www.pei.de

Variable Selection and Biomarker Discovery in High-Dimensional Data Sets

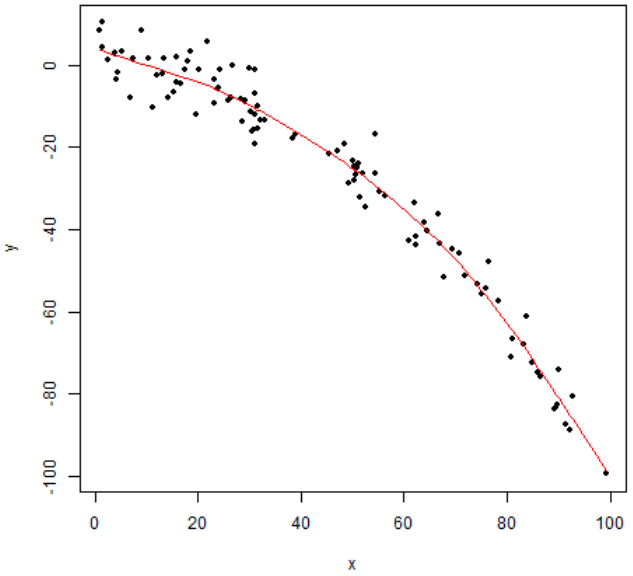
A (rather) Non-Technical Introduction

Dr. Benjamin Hofner
Section Biostatistics

02.09.2016

About me

- Studied Statistics at LMU Munich (Until 2008)
- Working at FAU Erlangen-Nürnberg
- PhD in Statistics at LMU Munich (2011)
- PostDoc in Erlangen
- Since April 2016 at PEI



(Part of) my research:

- Statistical modelling
  - Linear regression models
  - Additive regression models (> smooth effects)
- Variable selection
- Prediction models
- Implementation in open source software packages for R

Introduction

Situation:

- Data set with many *potential* predictors for a single outcome
  (e.g., omics data to predict health status of a patient)

Aims:

- E.g., derive risk profiles / biomarker signatures for patients

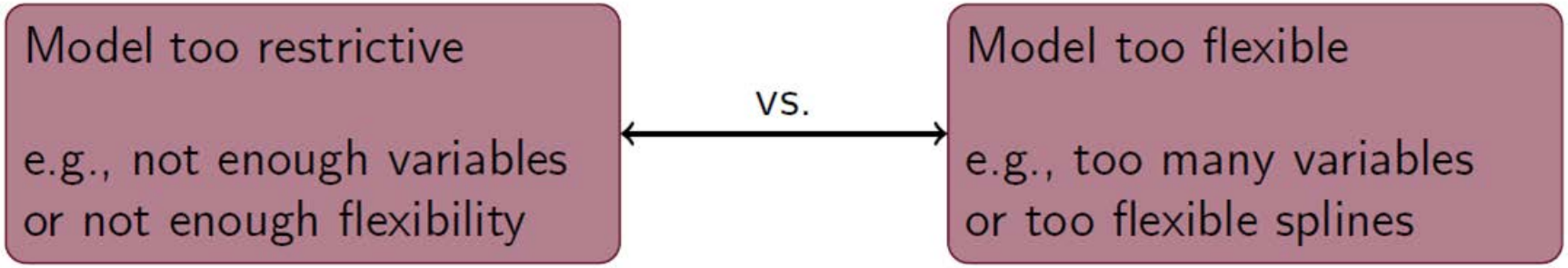
- Find optimal prognostic model,
  i.e., a model that allows to predict the outcome also for new data
  - Variable selection > Identify *relevant* predictors
  - Model choice > Identify type of influence, i.e., linear effect, non-linear effect, jumps, …
- Model should be interpretable,
  i.e., biologists and clinicians should be able to understand and interpret the results

Gene Expression for Diffuse Large-B-Cell Lymphoma

- Predict survival outcome after chemotherapy based on microarray gene-expression profiles of lymphoma samples
- 222 DLBL patients, 7399 genes per patient

Which combination of these genes allows to predict the survival time of patients?

