Module 12: Summing up and some cautionary notes

TMA4268 Statistical Learning V2021

Stefanie Muff, Department of Mathematical Sciences, NTNU

April 26, 2021

Overview

- Course content and learning outcome
- Overview of modules and core course topics
- Some cautionary notes

Some of the figures and slides in this presentation are taken (or are inspired) from G. James et al. (2013).

Learning outcomes of TMA4268

1. **Knowledge.** The student has knowledge about the most popular statistical learning models and methods that are used for *prediction* and *inference* in science and technology. Emphasis is on regression- and classification-type statistical models.
2. **Skills.** The student can, based on an existing data set, *choose a suitable statistical model, apply sound statistical methods*, and *perform the analyses using statistical software*. The student can present, interpret and communicate the results from the statistical analyses, and knows which conclusions can be drawn from the analyses, and what are the caveats.

And: you got to be an expert in using the R language and writing R Markdown reports.

Core of the course

Supervised and unsupervised learning:

- *Supervised*: regression and classification
  - examples of regression and classification type problems
  - how complex a model to get the best fit?
    $\rightarrow$ flexiblity/overfitting/underfitting.
  - the bias-variance trade-off
  - how to find a good fit - validation and cross-validation (or AIC-type solutions) Module 6
  - how to compare different solutions
  - how to evaluate the fit - on new unseen data

Modules 1-9 + 11

- *Unsupervised*: how to find structure or groupings in data? Module 10

and of course all **the methods** (with underlying models) to perform regression, classification and unsupervised learning.

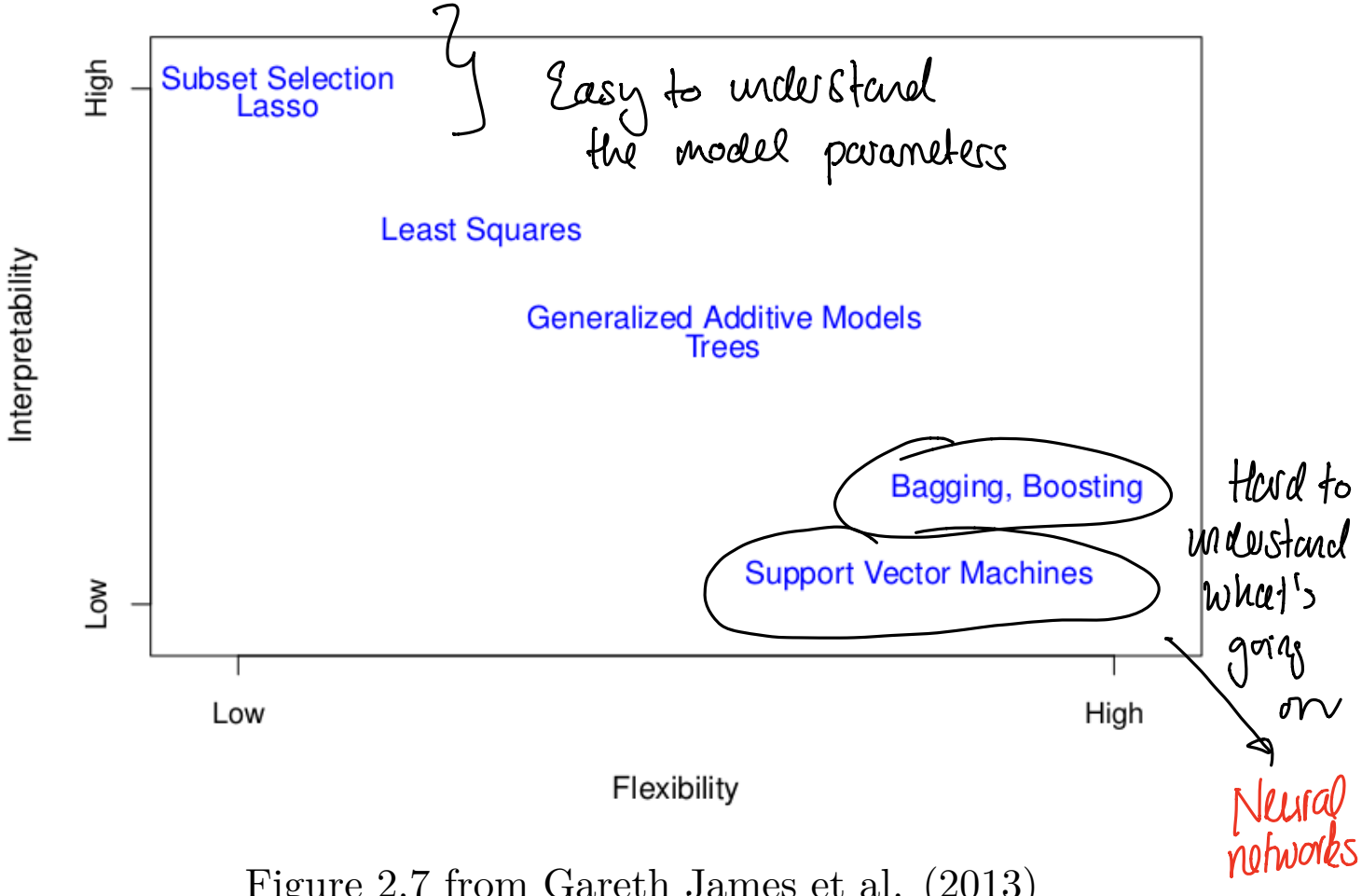


Figure 2.7 from Gareth James et al. (2013)

The modules

1. Introduction

- Examples, the modules, required background in statistics
- Introduction to R and RStudio via the online R-course

2. Statistical learning

- Model complexity
  - Prediction vs. interpretation (inference).
  - Parametric vs. nonparametric.
  - Flexible vs. inflexible.
  - Overfitting vs. underfitting
- Supervised vs. unsupervised.
- Regression and classification.
- Loss functions: quadratic and 0/1 loss. RSS, ME-error
- Bias-variance trade-off (polynomial example): mean squared error, training and test set.
- Vectors and matrices, rules for mean and covariances, the multivariate normal distribution.

3. Linear regression

- The classical normal linear regression model on vector/matrix form.
- Parameter estimators and distribution thereof. Model fit.
- Confidence intervals, hypothesis tests, and interpreting R-output from regression.
- Qualitative covariates, interactions.
- This module is a stepping stone for all subsequent uses of regression in Modules 6, 7, 8, and 11.

How to interpret qualitative covariates / interaction terms. !

Categorical / factor variables

very simple (naive?) idea

4. Classification (Mainly two-class problems)

- Bayes classifier: classify to the most probable class to minimize the expected 0/1 loss. We usually do not know the probability of each class for each input. The Bayes optimal boundary is the boundary for the Bayes classifier and the error rate (on a test set) for the Bayes classifier is the Bayes error rate.

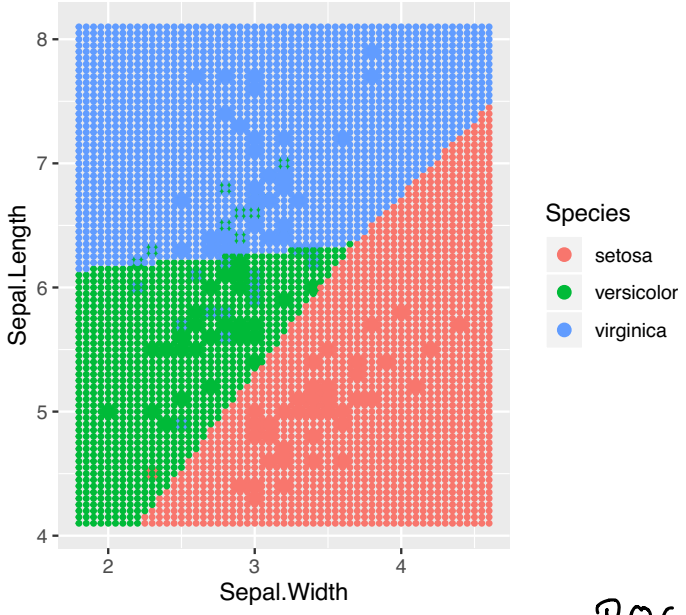
- Two paradigms (not in textbook): Ways to estimate the class probabilities
  - *Diagnostic* (directly estimating the posterior distribution for the classes). Example: KNN classifier, logistic regression.
  - *Sampling* (estimating class prior probabilities and class conditional distribution and then putting together with Bayes rule). Examples: LDA, QDA with linear or quadratic class boundaries.

