UNIVERSITÄT REGENSBURG
Fakultät für Mathematik

Masterarbeit:
Foundations of modern regression analysis and application to the analysis of telematics data

Markus Binder

Regensburg
12. März 2018

Zusammenfassung

In dieser Arbeit werden die Grundlagen von modernen Regressionsverfahren erarbeitet. Regressionsverfahren modellieren Beziehungen zwischen einer reellwertigen Zielvariablen und mehreren erklärenden Variablen. Neben Erweiterungen von klassischen Verfahren sind dabei in den letzten Jahren auch, unter anderem in Hinblick auf die größer werdenden zu untersuchenden Datenmengen, neue Methoden in einem anderen theoretischen Rahmen entstanden. Die vorgestellten Ansätze werden dann zur Analyse von Telematikdaten angewendet.

Zunächst werden dazu die Grundlagen der Linearen Regression und dabei die Theorie der Fehlerquadrate dargestellt. Darauf aufbauend werden dann verallgemeinerte lineare Modelle betrachtet, mit denen auch nichtlineare Beziehungen modelliert werden können. Dazu werden Exponentialfamilien eingeführt, Maximum-Likelihood-Schätzer betrachtet und numerische Lösungsverfahren für nichtlineare Gleichungen angewandt.

Dann werden die Grundlagen von alternativen Regressionsmethoden im Rahmen des maschinellen Lernens, einem Ansatz aus der Informatik, erarbeitet. Hier beruht die Theorie nicht wie in den ersten Kapiteln auf Verteilungsannahmen, sondern es werden ohne solche Annahmen Fehlergrenzen mittels Konzentrationsungleichungen und der Einführung einer Rademacher-Komplexität erarbeitet. Zur Umsetzung dieser Verfahren werden unter anderem quadratische Optimierungsprobleme unter Nebenbedingungen gelöst.
Außerdem erfolgt hier die Erweiterung zur Modellierung von nichtlinearen Beziehungen mittels spezieller Kernfunktionen, die einem Skalarprodukt auf einem Hilbertraum entsprechen.
Als vorgeschaltetes Instrument zur Regression wird dabei die Hauptkomponentenanalyse mittels Singulärwertzerlegung betrachtet. Dazu führen wir einen modernen Algorithmus zur Berechnung der Singulärwertzerlegung ein.

Im nächsten Kapitel werden dann stochastische Simulationsmethoden eingeführt, die bei der Validierung und Verbesserung von Regressionsmodellen helfen können. Zur Rechtfertigung dieser werden anhand des Berry-Esseen-Theorems Konvergenzraten im Zentralen Grenzwertsatz erarbeitet.

Im Anwendungskapitel werden die erarbeiteten Methoden schließlich angewandt, um 13 Milliarden Telematikdaten zu analysieren. Dabei werden mit Hilfe der Singulärwertzerlegung zunächst charakteristische Fahrprofile herausgearbeitet. Darauf aufbauend wird dann ein verallgemeinertes lineares Modell erstellt, dass einem beliebigen Fahrprofil eine Kraftfahrt-Haftpflicht Prämie zuordnet.

Contents

Introduction **iii**

1 Preliminaries: Linear Regression and Least Squares Theory **1**

2 Generalized Linear Models **7**
2.1 Exponential Families 7
2.2 Extension to a Generalized Linear Model 10
2.3 Maximum Likelihood Estimation 10
2.4 Fitting Generalized Linear Models 11

3 Regression models in the framework of Machine Learning **17**
3.1 Basic Framework 17
3.2 Generalization bounds - simplified case 19
3.3 Rademacher complexity 21
3.4 Generalization bounds 27
3.5 Kernel functions 29
3.6 Support Vector Regression 32
3.7 Principal Component Analysis 41

4 Further tools: Stochastical simulation methods **49**
4.1 Foundation of Monte-Carlo methods 50
4.2 Implementation of Monte-Carlo methods 53
4.3 Comparison with other methods 56
4.4 Polynomial Chaos Expansion 57

5 Application to the analysis of telematics data **61**
5.1 PCA of the telematics data 62
5.2 Constructing a GLM based on selected driving properties 65
5.3 Constructing a GLM based on the PCA 67
5.4 Outlook: Polynomial Chaos Expansion for broader model fitting 69

Bibliography **71**

Introduction

In this thesis we consider regression problems. Thereby we have sample observations of a real-valued response and corresponding given values of attributes that influence the response with a random component. Based on that sample we try to find a relationship between the attributes and the response.

Already 200 years ago, Carl Friedrich Gauß has developed the least-squares method, and showed that it is, under distribution assumptions for the response, an optimal approach for Linear Regression. This theory has been extended to Generalized Linear Models since then.

In recent years the alternative framework of Machine Learning has been developed which does not require assumptions on the specific underlying distribution. So rather than using the theory of Gauß, this framework concedes that it is not always possible to establish reasonable distribution assumptions and instead relies on establishing generalization bounds on the errors for large enough samples. Therefore, it is very popular in the modern context of "Big Data" problems, where big data sets have to be analyzed without having much structural information about them.
Moreover, in those problems it is often useful to preprocess the data. As a popular method to do that, we will consider the Principal Component Analysis, which allows a dimensionality reduction of the regression model.

Thereby we will focus on establishing the theoretical foundations for the considered methods. These will range from the classical Gauss-Markov theorem and Maximum Likelihood Estimation in the first statistical chapters to generalization bounds on the errors by using concentration inequalities and the Rademacher complexity notion in the Machine Learning framework.

Finally, simulation tools to validate and tweak the regression models will be considered. As a justification for them, we will establish convergence rates in the central limit theorem.

The concepts will then be applied to the analysis of a big data set of telematics data. That is, roughly 13 billion values about driving behavior of project participants are considered, based on which we try to estimate the expected claims expenditures of a driving profile for an insurance company.

Chapter 1

Preliminaries: Linear Regression and Least Squares Theory

To start, we will give a brief overview over the theory of ordinary linear models.

In this thesis we generally consider sample observations of a response, and factors, that could influence this response, in the first chapters called explanatory variables. In an ordinary linear model, the response is assumed to have normal distribution.
More precisely, we consider a response vector $(y_1,..,y_n)$ of n independent random observations, where we treat y_i as a realization $Y_i(\omega)$ of a normally distributed random variable $Y_i : (\Omega,\mathscr{F}) \to (\mathbb{R},\mathscr{B}(\mathbb{R}))$ here. Thereby the mean μ_i can vary for different Y_i, but the variance is assumed to be constant, i.e. $Var(Y_i) = \sigma^2 \ \forall i$.

Moreover, we denote the value of explanatory variable j for observation i by x_{ij}. This gives a $n \times p$ model matrix X, where p denotes the amount of regarded explanatory variables. Thereby one extra column is usually reserved for a constant term.
To estimate fitted values for the response, a linear model uses a parameter vector β that we estimate based on the realizations $y_1,..,y_n$.

Definition 1.1. An ordinary linear regression model consists of a model matrix X and a parameter vector β to obtain the mean vector

$$\mu = X\beta$$

of the response vector Y. Thereby $Y = (Y_1,..,Y_n) : (\Omega,\mathscr{F},P) \to (\mathbb{R}^n,\mathscr{B}(\mathbb{R}^n))$ is assumed to have normally distributed and independent components with constant variance, i.e.
$Y \sim N_n(\mu,\Sigma)$ with $\Sigma = \sigma^2 I_n$.

Example 1.2. Consider the claims expenditures of insured car theft per year in Germany in million euros, according to the German insurance association (GDV).

2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
211,4	178,3	175,7	219,0	257,2	259,6	242,4	263,9	262,0	291,3

Given the values of the last ten years 2006-2015 we want to predict the claims expenditures in 2016.
In ordinary linear regression this is being estimated by the model

$$\mu_i = \beta_0 + \beta_1 x_i,$$

$i = 1,...,10,\ (x_1,x_2,..,x_{10}) = (1,2,..,10)$. In matrix notation we can write this as

$$\begin{pmatrix} \mu_1 \\ \mu_2 \\ ... \\ \mu_{10} \end{pmatrix} = \begin{pmatrix} 1\ 1 \\ 1\ 2 \\ ... \\ 1\ 10 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}.$$

Remark 1.3. An alternative way to express the ordinary linear model is

$$y = X\beta + \varepsilon$$

with a normal-distributed error term ε having $\mathbb{E}(\varepsilon) = 0$ and covariance matrix $V = \sigma^2 I$.

This expression states directly, that we make assumptions on the error terms in an ordinary linear model. However, for generalizations of this ordinary model the simple additive structure for the error terms is not suited, so we will mainly stick to the model description $\mu = X\beta$.

The standard approach to estimate the parameter vector for observed data $y = (y_1,...,y_n)$ uses the least squares method. This determines the value of $\hat{\mu}$ that minimizes

$$||y - \hat{\mu}||^2 = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 = \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \hat{\beta}_j x_{ij})^2$$

on the vector space $C(X)$ formed by the image of X, i.e. it gives fitted values $\hat{\mu}$ with

$$||y - \hat{\mu}|| \leq ||y - \mu|| \quad \forall \mu \in C(X).$$

Another well-known approach is given by the Maximum-Likelihood estimation.
This method determines the parameter values that maximize the probability of making the observations given the parameters.
As we are considering independent observations that are normally distributed with constant variance in linear regression, we can define the joint density function for all observations as

$$f(y_1,...,y_n;\mu,\sigma^2) = \prod_{i=1}^{n} f_i(y_i;\mu_i,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-(y_i-\mu_i)^2/2\sigma^2}.$$

To maximize this term we apply the monotonic function log, which yields

$$log(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-(y_i-\mu_i)^2/2\sigma^2}) = constant - [\sum_{i=1}^{n} (y_i - \mu_i)^2]/2\sigma^2.$$

Therefore we have to minimize $\sum_{i=1}^{n}(y_i-\mu_i)^2$ in order to maximize the joint density function, which means that the Maximum Likelihood estimation corresponds to the least squares approach in the context of ordinary linear models.

Remark 1.4. *The function $logf(y_1,...,y_n;\mu,\sigma^2) =: log(L(\beta))$ is called the log-likelihood. Thereby we use the notation $L(\beta)$ to point out, that the means μ_i are determined by the parameter vector β. The ";" denotes the separation between the variable and the parameters of the density or mass function.*

The objective function $S(\beta) = ||y - X\beta||^2$ is convex, by composition of the convex function $u \to ||u||^2$ with the affine function $\beta \to y - X\beta$, and it is differentiable.
Thus, S admits a global minimum at β if and only if $\nabla S(\beta) = 0$.

Taking the partial derivatives of $S(\beta) = \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2$ gives

$$\sum_{i=1}^{n}(y_i-\mu_i)x_{ij} = 0, \quad j = 1,..,p.$$

Proposition 1.5 (normal equations)**.** *For the ordinary linear model $\mu = X\beta$ the least squares estimates satisfy the normal equations*

$$\sum_{i=1}^{n} y_i x_{ij} = \sum_{i=1}^{n} \hat{\mu}_i x_{ij}, \ j = 1,..,p.$$

To give a compact solution, we go back to matrix form, i.e.

$$X^T y = X^T X\hat{\beta}.$$

Now, provided the matrix X and consequently also X^TX has full rank, the extremum of $L(\beta)$ occurs at

$$\hat{\beta} = (X^TX)^{-1}X^Ty. \tag{1.1}$$

In Example 1.2 X^TX has full rank, so the unique least squares estimate of β is

$$\hat{\beta} = (X^TX)^{-1}X^Ty = \begin{pmatrix} 10 & 55 \\ 55 & 385 \end{pmatrix}^{-1} \begin{pmatrix} 1 & 1 & ... & 1 \\ 1 & 2 & ... & 10 \end{pmatrix} y = \begin{pmatrix} 175,46 \\ 11,0218 \end{pmatrix}.$$

Thus ordinary linear regression gives the prediction

$$\mu_{11} = 175,46 + 11,0218 \times 11 = 296,7,$$

i.e. predicted claims expenditures of 296,7 million euros due to insured car theft in 2016 in Germany.

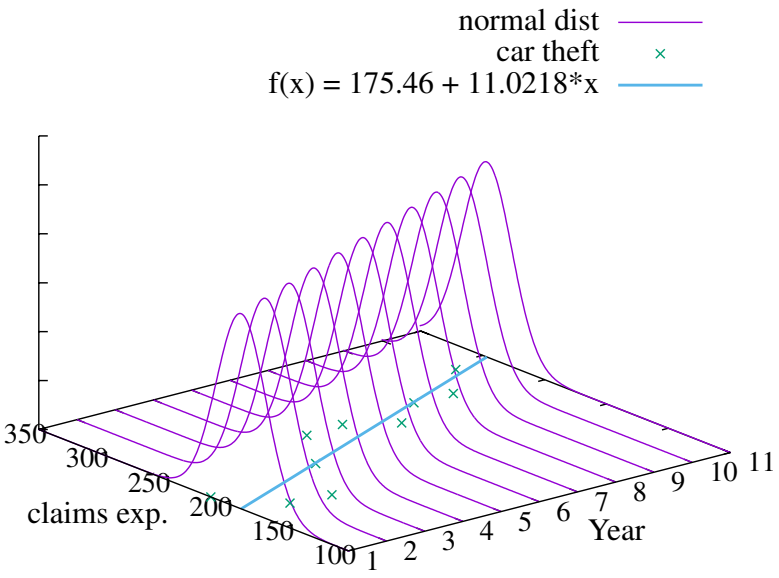

Figure 1.1: Ordinary Linear Regression

Remark 1.6. If X^TX does not have full rank, the equation admits a family of solutions that can be given in terms of the pseudo-inverse of matrix X^TX, which in turn can be obtained via the singular value decomposition of X^TX, for details see [2, A.2.2]. Considering the singular value decomposition can also be useful for dimensionality reduction in a regression model, as we will see in the application chapter.

To be able to state the well-known "BLUE" optimality result for the least squares estimator, we first have to recall the property of unbiasedness of an estimator. Here we only consider model matrices X with full rank.

Definition 1.7. An estimator $\hat{\beta}$ of a parameter β is called unbiased, if

$$\mathbb{E}[\hat{\beta}] = \beta,$$

i.e. if the expectation of the estimator matches the true parameter β.

Proposition 1.8. *The least squares estimator* $\hat{\beta} = (X^TX)^{-1}X^Ty$ *is unbiased.*

Proof. Taking the representation $y = X\beta + \varepsilon$ as in Remark 1.3, where β is the true parameter, we get

$$\hat{\beta} = (X^TX)^{-1}X^T(X\beta + \varepsilon) = \beta + (X^TX)^{-1}X^T\varepsilon.$$

Due to the linearity of the expectation, this yields

$$\mathbb{E}(\hat{\beta}) = \mathbb{E}(\beta + (X^T X)^{-1} X^T \varepsilon) = \beta + (X^T X)^{-1} X^T \underbrace{\mathbb{E}(\varepsilon)}_{=0} = \beta.$$

□

Theorem 1.9 (Gauss-Markov Theorem)**.** *Let $y = (y_1, .., y_n)^T$ be a realization of a random vector $Y : (\Omega, \mathscr{F}, P) \to (\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$. Suppose $\mathbb{E}(Y) = X\beta$, where X is a matrix with full rank, and that Y has covariance matrix $\Sigma = \sigma^2 I_n$.*
Then the least squares estimator $\hat{\beta} = (X^T X)^{-1} X^T y$ is the best linear unbiased estimator (BLUE) of β, in this sense: For any linear combination $a^T \beta$ of the estimators that are linear in y and unbiased, $a^T \hat{\beta}$ has minimum variance.

For a proof of this statement see [1, Section 2.7].

Remark 1.10. Note that we did not require the normality assumption for the Gauss-Markov Theorem, but using the normality assumption one can further show that the least squares estimator is a minimum variance unbiased estimator (MVUE). So the least squares operator is a minimum variance estimator under all unbiased estimators in an ordinary linear model, not only under all linear unbiased estimators.

The ordinary linear model for the problem Example 1.2 is a solid simple model but has some major restrictions and disadvantages. Firstly, the assumption of normal errors can clearly not be completely accurate, since the normal distribution has support $\mathbb{R}$ but in reality the claims expenditures will surely be nonnegative. Furthermore, in some cases it will be necessary to model nonlinear developments, for example an exponential growth may be more reasonable. We can try to model that by considering the logarithms of the claims expenditures but then we won't have control over the variance of the residuals in the model anymore in ordinary linear regression, called heteroscedasticity.

Therefore we will introduce more complex models in this thesis, which will give us opportunities to model more accurately.

Chapter 2

Generalized Linear Models

As we have seen in the preliminaries, an ordinary linear model has some major constraints. Therefore, nowadays Generalized Linear Models (GLMs) are used in statistics. They extend the standard linear regression model. Thereby a main extension is that they allow any distribution in an exponential family for the response variable.
To see how much of an advantage this is, we first have to precisely define the class of exponential families and determine which distributions actually belong to it.

2.1 Exponential Families

Definition 2.1 (exponential family)**.** Consider random variables $X : (\Omega, \Sigma, P) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$, i.e. random variables that take values in the one-dimensional Euclidean space equipped with the σ-field of Borel sets.
A family of distributions is an exponential family if its probability density or mass functions in the variable x can be written as

$$f_X(x;\theta,\phi) = exp\{[x\ \theta - b(\theta)]/a(\phi) + c(x,\phi)\} \tag{2.1}$$

with arbitrary, but fixed real-valued functions $a : \mathbb{R} \to \mathbb{R}$, $b : \mathbb{R} \to \mathbb{R}$ and $c : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$.
Thereby θ is called the natural parameter and ϕ is called the dispersion parameter of the exponential family.
Again, the ";" denotes the separation between the variable and the parameters of the density or mass function.

Remark 2.2. Note, that the support of a distribution family which builds an exponential family does not depend on its parameter values. The general Pareto distribution, for example, has varying minimum bound and therefore cannot build an exponential family.

Remark 2.3 (natural exponential family)**.** *Often $a(\phi) = 1$ and $c(x;\phi) = c(x)$. Then we get a natural exponential family*

$$f(x;\theta) = h(x)exp[x\ \theta - b(\theta)]. \tag{2.2}$$

Now we can verify that most of the common distribution families build an exponential family.

Example 2.4 (Poisson)**.** The probability mass function for a random variable with Poisson distribution is given by

$$f(k;\mu) = \frac{e^{-\mu}\mu^k}{k!} = \frac{1}{k!}\exp(k\ ln(\mu) - \mu) = \frac{1}{k!}\exp(k\theta - \exp(\theta)),\ k = 0,1,2...,$$

with the natural parameter $\theta = ln(\mu)$.
This has natural exponential form (2.2) with $h(k) = \frac{1}{k!}$ and $b(\theta) = \exp(\theta)$.

Example 2.5 (Binomial)**.** For $X \sim Bin(n,p)$ let $\theta = ln(p/(1-p))$.
This yields $p = \exp(\theta)/[1+\exp(\theta)]$ and consequently,

$$f(k;p) = \binom{n}{k} p^k (1-p)^{n-k} = \binom{n}{k} \exp[k\ \theta - n\ ln(1+exp(\theta))],$$

which has natural exponential form (2.2) with $h(k) = \binom{n}{k}$ and $b(\theta) = n\ ln(1+\exp(\theta))$.
The natural parameter $\theta = ln(p/(1-p))$ is called the logit.

Example 2.6 (Normal)**.** For the normal distribution we have the probability density function

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] = \exp\left[\frac{x\mu - \frac{1}{2}\mu^2}{\sigma^2} - \frac{1}{2}ln(2\pi\sigma^2) - \frac{x^2}{2\sigma^2}\right].$$

This has exponential form (2.1) with natural parameter $\theta = \mu$ and

$$b(\theta) = \frac{1}{2}\mu^2 = \frac{1}{2}\theta^2, \quad a(\phi) = \sigma^2, \quad c(x,\phi) = -\frac{1}{2}ln(2\pi\sigma^2) - \frac{x^2}{2\sigma^2}.$$

Having the exponential form of the probability density or mass function, one can directly determine the expected value and variance of that distribution:

Theorem 2.7. *For a random variable* $X : (\Omega, \mathscr{F}, P) \to (\mathbb{R}, \mathscr{B}(\mathbb{R}))$ *with probability density or mass function in exponential form* (2.1) *it holds*

$$\mathbb{E}[X] = b'(\theta) \ \ and \ \ Var(X) = b''(\theta)a(\phi).$$

Proof. Consider

$$\int \frac{\partial}{\partial\theta} f(x;\theta,\phi)dx = \int \frac{\partial}{\partial\theta} exp\{[x\ \theta - b(\theta)]/a(\phi) + c(x,\phi)\}dx.$$

According to [1], the regularity conditions that allow interchanging integration and differentiation here are generally fulfilled for distributions in an exponential family. A set of sufficient regularity conditions are: $\frac{\partial}{\partial\theta} f(x;\theta,\phi)$ is continuous in x and $\theta \in \Theta$ where Θ is an

open set, the integral $\int f(x;\theta,\phi)dx$ exists and $\int \left|\frac{\partial}{\partial\theta} f(x;\theta,\phi)\right| dx < M < \infty \; \forall\theta \in \Theta$.
Then the left side of the equation gives

$$\frac{\partial}{\partial\theta}\int f(x;\theta,\phi)dx = \frac{\partial}{\partial\theta}(1) = 0, \tag{2.3}$$

since f is a probability density or mass function. On the other hand, we have

$$\frac{\partial}{\partial\theta} f(x;\theta,\phi) = \frac{\partial}{\partial\theta} exp\{[x\,\theta - b(\theta)]/a(\phi) + c(x,\phi)\} = f(x;\theta,\phi)\;\frac{x-b'(\theta)}{a(\phi)},$$

so that (2.3) implies

$$\int f(x;\theta,\phi)\;\frac{x-b'(\theta)}{a(\phi)}dx = \underbrace{\int x\, f(x;\theta,\phi)dx}_{=E(X)} - b'(\theta)\underbrace{\int f(x;\theta,\phi)dx}_{=1} = 0.$$

This yields $E(X) = b'(\theta)$.