“Underfitting” / Overfitting



- Overly simplistic > Misses “true” effects, e.g. as the variables are not included in the model or due to confounding.
- Overly complex > Captures (amongst others) the “true” effects in the data but cannot be generalized (to new data) and is hard to interpret.
- Find optimal model in between extremes

Stepwise Regression as (Bad) Standard

- Forward / backward stepwise regression is often used to
  - select relevant predictors
  - estimate an interpretable model

- Forward stepwise regression:
  - Estimate all models with one predictor
  - Choose best predictor (e.g. based on p-value)
  - Now add each of the remaining predictors and refit model
  - Choose the second best predictor
  - …

- Backward stepwise regression:
  - Start with all variables and remove worst variables

Gene Expression for Diffuse Large-B-Cell Lymphoma

- Predict survival outcome after chemotherapy based on microarray gene-expression profiles of lymphoma samples
- 222 DLBL patients, 7399 genes per patient

Coefficient estimates

Step	Forward stepwise selection		
1	ID31242		
	1.169		
2	ID31242	ID27774	
	1.149	-0.587	
3	ID31242	ID27774	ID34344
	1.207	-0.556	0.450
⋮	⋮	⋮	⋮

Stepwise Regression as (Bad) Standard

- Problems:
  - Selection of variables very unstable
    (at least for high-dimensional data)
  - Often poor prediction performance
  - Sometimes even impossible to estimate the model
    (if $p > n$ or even $p \gg n$)

Model-based Boosting

- Model estimation via boosting (strongly simplified):
  1) Fit all variables *separately* to the outcome
  2) Select and update only the *best-fitting variable* (only use 10%)
  3) Compute the *residuals* (i.e., unexplained part of the outcome) and repeat process
  - ➢ Do this until no more improvement
- ➢ Model is estimated and variables are selected at the same time

- ➢ Better alternative to stepwise regression
  - ✓ Better variable selection properties
  - ✓ Optimizes prediction accuracy
  - ✓ Interpretable results
  - ✓ Feasible even if p ≫ n (GWAS, Microarrays, …)

Gene Expression for Diffuse Large-B-Cell Lymphoma

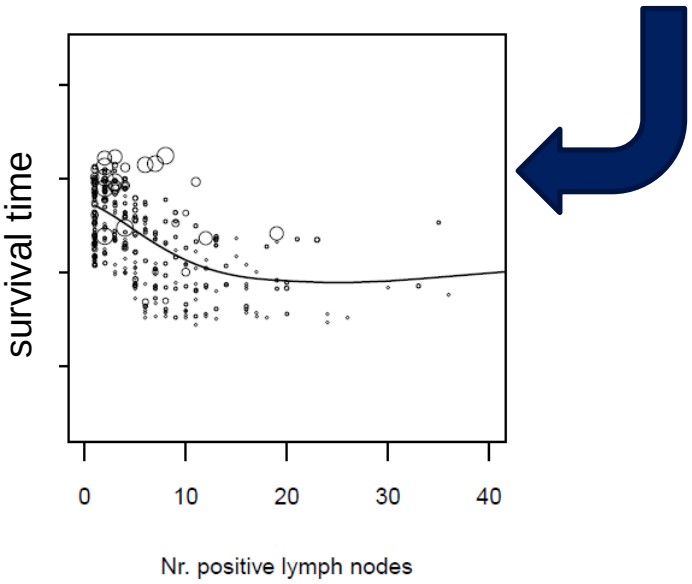
- Predict survival outcome after chemotherapy based on microarray gene-expression profiles of lymphoma samples
- 222 DLBL patients, 7399 genes per patient

