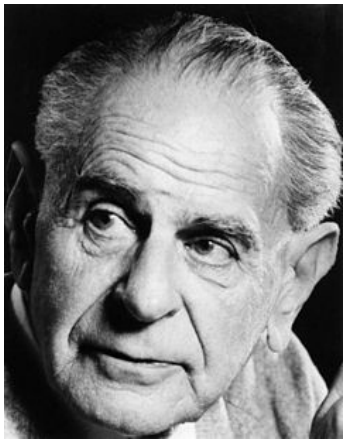
Statistical Methods for Bioinformatics

II-1: Bias and Variance trade-off, Cross-validation & Bootstrap

Rob Jelier

Statistics and the Philosophy of Science

"In so far as a scientific statement speaks about reality, it must be falsifiable; and in so far as it is not falsifiable, it does not speak about reality." Karl Popper



Popper believed scientific theories can be tested only indirectly, by evaluating their implications.

"The value for which P=0.05, is 1.96σ or nearly 2σ; it is convenient to take this point as a limit in judging whether a deviation ought to be considered significant or not."

RA Fischer

Neyman & Pearson proposed to use the p-value to formalize a decision making process. After your investigation, you either reject the null hypothesis and accept an alternative hypothesis, or vice versa.

The Higgs Boson discovery at 5σ

"The test statistic we use for looking at p-values is basically the likelihood ratio for the two hypotheses (H_0= Standard Model (S. M.) of Particle Physics, but no Higgs; H_1 = S.M with Higgs). A small p_0 (and a reasonable p_1) then implies that H_1 is a better description of the data than H_0. This of course does not prove that H_1 is correct, but maybe Nature corresponds to some H_2, which is more like H_1 than it is like H_0. Indeed in principle data will never prove a theory is true, but the more experimental tests it survives, the happier we are to use it – e.g. Newtonian mechanics was fine for centuries till the arrival of Relativity." Louis Lyons, CERN

The Role of Statistics in Science

- Statistics is
  - ... a formal way to deal with uncertainty in data
  - finding generalizable patterns in observations
  - collecting, visualizing, analyzing, finding and then testing hypotheses.
- Statistical tests are used to decide if a statement/hypothesis is supported by data
  - important paradigm in scientific communication
- Control of data quality
- Use statistical reasoning to optimally design experiments
- Statistical (also Machine) Learning approaches help predict the future

Content

Lecture 1: Linear regression and correlation
Lecture 2: Generalized linear models: Logistic Regression. Model building and model selection using Akaike Information Criterion (AIC)
Lecture 3: Multilevel Models: Longitudinal data
Lecture 4: Multilevel Models: Cluster data
Lecture 5: Missing data

Statistical Methods for Bioinformatics: Part II

23-3	Lecture 1: Bias and Variance trade-off, Cross-validation & Bootstrap
30-3&20-4	Lectures 2&3: High Dimensionality; Ridge, LASSO, PCR and Partial Least Squares
27-4&4-5	Lectures 4&5: Beyond Linearity; regression splines, smoothing splines, LOESS, Generalized Additive Models; Assignment handed out
11-5	Lecture 6: Trees, Bagging and Boosting
(18-5)	Remaining material, Revision and Exam preparation; Assignment due

Reading material

- Required books:
  - An introduction to statistical learning, G. James, D. Witten, T. https://www.statlearning.com/
  - Many of the examples and figures are from the book
- Recommended reading:
  - An introduction to generalized linear models, Annette J Dobson, CHAPMAN & HALL/CRC, 2002

Course rule book

- Keep up!
  - Later lessons build on earlier lessons.
  - It is a lot of material, waiting till the end may cause troubles
- The course will include both Theory and Practical Skills
- For each class there is a reading assignment. Up until the day before the class you can ask questions, that will then be discussed during class.
- The exercises will be in R. Let me know if you are unfamiliar with R!

Evaluation

- 1 Assignment, will be graded and counts for 4/20pts (for part II)
- Exam with theoretical questions and computer exercises

Planning today

- A general approach for statistical modelling
- Bias-Variance trade off
  - Relevant for choice of model. E.g. a non-linear vs a linear model
- The challenges of high dimensional datasets
  - In biology today, many datasets have observations for many actors. This causes specific challenges with modelling.
- The linear model and their assumptions
- Re-sampling approaches
  - Validation set
  - Cross-Validation
  - Bootstrap

