Overfitting, prediction error and trade-offs

Filippo Biscarini (CNR, Milan, Italy)

filippo.biscarini@cnr.it

Overfitting

What is overfitting?

We fitted a linear model on our dataset and made predictions; we then measured the “accuracy”of these predictions: **did we do it right?**

What is overfitting?

We fitted a linear model on our dataset and made predictions; we then measured the “accuracy”of these predictions: **did we do it right?**

- short answer: **NO!**
- main reason: **overfitting**

What is overfitting?

Overfitting:

Fitting too well the data: R^2 too large (≃1)

What is overfitting?

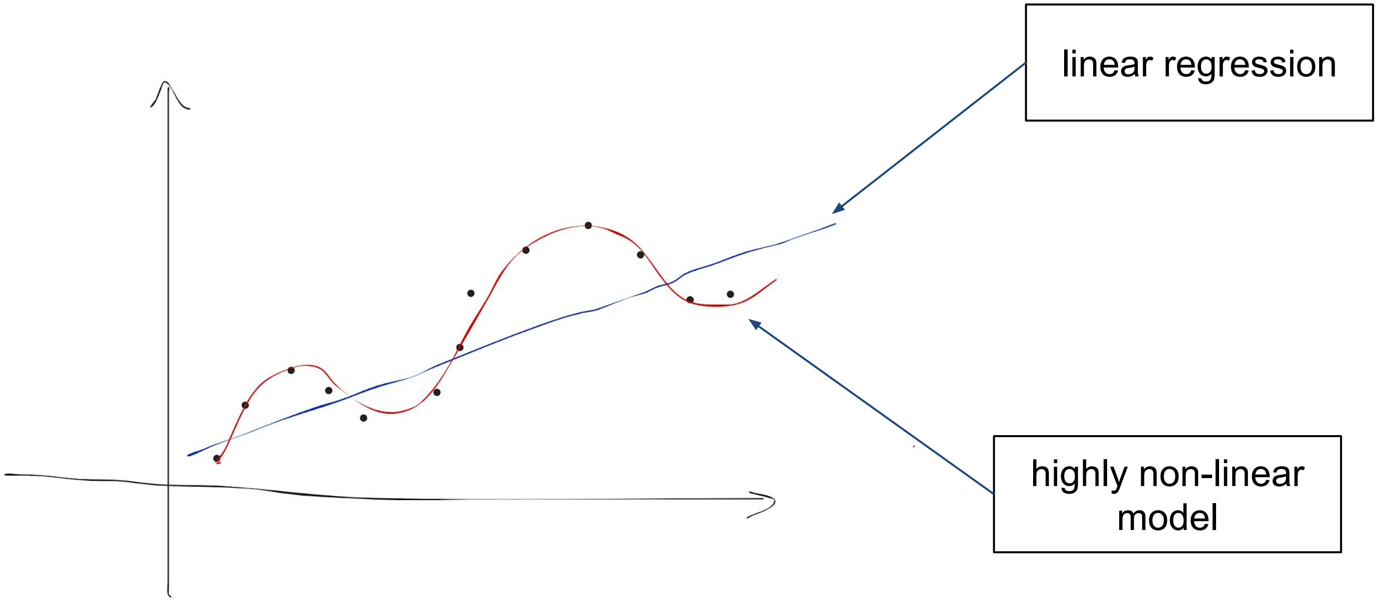
<u>Overfitting</u>:

Fitting too well the data: R^2 too large (≃1)

overfitting happens with:

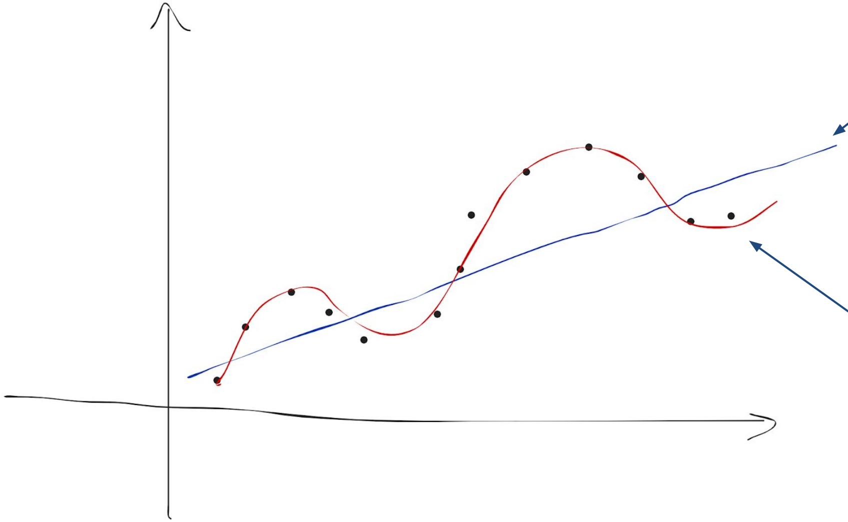
- using the same data to fit the model and make predictions
- overparameterization of the model (e.g. too many effects)
- flexible methods (e.g. polynomial functions, splines, classification trees etc.)