indirect

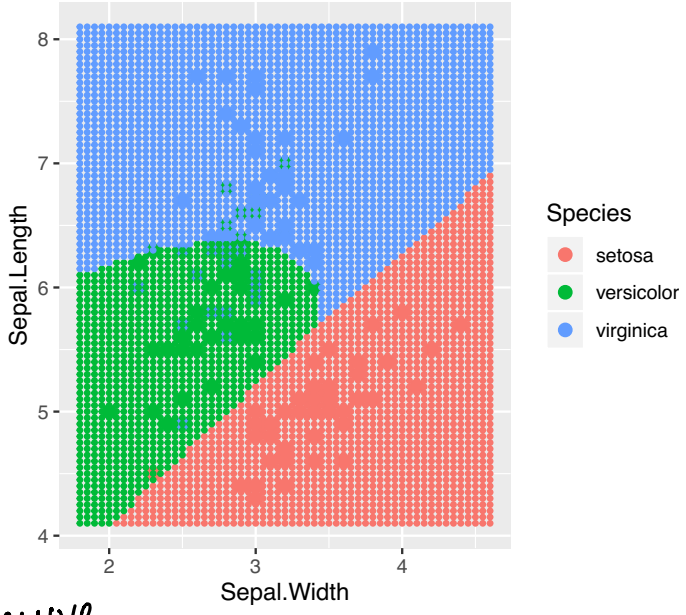
$$P(Y=k \mid X=x) \propto P(X=x \mid Y=k) \cdot P(Y=k)$$

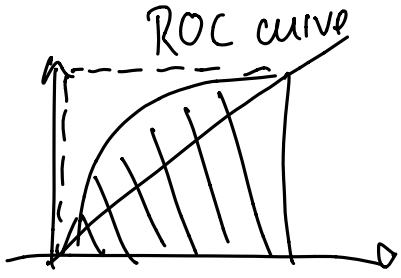
- ROC curves, AUC, sensitivity and specificity of classification methods.

