Stochastic Machine Learning
01 - Introduction

Thorsten Schmidt

Abteilung für Mathematische Stochastik

www.stochastik.uni-freiburg.de
thorsten.schmidt@stochastik.uni-freiburg.de

WS 2020/21

Generalized Linear Models

We already saw that transforming the input variables suitable might be helpful. This is the idea of a generalized linear model (GLM), see Casella & Berger (2002).

Definition

A GLM consists of three components:

1. Response variables (random) $Y_1, \dots, Y_n$,
2. a systematic component of the form $\alpha + \boldsymbol{\beta}^\top \boldsymbol{x}_i$, $i = 1, \dots, n$,
3. a link function g satisfying

$$\mathbb{E}[Y_i] = g(\alpha + \boldsymbol{\beta} \boldsymbol{x}_i), \quad i = 1, \dots, n.$$

Regularization of multiple linear regression

- One problem in practice is parsimony of a linear regression: suppose you have many covariates and you want to include only those which are relevant.
- It would be possible to iteratively throw out those parameters which are not significant. This procedure, however is not optimal. Many others have been proposed.
- We concentrate on **continuous** subset selection methods: it is better to introduce a penalty for including two many parameters, which we call regularization. This is moreover a standard procedure for ill-posed problems. We will consider a famous example: the **LASSO** introduced in R. Tibshirani (1996). „Regression Shrinkage and Selection via the Lasso“. In: **Journal of the Royal Statistical Society. Series B (Methodological)** 58.1, pp. 267–288.

LASSO

- The **least absolute shrinkage and selection operator** minimizes the following function

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \frac{1}{n} \parallel \boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\beta} \parallel_2^2 + \lambda \parallel \boldsymbol{\beta} \parallel_1 \right\}.$$

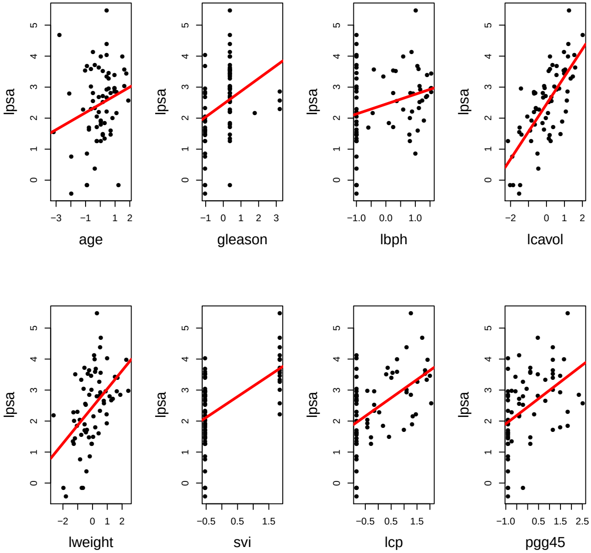
  The parameter λ has to be chosen and allows to vary the level of regularization. Clearly this model prefers to set non-significant parameters to zero.

- Let us illustrate the lasso with an example taken from Chris Franck, `http://www.lisa.stat.vt.edu/?q=node/5969`. The data stems from Stamey et.al.[12].

- The data describes clinical measures from 97 men about to undergo radical prostatectomy. It is of interest to estimate the relation between the clinical measures and the prostate specific antigen (measures are: lcavol - log (cancer volume), lweight - log(prostate weight volume), age, lbph - log (benign prostatic hyperplasia), svi - seminal vesicle invasion, lcp - log(capsular penetration), Gleason (score), ppg45 - percent Gleason scores 4 or 5, Y =lpsa - log(prostate specific antigen))

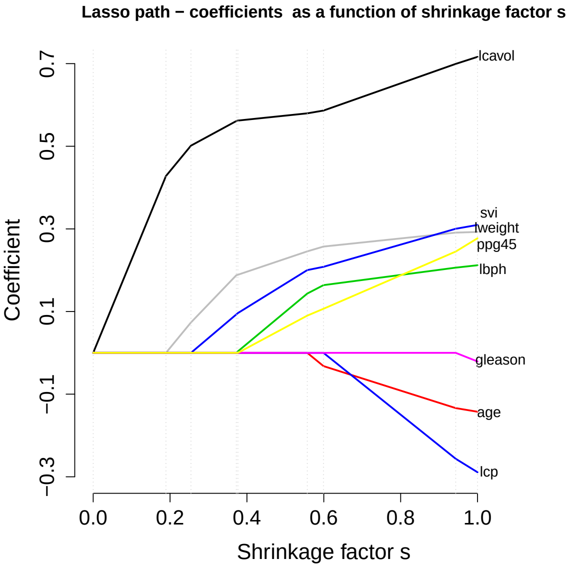
[12] T. A. Stamey et al. (1989). „Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II. Radical prostatectomy treated patients.“. In: **The Journal of urology** 141.5, pp. 1076–1083.

We start by examining bi-variate regressions.

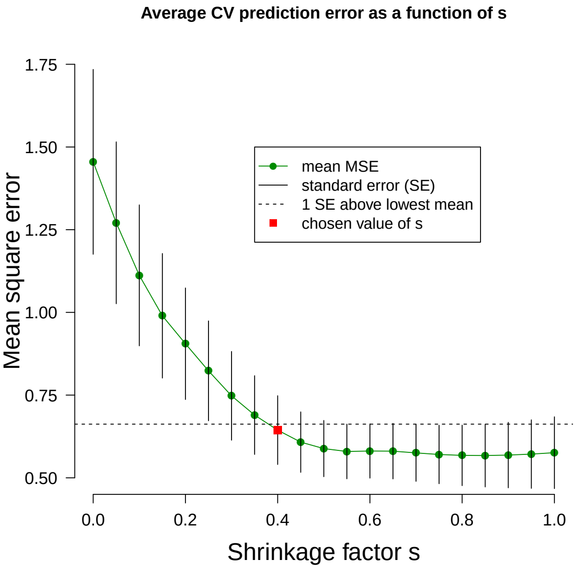


- It is obvious that some variables have fewer impact and some others seem to be more important. The question is how to effectively select those.
- We illustrate how cross-validation may be used in this case. This means we separate the data into a training set and a validation set. The tuning parameter λ is chosen based on the training set and validated on the validation set.
- We use a 10-fold cross validation, ie. the set is split into 10 pieces. Iteratively, each piece is chosen as the validation set while the remaining 9 sets are used to estimate the model.

This is the so-called lasso path. The shrinkage factor is antiproportional to λ.



This is the cross-validation result. A rule of thumb is to select that value of s that is within 1 standard error of the lowest value.



Remarks and Questions

- We see that the optimal choice of λ is far from trivial. Alternative approaches are at hand, compare the recent results by Johannes Lederer and coauthors, J. Lederer and C. Müller (Apr. 2014). „Don't Fall for Tuning Parameters: Tuning-Free Variable Selection in High Dimensions With the TREX“. In: **ArXiv e-prints**. eprint: `1404.0541` (stat.ME).
- What is a generalized linear model? Where are the differences to a linear model?
- What is the LASSO ?
- What are the differences to simple least squares ?
- What is an ill-posed problem ? Why do you regulate this ? Why is linear regression an ill-posed problem ?
- What is cross-validation ?

Please note that I encourage you to do research in the internet on words you don't know. Use the references, use google, google scholar, use the katalog at uni freiburg to find online resources for books and literature, use Wikipedia, use the mathematical encyclopdia ...