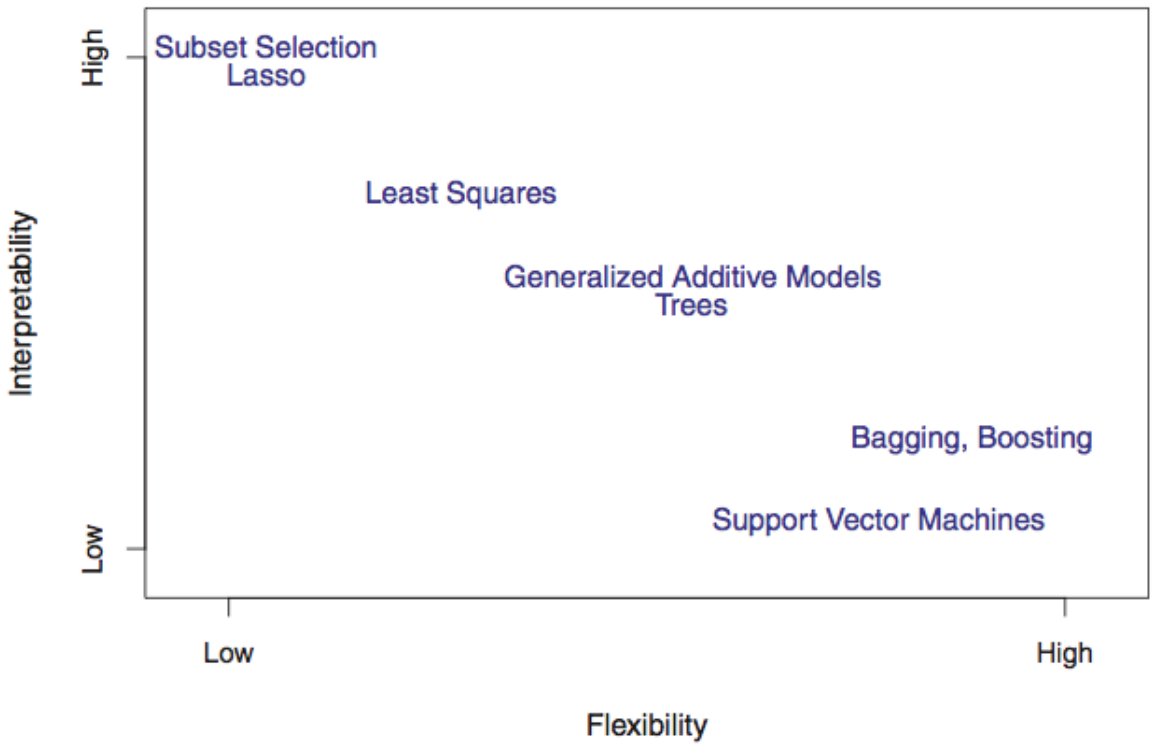
1. Statistical Modeling: survey the data

Any analysis of data should begin with a consideration of each variable separately, to check data quality but also to understand how a model could be formulated.

1. What kind of response and explanatory variables do you have?
   - Binary, Categorical, Ordinal
   - Continuous
   - Proportion
   - Count
   - Time at death
2. What is the shape of the distribution (e.g. look at histograms)
3. Do you see associations with other variables (e.g. look at scatter plot)

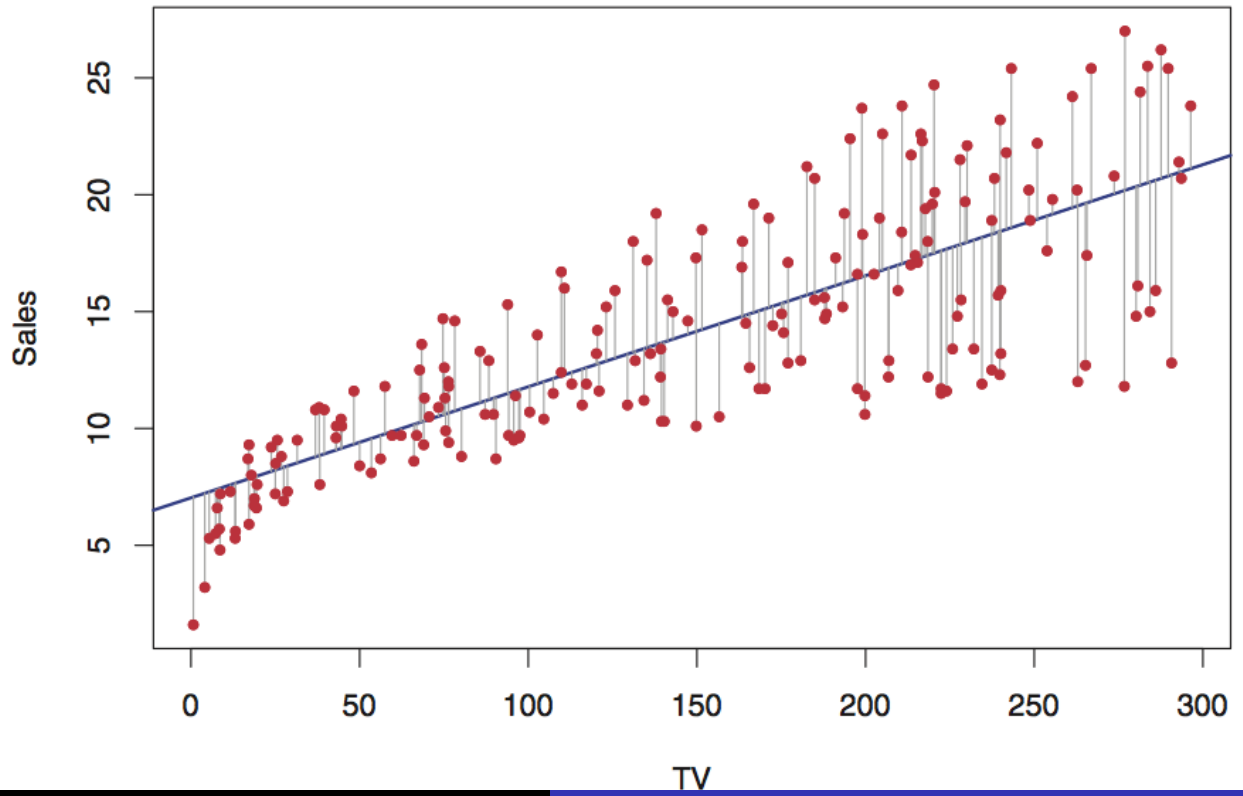
2. Statistical Modeling: choose a model

- Typical case: one response variable and several explanatory variables.
- There is no perfect method for all data
- A single perfect model is rare; different models can be fit with good performance.
  - Which level of complexity is adequate? Avoid overly complex models with limited benefit.