LDA → linear boundaries



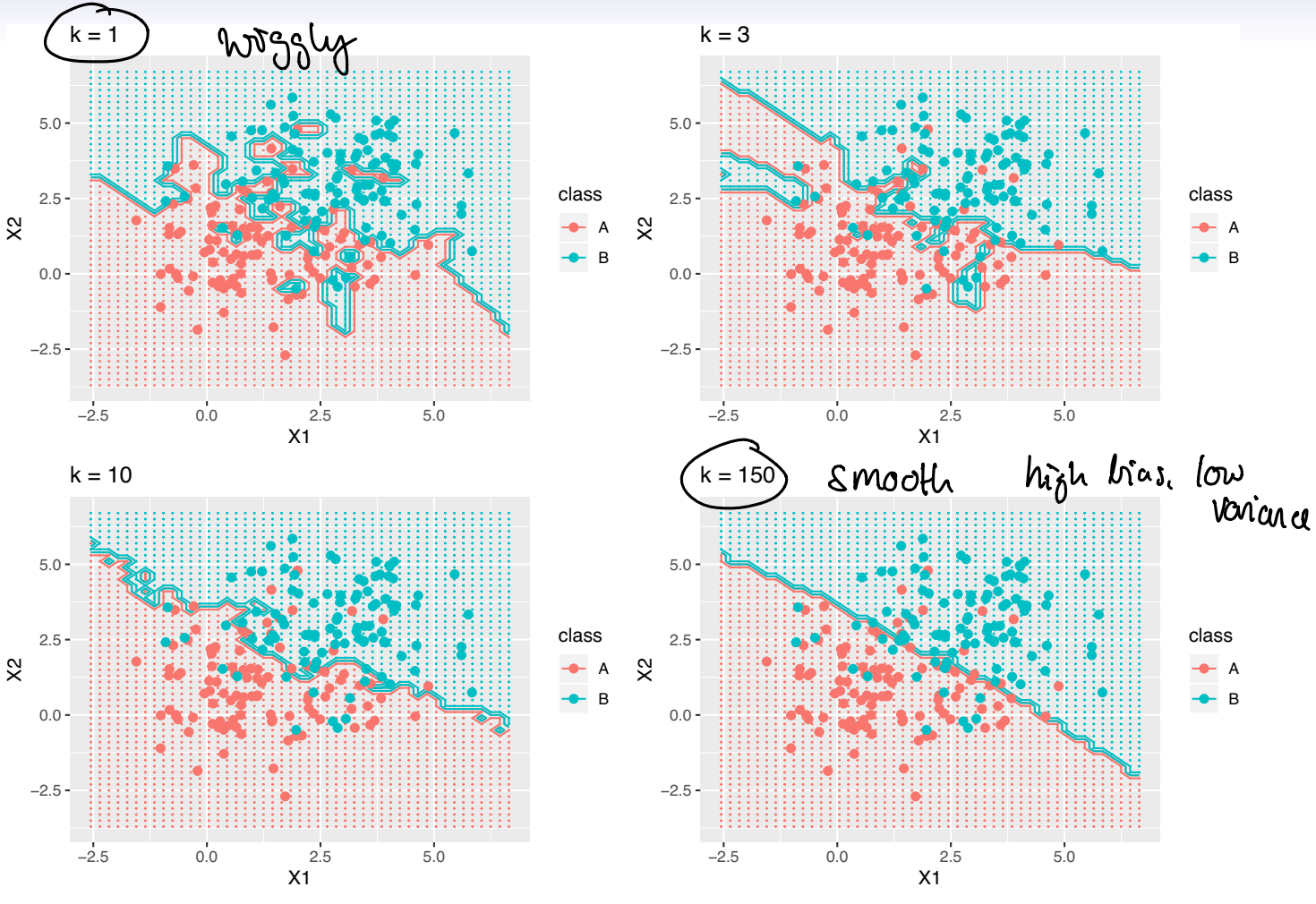
QDA → quadratic boundaries



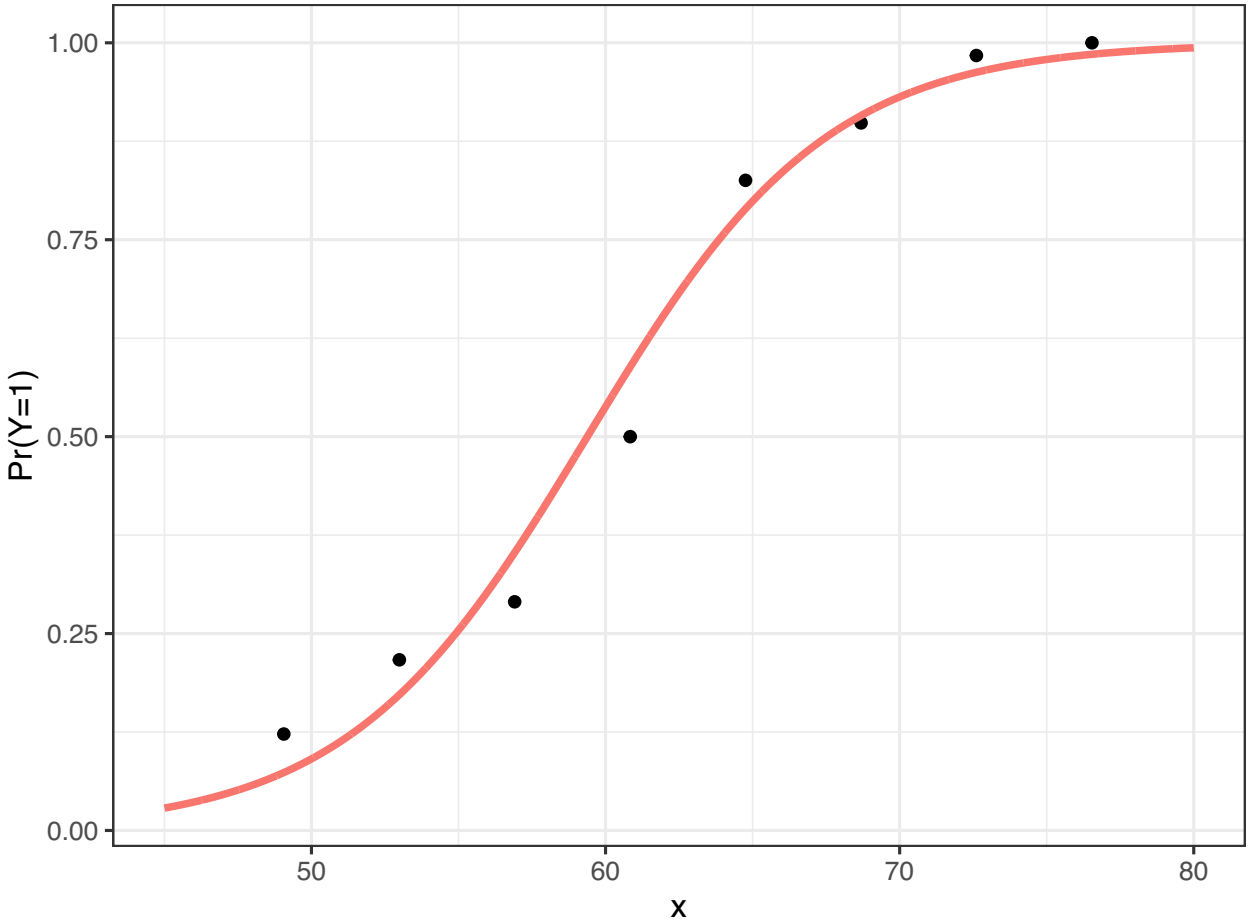


AUC > 0.5

low bias, high variance KNN



Logistic regression gives a probability, given a certain value of the covariats $P(Y = 1 \mid x)$.



5. Resampling methods

Cros-validation

- Data rich situation: Training, validation and test set.
- Validation set approach.
- Cross-validation for regression and for classification.
- LOOCV, 5 and 10 fold CV.
- Good and bad issues with validation set, LOOCV, 10-fold CV.
- Bias and variance for k-fold cross-validation.
- Selection bias – the right and wrong way to do cross-validation.
- Distinction between model selection and model assessment.

The Bootstrap

- Idea: Re-use the same data to estimate a statistic of interest by *sampling with replacement*.

6. Linear model selection and regularization:

Subset-selection. Discriminate:

- *Model selection*: estimate performance of different models to choose the best one.
- *Model assessment*: having chosen a final model, estimate its performance on new data.

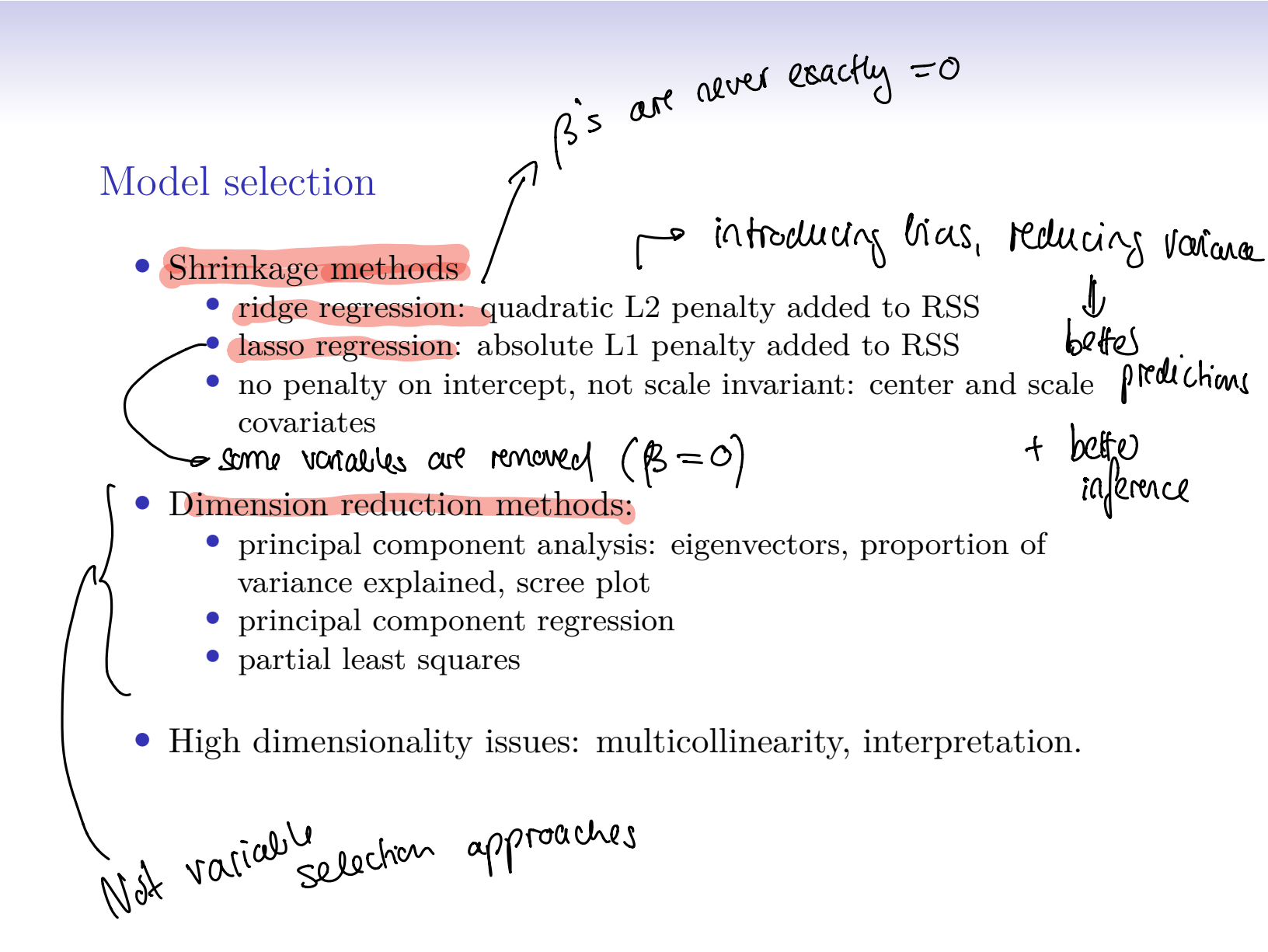
How?

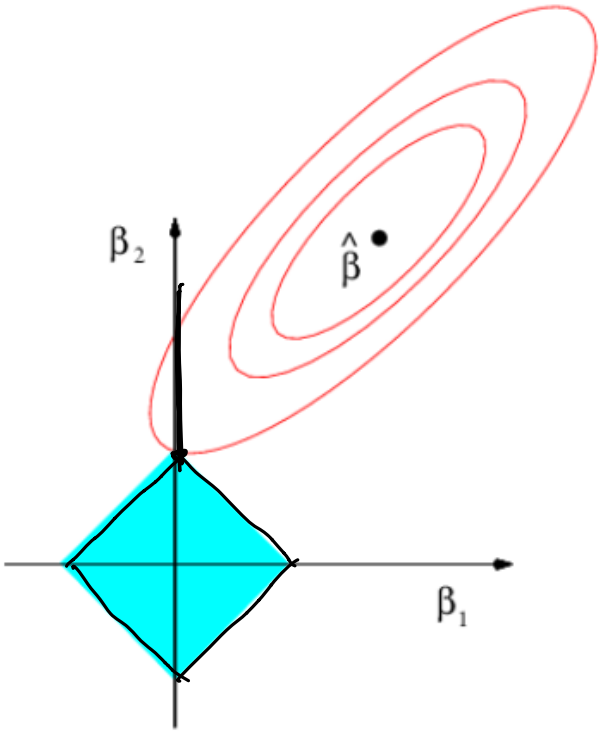
- Model selection by — Set of covariates $X_1, \ldots, X_p$
  - Subset selection (best subset selection or stepwise model selection)
  - Penalizing the training error: AIC, BIC, C_p, Adjusted R^2.
  - Cross-validation.
- Model assessment by
  - Cross-validation.

