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?

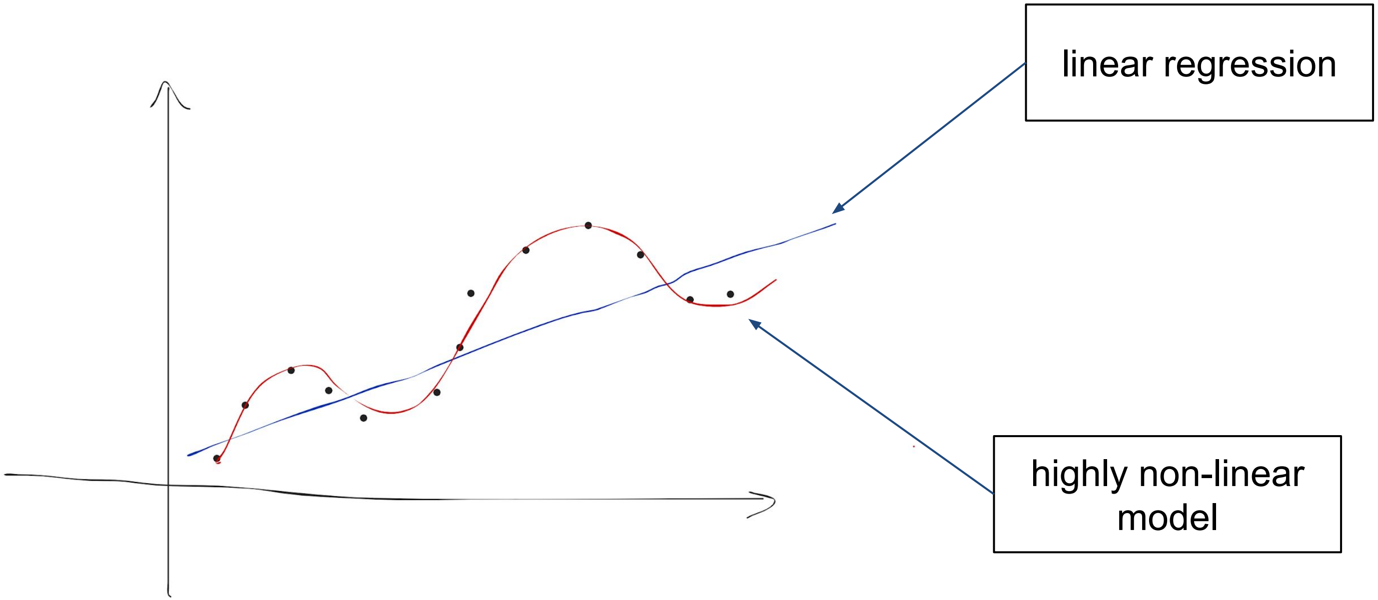
Overfitting:

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

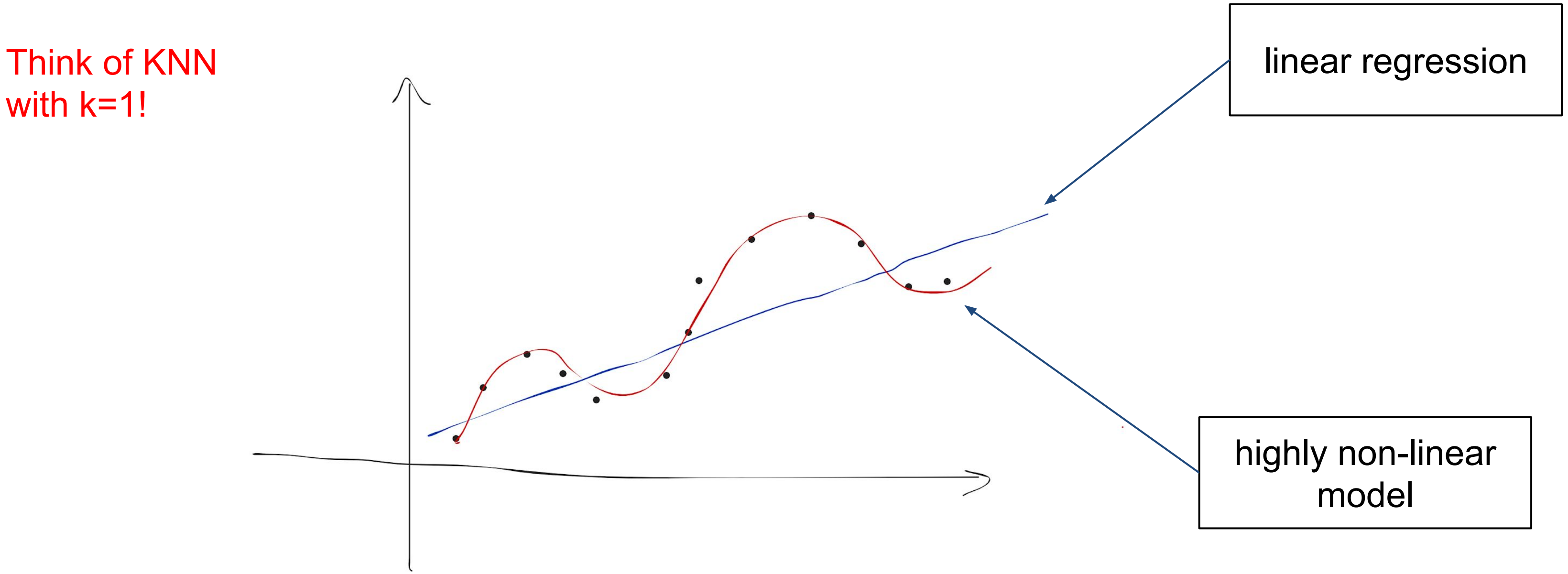
overfitting happens with:

- 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?



Prediction error

Prediction error

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

variance

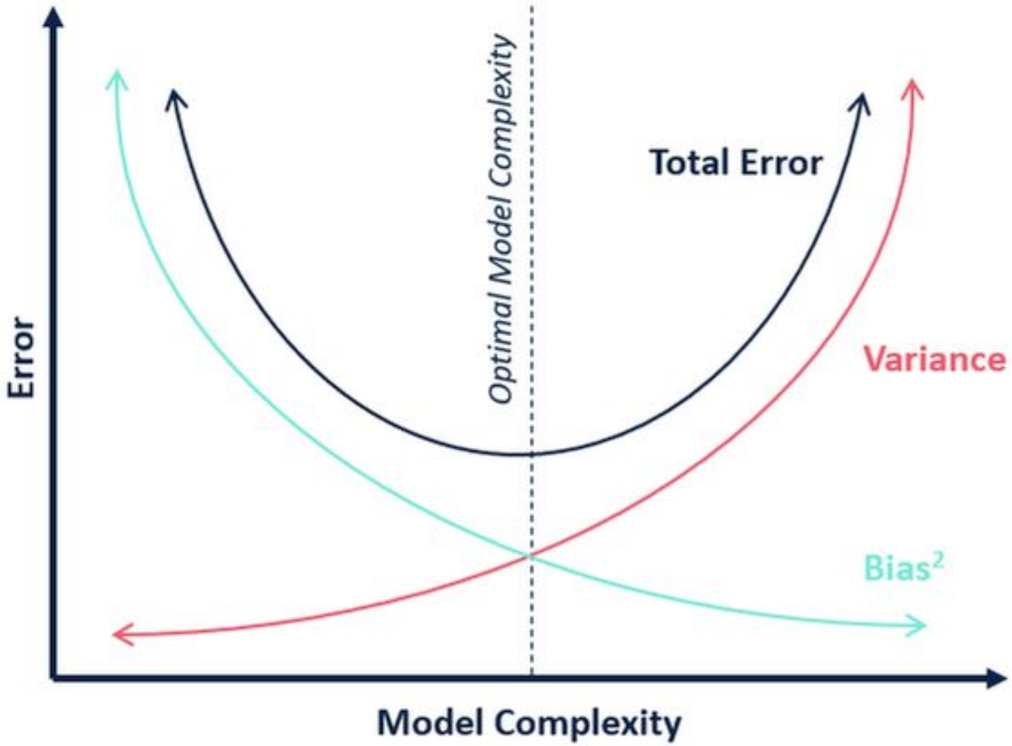
bias^2

Prediction error

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

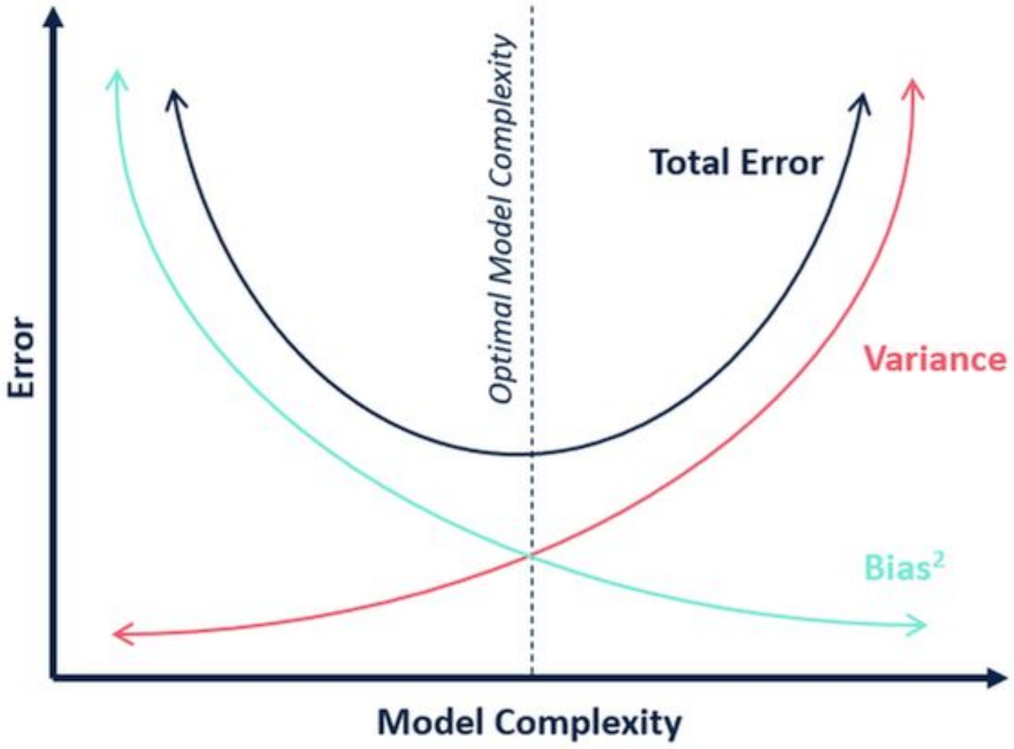
- **variance** refers to the change of the predictor if estimated using different training data
