Lecture 1 – Introduction

Martin Spindler

2016-04-26

Defintions | Taxonomy of Data Sets

- Larger data become more and more available.
- n: number of observations; p: number of variables
- "Tall data": big n, small p
  computational demanding
- "High-dimensional data" or "wide data": $n << p$ or small n, big p
  non-standard theory, computational demanding
- "Big Data": big n, small / big p
- Important concept: MapReduce and its software implementation hadoop, in particular for tall data

Defintions | Input and Output Variables

- Inputs X: measured or present variables. Synonyms: predictors, features or independent variables
- These inputs have some influence on one or more outputs.
- Output variable Y is also called response or dependent variable or outcome variables.
- $$Y = f(X) + \varepsilon$$
- f unknown function, $X = (X_1, \ldots, X_p)$ p predictor variables, ε random error term

Defintions | Supervised vs Unsupervised Learning

- Supervised Learning: Presence of the outcome variable to guide the learning process
  Goal: e.g. to use the inputs to predict the values of the outputs
  Methods: regression methods (linear, lasso, ridge, etc.), bagging, trees, random forests, ensemble learning, ...
- Unsupervised Learning: only features are observed, no measurements of the outcome variable
  Goal: insights how the data are organized or clustered
  Methods: Association Rules, PCA, cluster analysis

Definitions | Regression vs Classification

- Input variables X
- Quantitative output Y: *regression*
- Qualitative output (categorical / discrete) G: *classification*
- Also input variables can also vary in measurement type.
- Coding of qualitative variables: $0/1$, $-1/+1$, or in general case via dummy variables.

Basic Concepts | Prediction vs. Inference

- **Prediction**: Given inputs X, but not the output Y, we want to predict Y:

$$\hat{Y} = \hat{f}(X)$$

  We are interested in high quality predictions and not in the function f which is more or less considered as a black box.

- **Inference**: Here the goal is understanding the relationship between Y and X and the form of f. Related questions are which predictors are associated with the response (model selection) and is the relationship linear or nonlinear.

Basic Concepts | Trade-off between Prediction Accuracy and Model Interpretability

Some methods are less flexible or more restrictive, meaning that the range of shapes of f they can estimate is restricted. Other methods are more flexible in this regard.
Usually there is a tension between prediction accuracy and interpretability. This means that flexible models often deliver good prediction accuracy and give models which are harder to interpret.
This will become clearer in Part I.

Basic Concepts | The Bias-Variance Trade-off I

- The mean squared error (MSE) is defined as

$$MSE = 1/n \sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2$$

- Calculating the MSE for the sample used for estimation of f (training set) might lead to **overfitting**.
- Hence, MSE for a new unseen sample (testing set) is preferable:

$$MSE = Ave(y_0 - \hat{f}(x_0))^2$$

with a new observation x_0

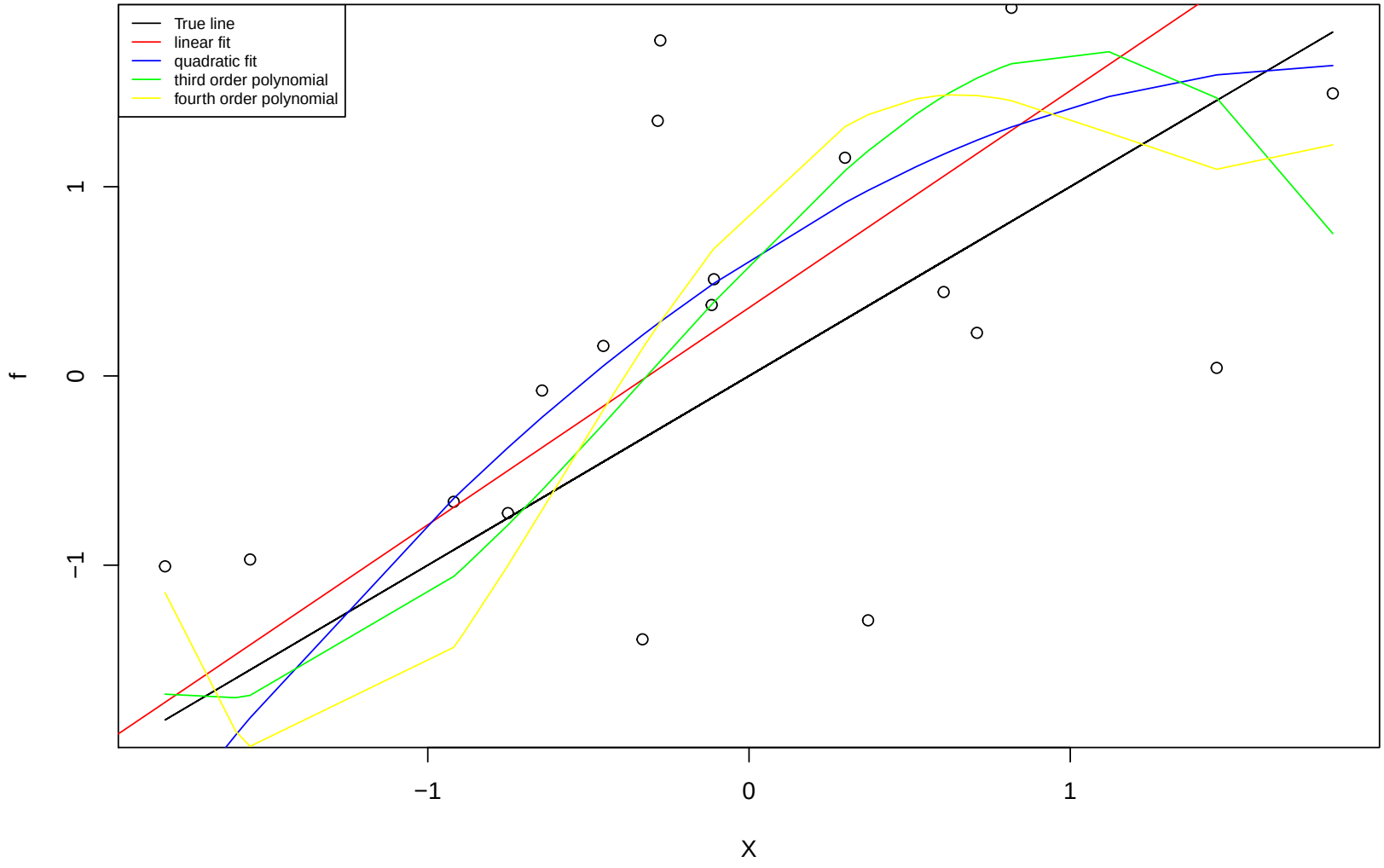
Basic Concepts | The Bias-Variance Trade-off II

- We have the following decomposition

$$\mathbb{E}(y_0 - \hat{f}(x_0))^2 = \mathbb{V}\mathfrak{D}\diagdown(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + \mathbb{V}\mathfrak{D}\diagdown(\varepsilon)$$

- Variance: amount by which $\hat{f}$ changes if estimated by using a different training data set
- Bias: error due to approximation the real relationship by a simpler model
- "Bias-Variance Trade-off"

Basic Concepts | The Bias-Variance Trade-off III (Illustration)



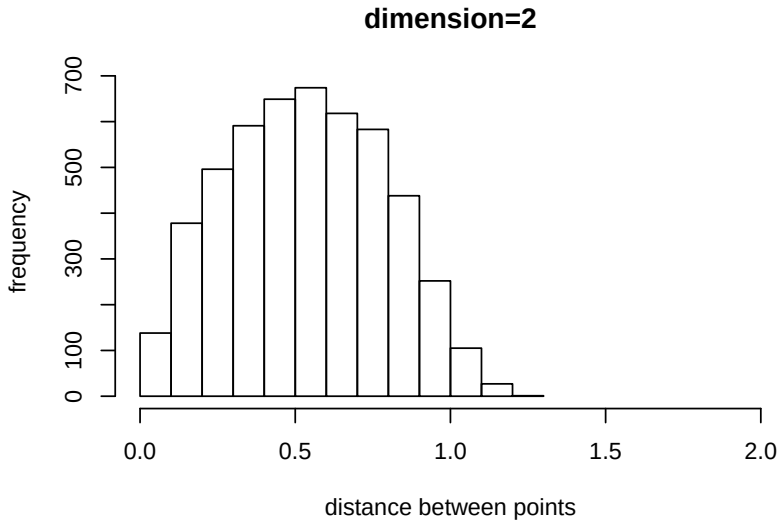
Problems / Challenges in High-Dimensions