Now we consider the second derivate to obtain

$$\int \frac{\partial^2}{\partial\theta^2} f(x;\theta,\phi)dx = \frac{\partial^2}{\partial\theta^2}\int f(x;\theta,\phi)dx = \frac{\partial^2}{\partial\theta^2}(1) = 0$$

and compare it with

$$\int \frac{\partial^2}{\partial\theta^2} f(x;\theta,\phi)dx = \int \frac{\partial}{\partial\theta} f(x;\theta,\phi)\;\frac{x-b'(\theta)}{a(\phi)}dx = \int f(x;\theta,\phi)\left[\left(\frac{x-b'(\theta)}{a(\phi)}\right)^2 - \frac{b''(\theta)}{a(\phi)}\right]dx$$

to get

$$\frac{(x-b'(\theta))^2}{a(\phi)^2} = \frac{b''(\theta)}{a(\phi)}.$$

In the first part of the proof we have shown that $b'(\theta) = E[X]$, so $(x-b'(\theta))^2 = Var[X]$ and the equation above yields

$$Var[X] = b''(\theta)a(\phi).$$

□

Example 2.8. [Poisson] For the Poisson distribution $f(k;\mu) = \frac{e^{-\mu}\mu^k}{k!}$ we have already established $\theta = ln(\mu)$ as the natural parameter, $a(\phi) == 1$ and $b(\theta) = \exp(\theta)$. So the theorem above yields $\mathbb{E}[X] = b'(\theta) = \exp(\theta) = \mu$ and $Var[X] = b''(\theta) = \exp(\theta) = \mu$ for a Poisson distributed random variable X.

Having established the exponential families, we can now exhibit how an ordinary linear model can be extended to a Generalized Linear Model.

2.2 Extension to a Generalized Linear Model

Recall, that the ordinary linear model consisted of a given model matrix X and a parameter vector β that we estimated based on the sample observations $y = (y_1, .., y_n)$.
Thereby we assumed that these observations are realizations of independent normal distributed random variables $Y_1, .., Y_n$ with constant variance.

Now in a GLM, we just require $Y_1, .., Y_n$ to belong to an exponential family. Thereby the exponential family, i.e. the functions a, b, c and the dispersion parameter are fixed over the Y_i, whereas the natural parameter varies, i.e. we have density functions

$$f_{Y_i}(y; \theta_i, \phi) = exp\{[y\ \theta_i - b(\theta_i)]/a(\phi) + c(y, \phi)\}.$$

Furthermore, constant variance of the response vector is not required in a GLM.

Moreover, the do not directly model the mean $\mu = \mathbb{E}(Y)$ linearly in a GLM, but rather an intermediate vector $\eta = g(\mu)$, where g is a monotonic and differentiable function, called link function. Overall we have

$$g(\mu_i) = \eta_i = \sum_{j=1}^{p} \beta_j x_{ij}, \quad i = 1, ..., n.$$

The link function g, that transforms μ_i to the natural parameter of Y_i for all i, is called the canonical link. As we saw in the section about exponential families, this natural parameter is the mean for a normal distribution, the log of the odds for a binomial distribution and the log of the mean for a Poisson distribution.

Remark 2.9. *This implies that for a normal distribution the canonical link function is just the identity, so we can extract ordinary linear models from the general definition by assuming normal distribution and taking the canonical link, i.e. the identity link function.*

Recall that for ordinary linear models it can also be shown that the least squares method provides the best possible estimator of model parameters in a certain sense and that the least squares method coincides with the Maximum Likelihood Estimation in that case. Likewise, for GLMs a general method for the estimation of parameters can be established.

2.3 Maximum Likelihood Estimation

Since the density or mass function of the response variable can be written in exponential form (2.1) in GLMs, we are also able to obtain a general expression for the Maximum Likelihood Estimation.

For n independent observations the log-likelihood is

$$L(\beta) = \log \prod_{i=1}^{n} f(y_i; \theta_i, \phi) = \sum_{i=1}^{n} log f(y_i; \theta_i, \phi) = \sum_{i=1}^{n} \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + \sum_{i=1}^{n} c(y_i, \phi).$$

To maximize the log-likelihood we will compute its partial derivatives.
In an ordinary linear model, the parameter vector β determined the means μ_i of the normally distributed response variables. In a GLM it determines the intermediate vector η, which in turn determines the means μ_i of the response vector and that the natural parameters θ_i in exponential family form. Therefore we have with $logf(y_i, \theta_i, \phi) =: L_i$, that

$$\frac{\partial L_i}{\partial \beta_j} = \frac{\partial L_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}.$$

by the chain rule. Since $\mu_i = b'(\theta_i)$ and $Var(Y_i) = b''(\theta_i)a(\phi)$ by Theorem 2.7, it holds

$$\frac{\partial L_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)} = \frac{y_i - \mu_i}{a(\phi)} \text{ and } \frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = \frac{Var(Y_i)}{a(\phi)}.$$

Furthermore we have $\partial \eta_i / \partial \beta_j = x_{ij}$ due to $\eta_i = \sum_{j=1}^{p} \beta_j x_{ij}$. Overall this yields

$$\frac{\partial L_i}{\partial \beta_j} = \frac{\partial L_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} = \frac{y_i - \mu_i}{a(\phi)} \frac{a(\phi)}{Var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} = \frac{(y_i - \mu_i) x_{ij}}{Var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i}. \tag{2.4}$$

Finally, summing over all n yields the desired likelihood equations.

Definition 2.10 (Likelihood equations for a GLM)**.** The expressions

$$\frac{\partial L(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i) x_{ij}}{Var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0, \ j = 1, 2, ..., p, \tag{2.5}$$

where $\eta_i = \sum_{j=1}^{p} \beta_j x_{ij} = g(\mu_i)$ for the link function g, are called the Likelihood equations.

Remark 2.11. It has to be verified, that the resulting extreme values are indeed maxima, but in most models this is given. For example, according to [1] it can generally be shown, that the log likelihood always is a concave function when we use the canonical link function. For a Poisson GLM, that is shown in an example at the end of the chapter.

2.4 Fitting Generalized Linear Models

The likelihood equations (2.5) are usually nonlinear in $\hat{\beta}$. A possible method to numerically solve these is the Newton-Raphson Method, i.e.

$$\beta^{(t+1)} = \beta^{(t)} - (H^{(t)})^{-1} u^{(t)},$$

with $u = (\partial L(\beta)/\partial \beta_1, ..., \partial L(\beta)/\partial \beta_p)^T$ and the Hessian matrix H with $H_{ab} = \partial^2 L(\beta)/\partial \beta_a \partial \beta_b$ evaluated at $\beta^{(t)}$, assuming $H^{(t)}$ is nonsingular at every step.
Recall that in general this method converges locally to a zero of $\partial L(\beta)/\partial \beta$.
Due to Remark 2.11, this is necessarily a global maximum of $L(\beta)$ if we use the canonical link function.

A more specific iterative method for solving likelihood equations is the Fisher Scoring. It is similar to the Newton-Raphson method but factors in the likelihood of all possible realizations instead of just the sample observations in the Hessian matrix, i.e. it uses a modification of the derivative of $\partial L(\beta)/\partial \beta$ for the Newton step.

Definition 2.12. The expected Fisher information matrix is defined as the negative of the expectation, integrating over y, of the Hessian matrix H as above, i.e.

$$J(\beta) = -\mathbb{E}_y(\partial^2 L(\beta)/\partial\beta_a\partial\beta_b)_{a,b=1,..,p}.$$

The formula for Fisher scoring is

$$\beta^{(t+1)} = \beta^{(t)} + (J^{(t)})^{-1}u^{(t)}.$$

Remark 2.13. The negative of the Hessian itself is sometimes called observed Fisher information. The Fisher information tells us how curved the log-likelihood $L(\beta)$ is (averaging out the sample in J) for a fixed parameter β. Therefore a small value of the Fisher information for the ML estimator indicates that we do not have much information about the parameter β, since the likelihood does not change much for varying parameter values at the peak.

Example 2.14. We illustrate the two methods in an example from [1, Chapter 4.5.3] for which the maximum of L can be seen and compared directly, a sample proportion y from a $bin(n,p)$ distribution.
Naturally, the maximum $\hat{p}$ of the likelihood function should match the sample proportion y. Let us verify that by considering the log likelihood, neglecting the binomial coefficient because it does not effect its derivatives,

$$L(p) = log(p^{ny}(1-p)^{n-ny}) = ny\ logp + (n-ny)log(1-p).$$

Taking the first two derivatives of $L(p)$ gives

$$u = (ny-np)/[p(1-p)],\ H = -(ny/p^2 + (n-ny)/(1-p)^2),$$

i.e. we indeed have $u = 0$ for $p = y$ ($p \neq 0$ and $p \neq 1$). Therefore the Newton-Raphson method is given by

$$p^{(t+1)} = p^{(t)} + \left[\frac{ny}{p^{(t)^2}} + \frac{n-ny}{(1-p^{(t)})^2}\right]^{-1} \frac{ny-np^{(t)}}{p^{(t)}(1-p^{(t)})}.$$

Note, that this correctly adjusts $p^{(t)}$ up if $y > p^{(t)}$ and reverse, since the only variable term for the sign of the addition is $n(y-p^{(t)})$. When $p^{(t)} = y$, no adjustment occurs and $p^{(t+1)} = y$, which is the correct answer for $\hat{p}$.
For Fisher scoring, we consider, with $k := ny$,

$$-\mathbb{E}_y(H) = -\sum_{k=0}^{n}\left(-\frac{k}{p^2} - \frac{n-k}{(1-p)^2}\right)\binom{n}{k}p^k(1-p)^{n-k} = ... = \frac{n}{p(1-p)}.$$

Therefore a step of Fisher scoring has the form

$$p^{(t+1)} = p^{(t)} + \left[\frac{n}{p^{(t)}(1-p^{(t)})}\right]^{-1} \frac{ny-np^{(t)}}{p^{(t)}(1-p^{(t)})} = p^{(t)} + (y-p^{(t)}) = y.$$

So in this example the Fisher scoring method gives the correct value of $\hat{p}$ after a single iteration, whereas you can check that the Newton-Raphson method needs more steps for most starting values.

Remark 2.15. Fisher scoring can be interpreted as an iteratively reweighted least squares (IRLS) method, see [1, Chapter 4.5.4] for details.

Simplifications with canonical link

Finally, we notice that fitting GLMs is simple if we use the canonical link function in our model. Recall that then the link function transforms μ_i to the natural parameter θ_i for all i, so we have

$$\theta_i = \eta_i = \sum_{j=1}^{p} \beta_j x_{ij},$$

and

$$\partial\mu_i/\partial\eta_i = \partial\mu_i/\partial\theta_i = \partial b'(\theta_i)/\partial\theta_i = b''(\theta_i),$$

due to $\mu_i = b'(\theta_i)$ by Theorem 2.7. Together with the second equation in this theorem, $Var(Y_i) = b''(\theta_i)a(\phi)$, this implies that (2.4) simplifies to

$$\frac{\partial L_i}{\partial \beta_j} = \frac{(y_i - \mu_i)}{Var(Y_i)} b''(\theta_i) x_{ij} = \frac{(y_i - \mu_i)x_{ij}}{a(\phi)}. \tag{2.6}$$

Moreover, taking the second partial derivative gives

$$\frac{\partial^2 L_i}{\partial \beta_h \partial \beta_j} = -\frac{x_{ij}}{a(\phi)} \left(\frac{\partial \mu_i}{\partial \beta_h} \right),$$

which does not depend on y_i. Therefore we get

$$\partial^2 L(\beta)/\partial\beta_h\partial\beta_j = \mathbb{E}[\partial^2 L(\beta)/\partial\beta_h\partial\beta_j]$$

and the Newton-Raphson method coincides with Fisher scoring when we use the canonical link function for our model.

Example 2.16. Let's look back at Example 1.2. We noticed, that the assumption of normally distributed responses cannot be completely accurate. Instead, an appropriate distribution to model claims expenditures has been found to be the Poisson distribution in practice. So we will build a GLM for Example 1.2 with Poisson responses and the canonical log-link function.
We have established that the Poisson distribution can be written in natural exponential form, i.e. $a(\phi) = 1$. So according to (2.6) the likelihood equations for the Poisson GLM with canonical link are

$$\sum_{i=1}^{10} (y_i - \mu_i) x_{ij} = 0, \ j = 1, 2.$$

in our example. For a normal model, as we have seen in the preliminaries, these are the normal equations. Now the difference is that we do not have $\mu = X\beta$, but $log(\mu) = X\beta$, i.e. $\mu = \exp(X\beta)$.
So we get the nonlinear likelihood equations

$$\sum_{i=1}^{10} (y_i - \exp(x_{i1}\beta_1 + x_{i2}\beta_2)) x_{ij} = 0, \ j = 1, 2.$$

To solve these nonlinear equations, we can use the Newton-Raphson or Fisher scoring algorithm, which, as we have just seen, coincide when we model with the canonical link function. We have

$$u = \left(\sum_{i=1}^{10} (y_i - \exp(\beta_1 + x_{i2}\beta_2)), \sum_{i=1}^{10} (y_i - \exp(\beta_1 + x_{i2}\beta_2)) x_{i2} \right)^T,$$

since $x_{i1} = 1 \; \forall i$ in our example. Furthermore,

$$H = \begin{pmatrix} \sum_{i=1}^{10} -\exp(\beta_1 + x_{i2}\beta_2) & \sum_{i=1}^{10} -\exp(\beta_1 + x_{i2}\beta_2) x_{i2} \\ \sum_{i=1}^{10} -\exp(\beta_1 + x_{i2}\beta_2) x_{i2} & \sum_{i=1}^{10} -\exp(\beta_1 + x_{i2}\beta_2) x_{i2}^2 \end{pmatrix}.$$

Since we only have a 2×2 matrix here, the inversion of the Hessian will be no challenge. The choice of the starting estimate will generally not be crucial either, because the log likelihood

$$L(\beta) = \sum_{i=1}^{n} \underbrace{y_i \theta_i - log(y_i!)}_{\text{linear in } \theta} \underbrace{-exp(\theta_i)}_{\text{concave in } \theta}$$

is a strictly concave function in the natural parameter θ here, so it has a unique global maximum. We choose

$$\beta^{(0)} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Now all that is left is to write a Newton-Raphson routine to get the iterates

$$\beta^{(t+1)} = \beta^{(t)} - (H^{(t)})^{-1} u^{(t)}.$$

Algorithm 1 Newton Algorithm for simple Poisson GLM fitting

```
while ||β − β̃|| > 10^−7 do
u, H = 0;
    for i = 0 to (time) T do
z = exp(β[0]+x[i]*β[1]);
u[0] += y[i] - z;
u[1] += (y[i] - z)*x[i];
H[0][0] -= z*x[i]*x[i];                    ▷ H will be inverted Hessian
H[0][1] += z*x[i];
H[1][1] -= z;
    end for
D = H[0][0]*H[1][1] - H[0][1]*H[0][1];     ▷ H symmetric
    if |D| < 10^−6 then
printf("Hessian matrix is not invertible in step .."); return 0;
    end if
H= H/D;
```

16: $\tilde{\beta} = \beta$;
17: $\beta[0]$ -= (H[0][0]*u[0] + H[0][1]*u[1]);
18: $\beta[1]$ -= (H[0][1]*u[0] + H[1][1]*u[1]);
19: ▷ here stop criterion in implementation
20: **end while**
21: return β;

After 10 iterations the algorithm converges to

$$\hat{\beta} = \begin{pmatrix} 5,19740 \\ 0,04686 \end{pmatrix}.$$

So the predicted claims expenditures in the Poisson GLM for the next year are

$$\exp(5,19740 + 0,04686 \times 11) = 302,7 \text{ [million euros]},$$

which are 6 million more than the prediction in the ordinary linear model.

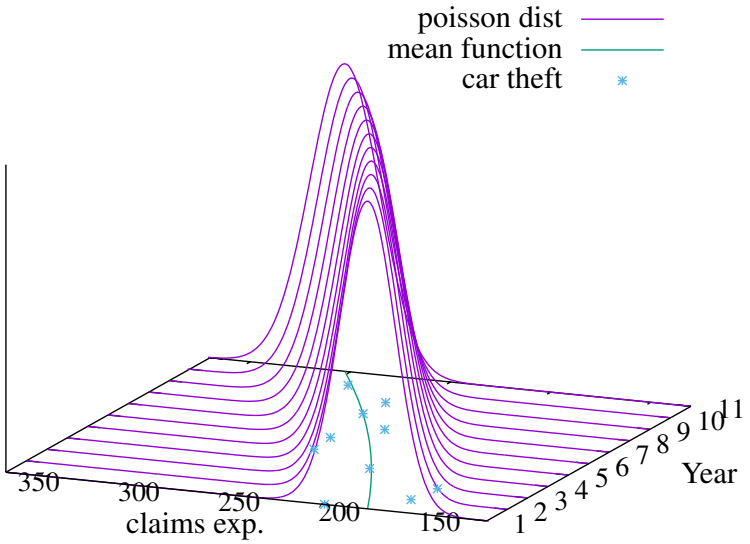

Figure 2.1: Poisson GLM

The graphic shows the exponential estimation function in a Poisson GLM with canonical link. Furthermore, one can see that the variance grows with the fitted expected values in the model, since we assume an underlying Poisson distribution.

As already stated, the Poisson distribution has shown to be suited in many applications for the estimation of claims expenditures. However, there is no mathematical evidence that this assumption is valid for a new situation. Therefore, we will look at an alternative framework for regression in the next chapter, that does not make assumptions on the specific underlying distribution.
The results will be theoretically weaker then, of course, but especially when we do not have a justification for assuming a specific distribution, this next chapter will give us an opportunity to still make profound conclusions. Especially in the modern context of Big Data, i.e. having large sample sizes of some new data without having much structural information about it, this "Brute Force" approach is quite helpful.

Chapter 3

Regression models in the framework of Machine Learning

In Generalized Linear Models the distribution of the response variable Y methods was taken out of an exponential family. However, sometimes it may not be possible to make any reasonable assumptions on the underlying distributions. In that cases, we can use the alternative Machine Learning framework, where no assumptions on the specific distribution of Y have to be made.

3.1 Basic Framework

The wording for the elements of a regression model is different in the Machine Learning framework. We will give the main correspondences.

Examples are the sample data we have given.
Features correspond to the explanatory variables.
Labels correspond to the values of the response variables.

Definition 3.1. Let $Y \subseteq \mathbb{R}$ be the set of labels and $X \subseteq \mathbb{R}^p$ the set of feature vectors in a regression problem.
Then a hypothesis is a mapping $h : \mathrm{X} \to \mathrm{Y}$.
A set of hypotheses H is simply called hypothesis set.

Definition 3.2. A loss function

$$L : Y \times Y \to \mathbb{R}_+$$

measures the difference between a predicted label and a true label.
Suited to the statistical approach, a common loss function is the squared loss

$$L(y, y') = (y - y')^2.$$

Remark 3.3. We only consider bounded loss functions in this chapter. So when we are using the squared loss, Y will typically be a bounded interval $I \subset \mathbb{R}$. In most practical applications one can establish a reasonable bound for the response variable. For instance, in our example of claims requirements due to insured car theft, we can take the total value of all insured cars as upper bound and 0 as lower bound.

Definition 3.4 (Generalization Error)**.** Given a hypothesis $h \in H$, a loss function L and an underlying distribution D with density f over $Z = X \times Y$, i.e. a random variable $\widetilde{Z} : (X \times Y, \mathscr{B}(X \times Y), P) \to (\mathbb{R}, \mathscr{B}(\mathbb{R})), (x,y) \mapsto L(h(x),y)$ with density f, the generalization error of h is defined by

$$R(h) = \mathbb{E}_{(x,y)\sim D}[L(h(x),y)] = \mathbb{E}[\widetilde{Z}] = \int_Z L(h(x),y)f(z)dz. \tag{3.1}$$

However, the generalization error is not accessible to the learner since the underlying distribution D is considered as completely unknown in this chapter.
Therefore we define the empirical error of a hypothesis on the labeled sample S.