3. Statistical Modeling: Fitting parameters

- The most commonly used estimation methods are maximum likelihood and least squares.
  - Maximum likelihood: given the data and the choice of model, what values of the parameters of the model make the observed data most likely?
  - Minimize least squares: find the fit for which $S = \sum_i^n (Y_i - \hat{Y}_i)^2$ is minimal



4. Statistical Modeling: Checking the model

- Have a look at the fit and the residuals (the difference between the predicted Y and the actual Y).
  - Is the fit good over the whole range? Evidence of non-linearity?
  - Do the residuals behave as expected, e.g. approximately normal, with mean 0?
- Could the model be simplified?
  - The law of parsimony (otherwise known as Occam's Razor) dictates that no more causes should be assumed than will account for the effect.

Thinking about modelling: a single predictor

- How do you decide which statistical learning method to choose?
- How do you choose how complex or flexible a model should be?
  - For example: is a variable significantly predictive in a linear regression?
- For a function to estimate the relationship (given as $f(x)$) between predictor x and the response variable Y:
  - The expected square error: $Err(x) = E[(Y - \widehat{f}(x))^2]$
  - The error has reducible and irreducible components: $Err(x) = (f(x) - \hat{f}(x))^2 + Var(\epsilon)$ with $Var(\epsilon)$ the variance by the irreducible error component. The reducible error can be split up further:

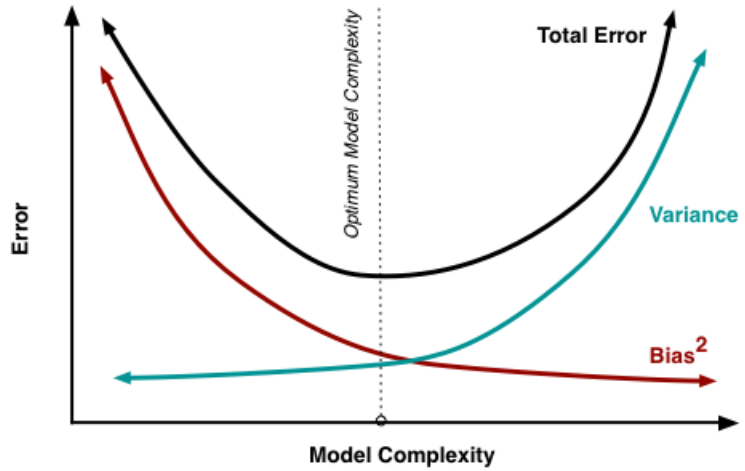
$$Err(x) = Var(\hat{f}(x_0)) + Bias(\hat{f}(x_0))^2 + Var(\epsilon)$$

The perennial trade-off: bias vs variance

- One wants a model that captures the regularities in training data, but also generalizes well to unseen data.

$$E(y_0 - \hat{f}(x_0))^2 = Var(\hat{f}(x_0)) + Bias(\hat{f}(x_0))^2 + Var(\epsilon)$$

- With Variance: $Var(\hat{f}) = E((E(\hat{f}) - \hat{f})^2)$
- Bias: $Bias(\hat{f})^2 = E((f - E(\hat{f}))^2)$

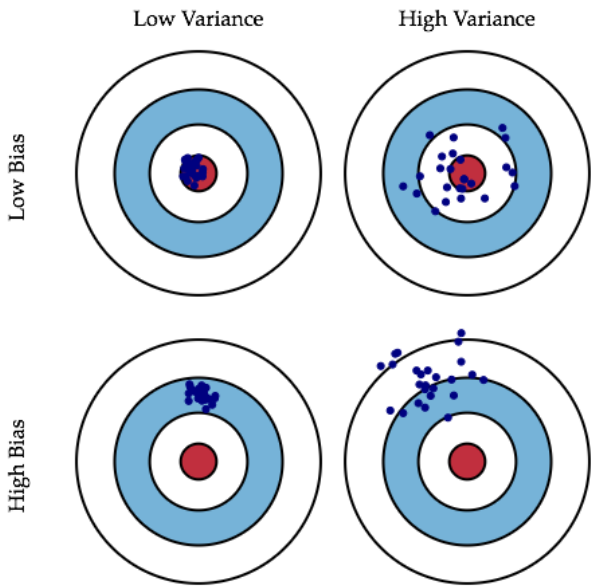


[1]

[1] From "Understanding the Bias-Variance Tradeoff" by S. Fortmann Roe

Bias Variance Trade-Off

- Models with high bias are intuitively simple models: restrictions on the kind of regularities that can be learned (e.g. linear classifiers).
  - These models tend to **underfit**, i.e. not learn the relationship between predicted (target) variables and features.
- Models with high variance are those that can learn many kinds of complex regularities
  - These models can learn noise in the training data, i.e. **overfitting**.



For example: a linear or a non-linear fit?

