Module 12: Summing up and some cautionary notes

TMA4268 Statistical Learning V2020

Stefanie Muff, Department of Mathematical Sciences, NTNU

April 17, 2020

Last update: April 17, 2020

Overview

- Course content and learning outcome
- Overview of modules and core course topics (with exam type questions)
- 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? flexiblity/overfitting/underfitting.
  - the bias-variance trade-off
  - how to find the perfect fit - validation and cross-validation (or AIC-type solutions)
  - how to compare different solutions
  - how to evaluate the fit - on new unseen data

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

and of cause all **the methods** (with underlying models) to perform regression, classification and unsupervised learning. We have gained some theoretical understanding, but in some cases deeper theoretical background and understanding of the models is provided in other statistics courses.

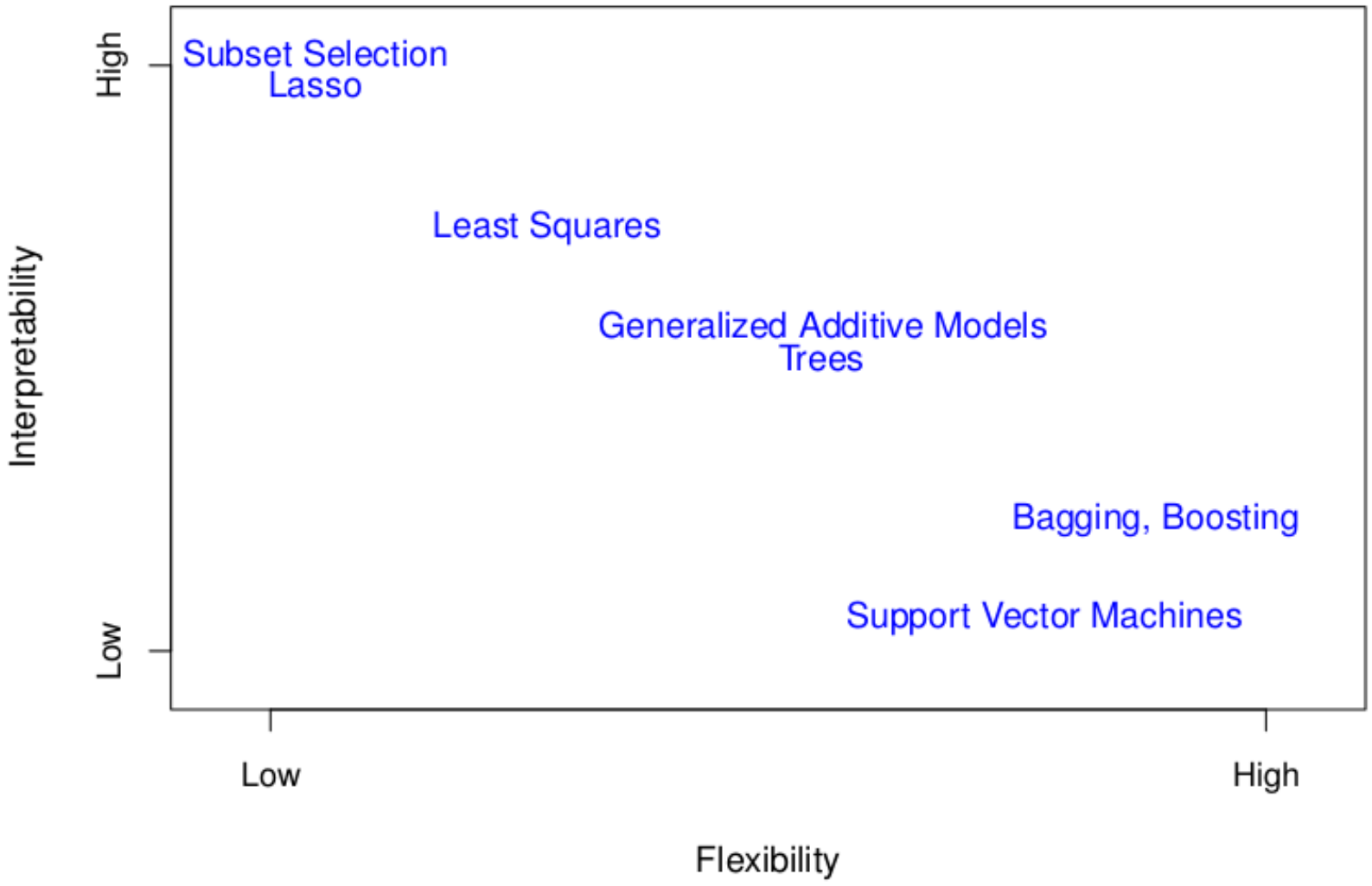


Figure 2.7 from Gareth James et al. (2013)

The modules

1. Introduction

- Examples, the modules, required background in statistics and
- Introduction to R

2. Statistical learning

- Model complexity
  - Prediction vs. interpretation.
  - Parametric vs. nonparametric.
  - Inflexible vs. flexible.
  - Overfitting vs. underfitting
- Supervised vs. unsupervised.
- Regression and classification.
- Loss functions: quadratic and 0/1 loss.
- 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.
- Model complexity and the bias-variance trade-off is important in "all" subsequent modules.