Definition 3.5 (Empirical Error)**.** Given a hypothesis $h \in H$, a loss function L and a sample $S = ((x_1,y_1),...,(x_m,y_m))$ drawn i.i.d. according to a distribution D, the empirical error of h is defined by

$$\hat{R}(h) = \frac{1}{m}\sum_{i=1}^{m} L(h(x_i),y_i). \tag{3.2}$$

We can directly check that the empirical error is unbiased.

Proposition 3.6.

$$\mathbb{E}[\hat{R}(h)] = R(h).$$

Proof. Consider random variables $Z_i^h : (X \times Y, \mathscr{B}(X \times Y), P) \to (\mathbb{R}, \mathscr{B}(\mathbb{R})), (x_i,y_i) \mapsto L(h(x_i),y_i)$ with densities $f_i = f \ \forall i = [1,m]$ and $\overline{Z}^h = 1/m \ \sum_{i=1}^m Z_i^h$. Then we have

$$\mathbb{E}_{S\sim D^m}[\hat{R}(h)] = \mathbb{E}[\overline{Z}^h] = \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[Z_i^h] = \mathbb{E}[Z_1^h] = \mathbb{E}_{(x,y)\sim D}\ [L(h(x),y)] = R(h)$$

due to the linearity of the expectation and since the sample is drawn i.i.d.

□

Remark 3.7. In the following we will mainly use the notations $\mathbb{E}_{(x,y)\sim D}$ and $\mathbb{E}_{S\sim D^m}$ instead of explicitly introducing corresponding random variables.

Given a hypothesis set H of functions $h : X \to Y$, the regression problem consists of using the labeled sample S to find a hypothesis $h \in H$ with small expected loss or generalization error $R(h)$.
As mentioned, we do not have direct access to the generalization error of a hypothesis. Nevertheless, it is possible to bound the generalization error in terms of the empirical error.

3.2 Generalization bounds - simplified case

First we consider the simplified case that the hypothesis set is finite. In that case we only need Hoeffding's inequality to derive generalization bounds.

Lemma 3.8 (Hoeffding's inequality)**.** *Let $X_1,...,X_m$ be independent random variables with X_i taking values in $[a_i,b_i]$ for all $i \in [1,m]$. Let $S_m = \sum_{i=1}^m X_i$. Then, for any $\varepsilon > 0$,*

$$Pr[S_m - \mathbb{E}[S_m] \geq \varepsilon] \leq \exp(-2\varepsilon^2 / \sum_{i=1}^m (b_i - a_i)^2),$$

$$Pr[S_m - \mathbb{E}[S_m] \leq -\varepsilon] \leq \exp(-2\varepsilon^2 / \sum_{i=1}^m (b_i - a_i)^2),$$

Proof. The proof mainly consists of applying Markov's inequality to $exp(t(S_m - \mathbb{E}[S_m]))$ for $t > 0$ and then using the following lemma.

Lemma 3.9 (Hoeffding's lemma)**.** *Let X be a random variable with $\mathbb{E}[X] = 0$ and $a \leq X \leq b$ with $b > a$. Then, for any $t > 0$, the following inequality holds:*

$$\mathbb{E}[e^{tX}] \leq \exp\left(\frac{t^2(b-a)^2}{8}\right).$$

Proof. The following proof of Hoeffding's lemma is given in [2, Appendix D]. Since $x \to e^{tx}$ is convex for all $t \in \mathbb{R}$, we have

$$e^{tx} \leq \frac{b-x}{b-a}e^{ta} + \frac{x-a}{b-a}e^{tb}$$

for all $x \in [a,b]$ and $t \in \mathbb{R}$. This yields

$$\mathbb{E}[e^{tX}] \leq \mathbb{E}[\frac{b-X}{b-a}e^{ta} + \frac{X-a}{b-a}e^{tb}] \overset{\mathbb{E}[X]=0}{=} \frac{b}{b-a}e^{ta} + \frac{-a}{b-a}e^{tb} = e^{\phi(t)}$$

with $\phi(t) = \log(\frac{b}{b-a}e^{ta} + \frac{-a}{b-a}e^{tb}) = ta + log(\frac{b}{b-a} + \frac{-a}{b-a}e^{t(b-a)})$.
Now calculating the first and second derivative of ϕ yields $\phi'(0) = 0$ and $\phi''(t) \leq \frac{(b-a)^2}{4}$ for all $t > 0$, for details see [2, Lemma D.1]. Therefore the second order Taylor expansion of ϕ implies that there exists $\theta \in [0,t]$, $t > 0$, such that

$$\phi(t) = \phi(0) + t\phi'(0) + \frac{t^2}{2}\phi''(\theta) \leq t^2\frac{(b-a)^2}{8}.$$

□

This lemma allows us to show Hoeffding's inequality. Since $x \to e^x$ is monotone, we have

$$Pr[S_m - \mathbb{E}[S_m] \geq \varepsilon] = Pr[e^{t(S_m - \mathbb{E}[S_m])} \geq e^{t\varepsilon}]$$

for all $t > 0$. Therefore we can apply Markov's inequality to get

$$Pr[S_m - \mathbb{E}[S_m] \geq \varepsilon] = Pr[e^{t(S_m - \mathbb{E}[S_m])} \geq e^{t\varepsilon}] \leq e^{-t\varepsilon}\mathbb{E}[e^{t(S_m - \mathbb{E}[S_m])}] = \prod_{i=1}^{m} e^{-t\varepsilon}\mathbb{E}[e^{t(X_i - \mathbb{E}[X_i])}],$$

because the X_i are independent. Using Lemma 3.9, this can be further bounded by

$$\prod_{i=1}^{m} e^{-t\varepsilon} e^{t^2(b_i-a_i)^2/8} = e^{-t\varepsilon} e^{t^2 \sum_{i=1}^{m}(b_i-a_i)^2/8} \leq e^{-2\varepsilon^2/\sum_{i=1}^{m}(b_i-a_i)^2}$$

by setting $t = 4\varepsilon/\sum_{i=1}^{m}(b_i - a_i)^2$.

$Pr[S_m - \mathbb{E}[S_m] \leq -\varepsilon] = Pr[\mathbb{E}[S_m] - S_m \geq \varepsilon]$ can be bounded analogously.

□

Remark 3.10. A well-known inequality to bound the probability of deviation from the mean is Chebyshev's inequality. It does not require fixed bounds on the values of the random variable but additionally to Hoeffding's inequality requires knowledge about its variance.

Note though, that even if we have that knowledge, Hoeffding's inequality can give sharper bounds in some cases.

Example 3.11. To see that, consider independent random variables $X_1, ..., X_m$ that map to $([0,1], \mathscr{B}[0,1])$, where $\mathbb{E}[X_i] = \mu$ and $Var[X_i] = \sigma^2$ for all $i \in 1, .., m$.
Since the X_i are independent, the variance of the sum $S_m = \sum_{i=1}^{m} X_i$ is given by $m\sigma^2$ and the variance of the empirical average $\overline{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$ by $\frac{\sigma^2}{m}$. Therefore Chebyshev's inequality yields the bound

$$Pr[|\overline{X} - \mu| \geq \varepsilon] \leq \frac{\sigma^2}{m\varepsilon^2}.$$

On the other hand, Hoeffding's inequality gives

$$Pr[|\overline{X} - \mu| \geq \varepsilon] = Pr[|S_m - \mathbb{E}[S_m]| \geq \varepsilon m] \leq 2e^{-2\varepsilon^2 m}.$$

For example, consider a sequence of unbiased coin tosses, i.e. $X_i : \{heads, tails\} \to \{0,1\}$ with $\mathbb{E}[X_i] = 0.5$ and $Var(X_i) = 0.25$ for all i. Furthermore, take $\varepsilon = 0.1$ and $m = 500$. Then Chebyshev's inequality gives $Pr[|\overline{X} - 0.5| \geq 0.1] \leq 0.05$, whereas Hoeffding's inequality gives the 500 times sharper bound $Pr[|\overline{X} - 0.5| \geq 0.1] \leq 0.0001$.
For a general sequence of (i.i.d.) biased coin tosses with $m = 500$ and $\varepsilon = 0.1$, we have the relationship

$$Pr.bound_{\text{Chebyshev}} = \frac{\sigma^2}{0.000454} Pr.bound_{\text{Hoeffding}},$$

i.e. the Chebyshev bound improves with lower variance in comparison.

As mentioned in Remark 3.3, we restrict ourselves to the case of bounded loss functions, that is, $L(y, y') \leq M$ for all $y, y' \in Y$. Under this assumption we can apply Hoeffding's inequality to establish generalization bounds for finite hypothesis sets.

Theorem 3.12. *Assume that the loss function is bounded by M and that the hypothesis set H is finite. Then, for any $\delta > 0$, with probability at least $1-\delta$, the following inequality holds for all $h \in H$:*

$$R(h) \leq \hat{R}(h) + M\sqrt{\frac{log|H| + log(1/\delta)}{2m}}.$$

Proof. We have already established that $\mathbb{E}[\hat{R}(h)] = R(h)$. Therefore we can apply Hoeffding's inequality with $X_i = \frac{1}{m}L(h(x_i), y_i)$ taking values in $[0, \frac{M}{m}]$, which yields

$$Pr_{S\in D^m}[R(h) - \hat{R}(h) \geq \varepsilon] = Pr_{S\in D^m}[\hat{R}(h) - R(h) \leq -\varepsilon] \leq exp(-2m\varepsilon^2/M^2). \tag{3.3}$$

for any $h \in H$. By the union bound, this implies

$$Pr[\exists h \in H|\ R(h) - \hat{R}(h) > \varepsilon] \leq \sum_{h\in H} Pr[R(h) - \hat{R}(h) > \varepsilon] \leq |H|exp(-2m\varepsilon^2/M^2). \tag{3.4}$$

Now we can set $|H|exp(-2m\varepsilon^2/M^2) =: \delta\ (>0)$ and therewith replace ε by δ in (3.4). After turning around the inequality this yields

$$Pr_{S\in D^m}\left[\forall h \in H \,\middle|\, R(h) - \hat{R}(h) \leq M\sqrt{\frac{log|H| + log(1/\delta)}{2m}}\right] \geq 1-\delta.$$

□

3.3 Rademacher complexity

The main task is now to extend the preceding theorem to infinite hypothesis sets. To do that, we will reduce the infinite set of hypotheses to the analysis of finite sets using the so-called Rademacher complexity notion.

To each $h : X \to Y$, we can associate a function g that maps $(x,y) \in X \times Y$ to $L(h(x),y)$, whereby L is some loss function that can be bounded by an interval $[0,M]$. So the following formal definitions of the empirical and average Rademacher complexity fit to our problem.

Definition 3.13. Let G be a family of (measurable) functions mapping from $Z = X \times Y$ to $[a,b]$ and $S = (z_1,...,z_m)$ a fixed sample of size m with elements in Z. Then, the empirical Rademacher complexity of G with respect to the sample S is defined as

$$\hat{\mathscr{R}}_S(G) = \mathbb{E}_\sigma[\sup_{g\in G} \frac{1}{m}\sum_{i=1}^{m} \sigma_i g(z_i)],$$

where $\sigma = (\sigma_1,..,\sigma_m)^T$ with σ_i being a independent uniform random variable taking values in $\{-1,+1\}$. The random variables σ_i are called Rademacher variables or random noise.

Remark 3.14. If we have a countable family of measurable functions G, it can easily be shown that the $\sup_{g\in G} g(z)$ is measurable as well, since any $\sigma-$algebra is closed under countable unions and intersections.

For families of functions with uncountable index set this can theoretically fail.
However, since according to [2] it tends to work for uncountable families of hypotheses and their associated loss functions as well, we do not exclude them from the definition, but just mention it here and assume measurable suprema in the following.

Proposition 3.15. *The empirical Rademacher complexity $\hat{\mathscr{R}}_S(G)$ is nonnegative.*

Proof. Consider the function $H(\sigma) = \sup_{g\in G} \frac{1}{m}\sum_{i=1}^{m} \sigma_i g(z_i)$. $H(\sigma)$ is convex as the supremum of linear functions. Thus Jensen's inequality yields

$$\mathbb{E}_\sigma[\sup_{g\in G} \frac{1}{m}\sum_{i=1}^{m} \sigma_i g(z_i)] \geq \sup_{g\in G} \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}[\sigma_i] g(z_i) = 0,$$

since $\mathbb{E}[\sigma_i] = 0 \ \ \forall i$.

□

Remark 3.16. An interpretation of the empirical Rademacher complexity is given in [2]: Let g_S denote the vector of values taken by function g over the sample S:

$$g_S = (g(z_1), ..., g(z_m))^T.$$

Then, the empirical Rademacher complexity can be rewritten as

$$\hat{\mathscr{R}}_S(G) = \mathbb{E}_\sigma[\sup_{g\in G} \frac{\sigma \cdot g_S}{m}].$$

Notice that, since $\mathbb{E}(\sigma) = 0$, the covariance $Cov(g_S, \sigma) = \mathbb{E}[g_S \cdot \sigma] - \mathbb{E}[g_S] \cdot \mathbb{E}[\sigma] = \mathbb{E}[g_S \cdot \sigma]$ corresponds to the inner product of g_S and σ, so the inner product $\sigma \cdot g_s$ measures the correlation of g_s with the vector σ. Thus, the supremum $sup_{g\in G}\ \sigma \cdot g_s/m$ measures the correlation of the function class G with σ over the sample S.
So overall, the empirical Rademacher complexity measures how well the function class G correlates with random noise over the sample S, on average.

Larger function classes tend to have a higher empirical Rademacher complexity since the supremum is taken over more functions. Furthermore, the empirical Rademacher complexity of a function class that contains only a single function is zero, because the terms in the expectation cancel each other out.

Definition 3.17. Let D denote the distribution according to which samples are drawn. For any integer $m \geq 1$, the Rademacher complexity of G is the expectation of the empirical Rademacher complexity over all samples of size m drawn according to D:

$$\mathscr{R}_m(G) = \mathbb{E}_{S\sim D^m}[\hat{\mathscr{R}}_S(G)].$$

To be able to establish generalization bounds based on Rademacher complexity, we will need to generalize Hoeffding's inequality. Therefore we first need to show Azuma's inequality, which is stated in the context of martingale difference sequences.

Definition 3.18. A stochastic process $\{V_n\}_{n\in\mathbb{N}}$ on $(\Omega,\mathscr{F},P)$ to $(\mathbb{R},\mathscr{B}(\mathbb{R}))$ is a martingale difference sequence with respect to a stochastic process $\{X_n\}_{n\in\mathbb{N}}:(\Omega,\mathscr{F})\to(\mathbb{R},\mathscr{B}(\mathbb{R}))$, if for all $i>0$, $V_i\in\sigma(X_1,..,X_i)$ and

$$\mathbb{E}[V_{i+1}|X_1,..,X_i]=0,$$

i.e. if V_i is $\sigma(X_1,..,X_i)$-measurable and the conditional expectation of V_{i+1} with respect to the sub-σ-field $\sigma(X_1,..,X_i)$ of $\mathscr{F}$ is zero for all $i>0$.

Remark 3.19. Note that this definition implies that if $\{Y_t\}_{t\in\mathbb{N}_0}$ is a martingale, then $\{W_t\}_{t\in\mathbb{N}}=\{Y_t-Y_{t-1}\}_{t\in\mathbb{N}}$ is indeed a martingale difference sequence with respect to the σ-algebra generated by $\{Y_t\}_{t\in\mathbb{N}_0}$, since for all $t>0$, $W_t=Y_t-Y_{t-1}\in\sigma(Y_0,..,Y_t)$, and

$$\mathbb{E}[W_{t+1}|Y_0,..,Y_t]=\mathbb{E}[Y_{t+1}|Y_0,..,Y_t]-\mathbb{E}[Y_t|Y_0,..,Y_t]=Y_t-Y_t=0.$$

Lemma 3.20 (Generalized Hoeffding lemma)**.** *Let V and $X_1,..,X_n$ be random variables satisfying $\mathbb{E}[V|X_1,..,X_n]=0$ and, for some $\sigma(X_1,..,X_n)$-measurable random variable Z and constants a and b, the inequalities:*

$$Z+a\le V\le Z+b.$$

Then, for all $t>0$, the following upper bound holds:

$$\mathbb{E}[e^{tV}|X_1,..,X_n]\le e^{t^2(b-a)^2/8}.$$

In particular, we get the usual Hoeffding lemma for $\sigma(X_1,..X_n)=\{\emptyset,\Omega\}$.

Proof. Note that we can follow the proof of Lemma 3.9 by taking $Z+a,\ Z+b$ instead of a,b and considering conditional expectations $\mathbb{E}(\cdot|X_1,..,X_n)$ instead of $\mathbb{E}(\cdot)$, since

$$\mathbb{E}[\frac{Z+b-V}{b-a}e^{t(Z+a)}+\frac{V-Z-a}{b-a}e^{t(Z+b)}|X_1,..,X_n]=\frac{Z+b}{b-a}e^{t(Z+a)}-\frac{Z+a}{b-a}e^{t(Z+b)}$$

because Z is $\sigma(X_1,..,X_n)$-measurable and $\mathbb{E}[V|X_1,..,X_n]=0$.

□

For a martingale $\{Y_t\}_{t\in\mathbb{N}_0}$, so a martingale difference sequence $\{W_t\}_{t\in\mathbb{N}}=\{Y_t-Y_{t-1}\}_{t\in\mathbb{N}}$, we have $\sum_{i=1}^m W_i=Y_m-Y_0$. In this aspect the following inequality establishes a bound on martingales with bounded differences, using the generalized Hoeffding lemma.

Lemma 3.21 (Azuma's inequality)**.** *Let $\{V_n\}_{n\in\mathbb{N}}$ be a martingale difference sequence with respect to $\{X_n\}_{n\in\mathbb{N}}$, and assume that for all $i>0$ there is a constant $c_i\ge 0$ and a random variable $Z_i\in\sigma(X_1,..,X_{i-1})$ that satisfies*

$$Z_i\le V_i\le Z_i+c_i.$$

Then, for all $\varepsilon>0$ and m, the following inequalities hold:

$$Pr\left[\sum_{i=1}^m V_i\ge\varepsilon\right]\le\exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^m c_i^2}\right),$$

$$Pr\left[\sum_{i=1}^m V_i\le-\varepsilon\right]\le\exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^m c_i^2}\right),$$

Proof. Define $S_k = \sum_{i=1}^{k} V_k$ for any $k \in [1,m]$. We have

$$Pr[S_m \geq \varepsilon] \leq e^{-t\varepsilon}\mathbb{E}[e^{tS_m}]$$

for all $t > 0$ as in the proof of Hoeffding's inequality. Due to the law of total expectation and since $e^{tS_{m-1}}$ is $\sigma(X_1,..,X_{m-1})$-measurable, this term equals

$$e^{-t\varepsilon}\mathbb{E}\left[\mathbb{E}[e^{tS_m}]|X_1,..,X_{m-1}\right] = e^{-t\varepsilon}\mathbb{E}\left[e^{tS_{m-1}}\mathbb{E}[e^{tV_m}|X_1,..,X_{m-1}]\right] \leq e^{-t\varepsilon}\mathbb{E}[e^{tS_{m-1}}]e^{t^2c_m^2/8}$$

due to Lemma 3.20 with Z_m, $a = 0$ and $b = c_m$.
Iterating that argument yields the bound

$$e^{-t\varepsilon}e^{t^2\sum_{i=1}^{m}c_i^2/8} = e^{-2\varepsilon^2/\sum_{i=1}^{m}c_i^2}$$

with $t = 4\varepsilon/\sum_{i=1}^{m}c_i^2$.
Again, $Pr\left[\sum_{i=1}^{m}V_i \leq -\varepsilon\right] = Pr\left[\sum_{i=1}^{m}-V_i \geq \varepsilon\right]$ can be bounded analogously, since $\{-V_n\}_{n\in\mathbb{N}}$ is also a martingale difference sequence with respect to $\{X_n\}_{n\in\mathbb{N}}$ and $-Z_i + c_i \leq -V_i \leq -Z_i$. □

Now we can use Azuma's inequality to show the desired generalization to Hoeffding's inequality.

Theorem 3.22 (McDiarmid's inequality)**.** *Let $X_1,..,X_m$ be independent random variables taking values in the set $\mathscr{X}$ and assume that there exist $c_1,...,c_m > 0$ such that $f : \mathscr{X}^m \to \mathbb{R}$ satisfies the following conditions:*

$$|f(x_1,..,x_i,..,x_m) - f(x_1,..,x_i',..,x_m)| \leq c_i, \tag{3.5}$$

for all $i \in [1,m]$ and any points $x_1,...,x_m,x_i' \in \mathscr{X}$. Let $f(S)$ denote $f(X_1,...,X_m)$, then, for all $\varepsilon > 0$, the following inequalities hold:

$$Pr[f(S) - \mathbb{E}[f(S)] \geq \varepsilon] \leq exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^{m}c_i^2}\right),$$

$$Pr[f(S) - \mathbb{E}[f(S)] \leq -\varepsilon] \leq exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^{m}c_i^2}\right).$$

Proof. Define random variables

$V = f(S) - \mathbb{E}[f(S)]$, $V_1 = \mathbb{E}[V|X_1] - \mathbb{E}[V]$ and $V_k = \mathbb{E}[V|X_1,..,X_k] - \mathbb{E}[V|X_1,..,X_{k-1}]$ for $k \in [2,m]$.