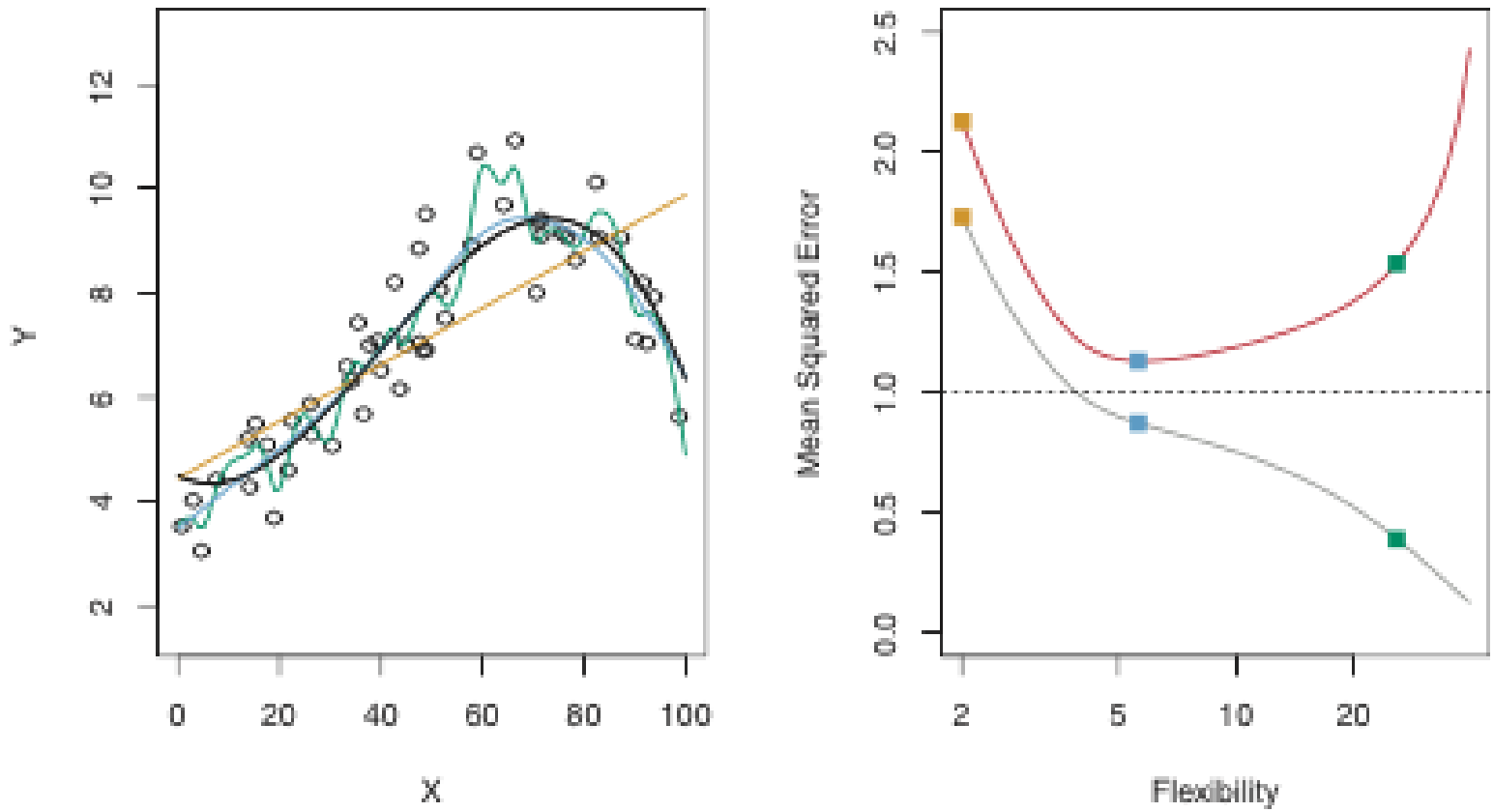
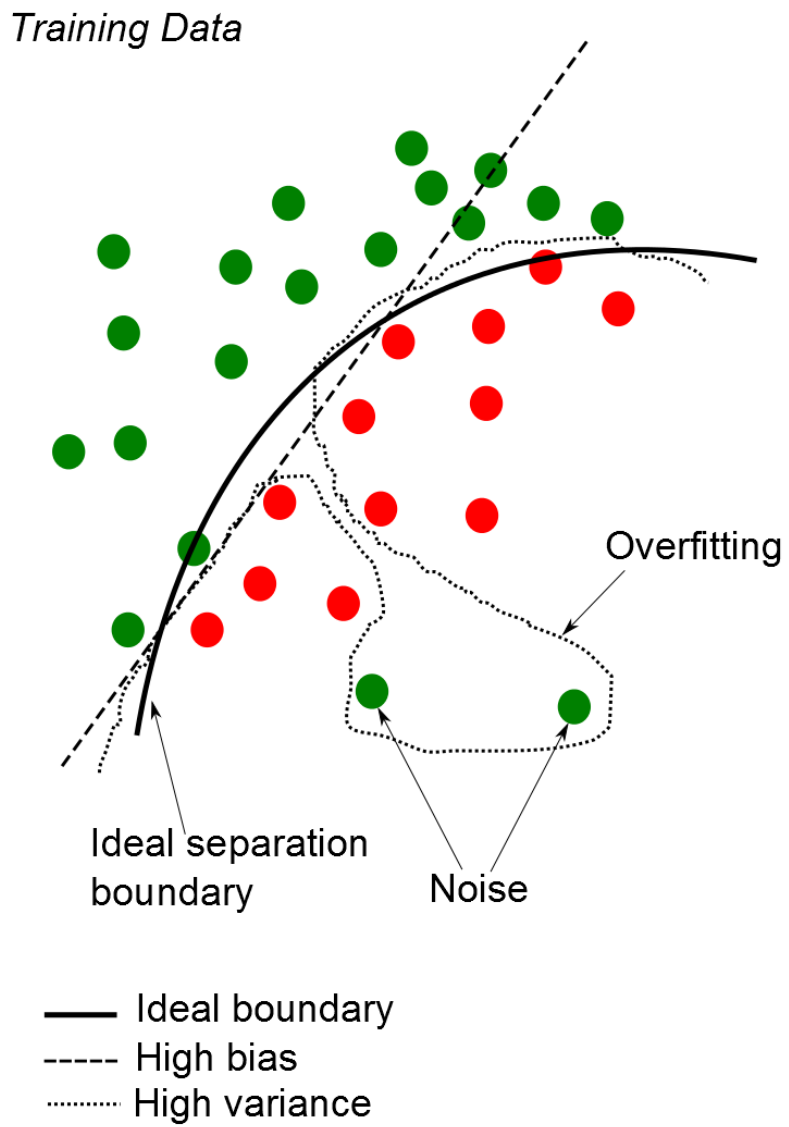
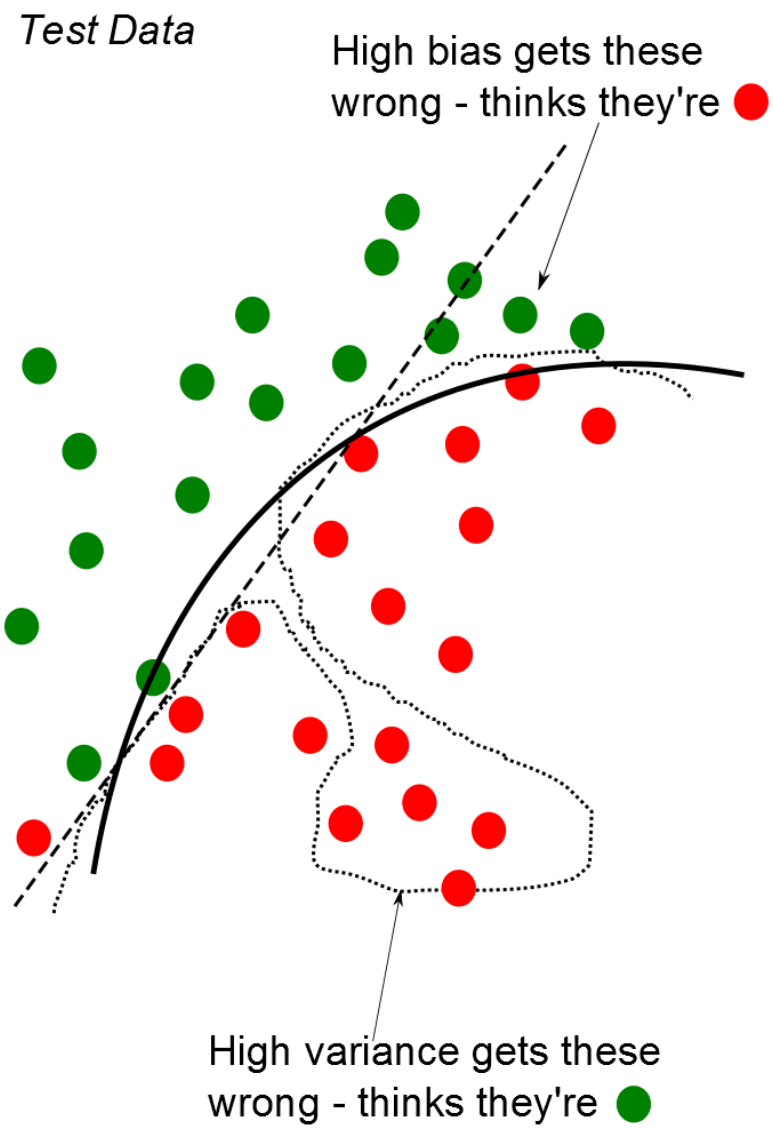