Cautious!

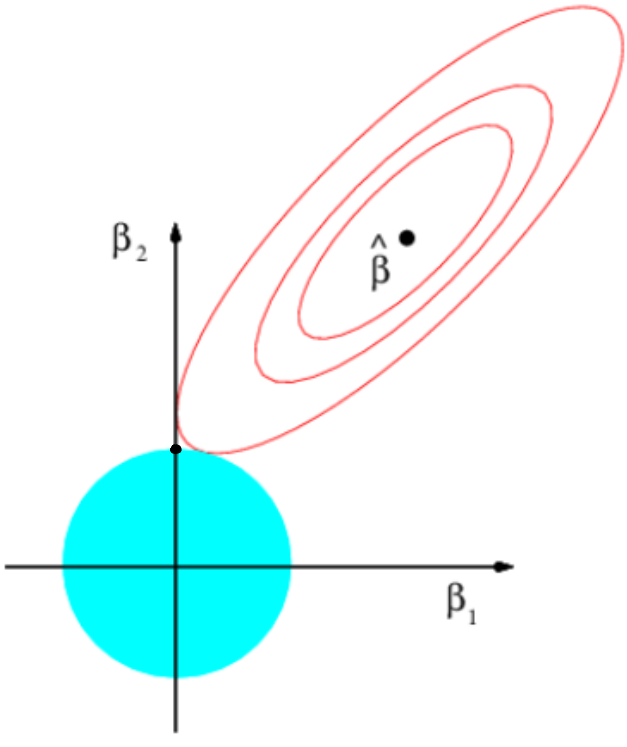
Model selection

- Shrinkage methods
  - ridge regression: quadratic L2 penalty added to RSS
  - lasso regression: absolute L1 penalty added to RSS
  - no penalty on intercept, not scale invariant: center and scale covariates
- Dimension reduction methods:
  - principal component analysis: eigenvectors, proportion of variance explained, scree plot
  - principal component regression
  - partial least squares
- High dimensionality issues: multicollinearity, interpretation.



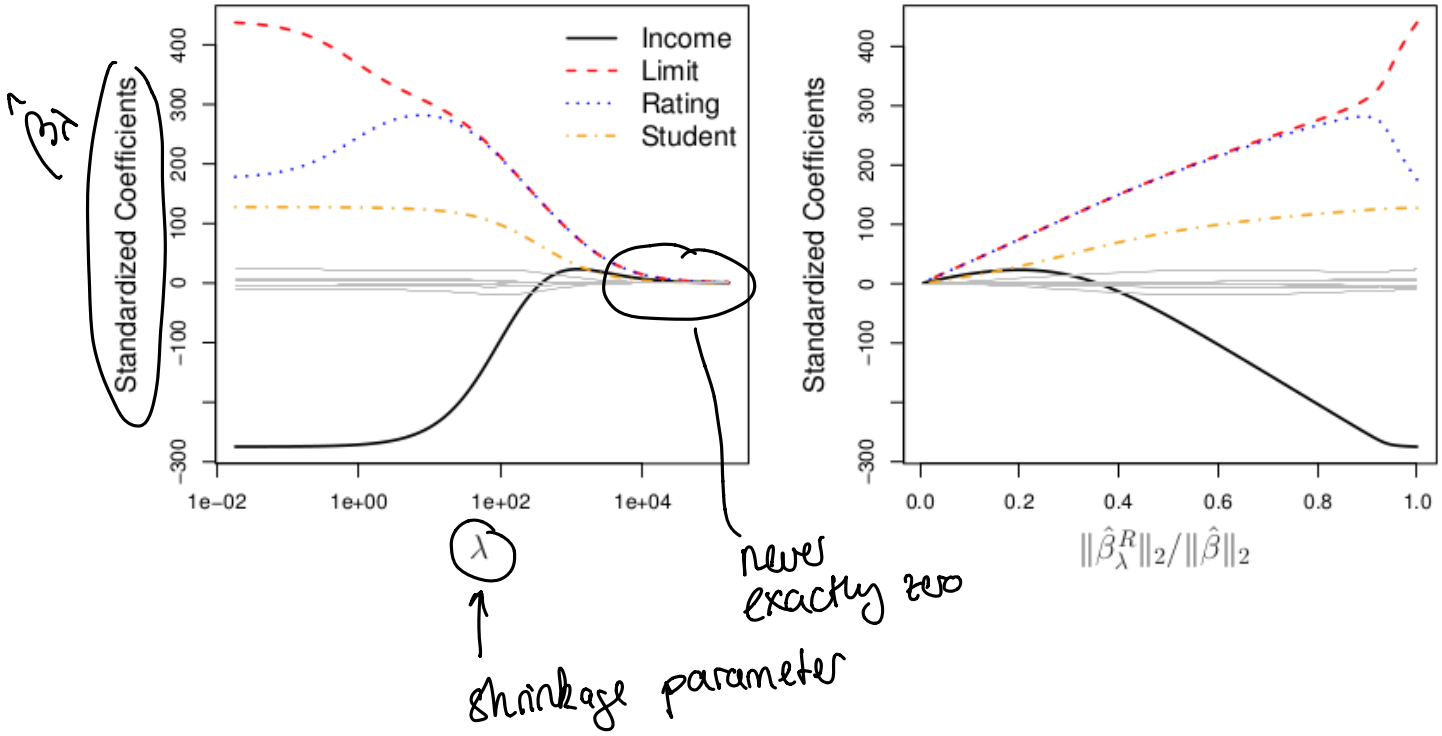


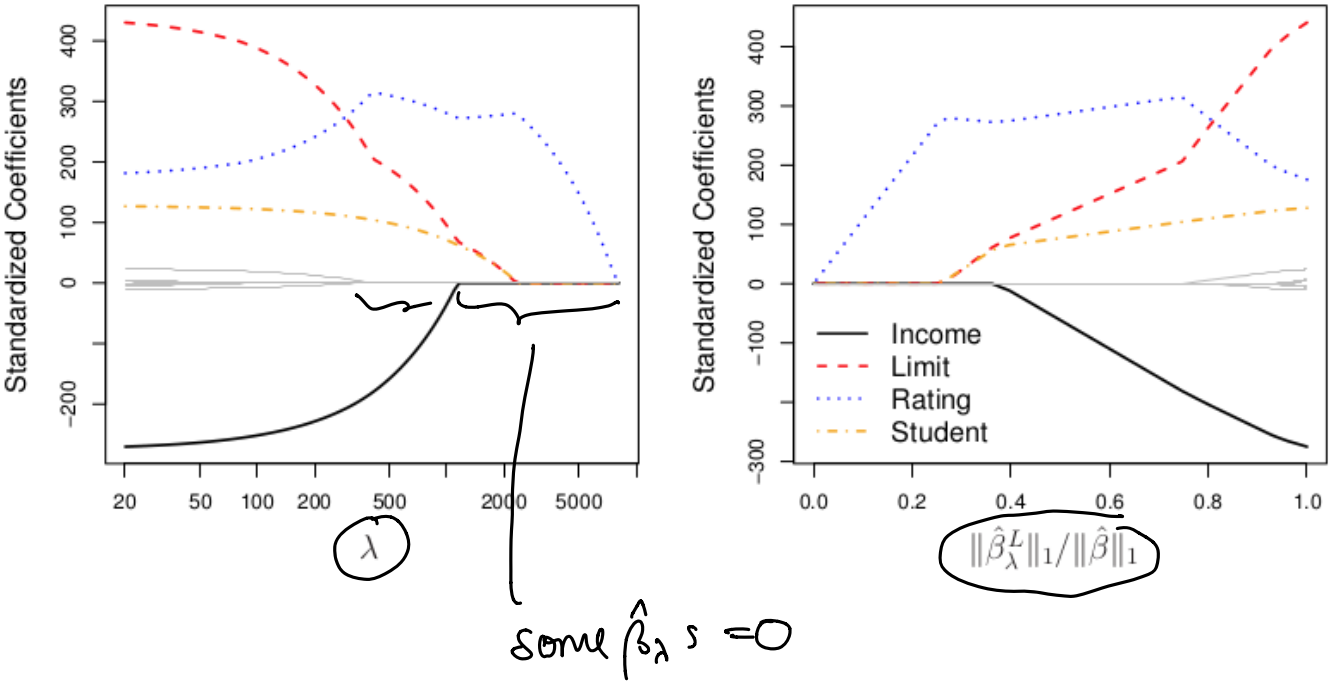
Lassa



Ridge

Regularization techniques (see also in module 11)