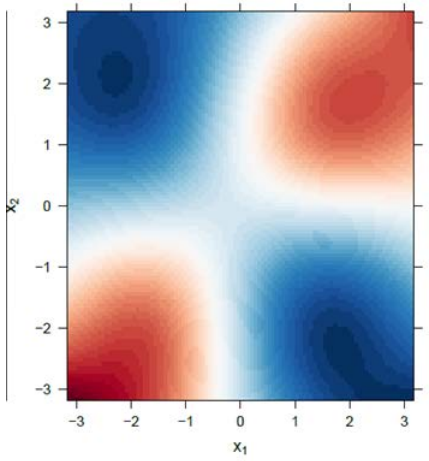
Coefficient estimates

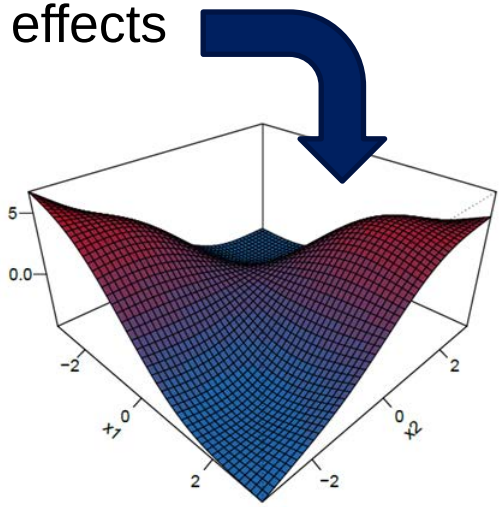
Step	Forward stepwise selection			Boosting		
1	ID31242 1.169			ID27774 -0.030		
2	ID31242 1.149	ID27774 -0.587		ID27774 -0.030	ID31981 0.060	
3	ID31242 1.207	ID27774 -0.556	ID34344 0.450	ID27774 -0.030	ID31981 0.060	ID24376 -0.011
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Further Benefits

- Boosting is highly flexible
  - Defined not only for Gaussian outcome but (almost) any outcome (e.g. binary data, count data, survival data, robust regression methods, …)
  - Defined for non-linear effects or spatial effects







Spatio-temporal variability of Acidobacteria in grassland soil

- Low-intensity managed grassland plot of 10 m x 10 m
- 358 sampling locations
- 6 sampling dates (in one year)