FIGURE 2.9. Left: *Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves).* Right: *Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.*

Another example: which decision boundary in a classifier?



Another example: which decision boundary in a classifier?



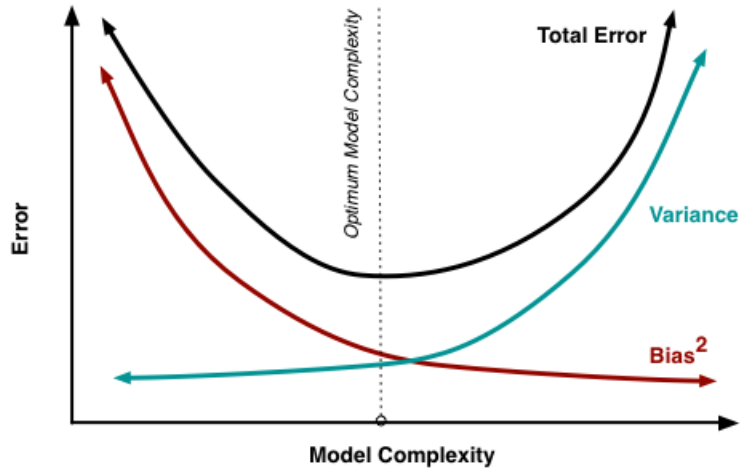
Images from Abhishek Ghose's blog

The perennial trade-off: bias vs variance

- One wants a model that captures the regularities in training data, but also generalizes well to unseen data.

$$E(y_0 - \hat{f}(x_0))^2 = Var(\hat{f}(x_0)) + Bias(\hat{f}(x_0))^2 + Var(\epsilon)$$

- With Variance: $Var(\hat{f}) = E((E(\hat{f}) - \hat{f})^2)$
- Bias: $Bias(\hat{f})^2 = E((f - E(\hat{f}))^2)$



[2]

[2] From "Understanding the Bias-Variance Tradeoff" by S. Fortmann Roe

Bias and variance trade-off: a crucial concept

- Returns throughout the course
- How complex or flexible should the model be?
  - Relevant e.g. in the part on non-linear models and Random Forests
  - Degrees of freedom in parametric flexible models, e.g. smoothing splines and non-parametric flexible models, e.g. kernel smoothing

Progress

- A general approach for statistical modelling
- Bias-Variance trade off
  - Relevant for choice of model. E.g. a non-linear vs a linear model
- The challenges of high dimensional datasets
  - In biology today, many datasets have observations for many actors. This causes specific challenges with modelling.
- The linear model and their assumptions
- Re-sampling approaches
  - Validation set
  - Cross-Validation
  - Bootstrap

Curse of dimensionality

- In this age of Big Data we often have many predictors and high-dimensional datasets
- High-dimensional datasets pose special challenges with respect to statistical learning
    - Curse of Dimensionality.