Standardized Coefficients
400
300
200
100
0
−100
−200
−300
20
50
100
200
500
2000
5000
λ
0.0
0.2
0.4
0.6
0.8
1.0
$\|\hat{\beta}^L_\lambda\|_1/\|\hat{\beta}\|_1$
Income
Limit
Rating
Student
some $\hat{\beta}_\lambda$ s = 0

7. Moving beyond linearity

- Modifications to the multiple linear regression model - when a linear model is not the best choice. First look at one covariate, combine in "additive model".
- Basis functions: fixed functions of the covariates.
- Polynomial regression: multiple linear regression with polynomial components as basis functions. $X, X^2, X^3, \ldots$
- Step functions - piecewise constants. Like our dummy variable coding of factors.
- Regression splines: regional polynomials joined smoothly - neat use of basis functions. Cubic splines very popular.

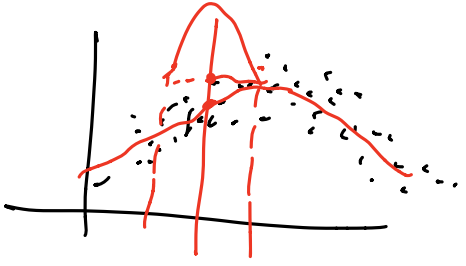
$$\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt$$

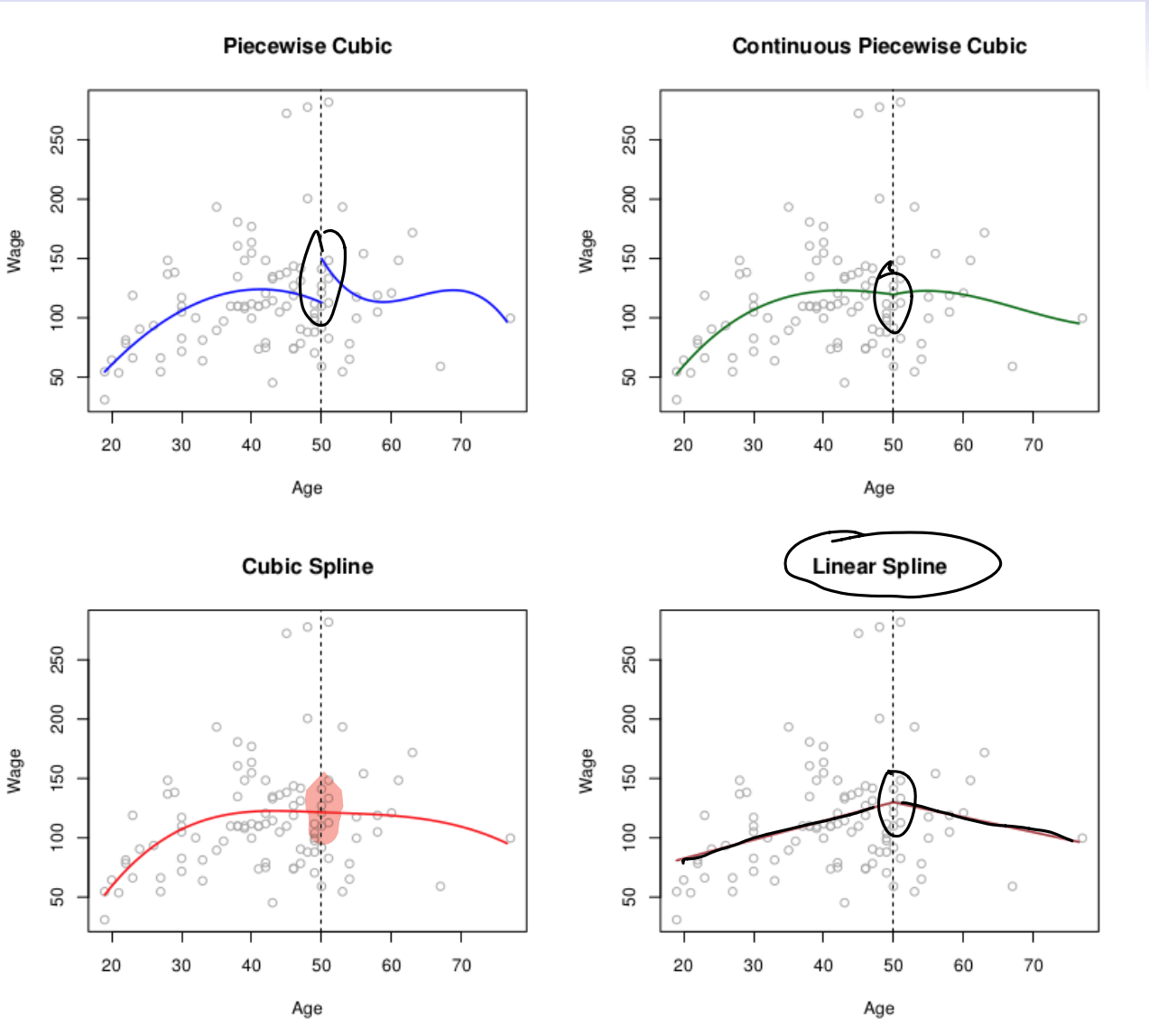
~ smoothing term

- Smoothing splines: smooth functions - minimizing the RSS with an additional penalty on the second derivative of the curve. Results in a natural cubic spline with knots in the unique values of the covariate.

- Local regressions: smoothed K-nearest neighbour with local regression and weighting. In applied areas `loess` is very popular.

- (Generalized) additive models (GAMs): combine the above. Sum of (possibly) non-linear instead of linear functions.




Piecewise Cubic
Continuous Piecewise Cubic
Cubic Spline
Linear Spline
Wage
Age
50
100
150
200
250
20
30
40
50
60
70

8. Tree-based methods

- Method applicable both to regression and classification (K classes) and will give non-linear covariate effects and include interactions between covariates.
- A tree can also be seen as a division of the covariate space into non-overlapping regions.
- Binary splits using only at the current best split: *greedy strategy*.
- Minimization criterion: residual sums of squares (RSS), Gini index or cross-entropy.
- Stopping criterion: When to stop: decided stopping criterion - like minimal decrease in RSS or less than 10 observations in terminal node.
- Prediction:
  - Regression: Mean in box R_j
  - Classification: Majority vote or cut-off on probabiity.

- *Pruning*: Grow full tree, and then prune back using pruning strategy: cost complexity pruning.

To improve prediction (but worse interpretation):

- *Bagging* (bootstrap aggregation): draw B bootstrap samples and fit one full tree to each, used the average over all trees for prediction.
- *Random forest*: as bagging but only m (randomly) chosen covariates (out of the p) are available for selection at each possible split.
- Out-of-bag estimation can be used for model selection - no need for cross-validation.
- Variable importance plots.
- *Boosting*: fit one tree with d splits, make residuals and fit a new tree, adjust residuals partly with new tree - repeat.

