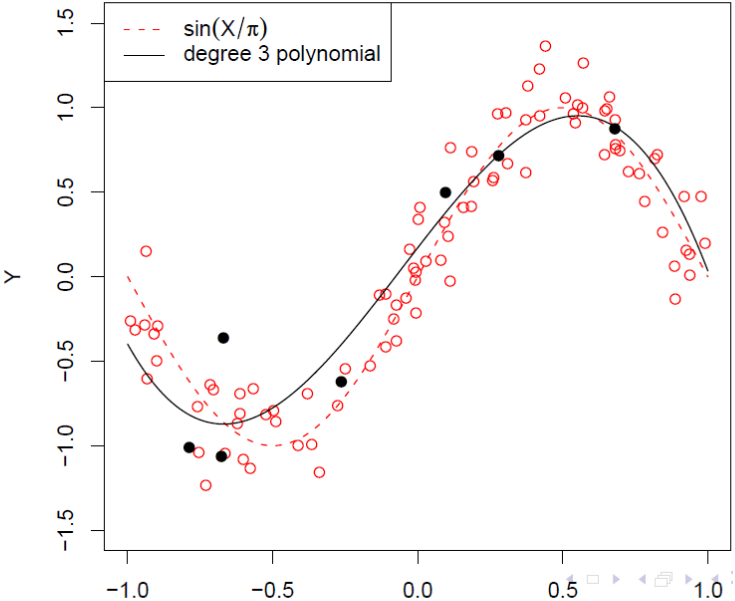
Example: Least square polynomial regression (1D)



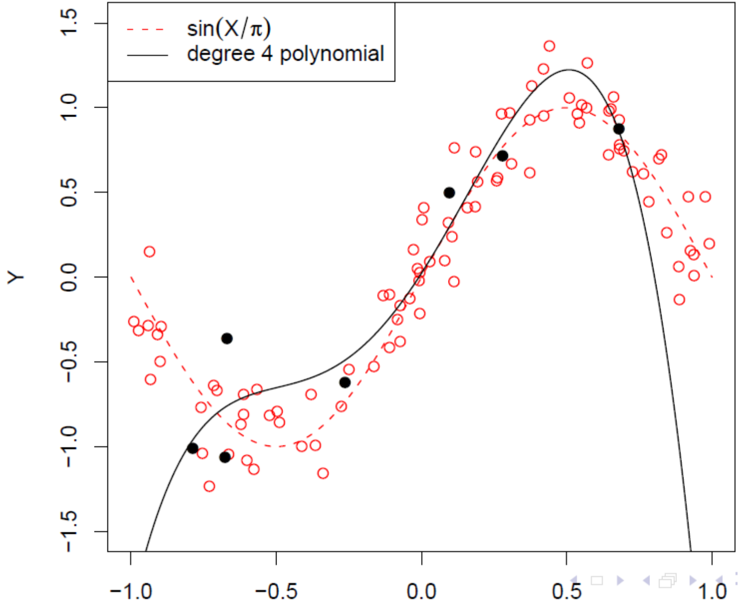
Example: Least square polynomial regression (1D)



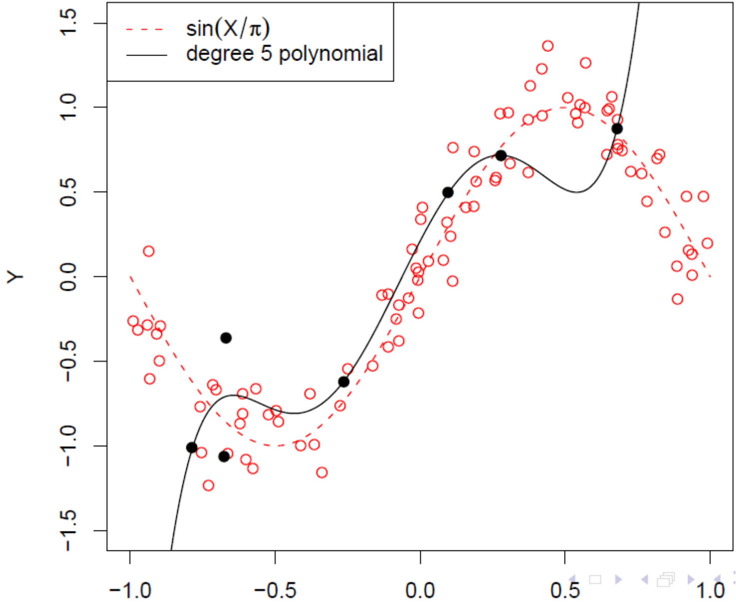
Example: Least square polynomial regression (1D)



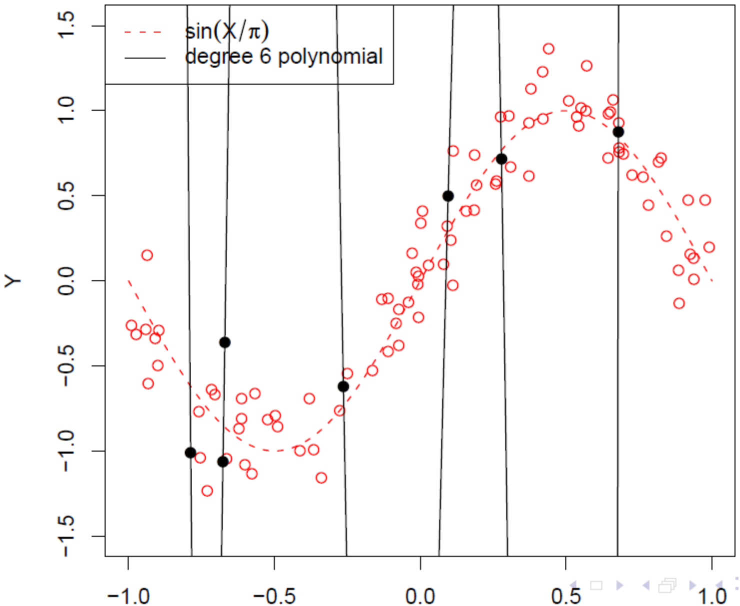
Example: Least square polynomial regression (1D)



Example: Least square polynomial regression (1D)



Example: Least square polynomial regression (1D)



Curse of dimensionality and overfitting

- Overfitted models fit training data well, but predict poorly for new data.
- In overfitting, predictions depend strongly on the choice of training data –--> the model has high variance over the choice (related to bias-variance dilemma)

- **The higher the data dimensionality, the more opportunities for overfitting!**
- E.g. classification: if there are more dimensions than samples, each sample can be separated from all others along some dimension.
- Ever more data needed to prevent overfitting

How to avoid the problems?

Many solutions - we'll show some of them on the next lecture!

References:

Michel Verleysen and Damien Francois. **The Curse of Dimensionality in Data Mining and Time Series Prediction.** In *Proceedings of IWANN 2005*, Springer, 2005.
http://perso.uclouvain.be/michel.verleysen/papers/iwann05mv.pdf

Robert Clarke, Habtom W. Ressom, Antai Wang, Jianhua Xuan, Minetta C. Liu, Edmund A. Gehan, and Yue Wang. **The properties of high-dimensional data spaces: implications for exploring gene and protein expression data.** *Nature Reviews Cancer*, 8(1): 37–49, January 2008.
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2238676/pdf/nihms36333.pdf

See also
https://en.wikipedia.org/wiki/Curse_of_dimensionality
and references therein.