The Challenges of High Dimensionality

- When your number of dimensions is on the order of the number of observations you have a problem fitting data, e.g. you get saturated systems, with limited fitting/generalization
  - a line fits every 2 points in 2D perfectly
  - a plane fits every 3 points in 3D perfectly
  - a hyperplane fits every 4 points in 4D perfectly
- Also in high dimensions the distance between observations tends to be larger
  - Sparse observations make it hard to make predictions
  - Complexity of functions of many variables can grow exponentially with D. For accurate fitting of the parameters an exponential increase in observations would be needed.

Curse of dimensionality

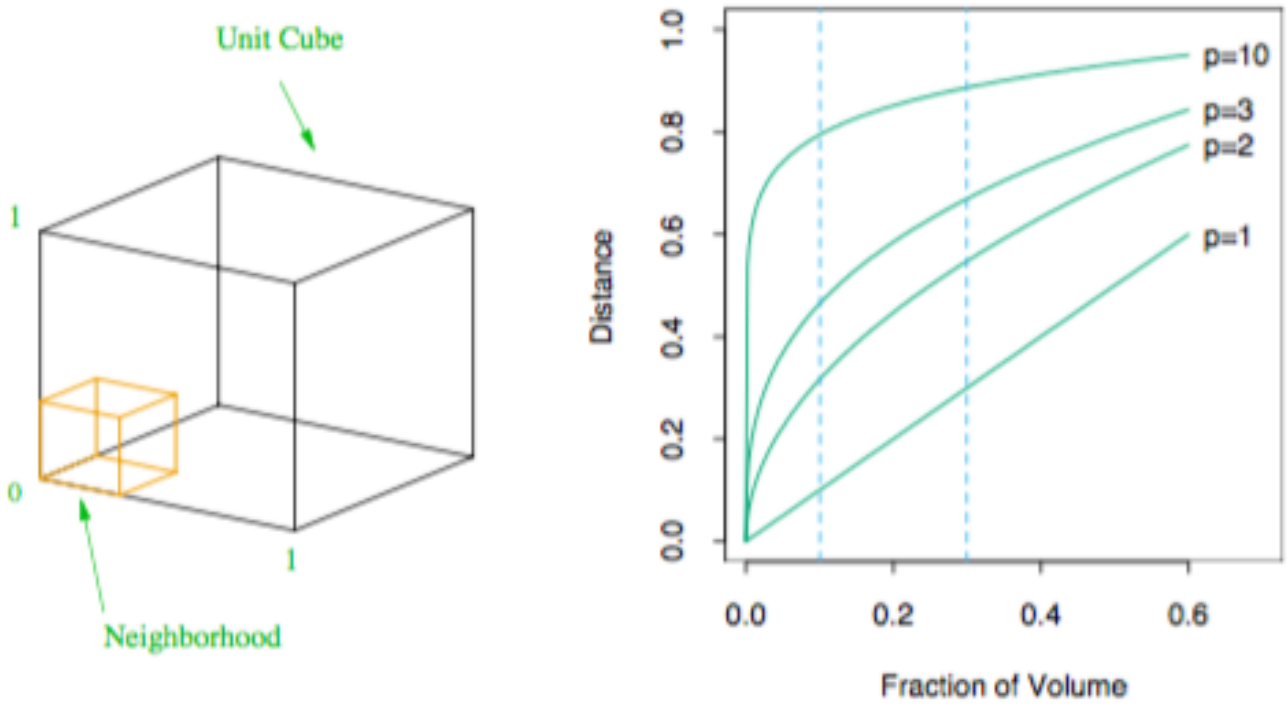


FIGURE 2.6. *The curse of dimensionality is well illustrated by a subcubical neighborhood for uniform data in a unit cube. The figure on the right shows the side-length of the subcube needed to capture a fraction r of the volume of the data, for different dimensions p. In ten dimensions we need to cover 80% of the range of each coordinate to capture 10% of the data.*

Curse of dimensionality

What are the consequences of high dimensionality for bias and variance?

High dimensional datasets

- How do you decide which (or all) predictors you will keep in your modelling?
- The methods discussed in the 2nd and 3rd classes deal properly with high dimensionality
- Considerations for interpreting analyses of high dimensional datasets in 3rd lecture.

Linear Models: powerful simplicity

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m + \varepsilon$$

- Easy interpretability: β_j is the average increase in Y when x_j increases by one, and the other variables are constant.
- Estimate parameters with e.g. Least Squares; $argmin_{\beta_o \ldots \beta_m} MSE = \frac{1}{n}\sum(Y_i - \hat{Y}_i)^2$ and $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_m x_m + \varepsilon$
- Measure of fit can be coefficient of determination, range $[0, 1]$, fraction explained variance $R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$

Linear Models for Essential Questions

Through a linear model you can test or evaluate the following questions:

- Do two variables have a relationship? Quantify the evidence!
- How strong is the relationship?
- Can we distinguish between variables? Which is most predictive?
- Is the relationship linear, or is there evidence for a non-linear effect?
- Are there interactions between variables?
- How well can we predict a variable knowing a single or set of other variables?

Testing if a coefficient is relevant

- For the estimated mean of a normal distribution we have: $SE(\hat{\mu})^2 = \frac{\sigma^2}{n}$, you can test significance with a t - test and calculate 95% confidence intervals.
  - For $\hat{Y} = \beta_0 + \beta_1 x_1$, we can write $SE(\beta_0)^2 = \sigma^2(\frac{1}{2} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2})$ and $SE(\beta_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2})$ with $\sigma^2 = var(\varepsilon)$
- The statistic follows from $t = \frac{\hat{\beta}_i}{S.E.(\hat{\beta}_i)}$
- Underlying assumption is the normality of the error distribution.