Then we have

$$V = \mathbb{E}[V|X_1,..,X_m] = \mathbb{E}[V|X_1,..,X_m] - \underbrace{\mathbb{E}[V]}_{=0} = \sum_{k=1}^{m} V_k.$$

Furthermore, by the tower property of the conditional expectation it holds that

$$\mathbb{E}[\mathbb{E}[V|X_1,..,X_k]|X_1,..,X_{k-1}] = \mathbb{E}[V|X_1,..,X_{k-1}].$$

This implies

$$\mathbb{E}[V_k|X_1,..,X_{k-1}] = \mathbb{E}[\mathbb{E}[V|X_1,..,X_k]|X_1,..,X_{k-1}] - \underbrace{\mathbb{E}[\mathbb{E}[V|X_1,..,X_{k-1}]|X_1,..,X_{k-1}]}_{=\mathbb{E}[V|X_1,..,X_{k-1}]} = 0,$$

so $\{V_k\}_{k\in[1,m]}$ is a martingale difference sequence with respect to $\{X_k\}_{k\in[1,m]}$.

Until now we have that $f(S) - \mathbb{E}[f(S)]$ can be expressed as a sum over a martingale difference sequence. Note, that we can express the summands as

$$V_k = \mathbb{E}[f(S)|X_1,..,X_k] - \mathbb{E}[f(S)|X_1,..,X_{k-1}],$$

because $\mathbb{E}[\mathbb{E}[f(S)]|X_1,..X_k] = \mathbb{E}[f(S)] = \mathbb{E}[\mathbb{E}[f(S)]|X_1,..X_{k-1}]$, so that terms cancel each other out. Therefore

$$W_k = \sup_x \mathbb{E}[f(S)|X_1,..,X_{k-1},x] - \mathbb{E}[f(S)|X_1,..,X_{k-1}]$$

is an upper bound for V_k and

$$U_k = \inf_x \mathbb{E}[f(S)|X_1,..,X_{k-1},x] - \mathbb{E}[f(S)|X_1,..,X_{k-1}]$$

is a lower bound for V_k.
By (3.5) it holds for any $k \in [1,m]$ that

$$W_k - U_k = \sup_{x,x'} \mathbb{E}[f(S)|X_1,..,X_{k-1},x] - \mathbb{E}[f(S)|X_1,..,X_{k-1},x'] \leq c_k,$$

i.e. $U_k \leq V_k \leq U_k + c_k$ with $U_k \in \sigma(X_1,..,X_{k-1})$.
Therefore we can apply Azuma's inequality to $V = f(S) - \mathbb{E}[f(S)] = \sum_{i=1}^m V_k$, which yields the statement.

□

Remark 3.23. Note, that Hoeffding's inequality is indeed a special case of McDiarmid's inequality by setting $f(x_1,..,x_m) = \sum_{i=1}^m x_i$.

Using McDiarmid's inequality we will now proof a generalization bound result for loss functions that can be bounded by $[0,1]$. For the general result in regression, we will have to rescale that bound to an interval $[0,M]$ with M arbitrary in the next section.

Theorem 3.24. *Let G be a family of functions mapping from Z with an underlying distribution D to $[0,1]$ and $\{z_1,..,z_m\}$ a fixed sample from Z. Then, for any $\delta > 0$, with probability at least $1-\delta$, each of the following holds for all $g \in G$:*

$$\mathbb{E}_{z\sim D}[g(z)] \leq \frac{1}{m}\sum_{i=1}^m g(z_i) + 2\mathscr{R}_m(G) + \sqrt{\frac{log(1/\delta)}{2m}} \text{ and}$$

$$\mathbb{E}_{z\sim D}[g(z)] \leq \frac{1}{m}\sum_{i=1}^m g(z_i) + 2\hat{\mathscr{R}}_S(G) + 3\sqrt{\frac{log(2/\delta)}{2m}}.$$

Proof. In [2, Chapter 3.1] the following proof is given.
For any sample $S = (z_1, ..., z_m)$ and any $g \in G$, we denote by $\hat{\mathbb{E}}^S[g]$ the empirical average of g over S, i.e.

$$\hat{\mathbb{E}}^S[g] = \frac{1}{m}\sum_{i=1}^{m} g(z_i).$$

The proof consists of applying McDiarmid's inequality to the function Φ defined for any sample S by

$$\Phi(S) = \sup_{g\in G} \mathbb{E}[g] - \hat{\mathbb{E}}^S[g].$$

Let S and S' be two samples differing by exactly one point, say z_m in S and z'_m in S'. Then, since the difference of suprema does not exceed the supremum of the difference, we have

$$\Phi(S') - \Phi(S) \leq \sup_{g\in G}\left(\hat{\mathbb{E}}^S[g] - \hat{\mathbb{E}}^{S'}[g]\right) = \sup_{g\in G}\frac{g(z_m) - g(z'_m)}{m} \leq \frac{1}{m},$$

because g is bounded by $[0,1]$. Similarly, we can obtain $\Phi(S) - \Phi(S') \leq \frac{1}{m}$, thus

$$|\Phi(S) - \Phi(S')| \leq \frac{1}{m}.$$

Therefore applying McDiarmid's inequality gives

$$Pr[\Phi(S) - \mathbb{E}[\Phi(S)] \geq \varepsilon] \leq \exp(-2m\varepsilon^2)$$

Setting the right hand side to be equal to $\delta/2$, analogous to the proof of Theorem 3.12, yields for any $\delta > 0$, with probability at least $1 - \frac{\delta}{2}$,

$$\Phi(S) \leq \mathbb{E}[\Phi(S)] + \sqrt{\frac{\log(2/\delta)}{2m}}. \tag{3.6}$$

The expectation on the right-hand side can be rewritten to

$$\mathbb{E}_S[\Phi(S)] = \mathbb{E}_S\left[\sup_{g\in G}\mathbb{E}_S(g) - \hat{\mathbb{E}}^S(g)\right] = \mathbb{E}_S\left[\sup_{g\in G}\mathbb{E}_{S'}[\hat{\mathbb{E}}^{S'}(g) - \hat{\mathbb{E}}^S(g)]\right],$$

because the empirical average is unbiased, so $\mathbb{E}_S[g] = \mathbb{E}_{S'}[\hat{\mathbb{E}}^{S'}(g)]$.
This can be bounded by

$$\mathbb{E}_{S,S'}[\sup_{g\in G}\hat{\mathbb{E}}^{S'}(g) - \hat{\mathbb{E}}^S(g)] = \mathbb{E}_{S,S'}[\sup_{g\in G}\frac{1}{m}\sum_{i=1}^{m}(g(z'_i) - g(z_i))],$$

because the supremum function is convex, so Jensen's inequality can be applied.
Now we can add Rademacher variables σ_i to the term,

$$\mathbb{E}_{\sigma,S,S'}[\sup_{g\in H}\frac{1}{m}\sum_{i=1}^{m}\sigma_i(g(z'_i) - g(z_i))],$$

which does not change the expectation, because the integration over S is equivalent to the one over S', so for $\sigma_i = -1$ just the order of the summands within the expectation is swapped. Since the supremum is sub-additive, this term can further be bounded by

$$\mathbb{E}_{\sigma,S'}[\sup_{g\in G}\frac{1}{m}\sum_{i=1}^{m}\sigma_i g(z_i')] + \mathbb{E}_{\sigma,S}[\sup_{g\in G}\frac{1}{m}\sum_{i=1}^{m}-\sigma_i g(z_i)] = 2\,\mathbb{E}_{\sigma,S}[\sup_{g\in G}\frac{1}{m}\sum_{i=1}^{m}\sigma_i g(z_i)] = 2\,\mathscr{R}_m(G),$$

because σ_i and $-\sigma_i$ are identically distributed.
So overall we have, with probability at least $1-\delta$,

$$\sup_{g\in G}\mathbb{E}[g] - \hat{\mathbb{E}}^S[g] = \Phi(S) \leq \mathbb{E}[\Phi(S)] + \sqrt{\frac{\log(1/\delta)}{2m}} \leq 2\,\mathscr{R}_m(G) + \sqrt{\frac{\log(1/\delta)}{2m}},$$

which implies the first inequality in the theorem.
To get to the second inequality, observe that changing one point in S changes $\hat{\mathscr{R}}_S(G)$ by at most $1/m$. Therefore we can apply McDiarmid's inequality again, to get, with probability at least $1-\delta/2$,

$$\mathscr{R}_m(G) \leq \hat{\mathscr{R}}_S(G) + \sqrt{\frac{log(2/\delta)}{2m}}.$$

Combined with (3.6) this yields, with probability at least $1-\delta$,

$$\sup_{g\in G}\mathbb{E}_S[g] - \hat{\mathbb{E}}^S[g] = \Phi(S) \leq 2\hat{\mathscr{R}}_S(G) + 3\sqrt{\frac{log(2/\delta)}{2m}}.$$

□

3.4 Generalization bounds

Now we apply the Rademacher complexity notion to generalize Theorem 3.12 to infinite hypothesis sets.
To do that, we first establish an upper bound on the Rademacher complexity of bounded L_p loss functions using the following lemma.
It tells us that if all hypotheses in a hypothesis set H get linked with a Lipschitz function, then the change of the Rademacher complexity can be bounded by the corresponding Lipschitz constant.

Lemma 3.25 (Talagrand's lemma)**.** *Let $\Phi : \mathbb{R} \to \mathbb{R}$ an L-Lipschitz function. Then, for any hypothesis set H of real-valued functions, the following inequality holds:*

$$\hat{\mathscr{R}}_S(\Phi \circ H) \leq L\,\hat{\mathscr{R}}_S(H).$$

Proof. See [2, Lemma 4.2]. □

Theorem 3.26. *Let $p \geq 1$ and $H_p = \{(x,y) \to |h(x)-y|^p| : h \in H\}$. Assume that $|h(x)-y| \leq M$ for all $(x,y) \in \mathrm{X} \times \mathrm{Y}$ and $h \in H$. Then, for any sample S, the following inequality holds:*

$$\hat{\mathscr{R}}_S(H_p) \leq pM^{p-1}\hat{\mathscr{R}}_S(H). \tag{3.7}$$

Proof. Consider the function $\phi_p : x \to |x|^p$. We have $H_p = \{\phi_p \circ h : h \in H'\}$, where $H' = \{(x,y) \to h(x)-y : h \in H\}$.
Note, that ϕ_p is Lipschitz over $[-M,M]$ with constant pM^{p-1} for $p \geq 1$, simply by the mean value theorem. Using Lemma 3.25 this gives

$$\hat{\mathscr{R}}_S(H_p) \leq pM^{p-1}\hat{\mathscr{R}}_S(H').$$

Furthermore,

$$\hat{\mathscr{R}}_S(H') = \frac{1}{m}\mathbb{E}_\sigma[\sup_{h\in H}\sum_{i=1}^{m}(\sigma_i h(x_i)-\sigma_i y_i)] = \frac{1}{m}\mathbb{E}_\sigma[\sup_{h\in H}\sum_{i=1}^{m}\sigma_i h(x_i)] - \frac{1}{m}\mathbb{E}_\sigma[\sum_{i=1}^{m}\sigma_i y_i] = \hat{\mathscr{R}}_S(H),$$

since $\mathbb{E}_\sigma[\sum_{i=1}^m \sigma_i y_i] = \sum_{i=1}^m \mathbb{E}_\sigma[\sigma_i] y_i = 0$.

□

This finally allows us to state Rademacher complexity bounds for regression with bounded L_p loss functions.

Theorem 3.27. *Let $p \geq 1$ and $H_p = \{(x,y) \to |h(x)-y|^p| : h \in H\}$. Assume that $|h(x)-y| \leq M$ for all $(x,y) \in \mathrm{X} \times \mathrm{Y}$ and $h \in H$. Then, for any $\delta > 0$, with probability at least $1-\delta$ over a sample S of size m, each of the following holds for all $h \in H$:*

$$\mathbb{E}[|h(x)-y|^p] \leq \frac{1}{m}\sum_{i=1}^{m}|h(x_i)-y_i|^p + 2pM^{p-1}\mathscr{R}_m(H) + M^p\sqrt{\frac{log(1/\delta)}{2m}}$$

$$\mathbb{E}[|h(x)-y|^p] \leq \frac{1}{m}\sum_{i=1}^{m}|h(x_i)-y_i|^p + 2pM^{p-1}\hat{\mathscr{R}}_S(H) + 3M^p\sqrt{\frac{log(2/\delta)}{2m}}.$$

Proof. Theorem 3.24 applied to $g(x,y) := (|h(x)-y|/M)^p$ and the upper bound (3.7) directly yield the result.

□

Regression algorithms

Since the generalization bounds in Theorem 3.27 are depended on the empirical errors, a direct approach for regression algorithms is to seek the hypothesis that minimizes the empirical error over all $h \in H$.
Moreover, we have also seen that the generalization bounds are better when we are using hypothesis sets with relatively low Rademacher complexity. A function class that typically fulfills this are linear functions.
So again, as a first algorithm, we consider the ordinary linear regression, now stated in the Machine Learning framework.

Example 3.28. (Linear regression) Consider the family of linear hypotheses,

$$H = \{x \to w \cdot x + b : w \in \mathbb{R}^p, b \in \mathbb{R}\}.$$

Now in linear regression we determine the hypothesis in H with the smallest empirical mean squared error. Thus, for a sample $S = ((x_1, y_1), ..., (x_n, y_n)) \in (X \times Y)^n$, we get the corresponding optimization problem

$$\min_{w,b} \frac{1}{n} \sum_{i=1}^{n} (w \cdot x_i + b - y_i)^2$$

or in matrix form,

$$\min_{W} F(W) = \frac{1}{n} ||XW - Y||^2,$$

where

$$X = \begin{bmatrix} 1 & x_1 \\ .. & .. \\ .. & .. \\ 1 & x_n \end{bmatrix}, \; W = \begin{bmatrix} b \\ w_1 \\ .. \\ w_p \end{bmatrix} \text{ and } \; Y = \begin{bmatrix} y_1 \\ .. \\ .. \\ y_n \end{bmatrix}.$$

We have already established the solution to this optimization problem in the preliminaries.

In the GLM chapter we used link functions to extend the linear regression and to model nonlinear relationships. In the framework of Machine Learning, this is done by using specific kernel functions.

3.5 Kernel functions

Definition 3.29 (Kernel)**.** A function $K : X \times X \to \mathbb{R}$ is called a kernel over X.

Now, rather that applying a non-linear function to the response as in the GLM chapter, we transform the input space with some non-linear mapping. Also mappings to spaces with much higher dimension will be possible. In those spaces we will then seek for linear relationships again.
Thereby, instead of explicitly constructing some non-linear mapping from the input space X to a possibly high-dimensional feature space and computing inner products in that, so-called PDS kernels will allow us to implicitly do that.

Definition 3.30 (PDS kernel)**.** A kernel $K : X \times X \to \mathbb{R}$ is said to be positive definite symmetric (PDS) if for any $(x_1, ..., x_m) \subseteq X$, the matrix $\mathbb{K} = [K(x_i, x_j)]_{i,j} \in \mathbb{R}^{m \times m}$ is symmetric positive semidefinite.

Lemma 3.31 (Cauchy-Schwarz inequality for PDS kernels)**.** *Let $K : X \times X \to \mathbb{R}$ be a PDS kernel. Then, for any $x, x' \in X$,*

$$K(x, x')^2 \leq K(x, x) K(x', x').$$

Proof. Since K is a PDS kernel, the matrix

$$\mathbb{K} = \begin{pmatrix} K(x,x) & K(x,x') \\ K(x',x) & K(x',x') \end{pmatrix}$$

is symmetric positive semidefinite for all $x,x' \in X$. Therefore the product of its eigenvalues, $det(\mathbb{K})$, is non-negative for all $x,x' \in X$, i.e.

$$det(\mathbb{K}) = K(x,x)K(x',x') - K(x,x')K(x',x) \overset{\mathbb{K} \text{ symmetric}}{=} K(x,x)K(x',x') - K(x,x')^2 \geq 0.$$

□

The following theorem shows that a PDS kernel does indeed correspond to an inner product in some Hilbert space and can therefore be used to implicitly calculate inner products in higher dimensional spaces. Of course, it will be crucial again in applications to use suited kernel functions that produce linear relationships.

Theorem 3.32. *Let $K : X \times X \to \mathbb{R}$ be a PDS kernel. Then, there exists a Hilbert space $\mathbb{H}$ and a mapping Φ from X to $\mathbb{H}$ such that:*

$$\forall x,x' \in X, \;\; K(x,x') = \langle \Phi(x), \Phi(x') \rangle.$$

$\mathbb{H}$ is called a feature space and Φ a feature mapping associated to K.

Proof. The following proof is given in [2, Theorem 5.2]. In that proof we construct a specific space and an operation on it, which is then shown to be an inner product.

For any $x \in X$, define $\Phi(x) : X \to \mathbb{R}$ via

$$\forall x' \in X : \Phi(x)(x') = K(x,x').$$

Let $\mathbb{H}_0$ be the set of finite linear combinations of such functions $\Phi(x)$, i.e.

$$\mathbb{H}_0 = \left\{ \sum_{i \in I} a_i \Phi(x_i) : \; a_i \in \mathbb{R}, \; x_i \in X, \; card(I) < \infty \right\},$$

which forms a vector space over $\mathbb{R}$.
Now we define an operation $\langle \cdot, \cdot \rangle$ on $\mathbb{H}_0 \times \mathbb{H}_0$ for all $f,g \in \mathbb{H}_0$ with $f = \sum_{i \in I} a_i \Phi(x_i)$ and $g = \sum_{j \in J} b_j \Phi(x_j)$ by

$$\langle f,g \rangle = \sum_{i \in I, j \in J} a_i b_j K(x_i,x'_j) = \sum_{j \in J} b_j f(x'_j) = \sum_{i \in I} a_i g(x_i).$$

At first, observe that $\langle \cdot, \cdot \rangle$ is symmetric. Next, the last two equations show that $\langle f,g \rangle$ does not depend on the particular representations of f and g, i.e. that $\langle \cdot, \cdot \rangle$ is well-defined. Moreover, they show that $\langle \cdot, \cdot \rangle$ is bilinear. Furthermore we have for any $f = \sum_{i \in I} a_i \Phi(x_i) \in \mathbb{H}_0$,

$$\langle f,f \rangle = \sum_{i,j \in I} a_i a_j K(x_i,x_j) \geq 0 \tag{3.8}$$

because K is a PDS kernel. Thus, $\langle\cdot,\cdot\rangle$ is a symmetric, positive semidefinite bilinear form. So what is left to show for $\langle\cdot,\cdot\rangle$ to be an inner product on $\mathbb{H}_0$ is $\langle f,f\rangle = 0 \Leftrightarrow f = 0$.

Inequality (3.8) actually also implies, combined with the bilinearity of $\langle\cdot,\cdot\rangle$, that

$$\forall f_1,..,f_m \in \mathbb{H}_0\ \forall c_1,..,c_m \in \mathbb{R} : \sum_{i,j=1}^{m} c_i c_j \langle f_i, f_j\rangle = \left\langle \sum_{i=1}^{m} c_i f_i, \sum_{j=1}^{m} c_j f_j \right\rangle \geq 0,$$

Thus, $\langle\cdot,\cdot\rangle$ is a PDS kernel on $\mathbb{H}_0$ and we can apply Lemma 3.31 on $\mathbb{H}_0$ to get

$$\forall f \in \mathbb{H}_0\ \ \forall x \in X : \langle f, \Phi(x)\rangle^2 \leq \langle f,f\rangle\langle \Phi(x),\Phi(x)\rangle. \tag{3.9}$$

Moreover, by the definition of $\langle\cdot,\cdot\rangle$, we have

$$\forall f = \sum_{i\in I} a_i \Phi(x_i) \in \mathbb{H}_0\ \ \forall x \in X : f(x) = \sum_{i\in I} a_i K(x_i,x) = \langle f, \Phi(x)\rangle. \tag{3.10}$$

This yields

$$\forall x \in X : |f(x)|^2 = |\langle f,\Phi(x)\rangle|^2 \overset{(3.9)}{\leq} \langle f,f\rangle\langle\Phi(x),\Phi(x)\rangle = \langle f,f\rangle K(x,x),$$

i.e. $\langle f,f\rangle > 0$ if there exists $x \in X$ such that $|f(x)| > 0$. Clearly, also $\langle f,f\rangle = 0$ for $f = 0$. So $\langle\cdot,\cdot\rangle$ is positive definite and defines an inner product on $\mathbb{H}_0$, which thereby becomes a pre-Hilbert space. To conclude, recall that the pre-Hilbert space $\mathbb{H}_0$ can be completed to form a Hilbert space $\mathbb{H}$ in which it is dense.