3. Linear regression

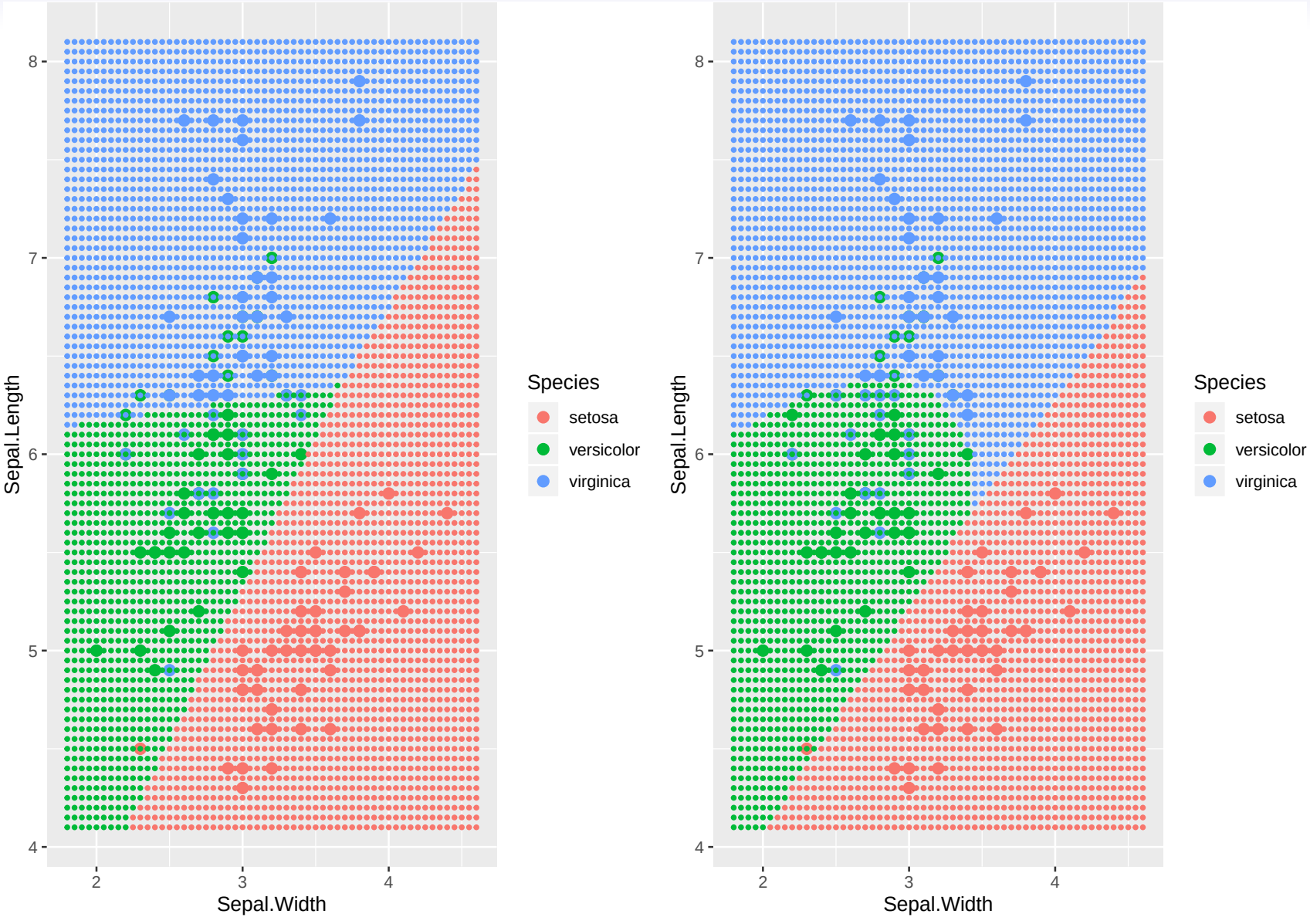
- The classical normal linear regression model on vector/matrix form.
- Parameter estimators and distribution thereof. Model fit.
- Confidence intervals, hypothesis tests, and interpreting R-output from regression.
- Qualitative covariates, interactions.
- This module is a stepping stone for all subsequent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

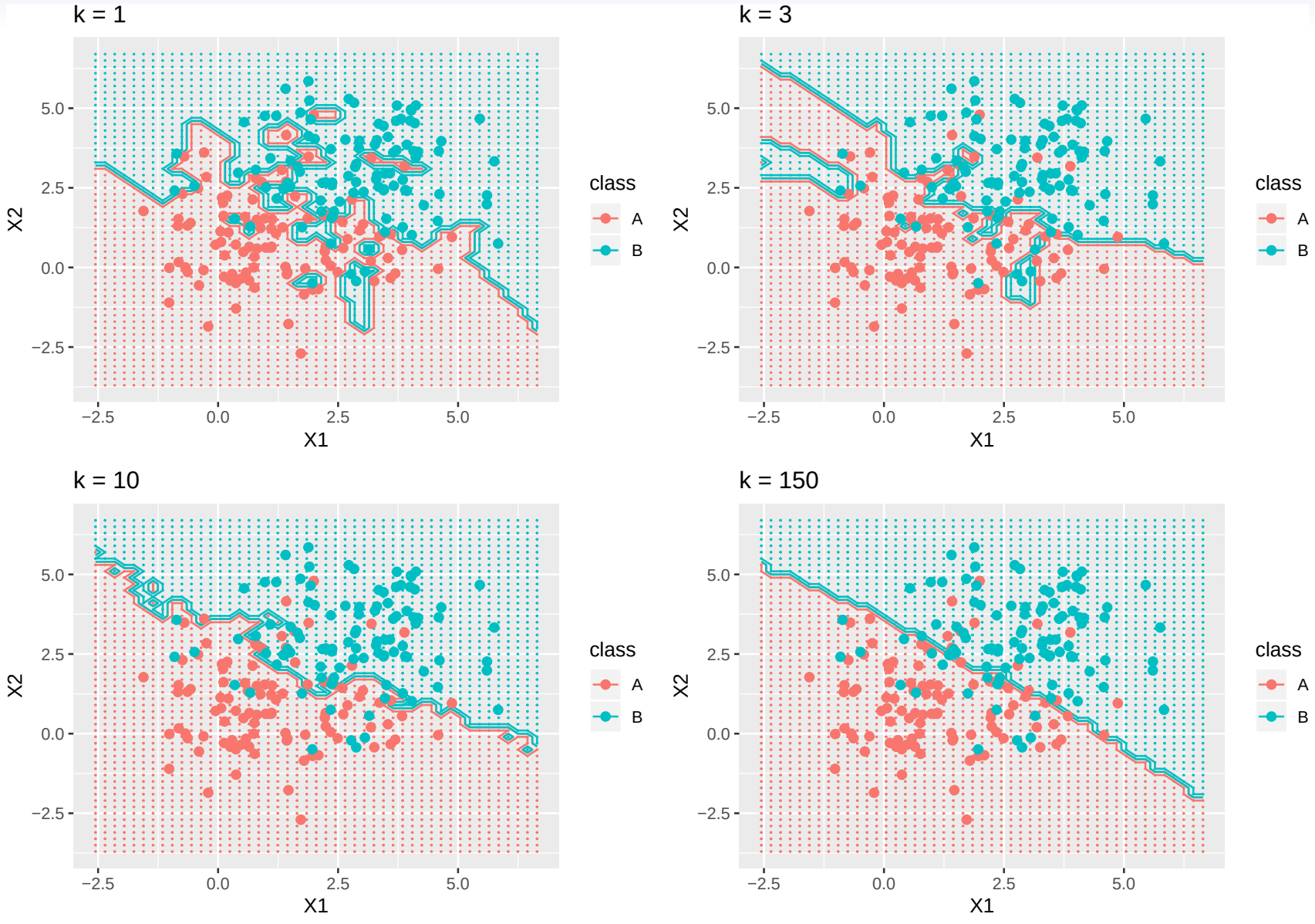
4. Classification (Mainly two-class problems)

- Bayes classifier: classify to the most probable class gives the 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.