- **bias** refers to the approximation of a real problem by a simpler model

Bias-variance trade-off



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

Bias-variance trade-off



- models with low bias and high variance (e.g. KNN with k=1)
- models 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

<u>Related trade-offs</u>

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

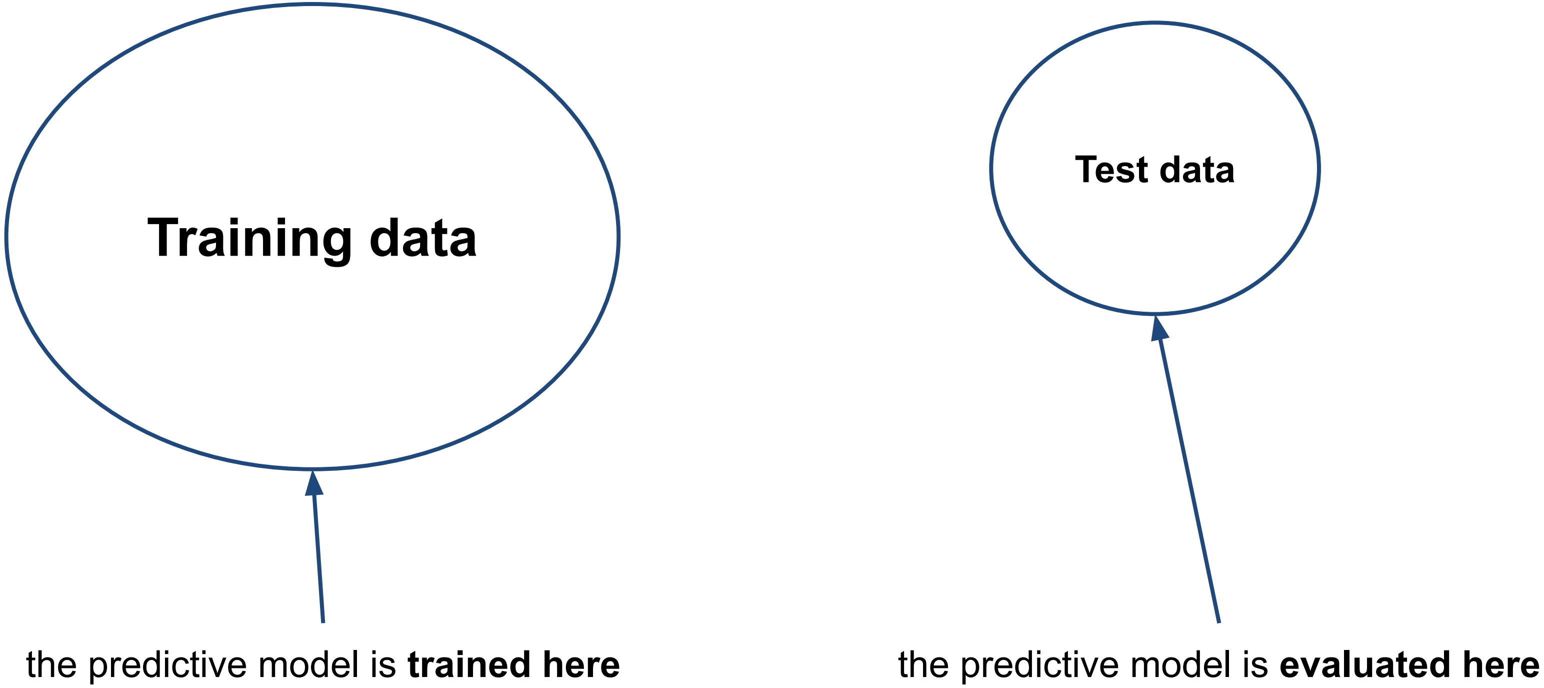
Important for:

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

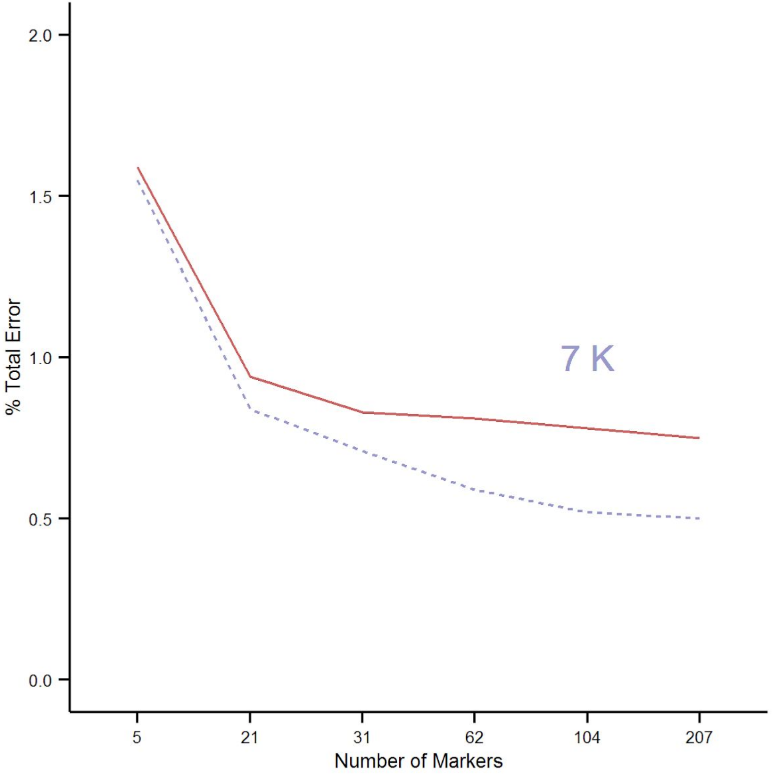
Training and testing sets

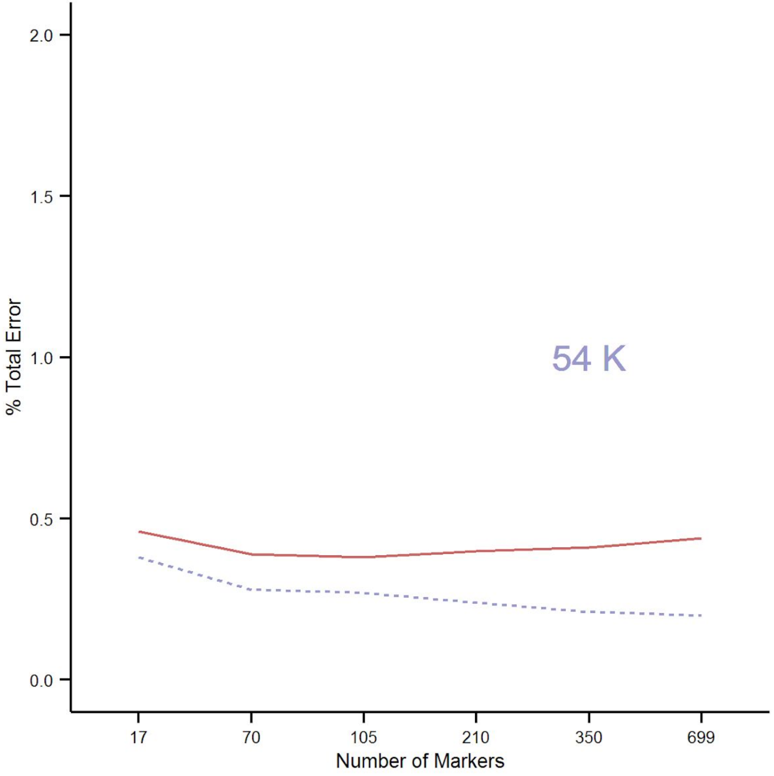


Training and testing sets

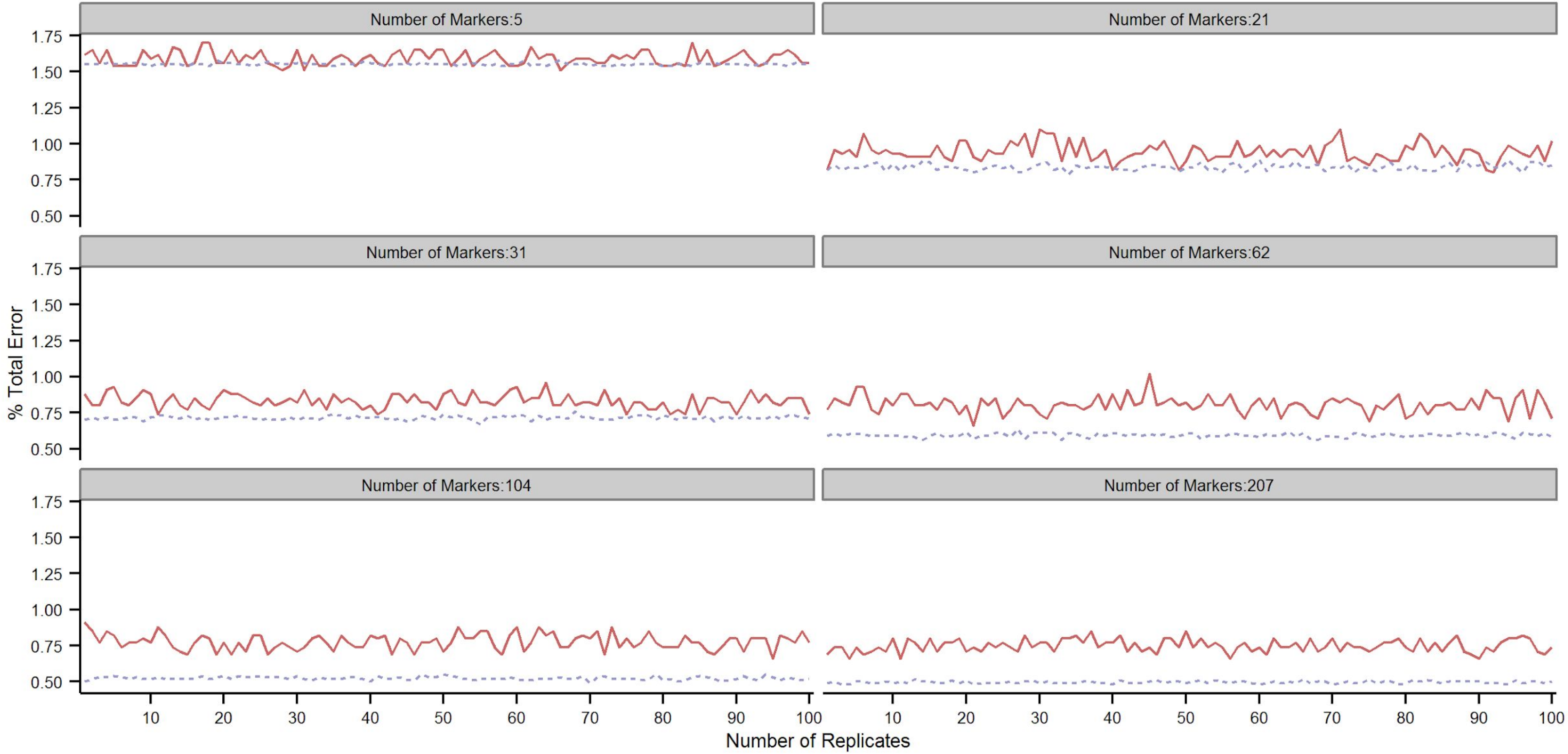
- accuracy (model performance) on the training set is “optimistic” (biased upward ← *overfitting*)
- a better estimate of model performance can be obtained from independent test data
- usually we are interested in the predictive performance on new data
- accuracy in the test set is usually lower than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

Training and testing sets





Training and testing sets



Overfitting - hands on!

→ 3.training_testing.ipynb
Exercise 3.1