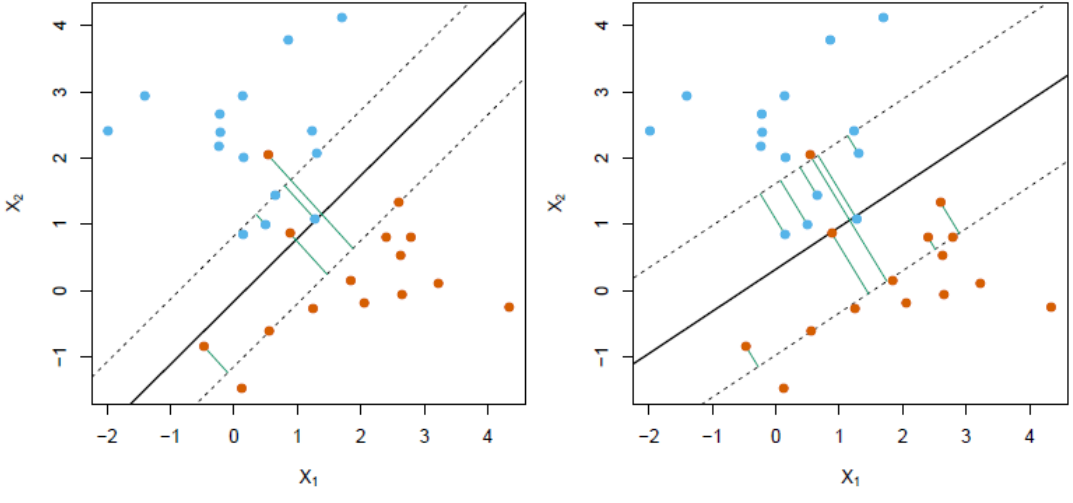
Making trees competitive!

overtaken by NN

9. Support vector machines

- SVM can be used both classification and regression, but we have only studied two-class classification.
- Aim: find high dimensional hyperplane that separates two classes $f(x) = \beta_0 + x^T\beta = 0$. If $y_i f(x_i) > 0$ observation x_i is correctly classified.

- Maximizing the distance (on both sides) from the class boundary to the closes observations (the margin M).
- Relaxed with slack variables (support vector classifiers), and to allow nonlinear functions of x (inner products).
- Support vectors: observations that lie on the margin or on the wrong side of the margin.



- Kernels: generalization of an inner product to allow for non-linear boundaries
- Most popular kernel is radial

defining a distance in ∞-dim space

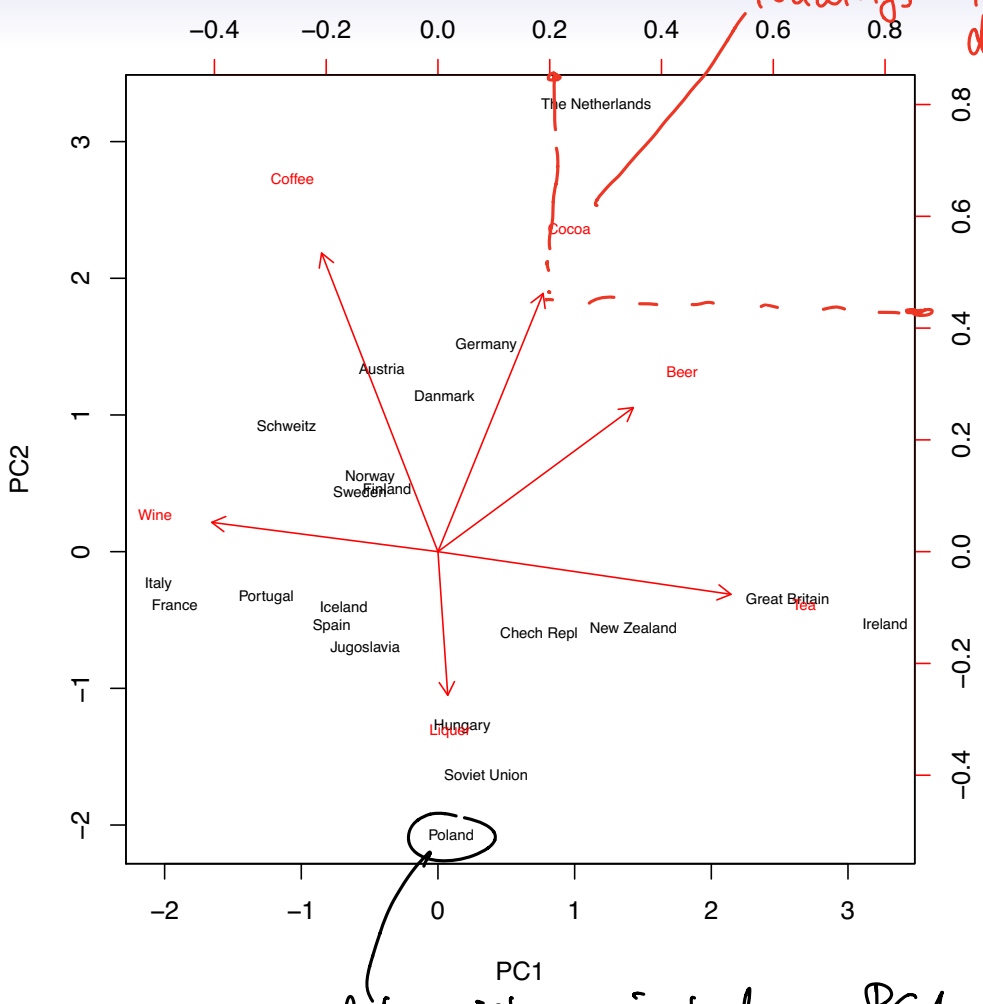
$$K(x_i, x_i') = \exp(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2) \ .$$

- Tuning parameters: cost and parameters in kernels - chosen by CV.
- Unfortunately not able to present details since then a course in optimization is needed.
- *SVMs have been overtaken largely by neural networks and tree-based methods.*

10. Unsupervised learning

- Principal component analysis:
  - Mathematical details (eigenvectors corresponding to covariance or correlation matrix) also in TMA4267.
  - Understanding loadings, scores and the biplot, choosing the number of principal components from proportion of variance explained or scree-type plots (elbow).
- Clustering:
  - k-means: number of clusters given, iterative algorithm to classify to nearest centroid and recalculate centroid
  - Hierarchical clustering: choice of distance measure, choice of linkage method (single, average, complete),

Biplot (PCA)