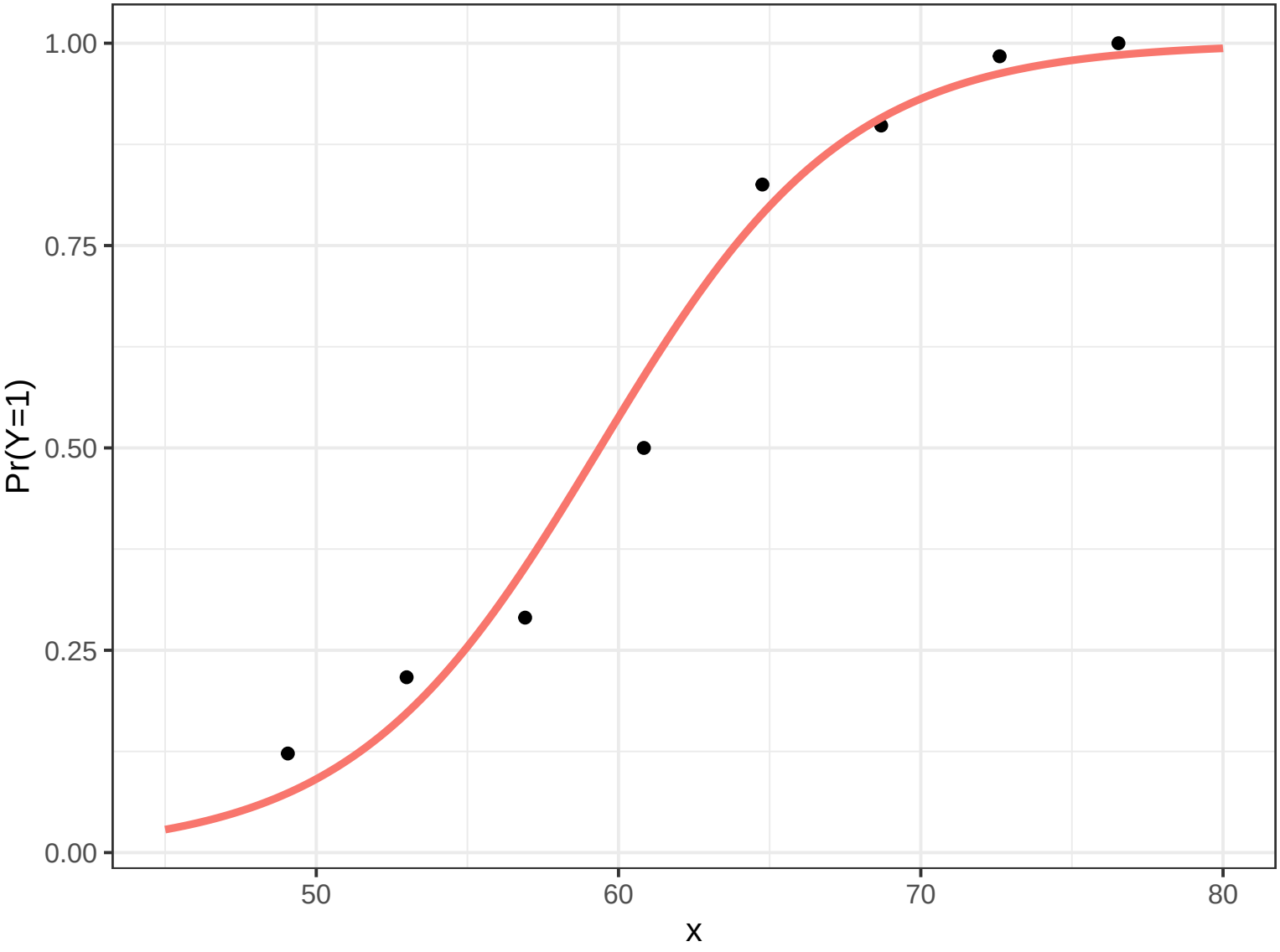
- Two paradigms (not in textbook):
  - *Diagnostic* (directly estimating the posterior distribution for the classes). Example: KNN classifier, logistic regression.
  - *Sampling* (estimating class prior probabilities and class conditional distribution and then putting together with Bayes rule). Examples: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

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


Sepal.Length
8
7
6
5
4
2
3
4
Sepal.Width
Species
setosa
versicolor
virginica
Sepal.Length
8
7
6
5
4
2
3
4
Sepal.Width
Species
setosa
versicolor
virginica