The assumptions of linear regression

1. Linearity of the relationship
   - Violations will cause problems when you try to make predictions
   - Does a transformation of the data produce a linear relationship?
2. Normality of the errors
3. Independence of the errors
   - Random effects can cause problems, e.g. plants grown in plots
   - Repeated measurements of the same subject, e.g. different times or conditions
4. Homoscedastity (equal variance over the predictions)
   - Affects confidence intervals

Potential Fit Problems

There are a number of possible problems that one may encounter when fitting the linear regression model. In addition to looking at the performance statistics RSE and R^2, one should analyze and plot the data. Graphical summaries can reveal problems with a model.

1. Non-linearity of the data
2. Dependence of the error terms
3. Non-constant variance of error terms
4. Outliers
5. High leverage points
6. Collinearity

Challenges with models

- How do you compare models?
  - Parameter selection
  - Random forests vs a GLM?
- How do you decide on significance of a test or the confidence intervals of a coefficient in a linear model if the assumptions underlying the statistical distributions are violated?

Re-sampling Methods

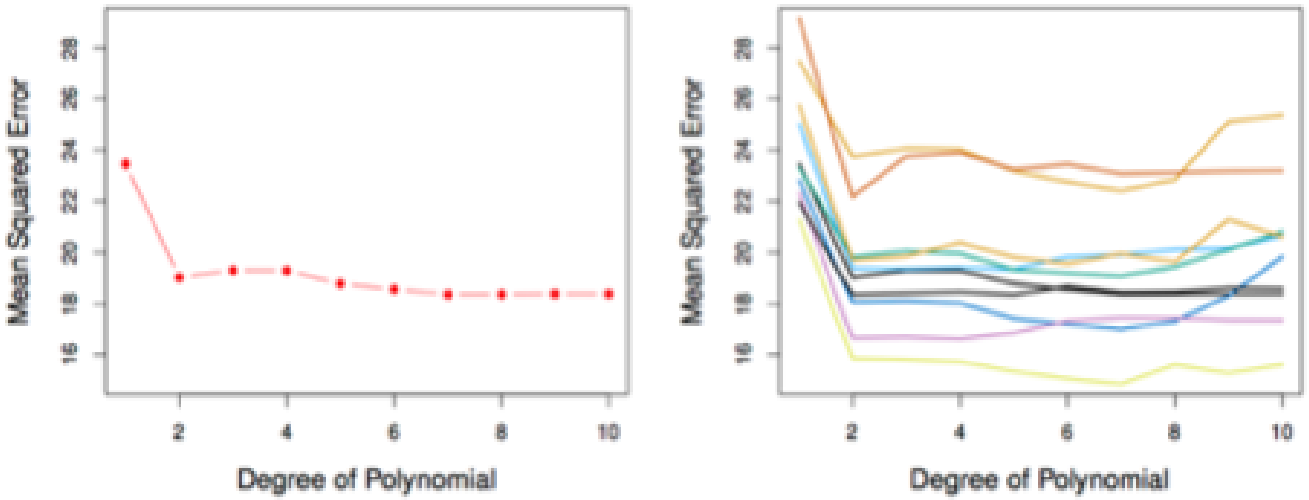
- Introduction
- Single validation set
- Cross Validation
  - Leave-one-out Cross Validation
  - K-fold Cross Validation
  - Bias-Variance Trade-off for k-fold Cross Validation
- Bootstrap

Re-sampling Methods

- Methods that draw a sample or samples from a training set and fit a model on each sample to obtain more information about the model's properties
  - Model Assessment: estimate test error rates
  - Model Selection: select the appropriate level of model flexibility
- Can be computationally expensive; but in exceptional cases nearly free
- Re-sampling methods:
  - Cross Validation
  - Bootstrapping

Classical validation set approach

- Find a set of variables that give lowest *test* (instead of training) error rate
- If we have a large data set, we can achieve this goal by randomly splitting the data into training and validation (testing) parts
- Builld models on the training part, choose model with lowest error rate when applied to the validation data



Validation set approach

- Advantages:
  - Simple
  - Easy to implement
- Disadvantages:
  - The validation performance estimate (e.g. Mean Squared Error) can be highly variable $MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$
  - Only a subset of observations are used to fit the model (training data). Statistical methods tend to perform worse when trained on fewer observations.

Leave-One-Out Cross Validation (LOOCV)

- Similar to the Validation Set Approach, but it tries to address the latter's disadvantages
- For each suggested model, do:
  - Split the data set of size n into
    - Training data set size: n -1
    - Validation data set size: 1
  - Fit the model using the training data
  - Validate model using the validation data, and compute the corresponding MSE
  - Repeat this process n times
- The MSE for the model is $CV = \frac{1}{n} \sum_{i=1}^{n} MSE_i$