What is overfitting?



What is overfitting?

Think of KNN
with k=1!



linear regression

highly non-linear
model

What is overfitting?

- we want to predict whether a new patient will die or survive from a disease.
- we have 1000 historical patients with that disease and observe that 900 survived and 100 died: probability of survival = 90%

simple, but biased (men/women, old/young, overweight, smokers etc.)

What is overfitting?

- we want to predict whether a new patient will die or survive from a disease.
- we have 1000 historical patients with that disease and observe that 900 survived and 100 died: probability of survival = 90%

- we refine our estimate to old (> 70 years) men: 300 patients, 100 died and 200 survived →prob. surv. = 66.6% (**less biased**)
- we can go on and make a less and less biased comparison: old men that are overweight, who smoke, don’t exercise, don’t drink alcohol, with high socio-economic status, independent professionals, who live in a residential neighbourhood, with one known comorbidity etc.
- we are left with only three historical patients who all died: shall we say that the probability of survival for our new patient is 0%?

What is overfitting?

- we refine our estimate to old (> 70 years) men: 300 patients, 100 died and 200 survived →prob. surv. = 66.6% (**less biased**)
- we can go on and make a less and less biased comparison: old men that are overweight, who smoke, don't exercise, don't drink alcohol, with high socio-economic status, independent professionals, who live in a residential neighbourhood, with one known comorbidity etc.
- we are left with only three historical patients who all died: shall we say that the probability of survival for our new patient is 0%?
- **the <u>bias was reduced</u>, but <u>the variability was increased dramatically</u> - the estimate based on only 3 observations has large variance →** $\frac{\sigma^2}{\sqrt{N}}$

Prediction error

Prediction error

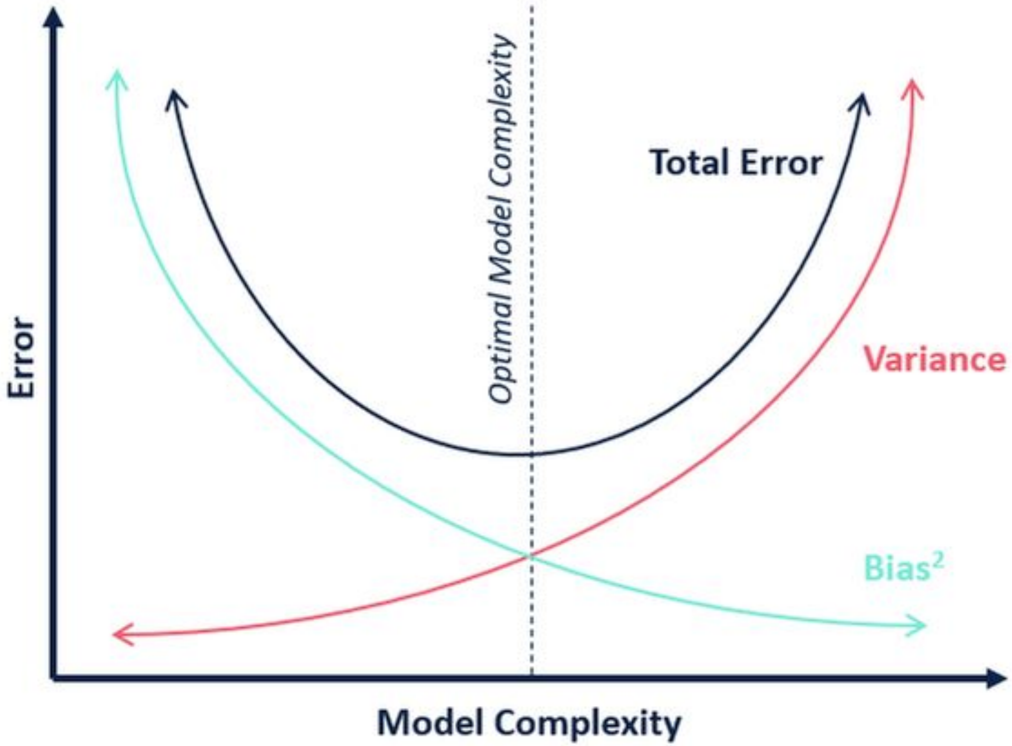
$$E\left(y - \hat{f}(x)\right)^2 = \underbrace{Var\left(\hat{f}(x)\right)}_{\text{variance}} + \underbrace{\left[\mathrm{Bias}(\hat{f}(x))\right]^2}_{\text{bias}^2} + Var(\epsilon)$$