□

Remark 3.33. In the preceding proof we constructed a feature space $\mathbb{H}$ as a Hilbert space of functions $X \to \mathbb{R}$. In that sense (3.10) also implies that the constructed feature space is a so-called reproducing kernel Hilbert space, that is $\forall f \in \mathbb{H}, \forall x \in X : f(x) = \langle f, K(x,\cdot)\rangle$.
Note, that in those Hilbert spaces of functions point evaluation $\delta_x(f) = f(x)$ is a continuous linear functional, which is a useful property for further studies of kernel methods.

Example 3.34. Let us consider the family of polynomial kernels $K(x,x') = (x^T x' + c)^d$ on $X = \mathbb{R}^p$, for which we get $\mathbb{H} \cong \mathbb{R}^N$, i.e. finite-dimensional feature spaces.
Note, that the polynomial kernel $K(x,x') = (x^T x' + 1)^2$ on $X = \mathbb{R}^2$ corresponds to the inner product in $\mathbb{R}^6$ with $\Phi(x) = (x_1^2, x_2^2, 1, \sqrt{2}x_1x_2, \sqrt{2}x_1, \sqrt{2}x_2)$. More generally, it can be shown that for a polynomial kernel the dimension of a feature space is $N = \binom{p+d}{p}$. So we can extend the ordinary linear regression in this framework by considering the hypothesis set

$$H = \{x \mapsto w \cdot \Phi(x) + b : w \in \mathbb{R}^N, b \in \mathbb{R}\}$$

with a feature mapping $\Phi : X \to \mathbb{R}^N$ associated to some polynomial kernel K and the optimization problem

$$\min_{w,b}\ \frac{1}{n}\sum_{i=1}^{n} (w \cdot \Phi(x_i) + b - y_i)^2.$$

Thereby we do not have to explicitly calculate the inner products $w \cdot \Phi(x)$ in $\mathbb{R}^N$, but can use the kernel function K to implicitly do that.

Remark 3.35. We clearly always have a positive generalization error for all hypotheses, when the labels y are not given deterministically through the features x.
A hypothesis h^* with a minimal generalization error over all $h \in H$ is called best-in-class hypothesis.
Of course, we do not have access to find h^*, or $\inf_{h\in H} R(h)$ if h^* does not exist, in general. However, for algorithms that minimize the empirical error over all hypotheses, like the linear regression above, we have for the solution h_S that $\hat{R}(h_S) \leq \hat{R}(h^*)$, so

$$R(h_S) - R(h^*) = R(h_S) - \hat{R}(h_S) + \hat{R}(h_S) - R(h^*) \leq R(h_S) - \hat{R}(h_S) + \hat{R}(h^*) - R(h^*)$$

$$\leq 2 \sup_{h\in H} |R(h) - \hat{R}(h)| \leq 2\left(2pM^{p-1}\hat{\mathscr{R}}_S(H) + 3M^p\sqrt{\frac{log(2/\delta)}{2m}}\right),$$

if we consider bounded L_p loss functions due to Theorem 3.27.

Note, that the complexity of the hypothesis set H plays quite an interesting role in the generalization bound above. On the one hand, we get better guarantees for the selected hypothesis h_s to have a generalization error close to the best one $R(h^*)$ over all $h \in H$, when the hypothesis set H has a small (empirical) Rademacher complexity. On the other hand, the minimum value $R(h^*)$ itself will tend to go down when we consider larger hypothesis sets H with higher Rademacher complexity.

It turns out, that in practice algorithms perform better, which do not just minimize the empirical error over all $h \in H$, but have an added regularization term that penalizes complex hypotheses with respect to the Rademacher complexity.
We will show this approach in one of the most popular algorithms in the Machine Learning framework, the support vector machine.

3.6 Support Vector Regression

In support vector regression the loss function is generally chosen so, that only points outside of an ε tube around the predicted function are penalized.
Consider the hypothesis set H of linear functions $H = \{x \to w \cdot \Phi(x) + b : w \in \mathbb{R}^n, b \in \mathbb{R}\}$, where Φ is the feature mapping corresponding some PDS kernel K. Now support vector regression does not only focus on minimizing the error on the training sample but has an regularization term $||w||^2$ added that tries to reduce the model complexity.
Furthermore, a parameter C is added that determines the trade-off between the minimization of $||w||^2$, i.e. the minimization of the model complexity, and the minimization of the training errors. All in all, the optimization problem for Support Vector Regression (SVR) can be written as follows:

$$\min_{w,b} \frac{1}{2}||w||^2 + C\sum_{i=1}^{n} |y_i - (w \cdot \Phi(x_i) + b)|_\varepsilon,$$

where $|\cdot|_\varepsilon$ denotes an ε-insensitive loss function. We will consider the linear ε-insensitive loss:

$$\forall y, y' \in Y, |y' - y|_\varepsilon = \max(0, |y' - y| - \varepsilon).$$

Remark 3.36. Other popular loss functions in support vector regression are the quadratic ε-insensitive loss and the Huber loss, which penalizes larger errors linearly and smaller ones quadratically.
As one can already see, we are quite free in the choice of loss functions and kernels in this framework. The main aspect we have to care about is that it leads to convex optimization problems, which is guaranteed here when we use PDS kernels and convex loss functions like the ones mentioned.

Using slack variables $\xi_i \geq 0$ and $\xi_i' \geq 0$, $i \in [i,n]$, the optimization problem can be equivalently written as

$$\min_{w,b,\xi,\xi'} \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i'),$$

subject to

$$w \cdot \Phi(x_i) + b - y_i \leq \varepsilon + \xi_i$$
$$y_i - (w \cdot \Phi(x_i) + b) \leq \varepsilon + \xi_i'$$
$$\xi_i \geq 0, \xi_i' \geq 0, \ \forall i \in [1,n].$$

Recall, that for optimization with equality constraints generally the method of Lagrangian multipliers is being applied to convert the problem into an unconstrained one. However, here we have inequality constraints.
Therefore we will introduce the Karush-Kuhn-Tucker (KKT) approach, which generalizes the method of Lagrangian multipliers to inequality constraints.

Insertion: Optimization with inequality constraints

To do that, we first define an associated Lagrange function to the general optimization problem, similar to the method of Lagrangian multipliers. The Lagrangian can be built for any constrained optimization problem, but we will then focus on convex problems.

Definition 3.37. The Lagrange function or the Lagrangian associated to the general optimization problem

$$\min_{x \in X \subseteq \mathbb{R}^N} f(x), \quad f : X \to \mathbb{R},$$

subject to

$$g_i(x) \leq 0 \ \forall i, \quad g_i : X \to \mathbb{R}, \quad i \in \{1,..,n\}$$

is the function L defined over $X \times \mathbb{R}_+^n$ by

$$\forall x \in X, \forall \alpha \geq 0, \quad L(x,\alpha) = f(x) + \sum_{i=1}^{n} \alpha_i g_i(x),$$

where $\alpha = (\alpha_1,...,\alpha_n)^T$. The variables α_i are called Lagrange or dual variables.

Note, that the inequality constraints formulation is indeed a generalization, since we can express an equality constraint by two opposite inequality constraints.
To be able to solve this optimization problem, we will also consider its dual problem.

Definition 3.38. The (Lagrange) dual function associated to the constrained optimization problem is defined by

$$\forall x \in X, \forall \alpha \geq 0, \quad F(\alpha) = \inf_{x \in X} L(x,a) = \inf_{x \in X} (f(x) + \sum_{i=1}^{n} \alpha_i g_i(x))$$

Definition 3.39. The dual problem associated to the constrained optimization problem is

$$\max_{\alpha} F(\alpha),$$

$$\text{subject to } \alpha \geq 0.$$

The solution of the dual problem of a convex optimization problem is equal to the original solution, when the constraints are qualified.

Definition 3.40. Assume that $int(X) \neq \emptyset$. Then, the strong constraint qualification or Slater's condition is defined as

$$\exists \bar{x} \in int(X) : g(\bar{x}) < 0.$$

Definition 3.41. Assume that $int(X) \neq \emptyset$. Then, the weak constraint qualification or weak Slater's condition is defined as

$$\exists \bar{x} \in int(X) : \forall i \in [1,n], (g_i(\bar{x}) < 0) \vee (g_i(\bar{x}) = 0 \wedge g_i \text{ affine}).$$

Now we first show that a saddle point of the Lagrangian is a solution of the problem and then give a statement which guarantees the existence of a saddle point (x, α), if x is a solution of the convex problem and if Slater's condition holds.

Theorem 3.42. *Let P be a constrained optimization problem over $X = \mathbb{R}^N$. If (x^*, α^*) is a saddle point of the associated Lagrangian, that is,*

$$\forall x \in \mathbb{R}^n, \forall \alpha \geq 0, \;\; L(x^*, \alpha) \leq L(x^*, \alpha^*) \leq L(x, \alpha^*),$$

then (x^*, α^*) is a solution of the problem P.

Proof. The first inequality implies

$$\forall \alpha \geq 0, \alpha \cdot g(x^*) \leq \alpha^* \cdot g(x^*). \tag{3.11}$$

This yields $g(x^*) \leq 0$, since (3.11) holds for arbitrary big α. On the other hand, (3.11) also implies $\alpha^* \cdot g(x^*) = 0$, because it holds for arbitrary small α as well.
Therefore, the second inequality in the theorem implies

$$\forall x, \;\; f(x^*) \leq f(x) + \alpha^* \cdot g(x),$$

i.e., for all x satisfying the constraints $g(x) \leq 0$,

$$f(x^*) \leq f(x).$$

□

Theorem 3.43. *Assume that f and g_i, $i \in [1,n]$, are convex functions and that Slater's condition holds. Then, if x is a solution of the constrained optimization problem, then there exists $\alpha \geq 0$, such that (x,α) is a saddle point of the Lagrangian.*

If the functions are even convex differentiable, then fulfilling the weak Slater's condition is enough for the statement.

Theorem 3.44. *Assume that f and g_i, $i \in [1,n]$, are convex differentiable functions and that the weak Slater's condition holds. Then, if x is a solution of the constrained optimization problem, then there exists $\alpha \geq 0$, such that (x,α) is a saddle point of the Lagrangian.*

The following last theorem connects the established implications and gives us concrete conditions that are equivalent to $\bar{x}$ being a solution of the problem.

Theorem 3.45 (Karush-Kuhn-Tucker's theorem)**.** *Assume that $f, g_i : X \to \mathbb{R}$ $\forall i$ are convex and differentiable and that the constraints are qualified. Then $\bar{x}$ is a solution of the constrained program if and only if there exists an $\overline{\alpha} \geq 0$, such that*

$$\nabla_x L(\bar{x},\overline{\alpha}) = \nabla_x f(\bar{x}) + \overline{\alpha} \cdot \nabla_x g(\bar{x}) = 0$$

$$\nabla_\alpha L(\bar{x},\overline{\alpha}) = g(\bar{x}) \leq 0$$

$$\overline{\alpha} \cdot g(\bar{x}) = \sum_{i=1}^{n} \overline{\alpha}_i g(\bar{x}_i) = 0.$$

These conditions are known are called the KKT conditions.
Note, that since $g(\bar{x}) \leq 0$, the last condition can be substituted by

$$\overline{\alpha}_i g_i(\bar{x}) = 0 \ \forall i \in 1,..,n.$$

These equalities are called complementarity conditions.

Proof. " $\Rightarrow$ ": Let $\bar{x}$ be a solution of the constrained program. Since the constraints are qualified, there exists $\overline{\alpha}$ such that $(\bar{x},\overline{\alpha})$ is a saddle point of the Lagrangian. That a saddle point fulfills the three conditions can be seen in the proof of Theorem 3.42 (the first one follows directly from the definition of a saddle point).

" $\Leftarrow$ ": Assume that the three conditions are fulfilled. Then we have for any x with $g(x) \leq 0$, that

$$f(x) - f(\bar{x}) \geq \nabla_x f(\bar{x}) \cdot (x - \bar{x}) \geq -\sum_{i=1}^{n} \overline{\alpha}_i \nabla_x g_i(\bar{x}) \cdot (x - \bar{x}),$$

because f is convex and by the first condition. Since also the g_i are convex, this can be further approximated by

$$f(x) - f(\bar{x}) \geq -\sum_{i=1}^{n} \overline{\alpha}_i |g_i(x) - g_i(\bar{x})| \geq -\sum_{i=1}^{n} \overline{\alpha}_i g_i(x) \geq 0,$$

due to the third and second condition. That is, $f(\bar{x})$ is the minimum of f over the set of points that satisfy the constraints, i.e. $\bar{x}$ is a solution of the constrained program.

□

Let us go back to our optimization problem in support vector regression,

$$\min_{w,b,\xi,\xi'} \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i'),$$

subject to

$$w\cdot\Phi(x_i) + b - y_i \leq \varepsilon + \xi_i$$
$$y_i - (w\cdot\Phi(x_i) + b) \leq \varepsilon + \xi_i'$$
$$\xi_i \geq 0, \xi_i' \geq 0,\ \forall i \in [1,n].$$

We see that we indeed have a convex and differentiable problem, because the norm, and thus also the objective function, is convex and differentiable and the constraints are affine. Since the constraints are affine, they are also qualified. Therefore we can apply Karush-Kuhn-Tucker's theorem and the KKT conditions hold at the optimum.

To use that, we formulate the associated Lagrange function

$$L(w,b,\xi,\xi') = \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i') + \sum_{i=1}^{n}\alpha_i[w\cdot\Phi(x_i) + b - y_i - \varepsilon - \xi_i]$$

$$+\sum_{i=1}^{n}\alpha_i'[y_i - (w\cdot\Phi(x_i) + b) - \varepsilon - \xi_i'] - \sum_{i=1}^{n}\beta_i\xi_i - \sum_{i=1}^{n}\beta_i'\xi_i',$$

where $\alpha_i, \alpha_i', \beta_i, \beta_i'$ are the Lagrangian variables.
Now the first KKT condition is obtained by setting the gradient of the Lagrange function with respect to the variables w, b, ξ_i, ξ_i' to zero.

$$\nabla_w L = w + \sum_{i=1}^{n}\alpha_i\Phi(x_i) - \alpha_i'\Phi(x_i) = 0 \Rightarrow w = \sum_{i=1}^{n}(\alpha_i' - \alpha_i)\Phi(x_i) \tag{3.12}$$

$$\nabla_b L = \sum_{i=1}^{n}(\alpha_i - \alpha_i') = 0 \tag{3.13}$$

$$\nabla_{\xi_i} L = C - \alpha_i - \beta_i = 0 \Rightarrow \alpha_i + \beta_i = C \tag{3.14}$$
$$\nabla_{\xi_i'} L = C - \alpha_i' - \beta_i' = 0 \Rightarrow \alpha_i' + \beta_i' = C.$$

Plugging into the complementarity conditions yields

$$\forall i,\ \alpha_i[(w\cdot\Phi(x_i) + b) - y_i - \varepsilon - \xi_i] = 0 \Rightarrow \alpha_i = 0 \vee (w\cdot\Phi(x_i) + b) - y_i = \varepsilon + \xi_i \tag{3.15}$$

$$\forall i,\ \alpha_i'[y_i - (w\cdot\Phi(x_i) + b) - \varepsilon - \xi_i'] = 0 \Rightarrow \alpha_i' = 0 \vee y_i - (w\cdot\Phi(x_i) + b) = \varepsilon + \xi_i'$$

$$\forall i,\ \beta_i\xi_i = 0 \Rightarrow \beta_i = 0 \vee \xi_i = 0 \tag{3.16}$$

$$\forall i,\ \beta_i'\xi_i' = 0 \Rightarrow \beta_i' = 0 \vee \xi_i' = 0.$$

By (3.12), for the solution the vector w is a linear combination of $\Phi(x_1), ..., \Phi(x_m)$, i.e. of the training set vectors mapped to $\mathbb{R}^p$.

A vector $\Phi(x_i)$ appears in that expansion iff $\alpha_i \neq \alpha_i'$. Those vectors are called support vectors.
For that α_i or α_i' has to be positive. Thereby $\alpha_i > 0$ implies $\alpha_i' = 0$ and reversed, as we will see shortly, because y_i cannot lie both over and under the predicted function. By (3.15), if $\alpha_i > 0$, we have $(w \cdot \Phi(x_i) + b) - y_i = \varepsilon + \xi_i$ (similar for α_i'). So support vectors are the vectors, s.t. y_i has at least distance ε to the predicted function.
If the distance is bigger than ε, (3.16) implies $\beta_i = 0$ (or $\beta_i' = 0$), thus $\alpha_i = C$ (or $\alpha_i' = C$) by (3.14). Finally, if no y_i has a distance of exactly ε to the predicted function, by (3.13) we get that the number of y_i over the ε-tube matches the number under the tube in the solution.

To derive the dual form of the problem, we reformulate the Lagrangian using (3.12), (3.13) and (3.14).

$$
\begin{aligned}
L = \frac{1}{2}||\sum_{i=1}^{n}(\alpha_i' - \alpha_i)\Phi(x_i)||^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i') + \sum_{i=1}^{n}\alpha_i\left[\left(\sum_{i=1}^{n}(\alpha_i' - \alpha_i)\Phi(x_i)\right)\cdot\Phi(x_i) + b - y_i - \varepsilon - \xi_i\right] \\
+ \sum_{i=1}^{n}\alpha_i'\left[y_i - \left(\left(\sum_{i=1}^{n}(\alpha_i' - \alpha_i)\Phi(x_i)\right)\cdot\Phi(x_i) + b\right) - \varepsilon - \xi_i'\right] - \sum_{i=1}^{n}\beta_i\xi_i - \sum_{i=1}^{n}\beta_i'\xi_i'. \\
= \frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i' - \alpha_i)(\alpha_j' - \alpha_j)\Phi(x_i)\Phi(x_j) + \sum_{i,j=1}^{n}\alpha_j(\alpha_i' - \alpha_i)\Phi(x_i)\Phi(x_j) + \sum_{i=1}^{n}\alpha_i b - \sum_{i=1}^{n}\alpha_i y_i - \sum_{i=1}^{n}\alpha_i\varepsilon \\
+ \sum_{i=1}^{n}\alpha_i' y_i - \sum_{i,j=1}^{n}\alpha_j'(\alpha_i' - \alpha_i)\Phi(x_i)\Phi(x_j) - \sum_{i=1}^{n}\alpha_i' b - \sum_{i=1}^{n}\alpha_i'\varepsilon \\
= -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i' - \alpha_i)(\alpha_j' - \alpha_j)\Phi(x_i)\Phi(x_j) + b\underbrace{\sum_{i=1}^{n}(\alpha_i - \alpha_i')}_{=0} + \sum_{i=1}^{n}(\alpha_i' - \alpha_i)y_i - \sum_{i=1}^{n}(\alpha_i + \alpha_i')\varepsilon.
\end{aligned}
$$

This leads to the following equivalent dual problem in terms of the kernel matrix $\mathbb{K}$:

$$\max_{\alpha,\alpha'} -\varepsilon(\alpha' + \alpha)^T\mathbb{1} + (\alpha' - \alpha)^T y - \frac{1}{2}(\alpha' - \alpha)^T\mathbb{K}(\alpha' - \alpha) \tag{3.17}$$

subject to

$$(0 \leq \alpha \leq C) \wedge (0 \leq \alpha' \leq C) \wedge ((\alpha' - \alpha)^T\mathbb{1} = 0),$$

since $\alpha_i \geq 0, \beta_i \geq 0$ is equivalent to $0 \leq \alpha \leq C$ due to (3.14).

This problem is called a convex quadratic program (QP) and can be solved by convex optimization techniques, for example using an interior point method.

Since the constraints are qualified, the solution of the dual problem is equal to the solution of the original problem. So, after determining α, α', we can directly formulate the hypothesis returned by SVR, using (3.12), as

$$\forall x \in X, \quad h(x) = w \cdot \Phi(x) + b = \sum_{i=1}^{n}(\alpha_i' - \alpha_i)K(x_i, x) + b, \tag{3.18}$$

where b can be determined using a point x_j with $0 < \alpha_j < C$. Then y_j lies exactly ε under the predicted function, so

$$b = -\sum_{i=1}^{n} (\alpha_i' - \alpha_i) K(x_i, x_j) + y_j + \varepsilon,$$

and if $0 < \alpha_j' < C$, we get similarly

$$b = -\sum_{i=1}^{n} (\alpha_i' - \alpha_i) K(x_i, x_j) + y_j - \varepsilon.$$

Example 3.46. Let us consider Example 1.2 again and suppose we have drawn the claims expenditures y_j for the year x_j, $(1, y_1), (2, y_2), .., (10, y_{10})$, whereby we shift the years in the data frame to start from 1 again.
We take a linear kernel, i.e. the identity as feature mapping. Furthermore, we fix $\varepsilon = 10$ [million euros] and $C = 100$ to demonstrate the method. (For bigger data sets than in this example, the choice of kernels and parameters is usually decided based on error reduction, e.g. via cross-validation, see Remark 3.47.)
Then using a QP solver for (3.17) gives the coefficients $\hat{\alpha}_i = (\alpha_i' - \alpha_i)$ as in the following table.

