Minimum Description Length Principle

With An Application For Credit Risk Models

Paul van Leeuwen (paul.vanleeuwen@devolksbank.nl)

16 May 2023

Introduction

What is MDL?

Approach

Introduction

Why this subject?

- Improve current approaches and modelling techniques.
- Stay in the forefront of statistical innovation.
- Connection with other realms of statistics.

What is MDL?

Introduction

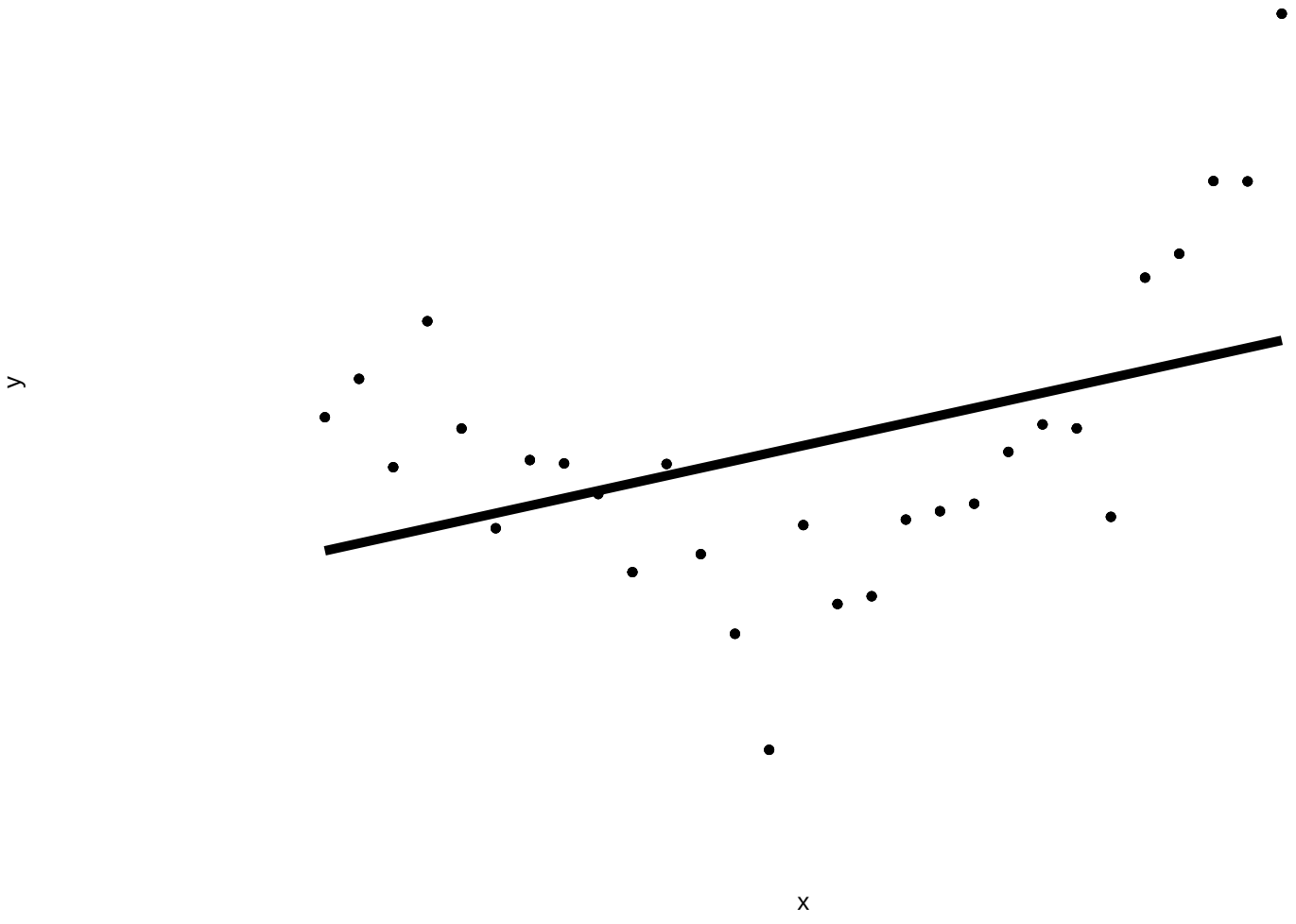
- The Minimum Description Length (MDL) principle aims to describe the data and its description mechanism with the smallest possible information 'length'.
- Applications (among else):
  - model selection (e.g. what order of the Markov model family do we want?);
  - deal with overfitting (e.g. how many explanatory variables to include);
  - exploratory data analysis (what prior knowledge can we confirm?).
- Close ties with frequentist statistics, Bayesian statistics, and machine learning.
- Why is MDL relatively unknown?
  - MDL is the intersection of advanced measure theory, information theory, and statistics.
  - For a decent introduction into MDL, see (Grünwald 2007).