k = 1
k = 3
k = 10
k = 150
X1
X2
class
A
B
5.0
2.5
0.0
−2.5

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

The Bootstrap

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

6. Linear model selection and regularization:

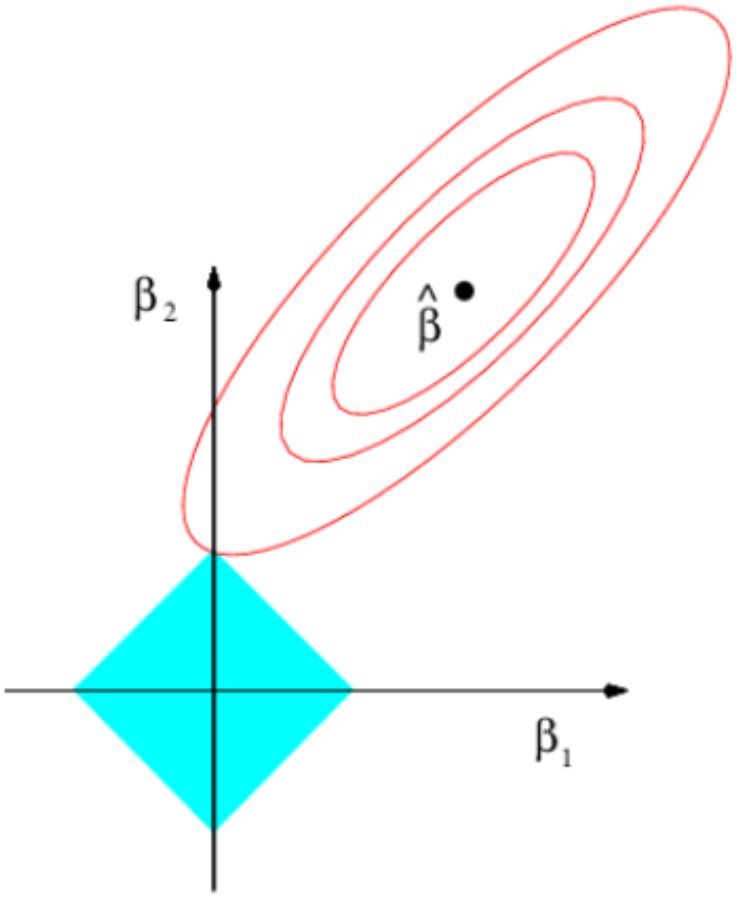
Subset-selection. Discriminate:

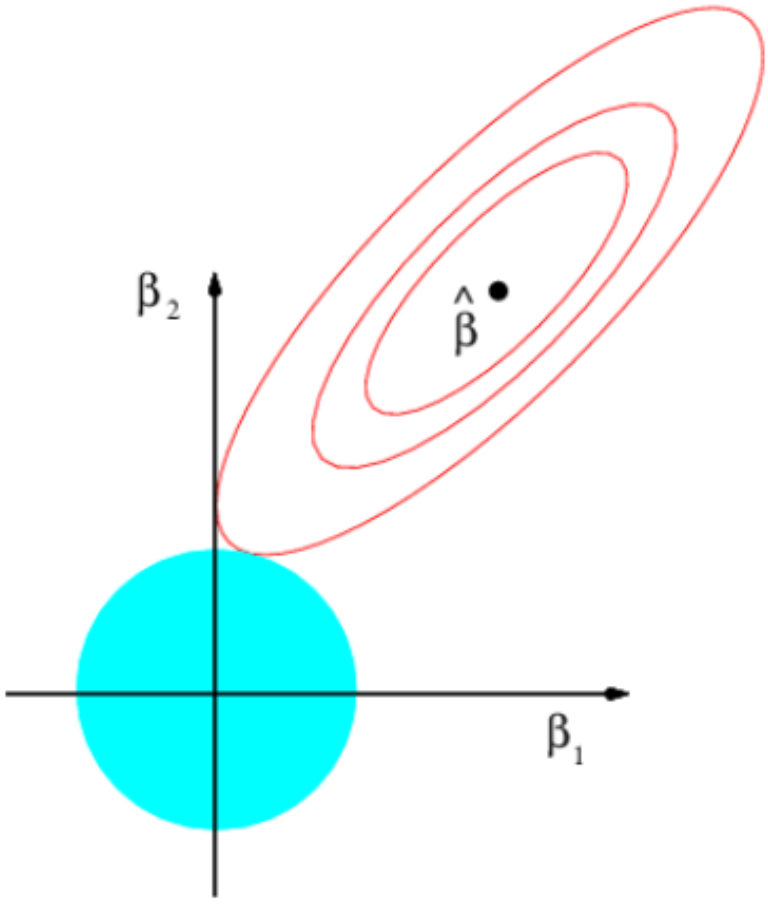
- *Model selection*: estimate performance of different models to choose the best one.
- *Model assessment*: having chosen a final model, estimate its performance on new data.

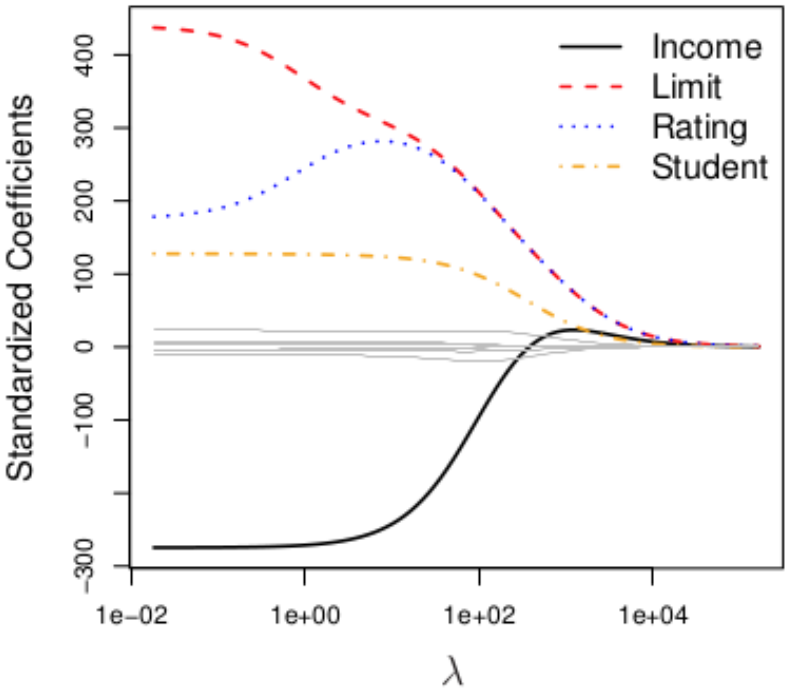
How?