$\hat{\alpha_1}$	$\hat{\alpha_2}$	$\hat{\alpha_3}$	$\hat{\alpha_4}$	$\hat{\alpha_5}$	$\hat{\alpha_6}$	$\hat{\alpha_7}$	$\hat{\alpha_8}$	$\hat{\alpha_9}$	$\hat{\alpha_{10}}$
100	-100	-100	0	100	100	-100	0	-9,3	9,3

Plugging these coefficients into (3.18) yields the hypothesis

$$h(x) = 188,3 + 9,3x,$$

which gives a predicted value for the claims expenditures in the next year of 290.6 million euros. So it is the first model that does not predict a new peak within the data frame for the next year.
In SVR with a linear kernel the slope will in general tend to be a bit smaller compared to ordinary linear regression due to the regularization term, as it is the case in this example.

Note, that in Figure 3.1 we indeed have $\hat{\alpha}_j = C$ for points under the tube, $\hat{\alpha}_j = -C$ for points over the tube, $0 < |\hat{\alpha}_j| < C$ for points on the margin and $\hat{\alpha}_j = 0$ for points within the tube.
So the support vectors in this example are all but x_4 and x_8.

Here we do not have a probabilistic model that includes properties like the distribution of the response. On the other hand, those properties were based on assumptions on the type of underlying distribution that were rather an educated guess than a mathematically profound one. So this SVR model can be seen as a purer mathematical model.

Another advantage of support vector regression is that various nonlinear extensions can be implemented very straightforwardly by using kernels, which leads to the same optimization problem, just with a different kernel matrix.

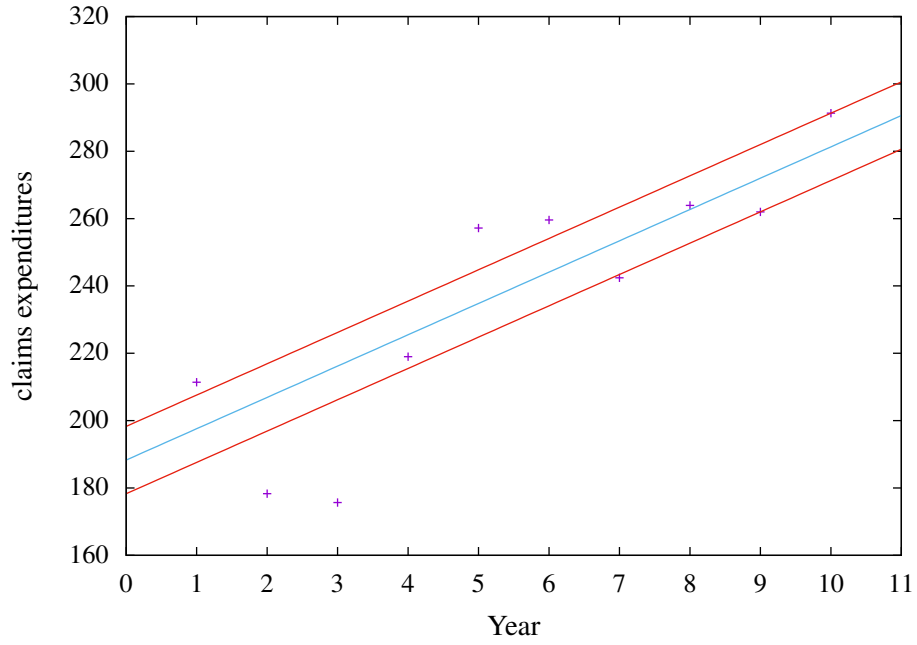

Figure 3.1: SVR with linear kernel, $\varepsilon = 10$, $C = 100$

A popular kernel in SVR is the radial basis function (RBF) kernel

$$K(x,x') = \exp(-\gamma||x-x'||^2), \quad \gamma > 0.$$

The choice of appropriate parameters is crucial for the model accuracy. However, this is not straightforward and therefore a disadvantage of SVR. One might even argue that this framework just shifts the need of model assumptions to the estimation of parameters.

Remark 3.47. A good approach to choosing the cost parameter C, or the free parameter γ when we use a RBF kernel, is the so-called k-fold cross validation. Thereby the data set is randomly split into k subsets. Then the algorithm is iteratively applied to all but one subset. On that subset the mean error based on the resulting model is measured. After iterating over all subsets and a grid of parameter values, the parameter with the lowest mean error is chosen.

Finally, we are also able to show generalization bounds for the SVR algorithm, using the theory we have already developed. Thereby we denote by $\hat{D}$ the empirical distribution defined by the training sample of size m.

Theorem 3.48. *Let $K : X \times X \to \mathbb{R}$ be a PDS kernel, let $\Phi : X \to \mathbb{H}$ be a feature mapping associated to K and let $H = \{x \to w \cdot \Phi(x) : ||w||_{\mathbb{H}} \leq \Lambda\}$. Furthermore, assume that there*

exists $r > 0$ such that $K(x,x) \leq r^2$ and $|y| \leq \Lambda r \ \forall y \in Y$. Fix $\varepsilon > 0$.
Then, for any $\delta > 0$, with probability at least $1-\delta$, each of the following inequalities holds for all $h \in H$:

$$\mathbb{E}_{(x,y)\in D}[|h(x)-y|_\varepsilon] \leq \mathbb{E}_{(x,y)\in \hat{D}}[|h(x)-y|_\varepsilon] + \frac{2r\Lambda}{\sqrt{m}}\left(1+\sqrt{\frac{\log(1/\delta)}{2}}\right),$$

$$\mathbb{E}_{(x,y)\in D}[|h(x)-y|_\varepsilon] \leq \mathbb{E}_{(x,y)\in \hat{D}}[|h(x)-y|_\varepsilon] + \frac{2r\Lambda}{\sqrt{m}}\left(\sqrt{\frac{Tr[\mathbb{K}]}{mr^2}}+3\sqrt{\frac{\log(2/\delta)}{2}}\right).$$

Proof. Define $H_\varepsilon = \{(x,y) \to |h(x)-y|_\varepsilon : h \in H\}$ and $H' = \{(x,y) \to h(x)-y : h \in H\}$.
Note that the function $\Phi_\varepsilon : x \to |x|_\varepsilon$ is 1-Lipschitz. Therefore we have $\hat{\mathscr{R}}_S(H_\varepsilon) \leq \hat{\mathscr{R}}_S(H')$ by Lemma 3.25. Furthermore, it holds that $\hat{\mathscr{R}}_S(H') = \hat{\mathscr{R}}_S(H)$, as shown in the proof of Theorem 3.26, so we also have

$$\hat{\mathscr{R}}_S(H_\varepsilon) \leq \hat{\mathscr{R}}_S(H).$$

Moreover, $\forall (x,y) \in X \times Y, \forall h \in H$,

$$|h(x)-y| \leq |h(x)| + |y| = |w \cdot \Phi(x)| + \Lambda r \leq \Lambda ||\Phi(x)|| + \Lambda r \leq 2\Lambda r,$$

because $||\Phi(x)|| = \sqrt{K(x,x)} \leq r$ as a feature mapping associated to K.
Next we have, by the definition of the empirical Rademacher complexity, that

$$\hat{\mathscr{R}}_S(H) = \frac{1}{m}\mathbb{E}_\sigma\left[\sup_{||w||\leq\Lambda}\langle w, \sum_{i=1}^m \sigma_i\Phi(x_i)\rangle\right] \leq \frac{\Lambda}{m}\mathbb{E}_\sigma\left[||\sum_{i=1}^m \sigma_i\Phi(x_i)||_{\mathbb{H}}\right]$$

due to the Cauchy-Schwarz inequality. Now Jensen's inequality yields the bound

$$\hat{\mathscr{R}}_S(H) \leq \frac{\Lambda}{m}\left[\mathbb{E}_\sigma\left[||\sum_{i=1}^m \sigma_i\Phi(x_i)||^2_{\mathbb{H}}\right]\right]^{1/2} = \frac{\Lambda}{m}\left[\mathbb{E}_\sigma\left[\sum_{i=1}^m |||\Phi(x_i)||^2_{\mathbb{H}}\right]\right]^{1/2}$$

since the Rademacher variables are independent, i.e. $\mathbb{E}_\sigma[\sigma_i\sigma_j] = \mathbb{E}_\sigma[\sigma_i]\mathbb{E}_\sigma[\sigma_j] = 0$ for $i \neq j$.
This yields

$$\hat{\mathscr{R}}_S(H) \leq \frac{\Lambda}{m}\left[\mathbb{E}_\sigma\left[\sum_{i=1}^m K(x_i,x_i)\right]\right]^{1/2} \leq \frac{\Lambda}{m}\left[\mathbb{E}_\sigma\left[mr^2\right]\right]^{1/2} = \sqrt{\frac{r^2\Lambda^2}{m}}.$$

Thus, we also have $\mathscr{R}_m(H) \leq \sqrt{\frac{r^2\Lambda^2}{m}}$.
Finally, Theorem 3.27 gives, for any $\delta > 0$, with probability at least $1-\delta$, that

$$\mathbb{E}_{(x,y)\in D}[|h(x)-y|_\varepsilon] \leq \mathbb{E}_{(x,y)\in \hat{D}}[|h(x)-y|_\varepsilon] + \frac{2r\Lambda}{\sqrt{m}}\left(1+\sqrt{\frac{\log(1/\delta)}{2}}\right).$$

The second statement follows similarly with $\sum_{i=1}^m K(x_i,x_i) =: Tr[\mathbb{K}]$.

□

Remark 3.49. In addition to linear regression and its kernel-based non-linear extension, as presented in this section, a popular approach to regression algorithms in the Machine Learning framework is the so-called decision tree learning.
Thereby a simple approach is to minimize the squared errors, that we can get by splitting the data into two parts through a value of a single feature and taking the average label in each part as estimator. After iterating that approach in each part up to a certain depth, one can visualize this model as a decision tree.

Remark 3.50. The machine learning framework can be also be applied to classification problems, i.e. to problems where Y is a finite set. Thereby ML methods have also already been successfully employed in many applications, for example in spam detection in emails.

3.7 Principal Component Analysis

We have already mentioned that in Big Data problems the data is usually unstructured and we do not have much information about it. Therefore it is popular to, before applying regression algorithms, first try to structure the data by using Data Mining techniques, which are called unsupervised Machine Learning methods in the context of ML.

The method we will consider is the Principal Component Analysis (PCA). This approach can give us essential information about the data and also help us to reduce the dimensionality of the problem. Thereby we start with a data matrix X with n samples in the rows and p respective properties in the columns.

This method performs an eigendecomposition of the sample covariance matrix, which always exists since the sample covariance matrix is real symmetric and therefore diagonalizable. Note, that after centering the properties in our data matrix X, the sample covariance matrix is up to a constant factor $1/n$ equal to $X^T X$.

Remark 3.51. In the field of statistics the constant factor is usually set to $1/(n-1)$ because it leads to an unbiased estimator for the variance. This is often called Bessel's correction in the literature.

Thereby centering does not change the sample covariance matrix since the variance measures the deviation from the mean.
Therefore, that eigendecomposition can be deduced from the Singular Value Decomposition (SVD) of the matrix X, which gives the eigenvalues and corresponding eigenvectors of $X^T X$.

To focus on the correlations only and to not have to worry about different units in the property set, one can also standardize the variance of the properties in the data matrix first. This gives a sample correlation matrix, on which the PCA can be performed similarly.

Recall that the SVD of a matrix X is given by

$$X = U\Sigma V^T,$$

where U and V are rotation matrices, i.e. have orthonormal columns, and Σ is diagonal with the singular values $\lambda_1 \geq \lambda_2 \geq ... \geq 0$ in decreasing order. Since

$$\frac{1}{n}X^T X = \frac{1}{n}(U\Sigma V^T)^T(U\Sigma V^T) = \frac{1}{n}V\Sigma U^T U\Sigma V^T = V\frac{\Sigma^2}{n}V^T$$

is an eigendecomposition of the sample covariance matrix, the columns of V are eigenvectors and the singular values are the square roots of the (non-zero) eigenvalues of $X^T X$. We will call the orthonormal eigenvectors of the sample covariance matrix, i.e. the columns of V, that are corresponding to non-zero eigenvalues, principal components in the following. The coefficients of the projections of the data onto them, which are given by

$$XV = U\Sigma V^T V = U\Sigma,$$

we will call principal component scores (PC scores).

It can be shown that the first principal component maximizes the variance of the PC scores amongst all linear combinations of properties. Equivalently, it minimizes the distances between the original data points and the projected ones. The second principal component does the same amongst all orthogonal linear combinations to the first PC and so on. Furthermore, the data points are uncorrelated in the new orthogonal system built by the principal components.

Moreover, since the singular values $\lambda_1 \geq \lambda_2 \geq ... \geq 0$ are in decreasing order, the matrix X can be approximated by considering the SVD only up to a certain singular value λ_k. It can be shown that this gives the best rank k approximation to the $n \times p$ matrix X in the Frobenius norm, i.e. it minimizes

$$\sum_{i=1}^{n}\sum_{i=1}^{p}(\hat{X}_{ij} - X_{ij})^2 : rank(\hat{X}) = k.$$

This result is known as the Eckart–Young–Mirsky theorem.

SVD algorithms

Now we will display how SVD routines can be implemented with consideration to optimized performance.
Recall that the singular values of a real matrix $X = U\Sigma V^T$ are the square roots of the non-zero eigenvalues of $X^T X$, and the corresponding columns of V are orthonormal eigenvectors. Therefore SVD algorithms seek eigenvalues and eigenvectors of $X^T X$. This is done without having to compute the full matrix $X^T X$.

More precisely, Jacobi rotations are performed to iteratively diagonalize the matrix. Furthermore, a QR-decomposition of the matrix X with a suited column pivoting method is performed as preconditioning. Thereby X denotes the $n \times p$ model matrix with p features in the columns and n samples in the rows again, $n \geq p$, so that we can apply a QR-procedure to X.

Remark 3.52. We will not consider the case $p < n$ here, since it is usually not meaningful to analyze a data matrix with very few samples. Theoretically we could apply the algorithm to X^T then though.

A Jacobi rotation matrix $\hat{V}$ generates zeros in two opposite off-diagonal entries H_{st} and H_{ts} of a real symmetric matrix H, i.e.

$$\begin{pmatrix} \hat{V}_{ss} & \hat{V}_{st} \\ -\hat{V}_{st} & \hat{V}_{ss} \end{pmatrix}^T \begin{pmatrix} H_{ss} & H_{st} \\ H_{st} & H_{tt} \end{pmatrix} \begin{pmatrix} \hat{V}_{ss} & \hat{V}_{st} \\ -\hat{V}_{st} & \hat{V}_{ss} \end{pmatrix} = \begin{pmatrix} H'_{ss} & 0 \\ 0 & H'_{tt} \end{pmatrix}$$

with $\hat{V}_{ss} = \hat{V}_{tt} = \cos(\phi)$ and $\hat{V}_{st} = -\hat{V}_{ts} = \sin(\phi)$.
Note, that therefore we have to have $0 = H'_{st} = H_{st}(\hat{V}_{ss}^2 - \hat{V}_{st}^2) + (H_{ss} - H_{tt})\hat{V}_{ss}\hat{V}_{st}$.

For the non-trivial case $H_{st} \neq 0$ we define $\tau = (H_{ss} - H_{tt})/(2H_{st})$, which yields with the condition above that (for $\hat{V}_{ss} \neq 0$) $q = \hat{V}_{st}/\hat{V}_{ss} = \tan(\phi)$ has to solve $q^2 + 2\tau q - 1 = 0$.
This gives the solutions $q = -\tau \pm \sqrt{1+\tau^2}$ and we get $\hat{V}_{ss} = 1/\sqrt{1+\tau^2}$, $\hat{V}_{st} = t\,\hat{V}_{ss}$.

Remark 3.53. According to [20], it is important to take the smaller of the two solutions for the performance of the algorithm because it guarantees $|\phi| \leq \pi/4$ and less changes in the remaining entries of the s-th and $t-$th row and column with respect to the Frobenius norm.

Note, that we cannot guarantee that in a Jacobi rotation already existing zeros in the s-th or t-th row or column are getting destroyed, but we do have that the summarized squared absolute values of all off-diagonal entries decrease with every non-trivial Jacobi rotation, since the Frobenius norm of H does not change by applying an orthogonal transformation.

A reasonable pivot strategy is to always create zeros in the off-diagonal entries with the largest absolute values in every iteration. However, to do that we always have to track all off-diagonal values in the algorithm, but in the SVD algorithm we will not even explicitly build the whole symmetric matrix X^TX (or later R^TR).
Instead, we can represent a sequence of rotations on $H^{(0)} = X^TX$ by applying it only to X itself, i.e.

$$X^{(k+1)} = X^{(k)}V^{(k)},$$

since we then have

$$H^{(k+1)} = V^{(k)^T}H^{(k)}V^{(k)} = V^{(k)^T}X^{(k)^T}X^{(k)}V^{(k)} = X^{(k+1)^T}X^{(k+1)}.$$

To calculate the four non-trivial entries $V_{ss}^{(k)}, V_{st}^{(k)}, V_{ts}^{(k)}$ and $V_{tt}^{(k)}$ of the rotation $V^{(k)}$ we just need to build the 2×2 Gram matrix of the s-th and t-th column of $X^{(k)}$.
Since we don't compute the full matrix X^TX, we will use a fixed row-cycling pivot strategy.
It can be shown the sequence of iterates $(H^{(k)})$, which we represent by $(X^{(k)})$, converges to a diagonal matrix Λ under suitable pivot strategies, including the ones mentioned above.
The accumulated product of the Jacobi rotations converges to a orthogonal matrix V of eigenvectors of H. Therefore we have $X^TXV = V\Lambda$. Furthermore, it can be shown that the sequence $(X^{(k)})$ converges to $U\Sigma$, s.t. the SVD of X is $U\Sigma V^T$.

The QR-preconditioning of the $(n \times p)$-matrix X turns out to be useful in several aspects. Firstly, it serves for dimensionality reduction because the resulting reduced R matrix is of dimension $p \times p$ and we have $X^T X = (QR)^T (QR) = R^T R$. Therefore we can also apply the rotations on R, i.e.

$$R^{(k+1)} = R^{(k)} \overline{V}^{(k)}$$

with $\overline{V}^{(k)}$ calculated by a 2×2 Gram matrix of $R^{(k)}$.

Remark 3.54. A comparison is given in [7]. A Householder QR-algorithm requires $2np^2 - 2p^3/3$ flops to calculate R. One full sweep, i.e. iterating once over all off-diagonal entries, of a Jacobi SVD algorithm with fast rotations requires $3np^2$ flops if the product of the rotations is not computed. Therefore, if only the singular values are needed, the QR preconditioning already pays off after one full sweep if $n > 7p/3$ and in two sweeps if $n > 4p/3$.

To take further advantages of the preconditioning we will use a QR-procedure with column pivoting to get

$$XP = Q \begin{pmatrix} R \\ 0 \end{pmatrix} \text{ such that } |R_{ii}| \geq \sqrt{\sum_{k=i}^{j} R_{kj}^2}, 1 \leq i \leq j \leq p.$$

This allows us to continue the algorithm with a smaller $(r \times p)$ matrix if $|R_{rr}|$ is sufficiently small for some $r < p$, which is called a rank-revealing QR-decomposition.
However, one has to be careful of the errors we produce by doing that speed-up step if high accuracy is needed.

Finally, it turns out that performing a second QR-decomposition $R^T = Q_1 R_1$ and then applying the algorithm to R_1^T instead of R_1 leads to further improvements, see [7] for details. Additionally, slight changes in the pivot strategy of Jacobi rotations can lead to faster convergence, as demonstrated in [8].

This yields the following pseudo-code for the computation of the singular values. In that algorithm, an initial row permutation on X and a pivoting in the second QR-factorization are added, which can at times be beneficial, see [7].
Thereby we denote matrices that do not have to explicitly be computed with $\langle \cdot \rangle$. Also, we use the notation M_∞ for the output of the Jacobi procedure after the rotations to the input matrix M and the notation $M_\infty(:,i)$ for its i-th column. Moreover, $||\cdot||$ denotes the vector Euclidean norm.

Algorithm 2 SVD: Computation of Σ

$(P_0 X)P = \langle Q \rangle \begin{pmatrix} R \\ 0 \end{pmatrix}; r = \text{rank}(R);$
$R(1:r, 1:p)^T P_1 = \langle Q_1 \rangle R_1;\ M = R_1^T;$
$M_\infty = M \langle V_M \rangle;$
$\sigma_i = ||M_\infty(:,i)||, i = 1,..,r;\ \text{diag}(\Sigma) = (\sigma_1,..,\sigma_r,0,..0);$

If we also need the right singular vectors, i.e. the principal components, of X, we can elegantly avoid calculating the accumulated product of the rotations by inputing the transposed matrix into the Jacobi procedure as well. In the corresponding algorithm, we pass on the second QR-decomposition for better running time here, but if high accuracy is required implementing it is still recommendable in many cases.