Example of dealing with overfitting

What polynomial generated by this dataset?

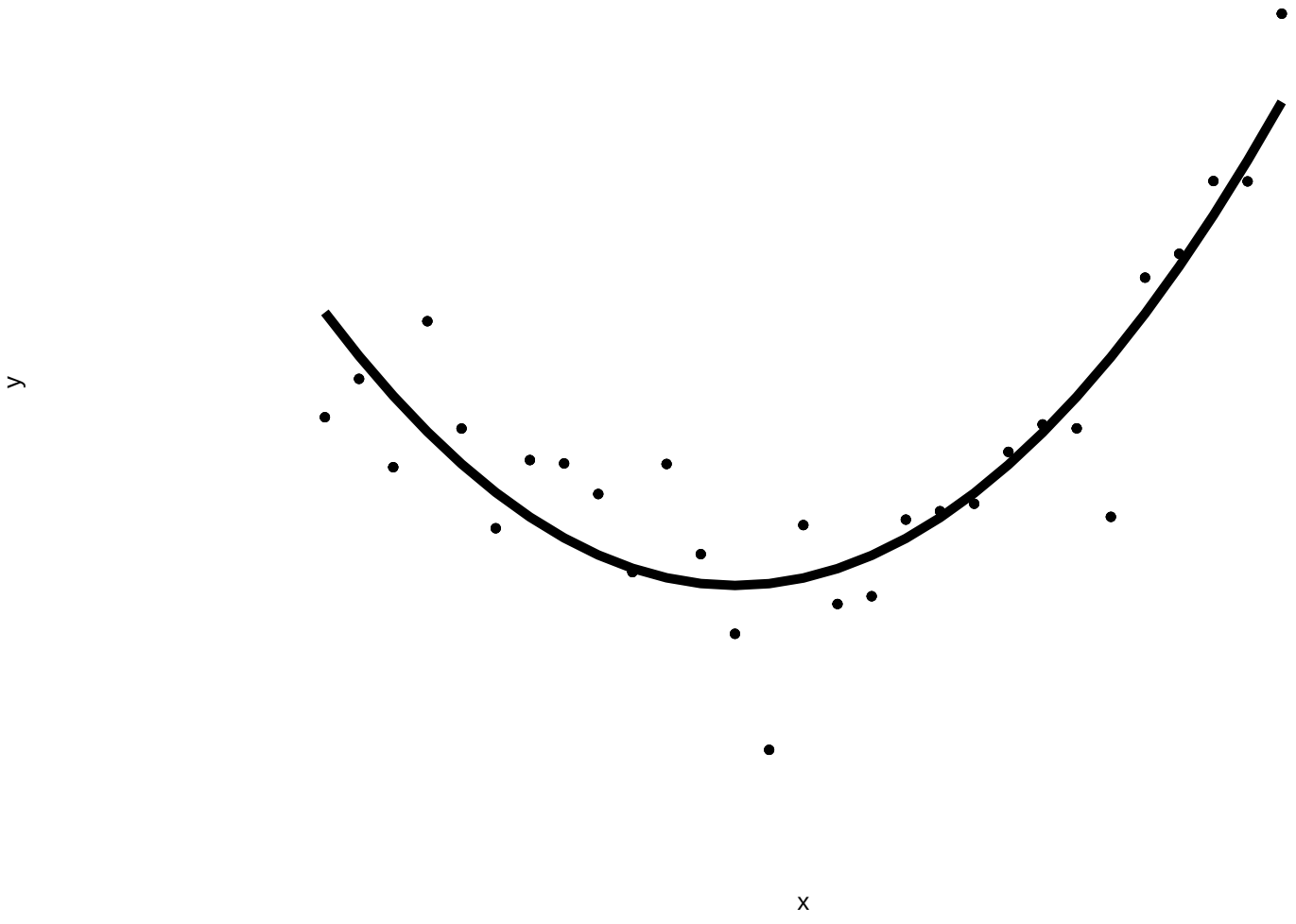
Example of dealing with overfitting

linear fit $\hat{y} = \beta_0 + \beta_1 x$



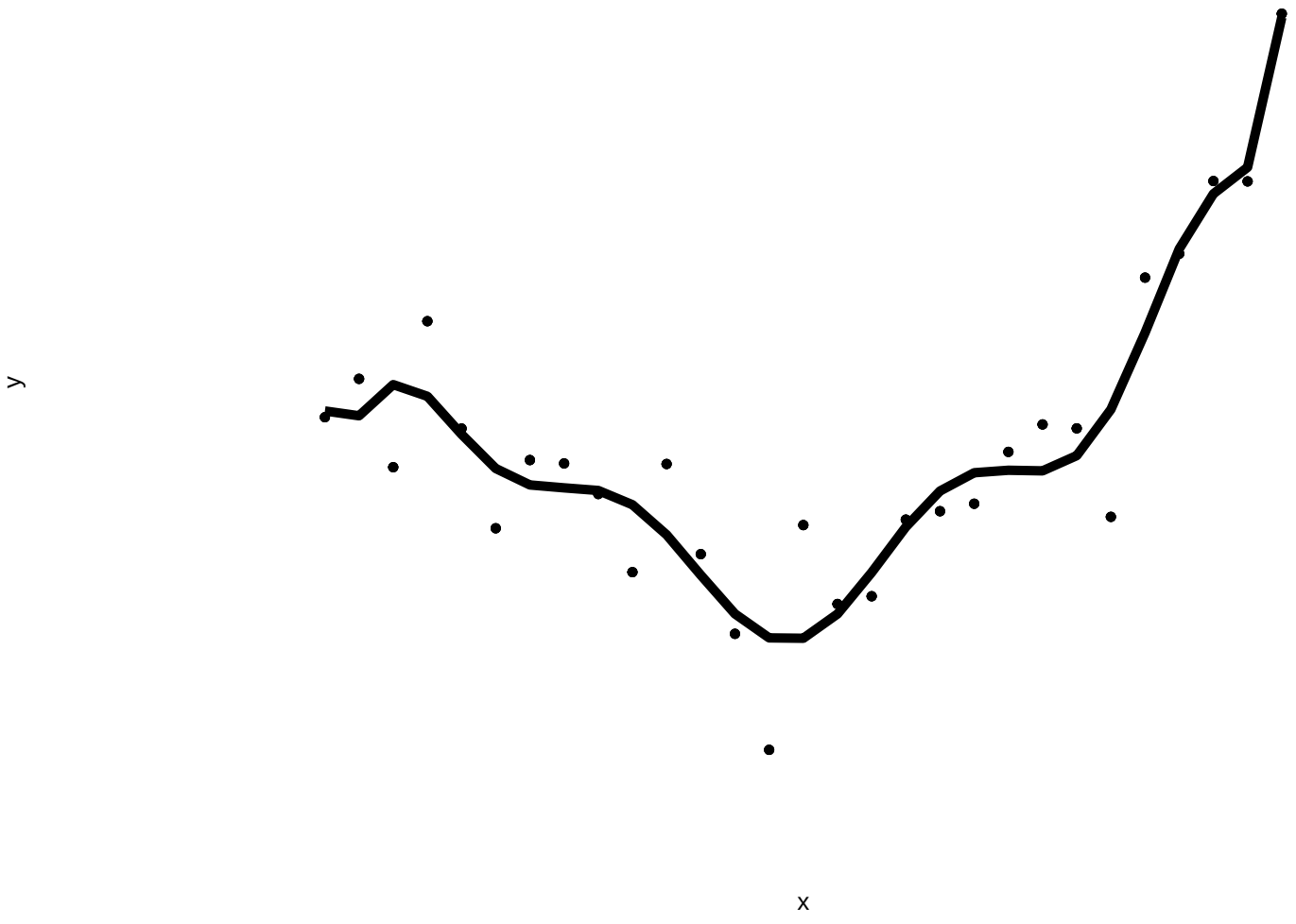
Example of dealing with overfitting

quadratric fit $\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2$