Prediction error

$$E\left(y - \hat{f}(x)\right)^2 = Var\left(\hat{f}(x)\right) + \left[\text{Bias}(\hat{f}(x))\right]^2 + Var(\epsilon)$$

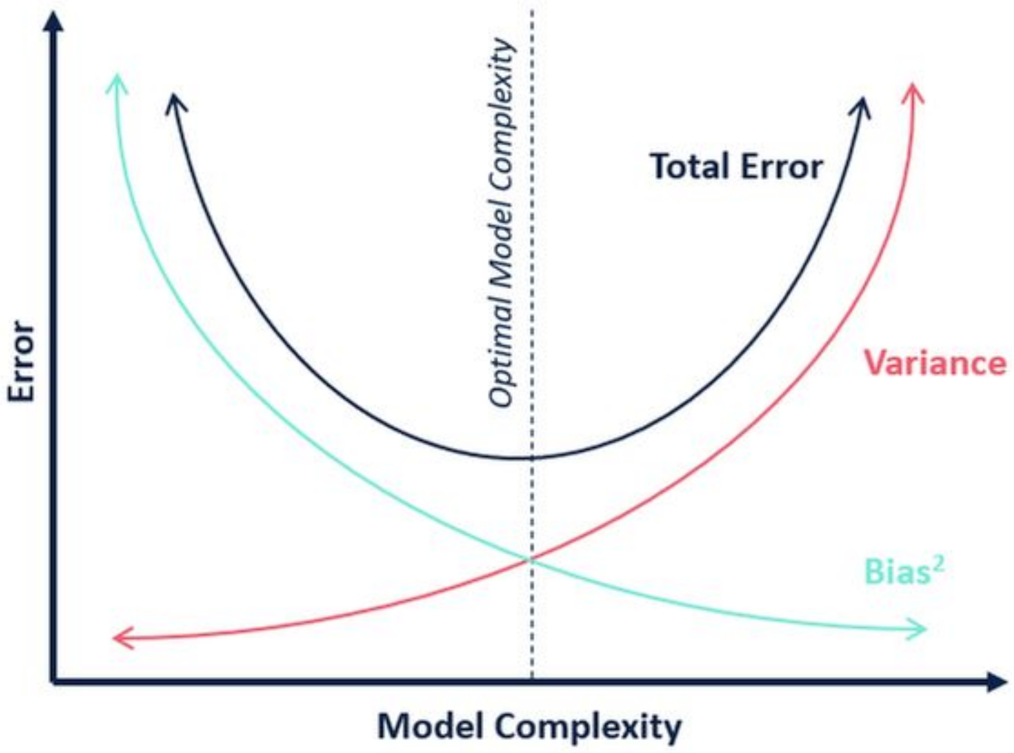
- **variance** refers to the change of the predictor if estimated using different training data
- **bias** refers to the approximation of a real problem by a simpler model

Bias-variance trade-off



Source: https://ai-pool.com/a/s/bias-variance-tradeoff-in-machine-learning

Bias-variance trade-off



- models/methods with low bias and high variance (e.g. KNN with k=1)
- models/methods with high bias and low variance (e.g. horizontal line crossing the data)
- → find models/methods with both low variance and low bias

Source: https://ai-pool.com/a/s/bias-variance-tradeoff-in-machine-learning

Bias-variance trade-off

Related trade-offs

1. Prediction accuracy vs model interpretability:
   - e.g. linear regression is easy to interpret, splines are not
2. Parsimony vs black-box:
   - e.g. variable selection, all-variable models (e.g. RF), Occam’s razor

Bias-variance trade-off

Important for:

1. Correctly estimating the performance of a predictive machine
2. Correctly estimating model parameters
3. Selecting between models

Bias-variance trade-off

Important for:

1. Correctly estimating the performance of a predictive machine
2. Correctly estimating model parameters
3. Selecting between models

So, how do we control for overfitting and the bias-variance trade-off?