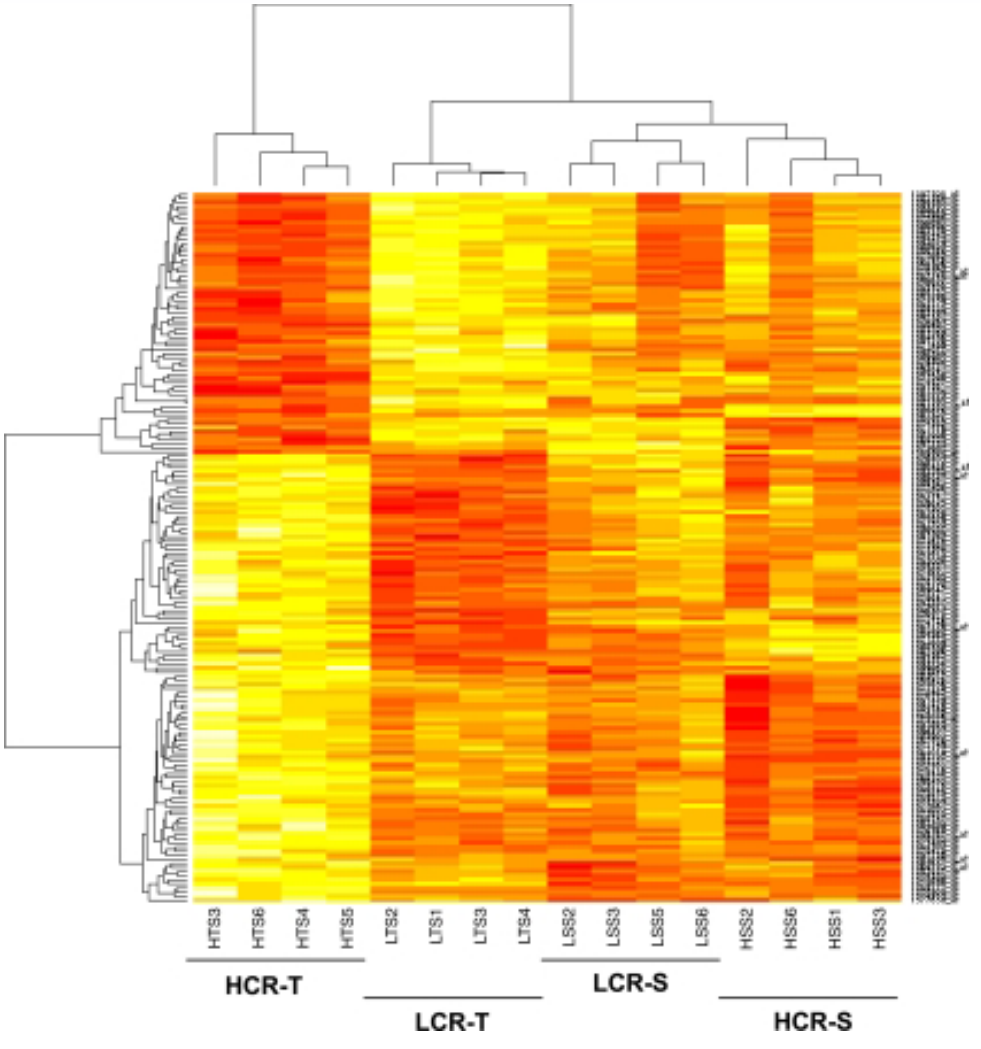


loadings: How much does this covariate contribute to PC1 / PC2

Look at Comp. 3 from 2020

data points, projected on PC1, PC2

Hierarchical clustering for visualization



11. Neural networks

- Feedforward network architecture: mathematical formula - layers of multivariate transformed (`relu`, `linear`, `sigmoid`) inner products - sequentially connected.
- Loss function to minimize (on output layer): regression (mean squared), classification binary (binary crossentropy), classification multiple classes (categorical crossentropy).
- Remember the correct choice of output activiation function: mean squared loss goes with linear activation, binary crossentropy with sigmoid, categorical crossentropy with softmax.
- Gradient based (chain rule) back-propagation - many variants.
- Technicalities: `nnet` in R
- `keras` in R. Use of tensors: Piping sequential layers, piping to estimation and then to evaluation (metrics).

Zoom-poll

Some cautionary words

- In most of the problems we looked at we could (or had to) choose a set of variables to explain or predict an outcome (y).
- Model selection was the topic of Module 6, but there is more to say about it, in particular in the regression context.
- Importantly, the approach to find a model **heavily depends on the aim** for which the model is built.

It is important to make the following distinction:

- The aim is to *predict* future values of y from known regressors.
- The aim is to *explain* y using known regressors. In this case, the ultimate aim is to find *causal relationships*.

→ Even among statisticians there is no real consensus about how, if, or when to select a model:

Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2016, **7**, 679–692 doi: 10.1111/2041-210X.12541

SPECIAL FEATURE: 5TH ANNIVERSARY OF *METHODS IN ECOLOGY AND EVOLUTION*

The relative performance of AIC, AIC_C and BIC in the presence of unobserved heterogeneity

Mark J. Brewer[1]*, Adam Butler[2] and Susan L. Cooksley[3]

[1]*Biomathematics and Statistics Scotland, Craigiebuckler, Aberdeen, AB15 8QH,UK;* [2]*Biomathematics and Statistics Scotland, JCMB, The King's Buildings, Edinburgh, EH9 3JZ,UK; and* [3]*The James Hutton Institute, Craigiebuckler, Aberdeen, AB15 8QH, UK*

Summary

1. Model selection is difficult. Even in the apparently straightforward case of choosing between standard linear regression models, there does not yet appear to be consensus in the statistical ecology literature as to the right approach.

Note: The first sentence of a paper in *Methods in Ecology and Evolution* from 2016 is: "Model selection is difficult."

Why is finding a model so hard?

A model is an approximation of the reality. The aim of statistics and data analysis is to find connections (explanations or predictions) thanks to simplifications of the real world.

Box (1979): *"All models are wrong, but some are useful."*

→ There is often not a "right" or a "wrong" model – but there are more and less useful ones.

→ Finding a model or the appropriate method with good properties is sometimes an art...

Predictive and explanatory models

When choosing a method or a model, you need to be clear about the scope:

- **Predictive models**: These are models that aim to predict the outcome of future subjects.

  Example: In the bodyfat example (module 3) the aim is to predict people's bodyfat from factors that are easy to measure (age, BMI, weight,..).

- **Explanatory models**: These are models that aim at understanding the (causal) relationship between covariates and the response.

  Example: The South African heart disease data (module 4) aims to identify important risk factors for coronary heart disease.

→ **The model selection strategy depends on this distinction.**

Prediction vs explanation

When the aim is ***prediction***, the best model is the one that best predicts the fate of a future subject (smallest test error rate). This is a well defined task and "objective" variable selection strategies to find the model which is best in this sense are potentially useful.

However, when used for ***explanation*** the best model will depend on the scientific question being asked, **and automatic variable selection strategies have no place.**

Chapters 27.1 and 27.2 in Clayton and Hills (1993)

AIC, BIC minimization, ... forward, backward...

This is missing in course book

Model selection with AIC, AIC_c, BIC, C_p, adjusted R^2

Given m potential variables to be included in a model. Remember from Module 6:

- Subset selection using forward, backward or best subset selection method.
- Use an "objective" criterion to find the "best" model.

Cautionary Note:
The coefficients of such an optimized "best" model should *not be interpreted* in a causal sense! Why?

- Subset selection may lead to **biased parameter estimates**, thus **do not draw (biological, medical,..) conclusions** from models that were optimized for prediction, for example by AIC/AICc/BIC minimization! See, e.g., Freedman (1983), Copas (1983).

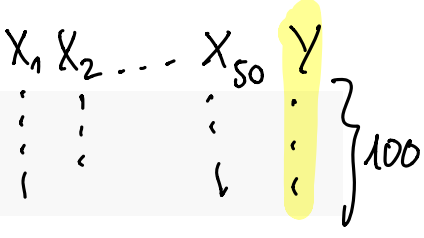
Illustration: Model selection bias

Aim of the example: To illustrate how model selection purely based on AIC can lead to biased parameters and overestimated effects.

Procedure:

1. Randomly generate 100 data points for 50 covariables $x^{(1)}, \ldots, x^{(50)}$ and a response y:

```
set.seed(123456)
data_aic <- data.frame(matrix(rnorm(51 * 100), ncol = 51))
names(data_aic)[51] <- "Y"
```



`data` is a 100× 51 matrix, where the last column is the response. The **data were generated completely independently**, the covariates do not have any explanatory power for the response!

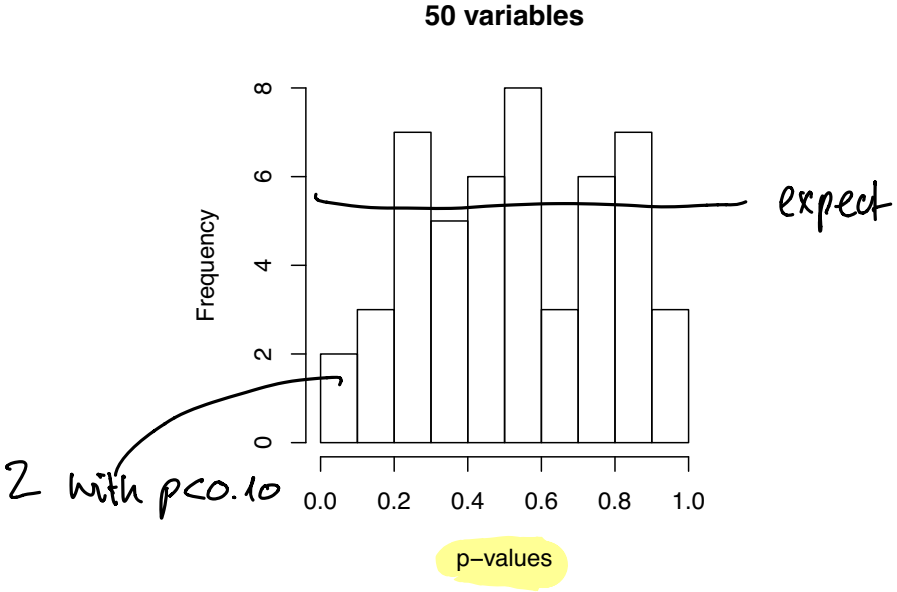
2. Fit a linear regression model of y against all the 50 variables

$$y_i = \beta_0 + \beta_1 x_i^{(1)} + ... + \beta_{50} x_i^{(50)} + \epsilon_i \ .$$

```
r.lm.aic <- lm(Y ~ ., data_aic)
```