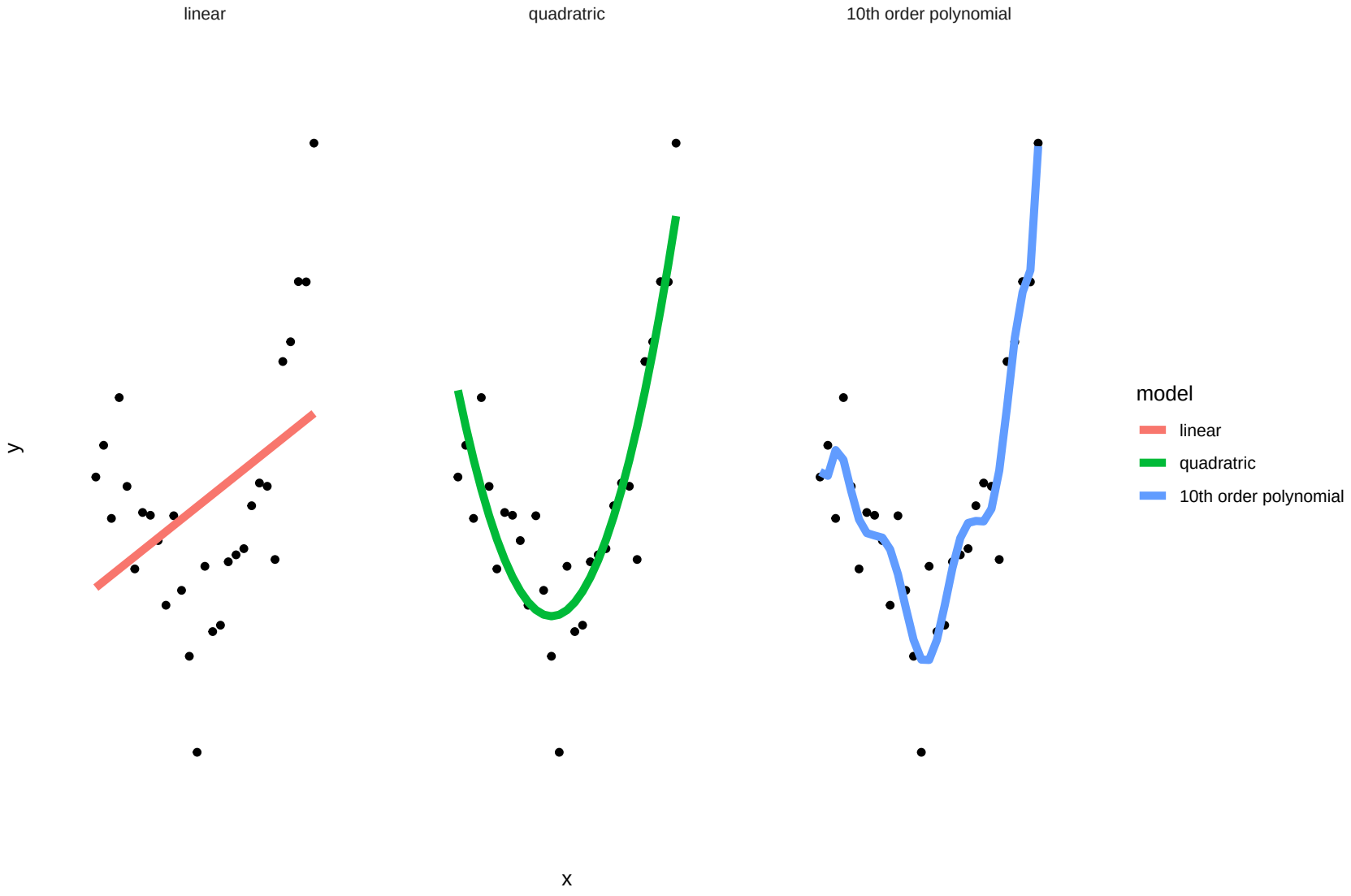
Example of dealing with overfitting

10th order polynomial fit $\hat{y} = \beta_0 + \beta_1 x + \ldots + \beta_{10} x^{10}$



Example of dealing with overfitting

data generated by $x^3 + 2x^2 + 5 + \varepsilon$, $\varepsilon \sim N(0, 1)$



How does MDL work?

- Patterns or regularities in the data can be described with less information 'length' compared to the data alone.
- less information 'length' = compression
- Choose the model that gives the shortest description of the data.
- Note that MDL is an approach, not an algorithm.
  - The modeller has to make choices to implement the MDL principle.