Training and test sets

Training and testing sets

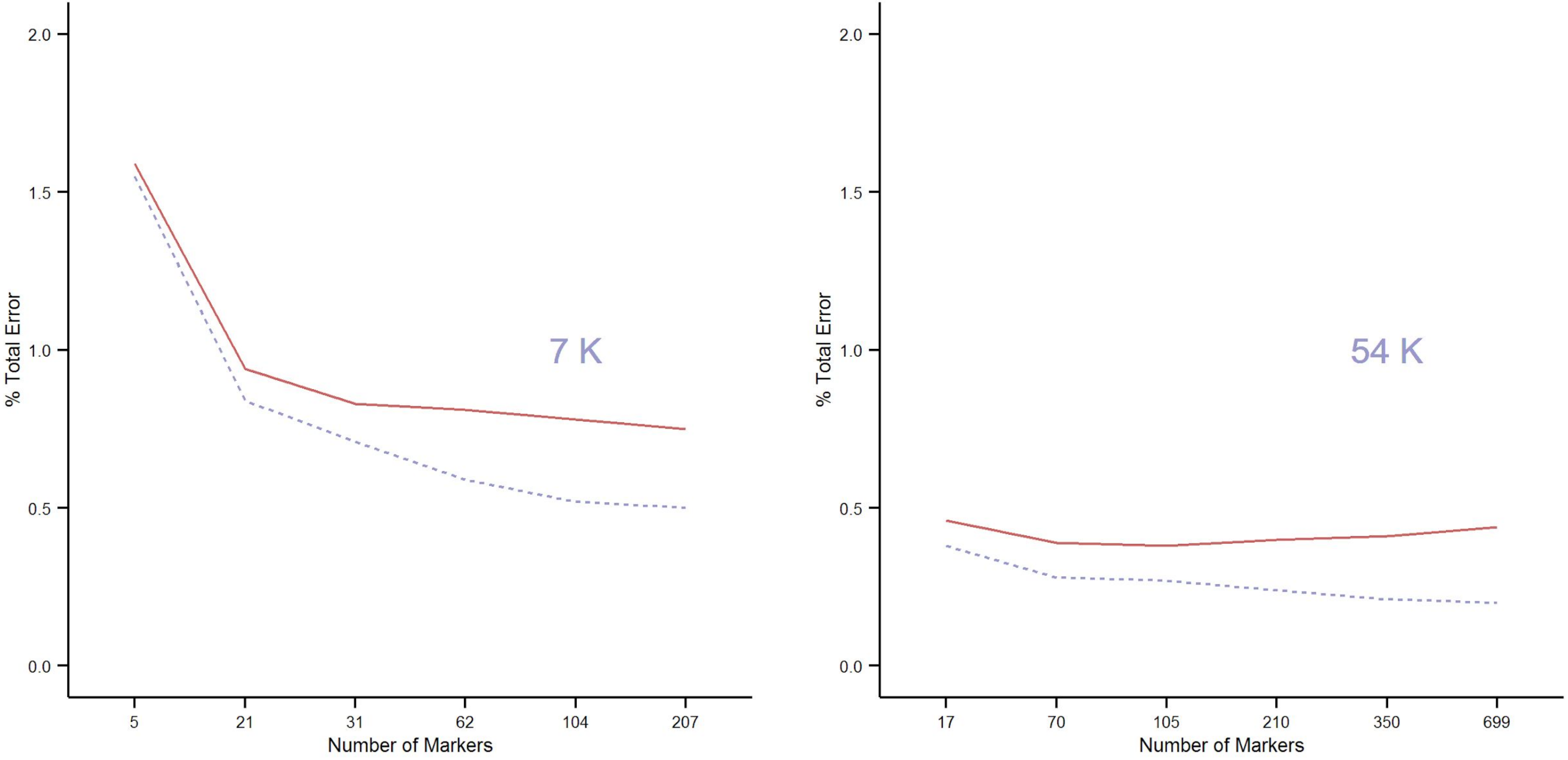
the predictive model is **trained here**