LOOCV vs Validation set approach

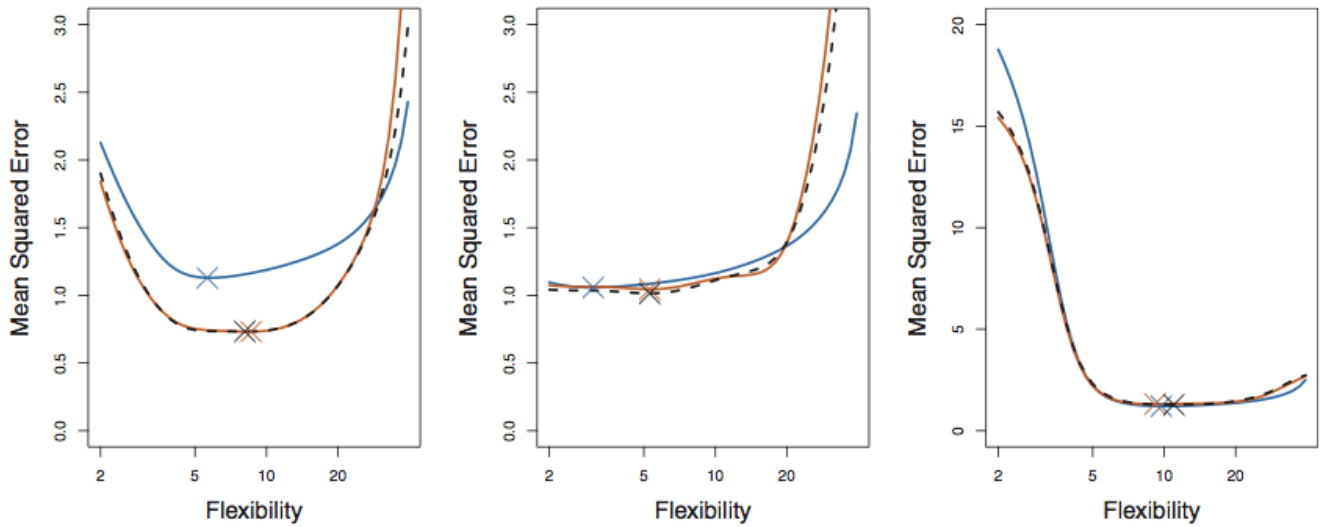
- LOOCV has less bias
  - almost all the data set is used
- LOOCV produces a less variable MSE estimate
  - effect randomness of splitting process reduced
- LOOCV can be computationally intensive
  - fitting model n times
  - Except with least squares linear or polynomial regression; here an shortcut makes LOOCV cost the same as the cost of single model fit. The following holds:

$$CV_n = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2$$

with h_i the leverage for this data point

k-fold Cross Validation

- We randomly divide the data set into k folds
  - each fold is used in turn as the validation set, with the remainder as training
- The estimated error rate is simply the average MSE
  - Stable, like LOOCV, but a.o. less computationally intensive



MSE for simulated data: true test MSE in blue, LOOCV as a black dashed line, 10-fold CV estimate in orange. Crosses indicate minimum of MSE curves.

Bias-Variance trade-off for CV

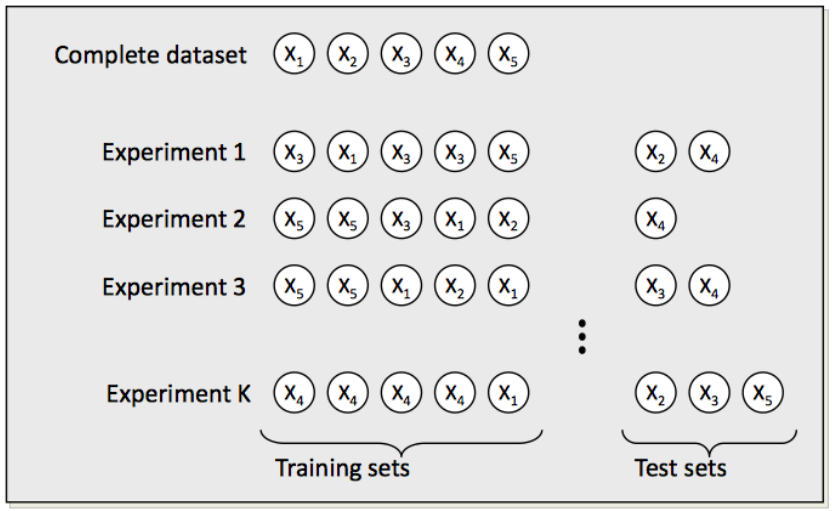
- LOOCV has less bias than k-fold CV (k<<n)
  - larger training sets
- LOOCV tends to have higher variance than k-fold CV (k<<n)
  - Learning sets are very similar!
  - Correlated learning sets inflate variance (a deep statistical truth)
- LOOCV can be useful but normally a trade-off is made: k-fold CV with $K = 5$ / $K = 10$
  - Empirical evidence to prevent excessively high bias and high variance in test error rate estimates.

Re-sampling Methods

- When considering learning models, where should we apply CV?
  - Parameter Selection: We select the most informative parameter(s) for a given classification problem
  - Model selection: Once we have chosen a set of parameters, how should we estimate the true error rate of a model?
    - The true error rate is the classifier's error rate when tested on the entire population / known production function

Bootstrap

- The bootstrap is a resampling technique with replacement
  - From a dataset with **n** examples
    - Randomly select (with replacement) **n** examples and use this set for training
    - The remaining examples that were not selected for training are used for testing
    - This value is likely to change from fold to fold
    - Repeat this process for a specified number of folds (**k**)
    - The true error is estimated as the average error rate on test data



Why Bootstrap

- Compared to basic CV, the bootstrap increases the variance that can occur in each fold
  - This is a desirable: a more realistic simulation of the real-life experiment from which our dataset was obtained
- Sampling with replacement preserves a-priori probabilities of the classes throughout the random selection process
- The bootstrap provides accurate measures of both the bias and variance of the estimator
  - The method is mostly used to assess uncertainty/standard errors/confidence intervals in estimators

Can Bootstrap estimate Prediction Error?

- In cross-validation, each of the K validation folds is distinct from the other $K - 1$ folds used for training: there is no overlap.
  - This is crucial for its success. (Why?)
- To estimate prediction error using the bootstrap, observations not selected by the bootstrap are used as a test set
  - But your test-set varies in size and you may see some observations in your test-sets more than others.

To do:

Reading for current class

Chapters 2 and 5

Preparation for next class (obligatory)

- Read part of chapter 6; 6.1 and 6.2
- Send in any questions day before class

Exercises

- Lab of chapter 5
- Chapter 5, exercises 1,4,5,6 & 8