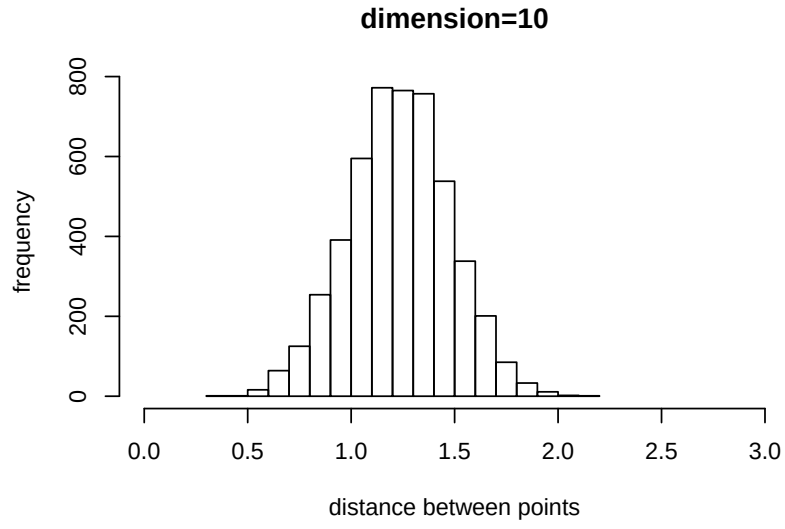
- Lost in the immensity of high-dimensional spaces
- Fluctuations cumulate.
- An accumulation of rare events may not be rare.
- Computational complexity