the predictive model is **evaluated here**

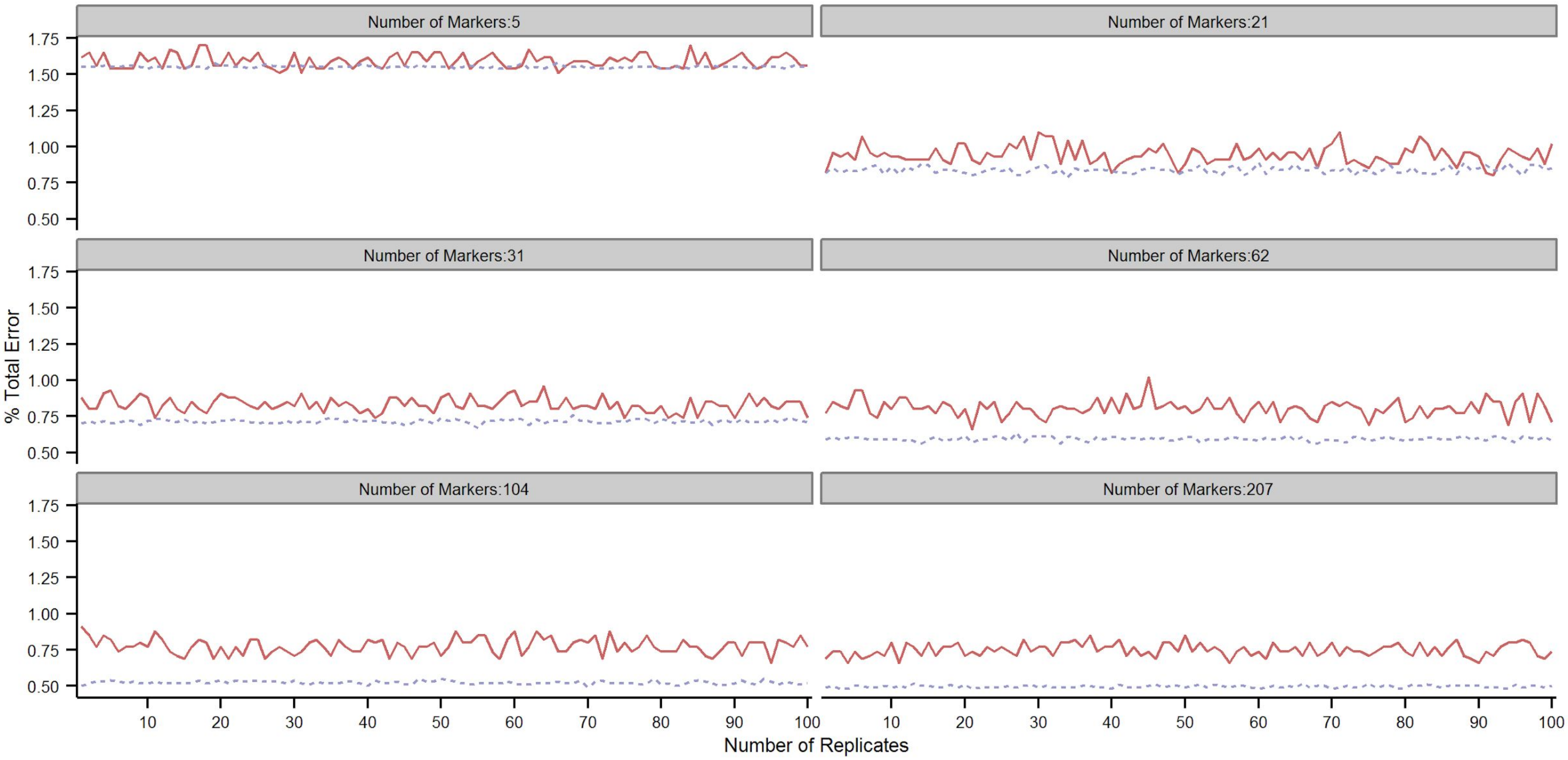
Training and testing sets

- accuracy (model performance) on the training set is “optimistic” (biased upward ← *overfitting*)
- a better estimate of model performance can be obtained from independent test data
- usually we are interested in the predictive performance on new data
- accuracy in the test set is usually lower than in the training set

Training and testing sets



Training and testing sets



Overfitting - hands on!

→ 3.training_testing.Rmd
Exercise 3.1