Algorithm 3 SVD: Computation of Σ and V

$(P_0X)P = \langle Q \rangle \begin{pmatrix} R \\ 0 \end{pmatrix}; r = \text{rank}(R);$
$M = R(1:r, 1:p)^T;$
$M_\infty = M\langle V_M \rangle;$
$\sigma_i = ||M_\infty(:,i)||, i = 1,..,r;\ \text{diag}(\Sigma) = (\sigma_1,..,\sigma_r,0,..0);$
$U_M(:,i) = \frac{1}{\sigma_i} M_\infty(:,i), i = 1,..,r;\ \ V = PU_M;$

Remark 3.55. Note, that we indeed get the right singular vectors of X by the matrix U_M in the computed SVD, since we inputed the transposed matrix $M = R^T$, which yields a shift between the meanings of the matrices U_M and V_M in the SVD.

For the error analysis of the presented algorithms we drop the permutation matrices, that is we assume that the matrix X is replaced with the permuted matrix $(P_0X)P$.
Our goal is to derive a backward stability result for the presented algorithms, i.e. for the calculation of the singular values or the singular values and the right singular vectors.
To do that, we first give a result which is based on the individual backward errors of the QR-decomposition and the Jacobi rotations and which states that there exists a nearby matrix to X, which equals an "almost SVD" given by the computed matrices Σ and U_M.

Proposition 3.56. *Let X be a real $n \times p$ matrix and assume that the SVD of X is computed by QR-preconditioning $X = Q \begin{pmatrix} R \\ 0 \end{pmatrix}$ and then applying the Jacobi SVD algorithm to $M = R^T$ in the IEEE standard, which guarantees a rounding error of $\varepsilon < 10^{-7}$ for single variables and $\varepsilon < 10^{-15}$ for double variables.*
Let $M \approx \tilde{U}_M \tilde{\Sigma} \langle \tilde{V}_M^T \rangle$ be the computed SVD. Then there exist a perturbation ΔX and orthogonal matrices $\hat{Q}, \hat{V}_M$ such that

$$X + \Delta X = \hat{Q} \begin{pmatrix} \hat{V}_M & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} \tilde{\Sigma} \\ 0 \end{pmatrix} \tilde{U}_M^T, \text{ where} \tag{3.19}$$

$$||\Delta X(:,i)|| \leq \tilde{\eta} ||X(:,i)||, i = 1,..,p,\ \tilde{\eta} = \varepsilon_{qr} + \varepsilon_J + \varepsilon_{qr}\varepsilon_J \tag{3.20}$$

with parameters $\varepsilon_{qr} \leq O(np)\varepsilon$ for Householder QR factorization and $\varepsilon_J \leq (1+6\varepsilon)^{s(2p-3)} - 1$ for the row or column cycling Jacobi algorithm that stops during the $s-$th sweep.

Proof. The following proof is given in [7, Proposition 6.1].
Let $\tilde{Q}$ and $\tilde{R}$ be the computed matrices in the numerical QR-decomposition with Householder transformations. Then there exist an orthogonal matrix $\hat{Q}$ and a backward perturbation δX

such that $X+\delta X = \hat{Q}\begin{pmatrix}\tilde{R}\\0\end{pmatrix}$, where we have $||\delta X(:,i)|| \le \varepsilon_{qr}||X(:,i)||$ with $\varepsilon_{qr} \le O(np)\varepsilon$ for all i, see [7, Proposition 2.2].
Let the cycling Jacobi SVD be applied to $M = \tilde{R}^T$. According to [7, Proposition 2.3], this further gives $M+F = \tilde{U}_M\tilde{\Sigma}\hat{V}_M^T$, where $||F(i,:)|| \le \varepsilon_J||M(i,:)||$ with $\varepsilon_J \le (1+6\varepsilon)^{s(2p-3)}-1$, so

$$X+\underbrace{\delta X+\hat{Q}\begin{pmatrix}F^T\\0\end{pmatrix}}_{=\Delta X} = \hat{Q}\begin{pmatrix}\hat{V}_M & 0\\0 & I\end{pmatrix}\begin{pmatrix}\tilde{\Sigma}\\0\end{pmatrix}\tilde{U}_M^T.$$

Moreover, in that equation the backward perturbation ΔX has column-wise bound

$$||\Delta X(:,i)|| \le \varepsilon_{qr}||X(:,i)|| + \varepsilon_J||\tilde{R}(:,i)|| \le \varepsilon_{qr}||X(:,i)|| + \varepsilon_J(1+\varepsilon_{qr})||X(:,i)||,$$

which yields the statement.

□

However, the right hand side in relation (3.19) is not a SVD yet, since the matrix $\tilde{U}_M$ is in general not exactly orthogonal. To actually get a SVD of a matrix that is close to X, we need to replace $\tilde{U}_M$ with a nearby orthogonal matrix $\hat{U}$.

Proposition 3.57. *In addition to the assumptions in Proposition 3.56, let* $\varepsilon_U = ||\tilde{U}_M^T\tilde{U}_M - I||_F < 1/(2\sqrt{2})$. *Then there exists a backward perturbation* $\mathscr{E}$ *and an orthogonal matrix* $\hat{U}$, *such that* $||\tilde{U}_M - \hat{U}||_F \le \sqrt{2}\varepsilon_u$ *and that the SVD of* $X+\mathscr{E}$ *is*

$$X+\mathscr{E} = \hat{Q}\begin{pmatrix}\hat{V}_M & 0\\0 & I\end{pmatrix}\begin{pmatrix}\tilde{\Sigma}\\0\end{pmatrix}\hat{U}^T, \text{ where for all } i$$

$$||\mathscr{E}(:,i)|| \le \hat{\eta}||X(:,i)||,\ \hat{\eta} = \tilde{\eta} + \sqrt{2p}\,\varepsilon_U(1+\tilde{\eta}) + O(\varepsilon_U^2).$$

Proof. Let P_d be the permutation matrix, so that the columns of XP_d are ordered in decreasing Euclidean norm. Let $P_d^T\tilde{U}_M = (I+G_0^T)\hat{U}_M$ be the RQ decomposition of $P_d^T\tilde{U}_M$, where G_0 is a lower triangular matrix and $\hat{U}_M$ orthogonal. Then we have

$$(I+G_0)(I+G_0)^T = \hat{U}_M\tilde{U}_M^T\tilde{U}_M\hat{U}_M^T = I+\underbrace{\hat{U}_M(\tilde{U}_M^T\tilde{U}_M - I)\hat{U}_M^T}_{:=A}.$$

Since $||A||_F = ||\tilde{U}_M^T\tilde{U}_M - I||_F = \varepsilon_U < 1/(2\sqrt{2})$ holds, it can be shown that $||G_0||_F \le \sqrt{2}\varepsilon_U$, see [9]. In particular, $||G_0||_F < 1$, so $I+G_0$ is invertible, whereby we denote $(I+G_0)^{-1} =: I+G$. Clearly, G is lower triangular, as $(I+G)$ is the inverse of a lower triangular matrix. Moreover, since $G = -G_0 + G_0^2(I+G_0)^{-1}$ [see the proof of Proposition 3.58], we have that

$$||G||_1 \le ||G_0||_1 + ||G_0||_1^2/(1-||G_0||_1).$$

Furthermore, with (3.19) and the orthogonal matrix $P_d\hat{U}_M = P_d(I+G)P_d^T\tilde{U}_M$ we get the SVD

$$(X+\Delta X)(I+P_dGP_d^T) = \hat{Q}\begin{pmatrix}\hat{V}_M & 0\\0 & I\end{pmatrix}\begin{pmatrix}\tilde{\Sigma}\\0\end{pmatrix}(P_d\hat{U}_M)^T, \tag{3.21}$$

Now, one can further verify that

$$XP_dG(:,i) = \sum_{k=i}^{p} X(:,\pi(k)),$$

where π denotes the permutation given by P_d. Since the columns are ordered to have decreasing Euclidean norms by that permutation, this yields

$$||XP_dG(:,i)|| \leq \sum_{k=i}^{n} |g_{ki}|\ ||X(:,\pi(k))|| \leq ||G||_1\ ||X(:,\pi,i)||.$$

Furthermore, the permutation matrix $P_d^T = P_d^{-1}$ rearranges the columns back to the original order, so we have

$$||(XP_dGP_d^T)(:,i)|| = ||(XP_dG)(:,\pi^{-1}(i))|| \leq ||G||_1\ ||X(:,i)||.$$

Therefore we also get with (3.20) that

$$||\Delta XP_dGP_d^T(:,i)|| \leq \tilde{\eta}||G||_1||X(:,i)||.$$

To conclude, note that the orthogonal matrix $\hat{U} := P_d\hat{U}_M$ satisfies $||\hat{U} - \tilde{U}_M||_F = ||G_0||_F \leq \sqrt{2}\varepsilon_U$, since $\tilde{U}_M = P_d\hat{U}_M + P_dG_0^T\hat{U}_M$ and since the Frobenius norm is invariant under orthogonal transformations, and that (3.21) defines a backward perturbation matrix $\mathscr{E}$ as in the statement because $||G_0||_1 \leq \sqrt{p}||G_0||_F \leq \sqrt{p}\sqrt{2}\varepsilon_U$, so

$$||G||_1 \leq ||G_0||_1 + ||G_0||_1^2/(1-||G_0||_1) \leq \sqrt{2p}\,\varepsilon_U + O(\varepsilon_U^2) \ \text{ (for } \varepsilon_U \to 0).$$

□

This proposition gives a desired backward stability result for variants of the presented algorithms, in which we don't exploit the rank revealing QR-decomposition to continue with a smaller matrix and only use one QR-preconditioning. For stability considerations regarding variants with two preconditionings, see [7, section 6.2].

In [7], also the forward errors are considered. For the singular values, the following estimate holds.

Proposition 3.58. *Let $X = U\Sigma V^T$ be the SVD of the $n \times p$ matrix X with full column rank, where $\Sigma = \begin{pmatrix} \overline{\Sigma} \\ 0 \end{pmatrix}$ with $\overline{\Sigma} = diag(\sigma_1,..,\sigma_p)$. Furthermore, let $\tilde{\sigma}_1 \geq .. \geq \tilde{\sigma}_p$ be the singular values of the perturbed matrix $X + \delta X = (I+\Gamma)X$, where $\Gamma = \delta XX^\dagger$ with the pseudo-inverse $X^\dagger = U\begin{pmatrix} \overline{\Sigma}^{-1} \\ 0 \end{pmatrix}V^T$. Finally, let $Sym(\Gamma) := \frac{1}{2}(\Gamma + \Gamma^T)$ and assume that $||\Gamma||_2 < 1$, where $||\cdot||_2 = \sigma_1(\cdot)$ denotes the spectral norm.*
Then it holds that

$$\max_{j=1,..,p} \frac{|\tilde{\sigma}_j - \sigma_j|}{\sqrt{\tilde{\sigma}_j\sigma_j}} \leq ||Sym(\Gamma)||_2 + \frac{1}{2}\frac{||\Gamma||_2^2}{1-||\Gamma||_2} \leq ||\Gamma||_2 + O(||\Gamma||_2^2).$$

Proof. Since $||\Gamma||_2 < 1$, $I+\Gamma$ is invertible. Therefore, according to [7], perturbation theory can be used to get

$$\max_{j=1,..,p} \frac{|\tilde{\sigma}_j - \sigma_j|}{\sqrt{\tilde{\sigma}_j \sigma_j}} \leq \frac{1}{2}||(I+\Gamma)^{-1} - (I+\Gamma)^T||_2 = \frac{1}{2}||-2Sym(\Gamma) + \Gamma^2(I+\Gamma)^{-1}||_2$$

by the self-referential form $(I+\Gamma)^{-1} = (I-\Gamma) + \Gamma^2(I+\Gamma)^{-1}$, which holds due to the convergent power series $(I+\Gamma)^{-1} = I - \Gamma + \Gamma_2 - \Gamma^3 + ...$
This yields the statement by applying $||(I+\Gamma)^{-1}||_2 \leq (1-||\Gamma||_2)^{-1}$.

□

For the forward errors of the right singular vectors, i.e. the principal components, recall Proposition 3.57, where we had the SVD

$$X + \mathscr{E} = \hat{Q} \begin{pmatrix} \hat{V}_M & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} \tilde{\Sigma} \\ 0 \end{pmatrix} \hat{U}^T \equiv \hat{U}_{X+\mathscr{E}} \begin{pmatrix} \tilde{\Sigma} \\ 0 \end{pmatrix} \hat{V}_{X+\mathscr{E}}^T. \tag{3.22}$$

Proposition 3.59. *Let $X = U\Sigma V^T$ be the SVD of the $n \times p$ matrix X and let* (3.22) *be the SVD of the perturbed matrix $X + \mathscr{E}$ with $||\mathscr{E}(:,i)|| \leq \hat{\eta}||X(:,i)||$ as in Proposition 3.57.*
Furthermore, let $\Phi = \mathscr{E}X^\dagger, \zeta = ||\Phi + \Phi^T + \Phi\Phi^T||_2$, and $\zeta \leq 2||Sym(\Phi)||_2 + ||\Phi||_2^2$. If we have $\zeta < \min\left\{2, \min_{j\neq i} \frac{|\sigma_j - \sigma_i|}{\sigma_i}\right\} := \rho_i$, then it holds that

$$\sin(\angle(V(:,i), \hat{V}_{X+\mathscr{E}}(:,i))) \leq \sqrt{2}\left\{\frac{\xi}{\rho_i - \zeta} + ||\Phi||_2\right\},$$

where $\xi \leq 2||Sym(\Phi)||_2 + O(||\Phi||_2^2)$ and $||\Phi||_2 \leq \sqrt{p}\,\hat{\eta}\,||X^\dagger\, diag(\frac{1}{||X^\dagger(:,i)||})_{i=1,..,p}||_2$.

Proof. See [7, Proposition 6.4], where also a suited reference for the proof is given. □

Using this proposition in combination with the backward result we can bound the errors of the computed right singular vectors, since for the computed numerically orthogonal matrix $\tilde{V}$ the additional angles $\angle(\tilde{V}(:,i), \hat{V}_{X+\mathscr{E}}(:,i)) = \angle(\tilde{U}_M(:,i), \hat{U}(:,i))$ with $||\tilde{U}_M - \hat{U}||_F \leq \sqrt{2}\varepsilon_U$, according to [7], are small and have sharper bounds.

Remark 3.60. Other popular data mining or unsupervised machine learning methods are often based on cluster analysis. These methods try so group the data set into clusters that contain similar elements.
For example in k-means clustering, an optimization problem is solved to get an optimal classifier from the data set to a fixed number of k clusters, so that the sum of the squared Euclidean distances of all elements in the data set to the mean of their respective cluster is minimized.

Chapter 4

Further tools: Stochastical simulation methods

The goal of this chapter is to introduce stochastical simulation methods to estimate expectations and establish the theoretical foundations for them, which will help us to tweak our regression models.
In the last chapter we have already shown generalization results that we can get out of a finite sample. Let us recall Example 3.11, where we also considered the simplified case that the label set consists of only two points $\{0,1\}$ (called classification problem).

Example 4.1. In the setting of Example 3.11, consider a biased coin toss ($1 :=$ heads$, 0 :=$ tails), where the probability of heads is p, consider the hypothesis $h \equiv 0$ that always predicts tails and the loss function $L(y,y') = |y-y'|$.
The generalization error in this case is $R(h) = p$. Let $\hat{p}$ be the empirical probability of heads based on some sample of size n, thus $\hat{R}(h) = \hat{p}$.
Hoeffding's inequality gives

$$\delta =: Pr[|p-\hat{p}| \geq \varepsilon] \leq 2e^{-2\varepsilon^2 n}.$$

Similarly to the end of the proof of Theorem 3.12, this yields, that for the hypothesis $h \equiv 0$ and for any $\delta > 0$, it holds with probability at least $1-\delta$ that

$$|R(h) - \hat{R}(h)| = |p-\hat{p}| \leq \sqrt{\frac{log(2/\delta)}{2n}}. \tag{4.1}$$

For example, let $\delta = 0.02$ and $n = 500$. Then, with probability at least 98%,

$$|p-\hat{p}| \leq \sqrt{\frac{log(100)}{1000}} \approx 0,068.$$

So we can already make a good assumption on the expectation of the biased coin toss based on 500 realizations. On the other hand, we can use (4.1) also to estimate the probability of the average over 500 realizations being in a specific interval around a given p.

In the following we will establish a more general foundation for simulation methods.
A more popular approach to estimate the probability of the empirical outcome being in an interval in Example 4.1 is by using the normal approximation, which itself is justified by the Berry-Esseen Theorem. This theorem will be shown in the following. We will see, that the convergence rate of $\frac{1}{\sqrt{n}}$ in (4.1) matches the one in the Berry-Esseen Theorem.
This theorem will also give us insights about the distribution of the errors in a finite sample.

These results allow us to justify estimating expectations by simulating a finite sample from a distribution. The methods to do that are usually called Monte-Carlo methods. To gain information about distributions, for example about the empirical distribution function of some data, it is often numerically convenient to be able to simulate realizations of that distribution and estimate expectations from that. This can allow us to construct a more accurate model.

4.1 Foundation of Monte-Carlo methods

In Monte-Carlo methods we produce realizations $y_1,...,y_n$ of a random variable Y by expressing it as a function $Y = f(X)$, where we are able to generate realizations $x_1,...,x_n$ of X. For the analysis of the realizations we consider corresponding sequences $Y_1,..,Y_n$ and $X_1,..,X_n$ of i.i.d. random variables with $Y_1 \sim Y$, $X_1 \sim X$.

Remark 4.2. In this chapter we use the notation X and Y for general random variables and the relation $Y = f(X)$ for simulation methods here should not be confused with the spaces X and Y in the regression problem itself.

The principle of Monte-Carlo methods lies in the strong law of large numbers.
If f satisfies $\mathbb{E}(|f(X)|) < \infty$, the strong law of large numbers gives

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} f(X_k) = \mathbb{E}(f(X)) \quad a.s.$$

So asymptotically the average $\frac{1}{n}(f(x_1)+..+f(x_n)) = \frac{1}{n}(y_1+..+y_n)$ over a big sample almost surely matches the expected value $\mathbb{E}(Y)$.

Convergence rate of Monte-Carlo methods

While the law of large numbers is an important result, it does not include information about the distribution of the errors

$$\varepsilon_n = \frac{1}{n}(Y_1 + ... + Y_n) - \mathbb{E}(Y)$$

that we are making by computing an expectation based on a finite sample.
The central limit theorem gives first insights to that. It states that if $(Y_i,\ i \geq 1)$ is a sequence of i.i.d. random variables such that $\mathbb{E}(Y_1^2) < \infty$ and σ^2 is the variance of Y_1, then

$$\left(\frac{\sqrt{n}}{\sigma}\varepsilon_n\right)$$

converges in distribution to a standard Gaussian random variable $N(0,1)$, i.e.

$$\forall t \in \mathbb{R}: \quad \lim_{n\to\infty} Pr[\frac{\sqrt{n}}{\sigma}\varepsilon_n \le t] = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

Proposition 4.3. *From the central limit theorem it follows that for all $c_1 < c_2$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\sigma}{\sqrt{n}}c_1 \le \varepsilon_n \le \frac{\sigma}{\sqrt{n}}c_2\right) = \lim_{n\to\infty} \mathbb{P}\left(c_1 \le \frac{\sqrt{n}}{\sigma}\varepsilon_n \le c_2\right) = \frac{1}{\sqrt{2\pi}} \int_{c_1}^{c_2} e^{-\frac{1}{2}x^2} dx.$$

For centralized and normalized random variables, i.e. random variables with zero mean and unit variance we have

$$\frac{\sqrt{n}}{\sigma}\varepsilon_n = \sum_{j=1}^{n} y_j/\sqrt{n}.$$

Using the central limit theorem we get the asymptotic distribution of the errors. Yet we still do not have a quantitative result for finite samples. For distributions that have finite third moments, we can establish general convergence rates by the Berry-Esseen theorem.

Theorem 4.4 (Berry-Esseen theorem)**.** *Let $\{Y_n\}$ be a sequence of independent and i.i.d. random variables with zero mean and unit variance. Let $Z = \sum_{j=1}^{n} Y_j/\sqrt{n}$, and $F_Z(z) = P[Z \le z]$. Suppose further that $\rho = \mathbb{E}[|Y_1|^3] < \infty$. Then*

$$\sup_z |F_Z(z) - \Phi(z)| \le C\rho/\sqrt{n},$$

where C is a constant independent of the distribution of the variables Y_n.