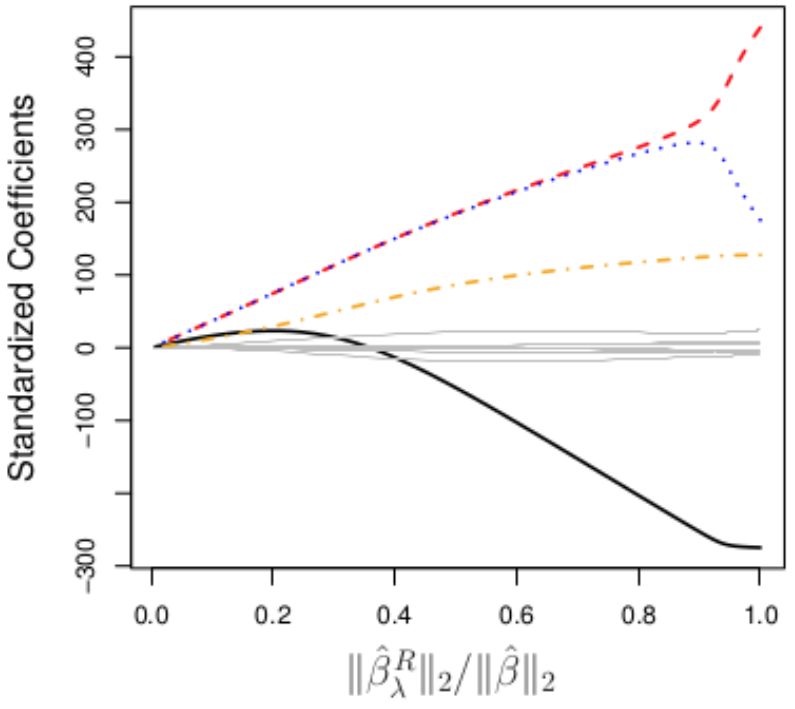
- Model selection by
  - Subset selection (best subset selection or stepwise model selection)
  - Penalizing the training error: AIC, BIC, C_p, Adjusted R^2.
  - Cross-validation.
- Model assessment by
  - Cross-validation.

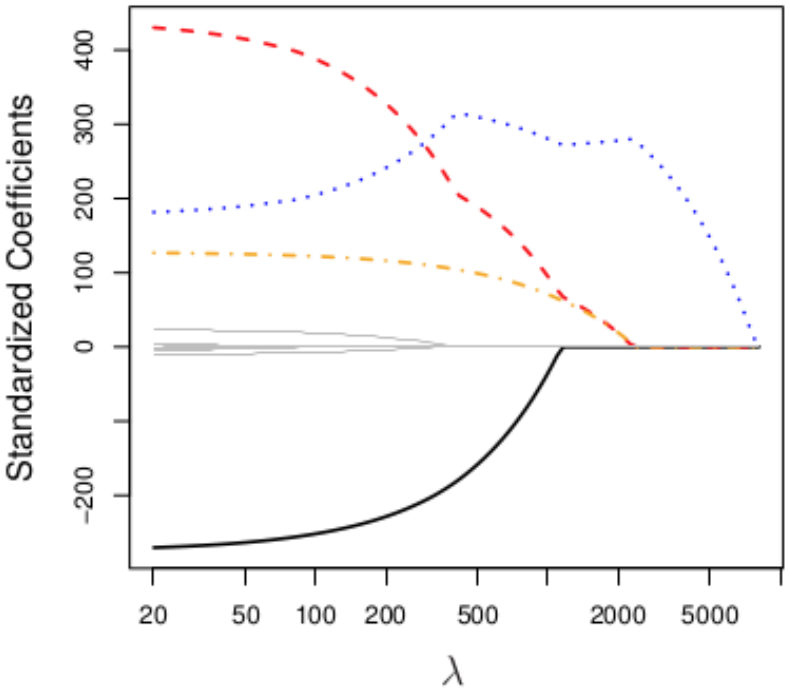
- Shrinkage methods
  - ridge regression: quadratic L2 penalty added to RSS
  - lasso regression: absolute L1 penalty added to RSS
  - no penalty on intercept, not scale invariant: center and scale covariates
- Dimension reduction methods:
  - principal component analysis: eigenvectors, proportion of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

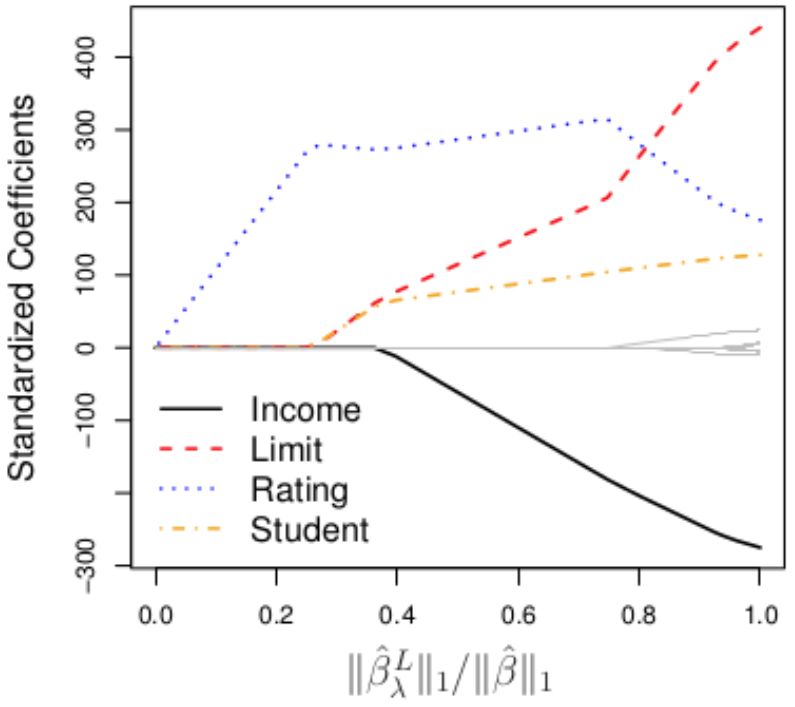

β_2
$\hat{\beta}$
β_1


β_2
$\hat{\beta}$
β_1


Standardized Coefficients
400
300
200
100
0
−100
−300
1e−02
1e+00
1e+02
1e+04
λ
Income
Limit
Rating
Student


Standardized Coefficients
400
300
200
100
0
−100
−300
0.0
0.2
0.4
0.6
0.8
1.0
$\|\hat{\beta}_\lambda^R\|_2/\|\hat{\beta}\|_2$