As expected, the distribution of the p-values is (more or less) uniform between 0 and 1, with none below 0.05:

```
hist(summary(r.lm.aic)$coef[-1, 4], freq = T, main = "50 variables",
    xlab = "p-values")
```

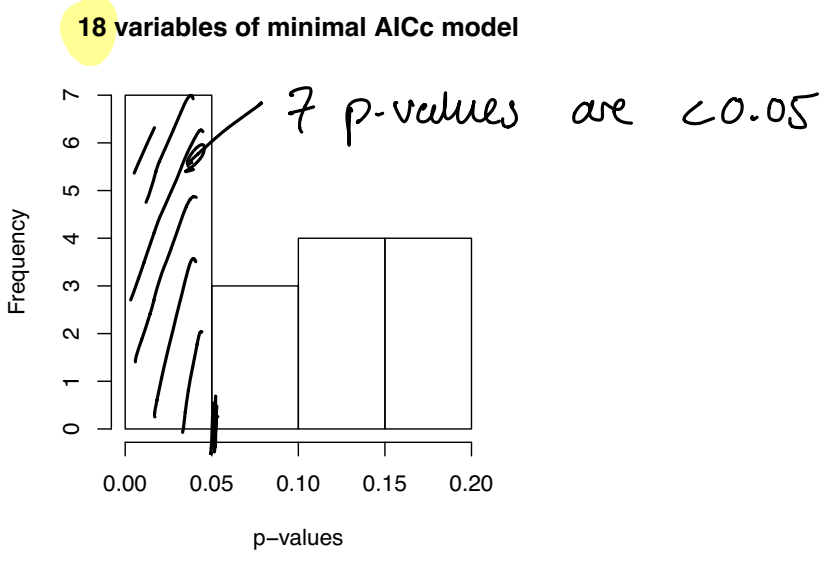


3. Then use AICc minimization to obtain the objectively "best" model:

Systematic, automatic

```
library(MASS)
r.AICmin <- stepAIC(r.lm.aic, direction = c("both"), trace = FALSE, AICc = TRUE)
```

```
hist(summary(r.AICmin)$coef[-1, 4], freq = T, main = "18 variables of minimal AICc model",
    xlab = "p-values")
```



The distribution of the p-values is now skewed: many of them reach rather small values (7 have $p < 0.05$). This happened *although none of the variables has any explanatory power!*

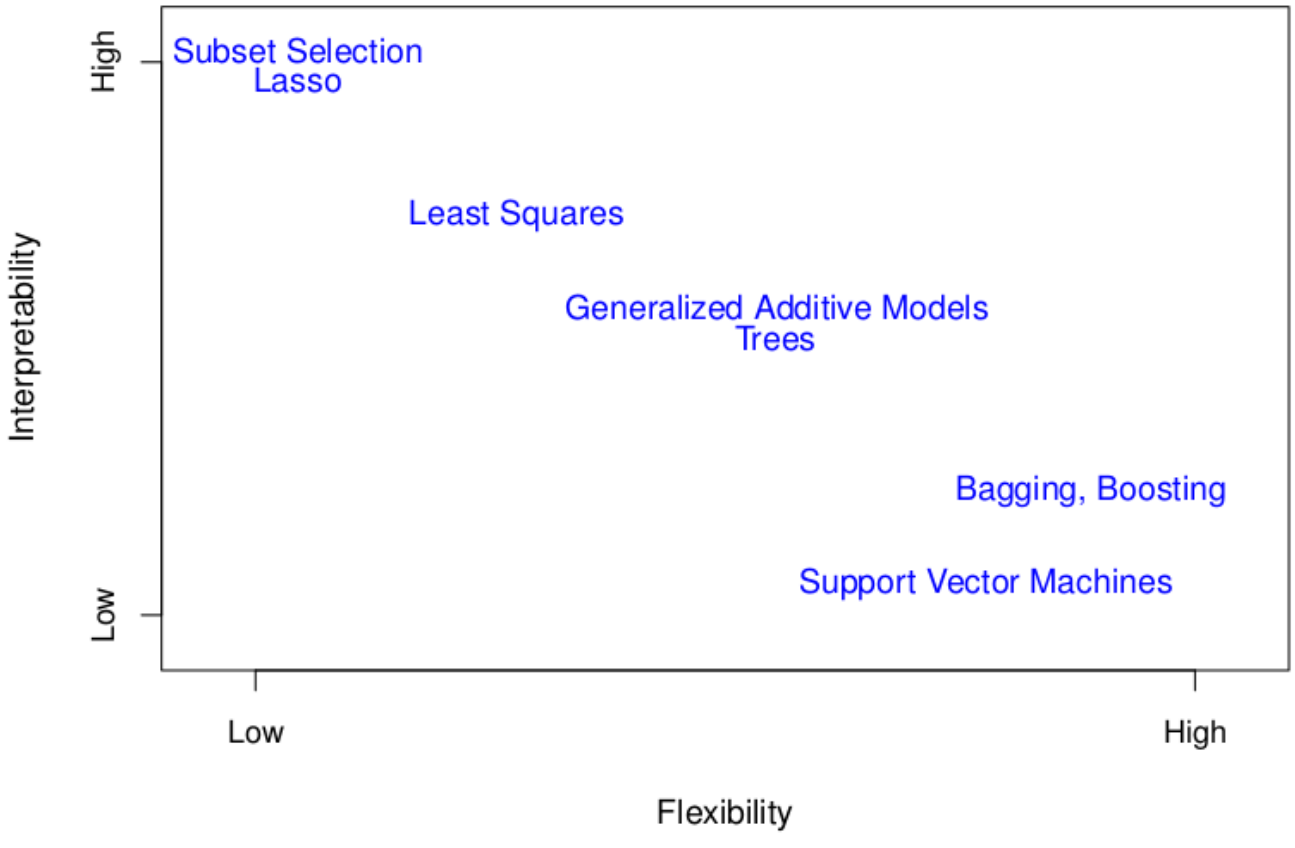
Summary: Main problem with model selection

When model selection is carried out based on objective criteria, the effect sizes will to be too large and the uncertainty too small. So you end up being too sure about a too large effect.

Exception: Shrinkage methods!

In particular Lasso

Which methods are suitable for explanation (inference)?



Bottomline

- *Less flexible* models tend to have *better interpretability* and are therefore better suited *for explanation* than more flexible models.
- The *more flexible* a model, the better it tends to perform *for prediction*.
- Some models can be used for both, prediction and explanation, but then the approach how to build the model (e.g., how to select variables) should still depend on the aim.

The exam

- **Digital home exam** May 26th, 9-13h. +30'

Tentative plan:

- Warm up section
- Two examples with real data
- A conceptual/theoretical section testing your understanding.
- A section with single/multiple choice questions

Thank you for attending this course - good luck for the exam and let's hope we are soon back into normal post-Covid teaching mode!

References

Clayton, D., and M. Hills. 1993. *Statistical Models in Epidemiology.* Oxford: Oxford University Press.

Copas, J. B. 1983. "Regression, Prediction and Shrinkage." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 45: 311–54.

Freedman, D. A. 1983. "A Note on Screening Regression Equations." *The American Statistician* 37: 152–55.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning.* Vol. 112. Springer.

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. *An Introduction to Statistical Learning with Applications in R.* New York: Springer.