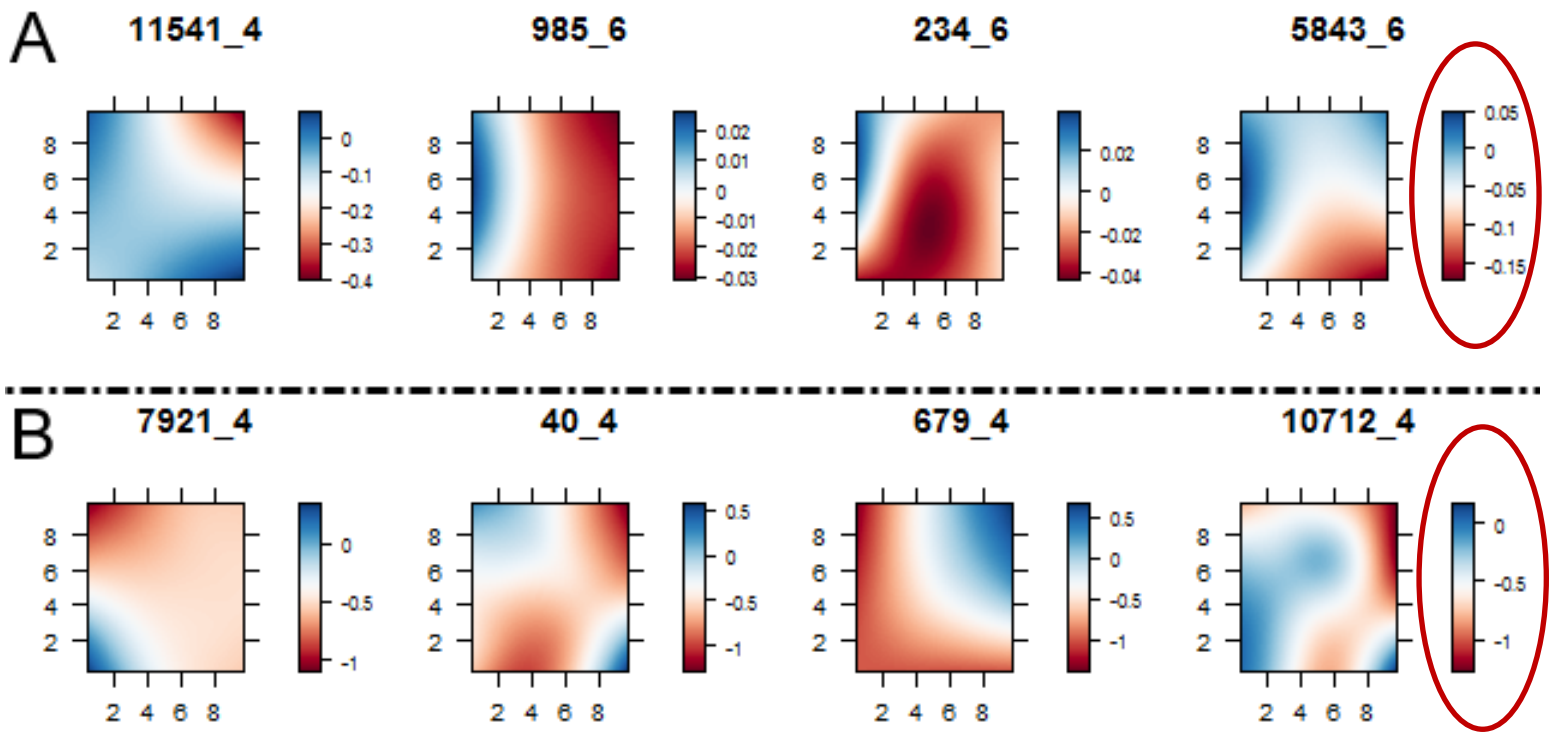
➢ Model abundance of acidobacteria

- Illumina sequencing of 16S rRNA > 1208 OTUs (Operational taxonomic unit; Species)
  - ➢ Count data (# of occurrences / # of rRNA measurements)
  - ➢ 24 environmental variables (e.g. pH, soil moisture, phosphate, clay (%), …)
  - ➢ Space & Time

➢ Abundance of each OTU was modelled separately by a boosted negative binomial model with smooth effects for environmental variables and additional temporal and spatial effects.

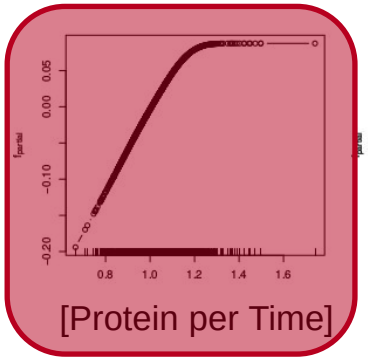
➢ Similarities over OTUs were assessed

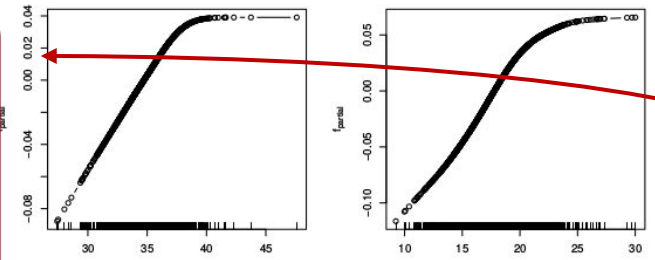
Acidobacteria in grassland soil
Example for results

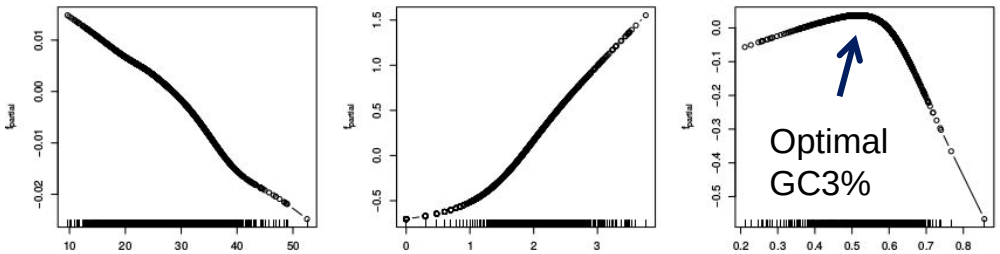


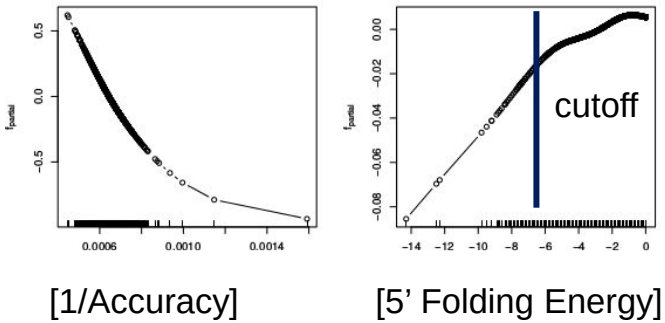
Panel A shows the most abundant OTUs, with comparably small gradients, i.e., small spatial variability. Panel B demonstrates the larger spatial variability of rare OTUs

Boosting already used at the PEI: Prediction of protein abundance produced from mRNA









- Multivariate model taking into account simulated translation and additional sequence features
- Explained variance (verified in new experiments): $R^2 = 0.55 – 0.66$
- Boosted GAMs had good prediction accuracy and good interpretability