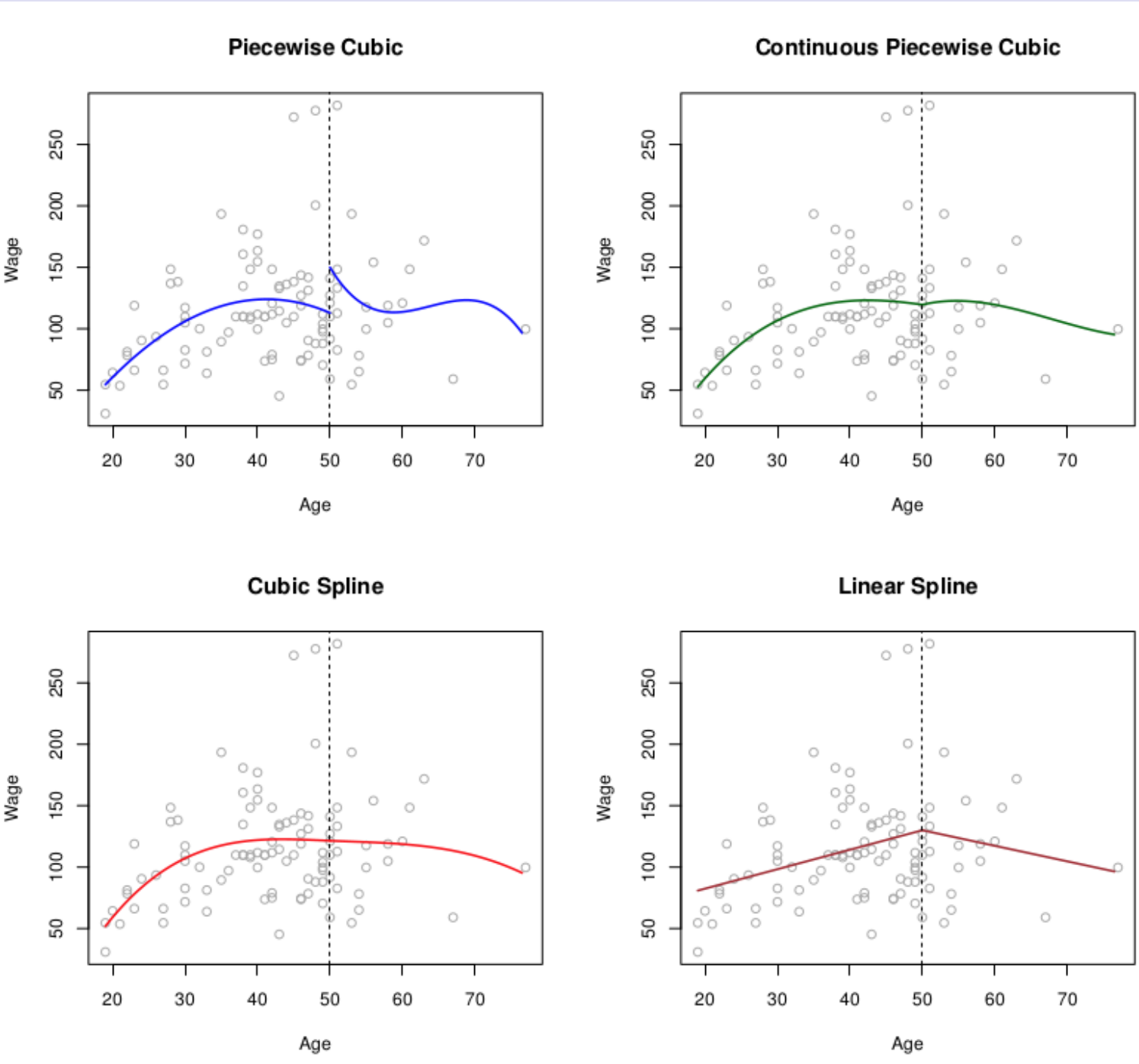

Standardized Coefficients
400
300
200
100
0
−200
20
50
100
200
500
2000
5000
λ


Standardized Coefficients
400
300
200
100
0
−100
−300
Income
Limit
Rating
Student
0.0
0.2
0.4
0.6
0.8
1.0
$\|\hat{\beta}_\lambda^L\|_1/\|\hat{\beta}\|_1$

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 (no parameters to estimate).
- Polynomial regression: multiple linear regression with polynomials as basis functions.
- Step functions - piece-wise constants. Like our dummy variable coding of factors.
- Regression splines: regional polynomials joined smoothly - neat use of basis functions. Cubic splines very popular.

- 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
Wage
Age
Continuous Piecewise Cubic
Wage
Age
Cubic Spline
Wage
Age
Linear Spline
Wage
Age

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. Rule of thumb for m is $\sqrt{p}$ for classificaton and $p/3$ for regression.
- Out-of-bag estimation can be used for model selection - no need for cross-validation.
- Variable importance plots: give the total amount of decrease in RSS or Gini index over splits of a predictor - averaged over all trees.
- *Boosting*: fit one tree with d splits, make residuals and fit a new tree, adjust residuals partly with new tree - repeat.

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.
- Central: maximizing the distance (on both sides) from the class boundary to the closes observations (the margin M). This was relaxed with slack variables (support vector classifiers), and to allow nonlinear functions of x by extending an inner product to kernels (support vector machine).
- 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 and to speed up calculations due to inner products only involve support vectors. Most popular kernel is radial