Approach

From Data to Model Selection

- From data to code.
- From code to code length.
- From code length to probabilities.
- From probabilities to model selection.

From Data to Code

- Examples:
  - `Hello, world` could map to 0.
  - `aabbaccdaaadd` could map to 110100.
- In general, a description method maps a sequence of symbols to a binary sequence.
  - The coding alphabet $\mathbb{B}$ can be binary ($\mathbb{B} = \{0, 1\}$), the Western alphabet ($\mathbb{B} = \{\mathrm{a}, \mathrm{b}, \ldots, \mathrm{z}\}$), etc.
- Mathematically: a dataset $D = (x_1, \ldots, x_n)$ with $x_i \in \mathbb{B}$ from a sample space $\mathcal{X}^n$ is mapped to $\{0, 1\}^m$ by $C \colon \mathcal{X}^n \mapsto \{0, 1\}^m$.
  - In the first examples above, we could have $\mathcal{X}^n = \{x_1\} = \{'\mathrm{Hello, world}'\}$ and $C(x_1) = 0$.
- We demand the mapping C to be *uniquely* decodable.
  - No multiple interpretations allowed.

From Data to Code

- Suppose we would like to encode a binary data sequence of length 2.
- We are not sure what outcome we observe.
- Let $X_i \in \{0, 1\}$ be the random variable at position $i = 1, 2$.
  - The complete data sequence becomes $X_1 X_2 \in \{00, 10, 01, 11\}$.
- Every sequence in $\{00, 10, 01, 11\}$ is assigned a code.
- In general, for the binary alphabet, for n positions there are 2^n possible data sequences.
  - For example, when $n = 3$ we have the data sequences $\{000, 100, 010, \ldots, 111\}$.
  - Without loss of generality, every non-binary alphabet can be mapped to the binary alphabet $\{0, 1\}$.

From Code to Code Length

- Given a data sequence x_i and its corresponding code $C(x_i)$, then we are interested in the code length $L(x_i)$ with $L\colon \mathcal{X}^n \mapsto \mathbb{R}_+$.
- For example, to map an integer from $\{1, \ldots, n\}$ in a uniform way, we need $\lceil \log_2(n) \rceil$ bits.
  - Note that n has to be known in advance.
  - For example, take $n = 64$. Then we have 64 binary data sequences of length $\lceil \log_2(64) \rceil = 6$.
  - Or take $n = 10$. Then we need $\lceil \log_2(10) \rceil = 4$ bits.
    - Note that $2^4 = 16$ data sequences are possible while we only use 10 of them.
- What coding scheme results in the smallest number of *expected* bits?

From Code to Code Length

- Recall the data sequences $\{00, 10, 01, 11\}$.
- Uniform method: the data sequence is the code.
  - $C(00) = 00$, $C(10) = 10$, $C(01) = 01$, and $C(11) = 11$.
  - Expected number of bits: 2.
- In general we can do better!
  - That is, $\mathbb{P}[X_i = 1] \neq \frac{1}{2}$ for at least one $i \in \{1, \ldots, n\}$.

From Code Length to Probabilities

- Suppose $\mathbb{P}[X_i = 1] = \frac{1}{4}$.
- We reserve code 0 for $X_1X_2 = 00$ so $C(00) = 0$, $C(10) = 100$, $C(01) = 110$, and $C(11) = 1110$.
- Then the expected number of bits is

$$1 \cdot \frac{3}{4}^2 + 3 \cdot \frac{3}{4} \cdot \frac{1}{4} + 3 \cdot \frac{1}{4} \cdot \frac{3}{4} + 4 \cdot \frac{1}{4}^2 = 1.9375 < 2$$

- This the Shannon-Fano coding scheme.
  - More general, reserve $\lceil -\log_2(p_i) \rceil$ bits for probability p_i corresponding to data sequence i.
  - The Huffman code is optimal and improves on the Shannon-Fano code.
- Main message:

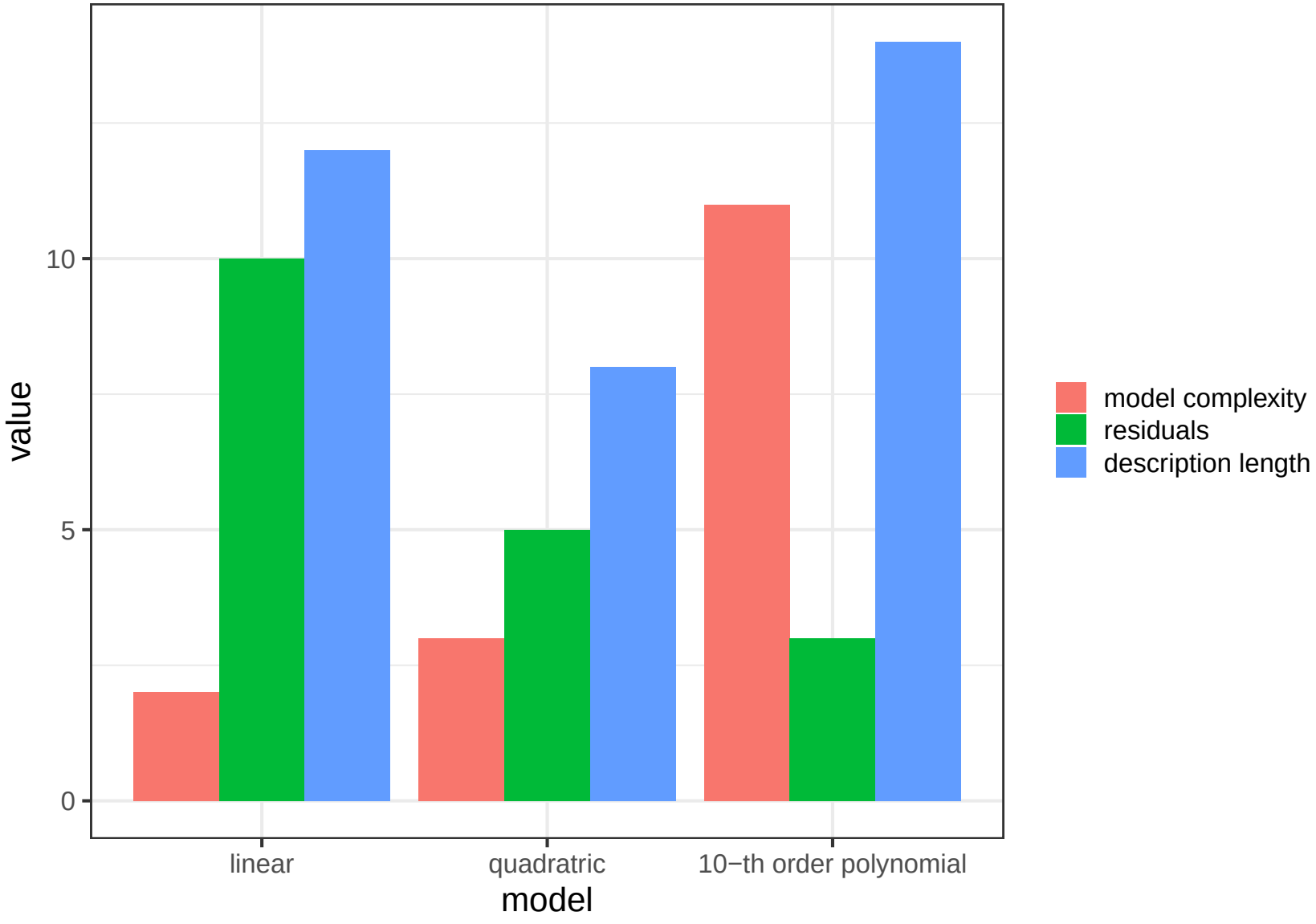
higher probability = smaller code length = less bits required

From Probabilities to Model Selection

- To describe any dataset we need $L(D, H)$ bits.
  - Both the description method H and the data D require bits.
- The MDL principle employed for model selection is to minimise the sum of
  - the number of bits to encode the description mechanism $L(H)$ and
  - the number of bits to encode, with the description mechanism H, the data observed $L(D|H)$.
- Information 'length' is this sum $L(H) + L(D|H)$.
  - $L(H)$ is the *model complexity*, $L(D|H)$ is the *fit of the data*.
- The MDL principle is to choose the model as to minimise this sum.
- Main message:

smaller description length = better model selection

Example of dealing with overfitting (continued)



Bibliography

Grünwald, Peter D. 2007. *The Minimum Description Length Principle*. MIT press.