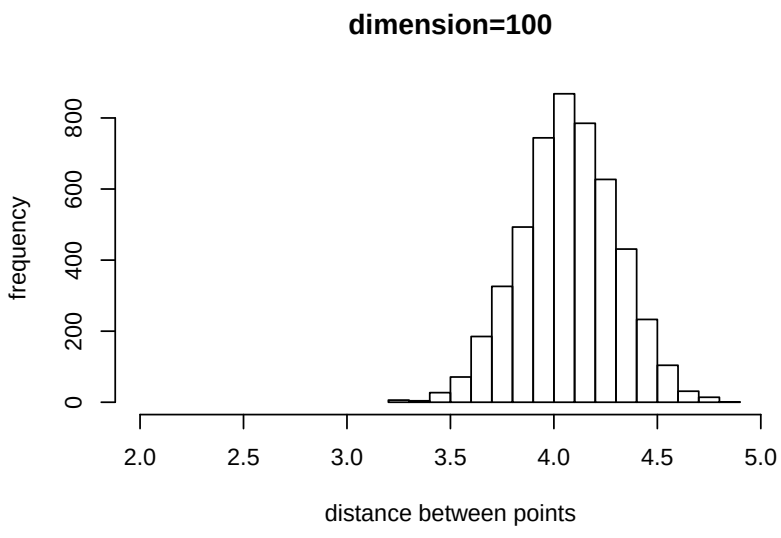
Immensity of High-Dimensional Spaces I

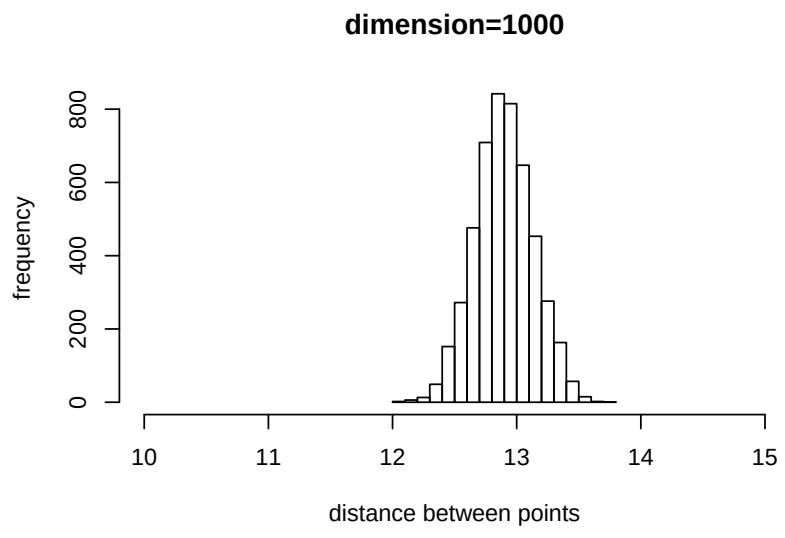
When the dimension p increases, the notion of "nearest points" vanishes. Below the histograms of the pairwise distances of $n = 100$ points randomly drawn (uniformly) from the unit cube are given.

Immensity of High-Dimensional Spaces II









Immensity of High-Dimensional Spaces III

How many points are needed in order to fill the hypercube $[0,1]^p$ in such a way that at any $x \in [0,1]^p$ there exists at least one point at distance less than 1 from x?

p	20	30	50	100	150	200
n	39	45630	$5.7 * 10^{12}$	$42 * 10^{39}$	$1.28 * 10^{72}$	Inf

Fluctuations accumulate.

In the linear regression model $Y = X\beta + \varepsilon$ for the OLS estimate $\hat{\beta} = (X^T X)^{-1} X^T Y$ we have

$$\mathbb{E}[||\hat{\beta} - \beta||] = \mathbb{E}[||((X^T X)^{-1} X^T \varepsilon)||^2] = Tr((X^T X)^{-1})\sigma^2.$$

In the case of orthogonal design:

$$\mathbb{E}[||\hat{\beta} - \beta||] = p\sigma^2$$

with $\mathbb{V}\mathfrak{D}\diagdown\varepsilon = \sigma^2$.
Hence the estimation error grows with the dimension p of the problem.

Fluctuations accumulate.