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

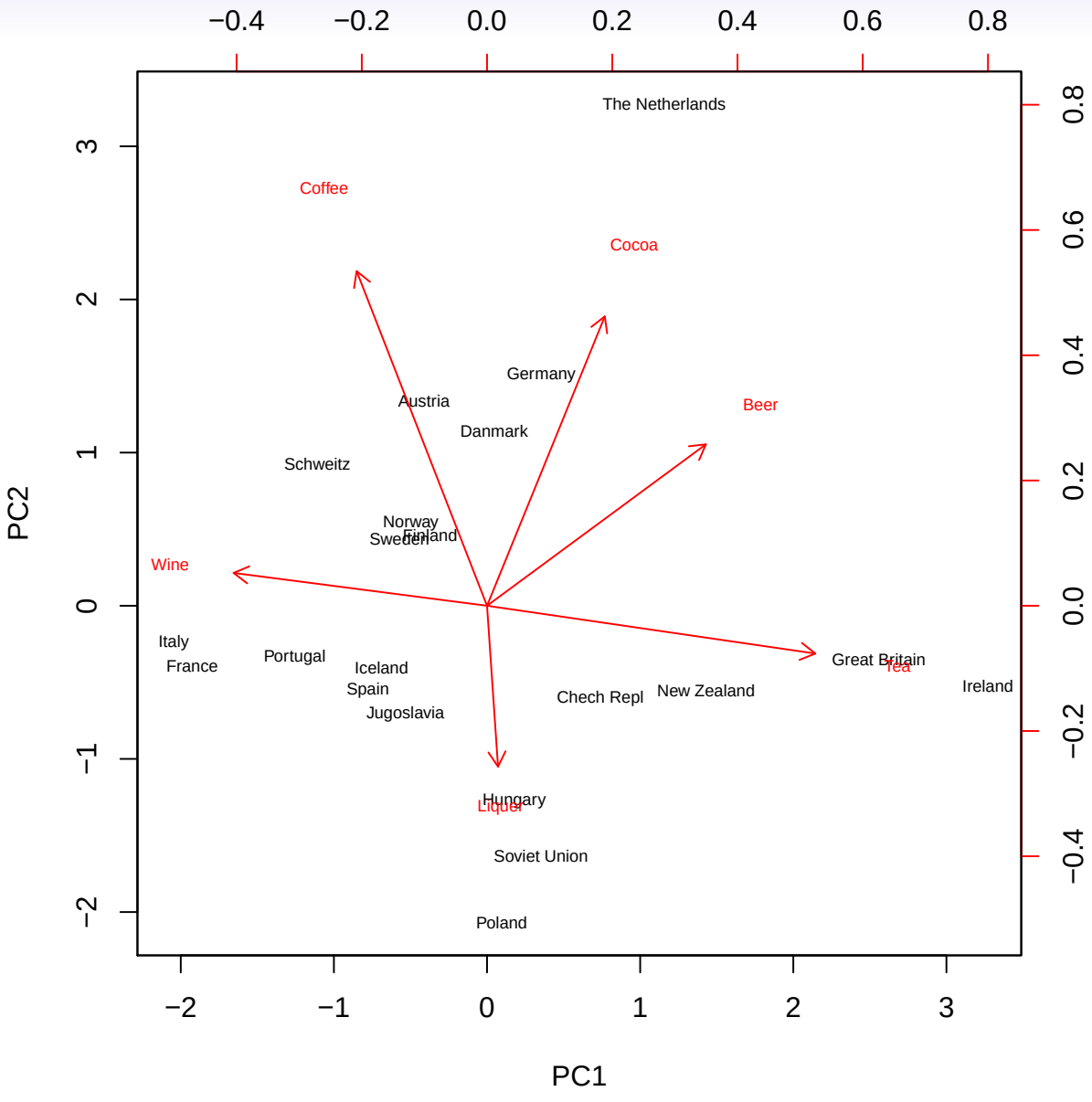
- Tuning parameters: cost and parameters in kernels - chosen by CV.

- Unfortunately not able to present details since then a course in optimization is needed.

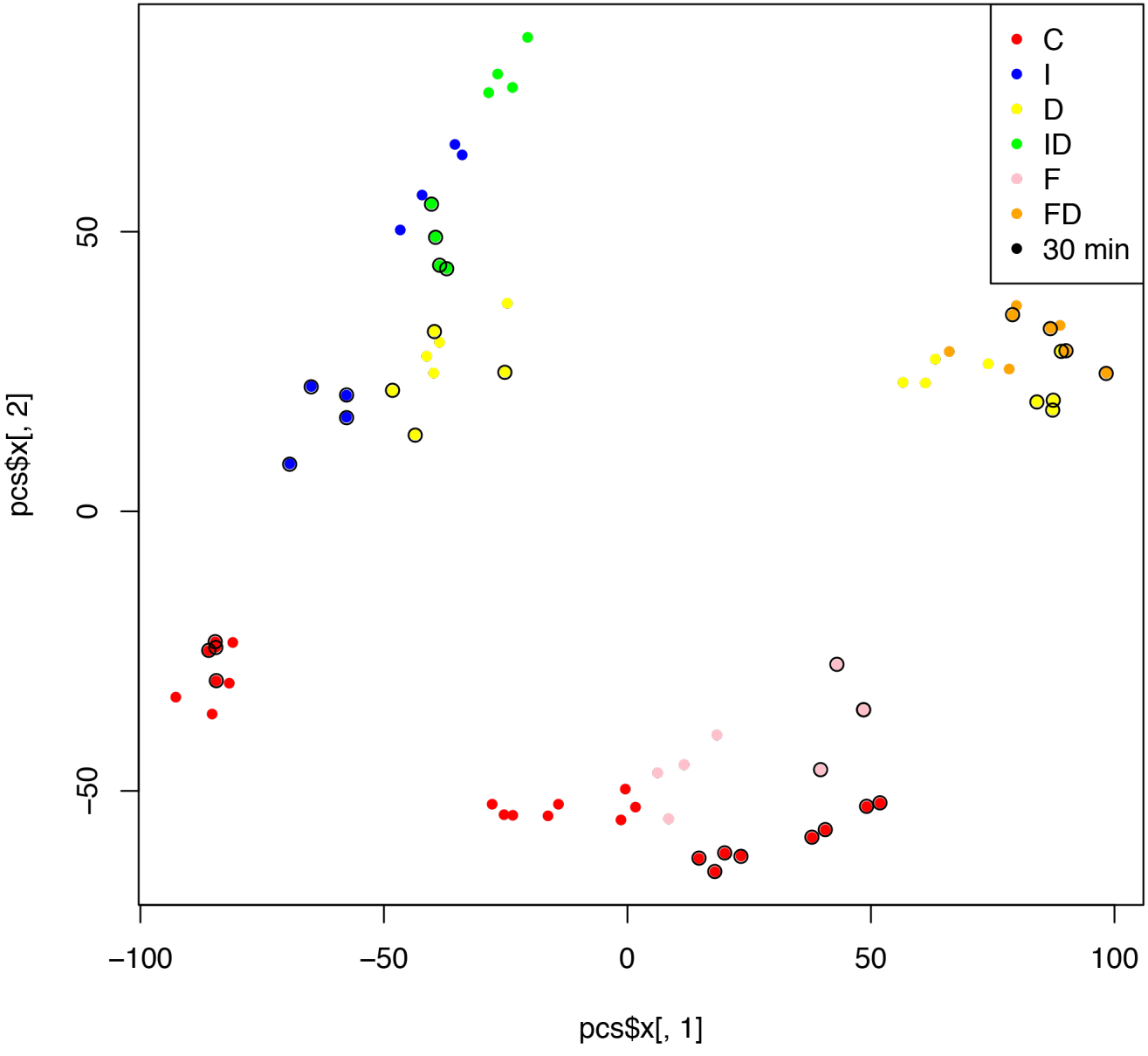
- Nice connection to non-linar and ridged version of logistic regression - comparing hinge loss to logistic loss - but then without the computational advanges of the kernel method.