Comparison

Stepwise Regression vs **Boosting**

- Refit whole model
- Only if p < n
- Selection based e.g. on p-values
- Difficult to extend
- P-values exist
  (but are wrong)

- Refit residuals
- Also for p ≫ n
- Selection based on prediction accuracy (i.e., relevance)
- Easy to extend
  - to complex effects
  - different types of outcome
  - …
- No p-values

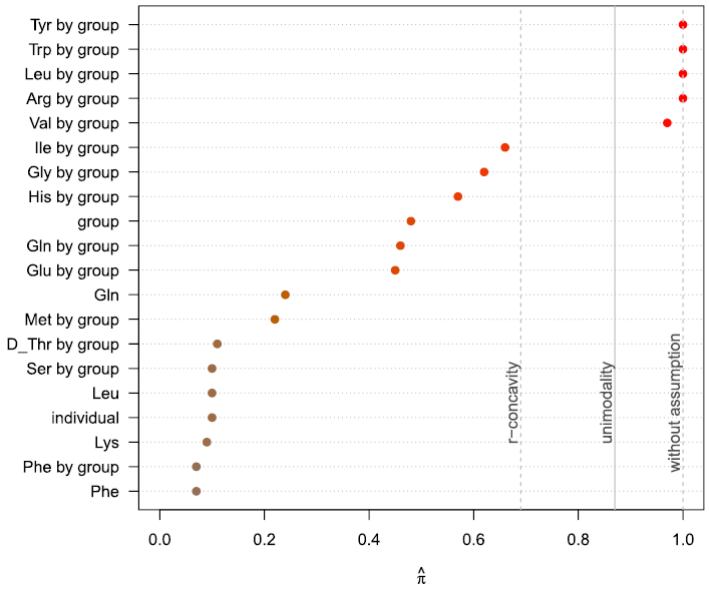
➢ Similar model

➢ Same interpretation

Stability Selection

- How much does the selection of variables depend on observed data and potential artifacts in the data?

- Use 100 random samples of the data
- Check in how many of these samples a variable was selected (e.g. via stepwise regression, boosting, …)
- Only use variables which are often selected
  - ➢ One additionally gets an error control (similar to „p-values“)



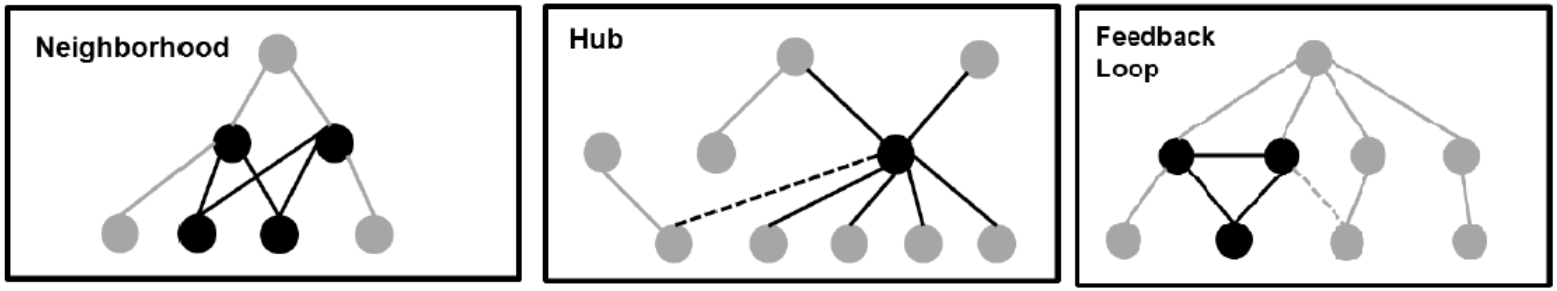
Differential expression of phenotypes of autism spectrum disorder patients compared to healthy controls. At maximum one false positive selection is accepted ($\alpha \approx 0.02$)

selected:
tyrosine (Tyr), tryptophan (Trp), leucine (Leu), arginine (Arg), valine (Val)

Summary & Outlook

- Boosting allows fitting of (complex) models in a single framework
- Variable selection is one of the major goals
- Resulting models are interpretable
- Models optimize prediction accuracy

- Current research:
  - GWAS studies:
    Incorporation of genetic networks (from KEGG) into boosting



  - I am looking forward to interesting and stimulating problems from PEI researchers