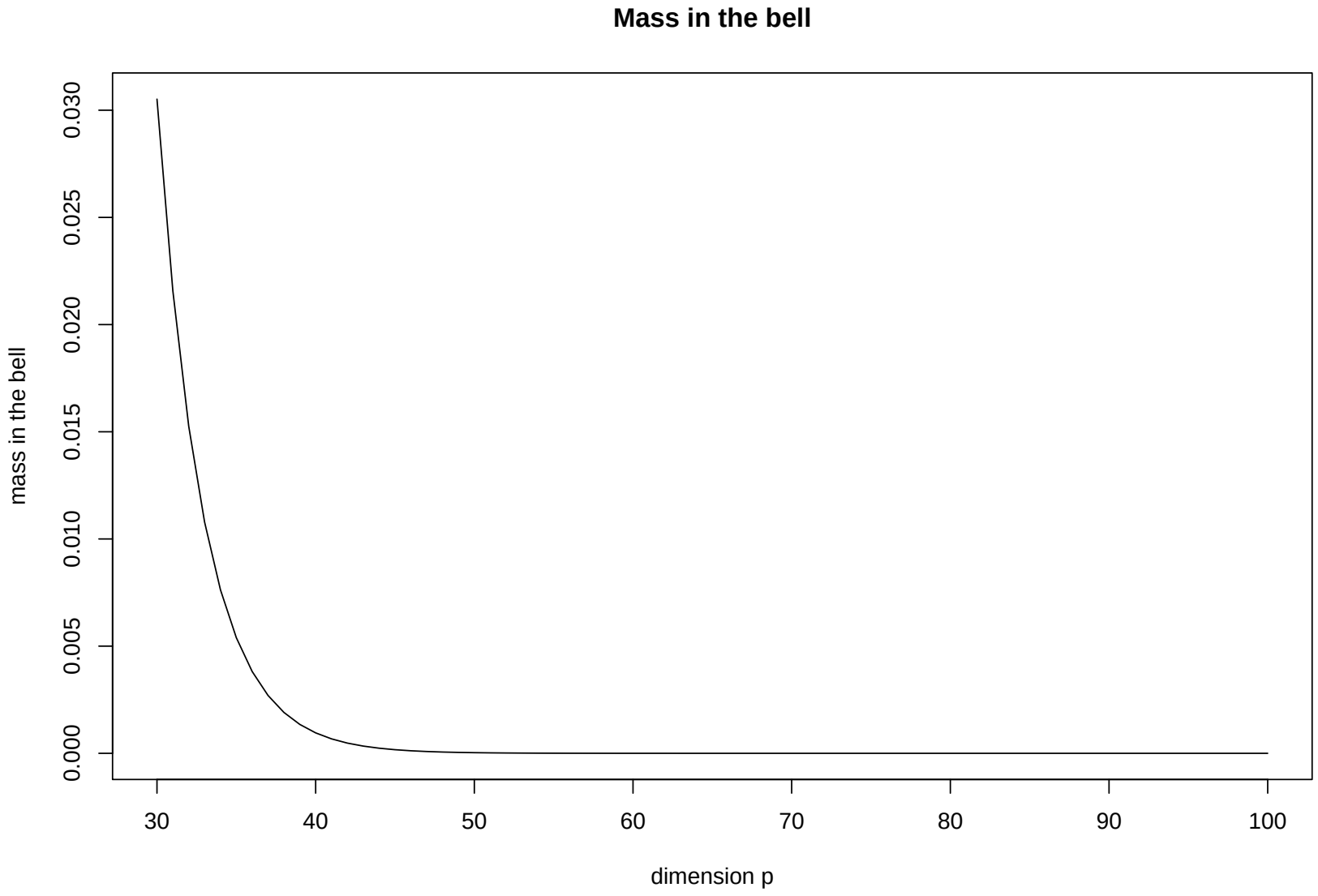
We consider a standard Gaussian distribution $\mathcal{N}(0, I_p)$ with density $f_p(x) = (2\pi)^{-p/2} \exp(-||x||^2/2)$. We are interested in the mass of the distribution in the "bell"

$$B_{p,\delta} = \{x \in \mathbb{R}^p : f_p(x) \geq \delta f_p(0)\} = \{x \in \mathbb{R}^p : ||x||^2 \leq 2\log(\delta^{-1})\}.$$

The Markov Inequality gives us:

$$\mathbb{P}(X \in B_{p,\delta}) = \mathbb{P}(e^{-||X||^2/2} \geq \delta) \leq 1/\delta \mathbb{E}[e^{-||X||^2/2}] = \frac{1}{\delta 2^{p/2}}$$

Fluctuations accumulate.



Accumulation of Rare Events

Suppose an error ε is Gaussian distributed with $\mathcal{N}(0,1)$. Then with probability at least $1-\alpha$, the noise ε has an absolute value smaller than $(2\log(1/\alpha))^{1/2}$. This follows from the inequality $\mathbb{P}(|\varepsilon| \geq x) \leq \exp(-x^2/2)$.
When we observe p noise variables $\varepsilon_1, \ldots, \varepsilon_p$ which are i.i.d. and standard normal, we have

$$\mathbb{P}(\max_{j=1,\ldots,p} |\varepsilon_j| \geq x) = 1 - (1 - \mathbb{P}(|\varepsilon_1| \geq x))^p \approx p\mathbb{P}(|\varepsilon_1| \geq x).$$

This means that if we want to bound the max of the absolute values with probability $1-\alpha$, then we can only guarantee that the maximum is smaller than $(2\log(p/\alpha))^{1/2}$.

Computational Complexity

With increasing dimension, numerical computations can become very demand and exceed the available computing resources.
Example: When we have p potential regressors, than the number of submodels is 2^p which grows exponentially with the number of regressors.