10. Unsupervised learning

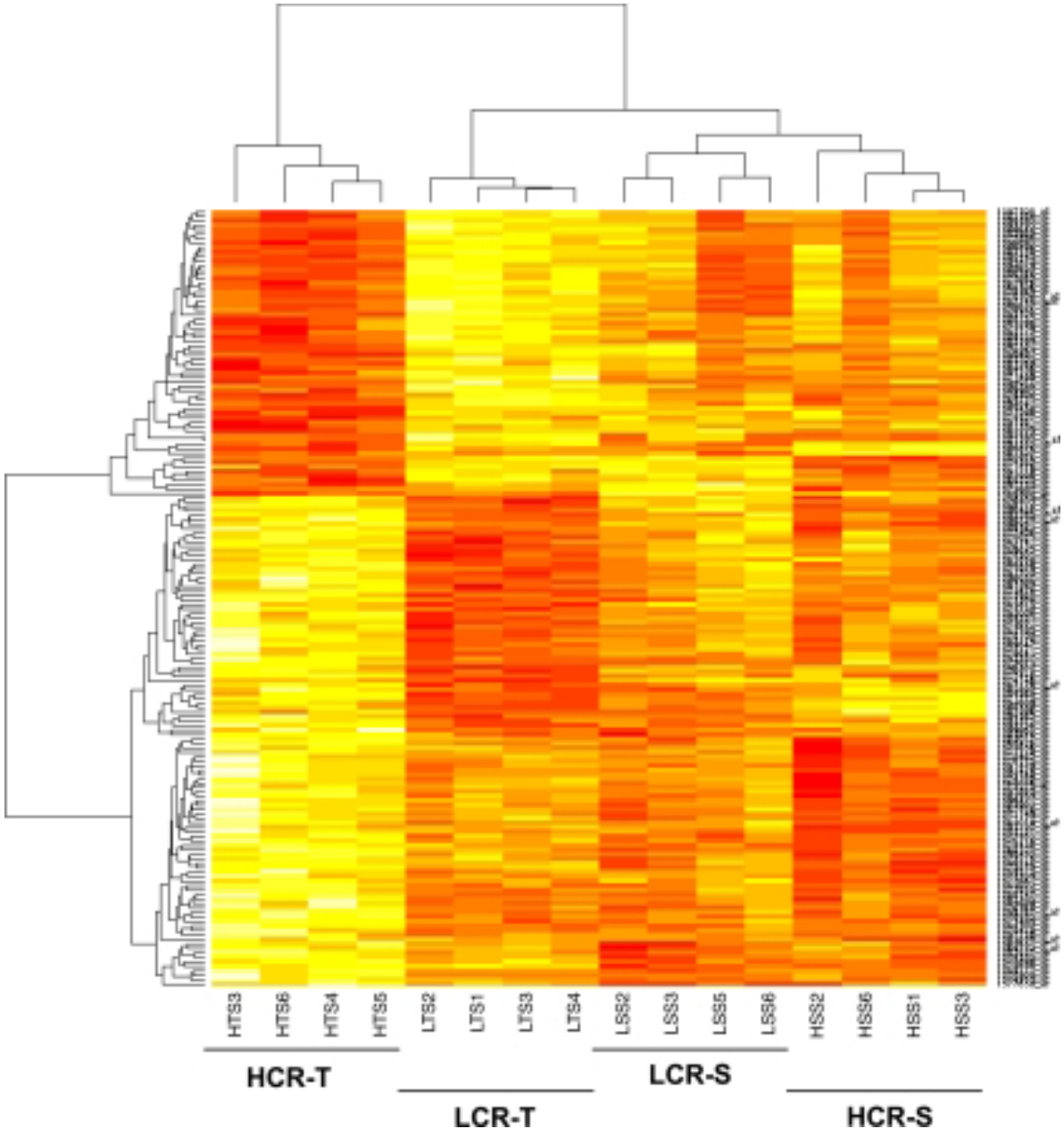
- Principal component analysis:
  - Mathematical details (eigenvectors corresponding to covariance or correlation matrix) also in TMA4267.
  - Understanding loadings, scores and the biplot, choosing the number of principal components from proportion of variance explained or scree-type plots (elbow).
- 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),


−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
The Netherlands
Coffee
Cocoa
Germany
Austria
Beer
Danmark
Schweitz
Norway
Sweden
Finland
Wine
Italy
France
Portugal
Iceland
Spain
Jugoslavia
Chech Repl
New Zealand
Great Britain
Tea
Ireland
Hungary
Liquer
Soviet Union
Poland
PC2
PC1
3
2
1
0
−1
−2
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

PCA for quality control



Hierarchical clustering for visualization



11. Neural networks

Topics in Module 11:

- Feedforward network architecture: mathematical formula - layers of multivariate transformed (`relu`, `linear`, `sigmoid`) inner products - sequentially connected.
- What is the number of parameters that need to be estimated? Intercept term (for each layer) is possible and is referred to as "bias term".
- Loss function to minimize (on output layer): regression (mean squared), classification binary (binary crossentropy), classification multiple classes (categorical crossentropy) — and remember to connect to the correct choice of output activiation function: mean squared loss goes with linear activation, binary crossentropy with sigmoid, categorical crossentropy with softmax.
- How to minimize the loss function: gradient based (chain rule) back-propagation - many variants.

- Technicalities: `nnet` in R
- Optional (not on reading list): `keras` in R. Use of tensors. Piping sequential layers, piping to estimation and then to evaluation (metrics).

After TMA4268 - what is next?

What are the statistical challenges we have not covered?

Do you want to learn more about the methods we have looked at in this course? And also methods that are more tailored towards specific types of data? Then we have many statistics courses that you may choose from.

An overview of statistics courses and also information on the statistics staff (for bachelor and master supervision)
https://folk.ntnu.no/mettela/Talks/3klinfo20190325.html

On behalf of the teaching staff - Michail, Andreas, Thiago and Mette-

thank you for attending this course - hope to see you for the exam supervision - and good luck on May 23!

References

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.