Proof. Recall that the characteristic function for a random variable X taking values in $\mathbb{R}$ is defined as

$$\zeta_X(\beta) = \mathbb{E}[\exp(i\beta X)].$$

We will need the following result for the proof, which was presented in [13].

Lemma 4.5. *Suppose a random variable with cumulative distribution function H, with expectation 0 and characteristic function ζ. Further suppose that G is differentiable, its Fourier transform ξ has a derivative ξ' which takes the value 0 at 0, and G has the limits 0 and 1 as its argument approaches $\pm\infty$ respectively. Then*

$$\forall x\ \forall \theta > 0 : |H(x) - G(x)| \le \frac{1}{\pi} \int_{-\theta}^{\theta} |\xi(\beta) - \zeta(\beta)| / |\beta| d\beta + 24 \max(|G'|)/(\pi\theta).$$

Note, that also $\zeta'(0) = 0$, since the distribution has expectation 0, so G somewhat fits to the cumulative distribution function H.
Denoting the characteristic function of Y_1 as $\hat{\xi}$ and using Lemma 4.5, we get

$$\pi|F_Z(z) - \Phi(z)| \le \int_{-\theta}^{\theta} |\hat{\xi}^n(\beta/\sqrt{n}) - \exp(-\beta^2/2)| / |\beta| d\beta + 24 \max(|\phi|)/\theta, \tag{4.2}$$

since it holds that

$$\xi(\beta) = \mathbb{E}(\exp(i\beta Z)) = \mathbb{E}(\exp(i(\beta/\sqrt{n})Y_1)^n) = \hat{\xi}^n(\beta/\sqrt{n})$$

and since the standard normal distribution is a fixpoint (neglecting the constant) for the Fourier transformation, i.e. it has characteristic function $\exp(-\beta^2/2)$.
Now the proof and construction of a preferably small constant C becomes a rather technical task. Clearly we have $\max(|\phi|) = 1/\sqrt{2\pi} \leq 0.4$ and using the result

$$|\exp(iy) - \sum_{k=0}^{l} (iy)^k/k!| \leq \frac{|y|^{l+1}}{(l+1)!} \quad \forall y \in \mathbb{R},$$

one can show that

$$|\hat{\xi}(\beta) - 1 + \beta^2/2| \leq \mathbb{E}|\exp(i\beta Y) - 1 - i\beta Y + \beta^2 Y^2/2| \leq |\beta^3|\rho/6,$$

which yields

$$|\hat{\xi}(\beta/\sqrt{n}) - 1 + \beta^2/(2n)| \leq |\beta^3|\rho\frac{n^{-3/2}}{6}.$$

Choose $\theta = (4/3)\sqrt{n}/\rho$.
By Jensen's inequality $\mathbb{E}(|Y|^2)^{3/2} \leq \mathbb{E}(|Y|^3)$, thus we have $\rho > 1$ and $\theta < (4/3)\sqrt{n} < \sqrt{2n}$ then.
Therefore we can further approximate

$$|\hat{\xi}(\beta/\sqrt{n})| \leq 1 - \beta^2/(2n) + |\beta^3|\rho\frac{n^{-3/2}}{6}, \text{ if } |\beta| \leq \theta.$$

Hence, for $\beta \leq \theta$,

$$|\hat{\xi}(\beta/\sqrt{n})| \leq 1 - \beta^2/(2n) + |\beta^2|\frac{4}{18n} = 1 - \frac{5}{18}\frac{\beta^2}{n} \leq \exp(-\frac{5}{18}\frac{\beta^2}{n})$$

and

$$|\hat{\xi}(\beta/\sqrt{n})|^{n-1} = \exp(-\frac{5(n-1)}{18}\frac{\beta^2}{n}) \leq \exp(-\beta^2/4) \text{ if } n \geq 10.$$

It can be estimated that the theorem holds for $n < 10$.
For α and η complex, with $|\alpha| \geq |\eta|$, it holds that

$$|\alpha^n - \eta^n| = |\alpha - \eta|\left|\sum_{j=0}^{n-1} \alpha^j \eta^{n-1-j}\right| \leq |\alpha - \eta| \sum_{j=0}^{n-1} |\alpha^j||\eta^{n-1-j}| = |\alpha - \eta|\ n\ |\alpha|^{n-1}.$$

Therefore the integrand in (4.2) is bounded by

$$|\hat{\xi}(\beta/\sqrt{n}) - \exp(-\beta^2/2n)|\exp(-\beta^2/4)\frac{n}{|\beta|}$$

and

$$|\hat{\xi}(\beta/\sqrt{n}) - \exp(-\beta^2/2n)| \leq |\hat{\xi}(\beta/\sqrt{n}) - 1 + \frac{\beta^2}{2n}| + |1 - \frac{\beta^2}{2n} - \exp(-\frac{\beta^2}{2n})| \leq |\beta^3|\frac{\rho}{6\sqrt{n}} + \frac{|\beta^4|}{8n}.$$

Thus the integrand can further be bounded by

$$n\left(\frac{|\beta|^2\rho}{6\sqrt{n}}+\frac{|\beta|^3}{8n}\right)\exp(-\frac{\beta^2}{4})\frac{4}{3\sqrt{n}}\frac{1}{\theta}\leq\left(\frac{2}{9}\beta^2+\frac{1}{18}|\beta|^3\right)\exp(-\frac{\beta^2}{4})\frac{1}{\theta}.$$

Partial integration over $(-\theta,\theta)$ yields

$$\pi\theta|F_Z(z)-\Phi(z)|\leq(8/9)\sqrt{\pi}+8/9+10<4\pi, \text{ i.e.}$$

$$|F_Z(z)-\Phi(z)|<4\frac{3}{4}\frac{\rho}{\sqrt{n}}.$$

Therefore the theorem holds with $C=3$.

□

The best estimate for the constant C, or at least the best that was broadly published so far, was given in [11], where it was proved that the Berry-Esseen theorem holds for $C=0.4748$. Esseen himself found the still current lower bound $C_0\geq(\sqrt{10}+3)/(6\sqrt{2\pi})=0.4097...$

Remark 4.6. In practice, one further deduces the normal approximation rule from the central limit theorem: for n large enough, the distribution of ε_n is approximately a Gaussian random variable with mean 0 and variance $\frac{\sigma^2}{n}$. It can be estimated via the Berry-Esseen theorem and estimates for the third moment, if a specific n is sufficiently large to allow that approximation.

The preceding rule allows one to define a confidence interval:

Example 4.7. A confidence level often used in practice is the 99% level.
About 99% of the area under a standard normal curve lies within $[-2.58,2.58]$. Therefore we get with Proposition 4.3 for large n, with a probability close to 99%, that

$$|\varepsilon_n|\leq 2.58\frac{\sigma}{\sqrt{n}}.$$

4.2 Implementation of Monte-Carlo methods

To implement this method on a computer, we can proceed as follows.
Most programming languages, or more precisely their corresponding compilers, have already implemented a procedure for creating a sequence of realizations of uniformly distributed random variables. Creating such a sequence directly for arbitrary distributions is generally not possible. Therefore the idea is to find a representation of the desired distribution as a function of a uniform distribution U on $[0,1]$. If the desired distribution X has an invertible cumulative distribution function F, a possible representation is

$$X=F^{-1}U.$$

To verify that, consider

$$P(F^{-1}U\leq a)=P(U\leq F(a))=F(a)=P(X\leq a)$$

since F is a monotone function and U is uniform on $[0,1]$.

Simulation of a uniform law on [0,1]

The simplest method to build a random number generator is to use the linear congruential generator. A sequence $(x_n)_{n\geq 0}$ of integers between 0 and $m-1$ is generated as follows:

$$\begin{cases} x_0 = \text{initial value } \in \{0,1,...,m-1\} \\ x_{n+1} = ax_n + c \ \text{ (modulo m)}, \end{cases}$$

a, c, m being integers to be chosen cautiously in order to obtain a good random number generator. The Hull-Dobell Theorem states that a linear congruential generator has full period m if and only if

$$\begin{cases} \text{m and c are relatively prime,} \\ \text{a-1 is divisible by all prime factors of m} \\ \text{and a-1 is divisible by 4 if m is divisible by 4.} \end{cases}$$

For example, Microsoft Visual C++ uses

$$m = 2^{32},\ a = 214013, \ \text{ and } c = 2531011.$$

Dividing this integer values by m gives pseudo-random real-valued numbers in $[0,1)$.

Linear congruential generators are not suitable if high-quality randomness is critical, one reason being that they produce serial correlation, i.e. corresponding tests yield a dependence of the value at some time or iterate t with the value at some time s.

Simulation of an exponential distribution

Since the exponential distribution has an invertible distribution function $F = 1 - e^{-\lambda x}$, $x \geq 0$, we can simulate a random variable X following an exponential distribution with parameter λ via

$$X = F^{-1}(U) = -\log(1-U)/\lambda.$$

Note, that $1-U$ is also uniformly distributed on $[0,1]$, so we can set

$$X = -\log(U)/\lambda.$$

Simulation of a Gaussian distribution

For the Gaussian distribution the inversion method above is not straightforwardly applicable, since the cumulative distribution function is not an elementary function. Therefore we take a different approach.

Consider two independent standard Gaussian variables X and Y. Their joint density is given by

$$f_{X,Y}(x,y) = f_X(x) f_Y(y) = \sqrt{\frac{1}{2\pi}} \exp(-\frac{x^2}{2}) \sqrt{\frac{1}{2\pi}} \exp(-\frac{y^2}{2}) = \frac{1}{2\pi} \exp(-\frac{x^2+y^2}{2}).$$

In polar variables $R = \sqrt{X^2+Y^2}$ and $\Theta = \arctan(Y/X)$, that is

$$f_{R,\theta}(r,\theta) = \frac{1}{2\pi} r \exp(-\frac{r^2}{2}),\ r \geq 0,$$

because the Jacobian determinant of the transformation is r.
This further yields

$$f_{R^2,\Theta}((\tilde{r},\theta)) = \frac{1}{2\pi}\frac{1}{2}\exp(-\frac{\tilde{r}}{2}) = f_\Theta(\theta) f_{R^2}(\tilde{r}),\ \tilde{r} \geq 0,$$

since in informal notation $dr^2 = 2rdr$ and by considering $f_{R^2}(\tilde{r})$ as a chi-squared distribution with two degrees of freedom.
So $R^2 \sim \exp(\frac{1}{2})$, $\Theta \sim U(0,2\pi)$ and R^2 is independent of Θ. Furthermore, we just saw that $-\log(U)/\lambda$ gives a simulation of an $exp(\lambda)$ distributed random variable. Hence we can set

$$R = \sqrt{-2\log(U_1)},$$

$$\Theta = 2\pi U_2,$$

which yields that

$$X = R\cos(\Theta) = \sqrt{-2\log(U_1)}\cos(2\pi U_2)$$
$$Y = R\sin(\Theta) = \sqrt{-2\log(U_1)}\sin(2\pi U_2)$$

are independent standard Gaussian variables.
Now, if we want to simulate an arbitrary Gaussian law with mean m and variance σ^2, we can set

$$X = m + \sigma g,$$

where $g \sim N(0,1)$.

Simulation of Gaussian vectors

In multidimensional normal models Gaussian vectors are considered. Having established a procedure for simulating a Gaussian variable, we can extend that to simulate a Gaussian vector $X = (X_1,...,X_n)$ with means $m = (\mathbb{E}(X_1),...,\mathbb{E}(X_n))$ and covariance matrix $\Gamma = (\Gamma_{ij})_{1\leq i,j\leq n}$ where $\Gamma_{ij} = \mathbb{E}(X_iX_j) - \mathbb{E}(X_i)\mathbb{E}(X_j)$.

To do that, derive the square root A of the positive semi-definite matrix Γ via Cholesky decomposition. If an eigenvalue of Γ is very close to 0, one has to be careful of numerical instability. Otherwise the Cholesky decomposition has a unique stable solution and A is invertible. Now it can be checked, that the vector $Z = A^{-1}(X-m)$ is a Gaussian vector with zero mean and the identity matrix as covariance matrix, i.e. the coordinates of Z are independent standard normal variables.
Therefore simulating n independent standard normal variables $G = (g_1,...,g_n)$ and computing $m + AG$ gives a realization of a Gaussian vector with mean m and covariance matrix $\Gamma = AA^T$.

Simulation of a Poisson distribution

Recall the definition of a Poisson process.

Definition 4.8. Let $(T_i)_{i\geq 1}$ be a sequence of i.i.d. random variables with distribution $\exp(\lambda)$. Let $\tau_n = \sum_{i=1}^{n} T_i$. We call the Poisson process with intensity λ the process N_t defined by

$$N_t = \sum_{n\geq 1} 1_{\{\tau_n \leq t\}} = \sum_{n\geq 1} n 1_{\{\tau_n \leq t < \tau_{n+1}\}}.$$

Also recall that N_t follows a Poisson law with parameter λt, i.e.

$$\mathbb{P}(N_t = n) = e^{-\lambda t}\frac{(\lambda t)^n}{n!}, n \geq 0.$$

This relationship between a Poisson process and the exponential distribution allows us to simulate a Poisson distribution.
N_1 has the desired distribution with parameter λ. As established before, we can express an exponential variable as $-\log(U)/\lambda$. This yields

$$\tau_n = T_1 + ... + T_n = -\frac{1}{\lambda}\log(U_1 U_2 \cdots U_n).$$

So adding and counting realizations of $\exp(\lambda)$ one at a time, until the term exceeds the value 1, gives a realization of a $Poi(\lambda)$ variable. To not have to compute a logarithm at every step, it is numerically more efficient to use the fact

$$T_1 + ... + T_n \leq 1 \Leftrightarrow \prod_{i=1}^{n} U_i \geq e^{-\lambda}.$$

4.3 Comparison with other methods

Note, that Monte Carlo simulation to estimate the expected value of a distribution $Y = f(X)$ with $X \sim U(0,1)$ corresponds to numerically estimating the integral

$$\mathbb{E}(f(X)) = \int f(x) p_X(x) dx = \int f(x) dx \approx \frac{1}{N}\sum_{i=1}^{N} f(x_i)$$

at random points $\{x_i\}$ in $[0,1]$.
The Monte Carlo method can theoretically be applied without having any information about the underlying distribution. When we know the distribution of X, as it is the case in most applications, where we have $X \sim U(0,1)$, we can consider different methods to estimate the integral above.
If we consider the simplest approach, using a staircase function, we get

$$\mathbb{E}(f(X)) \approx \sum_{i=1}^{N} f(x_i)\Delta x$$

with $x_i = i/N$ and $f : \mathbb{R}^m \to \mathbb{R}$.
It can be shown that for in the one dimensional case $f : \mathbb{R} \to \mathbb{R}$ the convergence rate is better when we just pick evenly distributed points from $U(0,1)$ as above, whereas for high dimensions it is much worse compared to simulating random points.

One can also consider more precise numerical integration methods, but those are sometimes difficult and expensive to implement.

Finally, instead of simulating with pseudo-random numbers in Monte Carlo methods one can also consider low discrepancy sequences. These are also deterministic sequences and in general more evenly distributed than the pseudo-random ones. This can lead to better convergence rates than $O(1/\sqrt{N})$ as in the regular Monte-Carlo methods even for higher dimensions. Consider though, that these so-called Quasi-Monte Carlo methods do not always lead to better results and their convergence rates are more difficult to estimate.

4.4 Polynomial Chaos Expansion

A fairly modern approach is to represent the desired distribution by using a so-called polynomial chaos expansion (PCE), i.e. for example to expand the relationship $X = F^{-1}(U)$ in a polynomial series.
Additionally, in polynomial chaos theory also other base distributions than the uniform distribution are considered, generally called germ distribution θ. Typically the choice is between uniform, normal and exponential distributions. In the case that the distribution function is invertible we have

$$X = f(\theta) = F_X^{-1} F_\theta(\theta),$$

since $F_\theta(\theta)$ is a uniform random variable on $[0,1]$.
Now a PCE takes that representation and expands it in a polynomial series. The specific polynomial basis used is a set of orthogonal polynomials with respect to the distribution of the germ. More precisely, consider the inner product of functions g_1, g_2 with respect to the probability density function p_θ of θ, defined by

$$\langle g_1, g_2 \rangle = \int g_1(\xi) g_2(\xi) p_\theta(\xi) d\xi.$$

Then the polynomial basis comprises polynomials $\Psi_0 = 1, \Psi_1, \Psi_2, ...$, where Ψ_j is a polynomial of order j and where they satisfy the orthogonality condition that

$$\langle \Psi_j, \Psi_k \rangle = 0 \quad \forall j \neq k. \tag{4.3}$$

The expected value of the polynomial function, hence continuous and on $\mathbb{R}$ measurable function Ψ_j of θ is given by

$$\mathbb{E}[\Psi_j(\theta)] = \int \Psi_j(\xi) p_\theta(\xi) d\xi.$$

Therefore, by construction, $\mathbb{E}[\Psi_j(\theta)] = 0$ for $j \geq 1$, since they are orthogonal to $\Psi_0 = 1$.

This further implies that $\langle \Psi_j, \Psi_j \rangle$ is the variance of $\Psi_j(\theta)$ for all $j \geq 1$ and $\langle \Psi_i, \Psi_j \rangle$ is the covariance of $\Psi_j(\theta)$ and $\Psi_k(\theta)$ for all $j,k \geq 1$.
Consequently, since $\langle \Psi_j, \Psi_k \rangle = 0$ for $j \neq k$, the $\Psi_j(\theta)$ are uncorrelated.
Also note, that since we always have $\Psi_0 = 1$,

$$x_0 := \langle f, \Psi_0 \rangle / \langle \Psi_0, \Psi_0 \rangle = \langle f, \Psi_0 \rangle = \int f(\xi) p_\theta(\xi) d\xi = \mathbb{E}(f(\theta)) = \mathbb{E}(X).$$

Furthermore, it can be shown that

$$Var(X) = \sum_{i \geq 1} (\langle f, \Psi_i \rangle / \langle \Psi_i, \Psi_i \rangle)^2 =: \sum_{i \geq 1} x_i^2.$$

So estimating the expected value and variance of X comes down to approximating the x_i.

A popular example are the Hermite polynomials.

Definition 4.9. The Hermite polynomials are defined by

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n} e^{-x^2}, n = 0, 1, 2, ...$$

Proposition 4.10. *If the germ distribution $\theta \sim N(0, \frac{1}{2})$, the Hermite polynomials satisfy the orthogonality condition* (4.3)*.*

Proof. The definition of the Hermite polynomials implies

$$\int \frac{1}{\sqrt{\pi}} e^{-x^2} H_m(x) H_n(x) dx = (-1)^n \frac{1}{\sqrt{\pi}} \int H_m(x) \frac{d^n}{dx^n} e^{-x^2} dx.$$

Partial integration, applied n times, yields that the integral vanishes for $m < n$. □

Now using the orthogonal basis polynomials we write

$$X = f(\theta) = \sum_{j=0}^{\infty} x_j \Psi_j(\theta). \tag{4.4}$$

The x_j are called mode strengths, the Ψ_j mode functions and their combination $x_j \Psi_j$ is called the j-th mode. Given f and the Ψ_j's there is a unique expansion (4.4) in which the mode strengths are given by

$$x_j = \langle f, \Psi_j \rangle / \langle \Psi_j, \Psi_j \rangle. \tag{4.5}$$

Thereby, the numerator usually has to be computed numerically, whereas the denominator will in general be known from the construction of the orthogonal polynomials.

Remark 4.11. In practice, PCEs are truncated to a finite number of terms, hence we consider

$$X_p = f_p(\theta) = \sum_{j=0}^{p} x_j \Psi_j(\theta).$$

For detailed insights, it is referred to the publications of Dongbin Xiu, see e.g. [16].

Remark 4.12. Another simulation approach in the context of regression problems is the so-called Bootstrap aggregation, where the empirical distribution function $\hat{D}$ over $X \times Y$ is being simulated by a Monte-Carlo method. That is, a number of equal-sized resamples with replacement from the empirical distribution function are being produced.
Then modeling from the resamples separately and averaging over the individual results can lead to better model fits than only considering the original data set.

Chapter 5

Application to the analysis of telematics data

A modern approach in German car insurance is to track the drives of the policyholders in order to get more information about high-risk driving profiles and to lower the amount of accidents by rewarding a safe driving style.
The analysis of this so-called telematics data is still quite unexplored. Therefore a project has been started by the mathematical association VM4K to further the research in this area. The first results are shown in this application chapter. The utilized methods have been established throughout this thesis.

Telematics data is recorded during the car drives of the project participants and tracks properties like current speed, acceleration and route type. As a first step, the data has been enriched by generating additional properties like the exact mileage and indicators for speeding and braking.

To be able to estimate the claims expenditures of a driving profile, we aggregate the data to a driver level, that is we build a $n \times p$ model matrix X consisting of n different drivers and p considered driving properties. Thereby we consider different property sets, which are being standardized as preparation for a Principal Component Analysis.
As introduced in section 3.7, this data mining approach allows us to extract principal components and to reduce the dimensionality of the respective regression models.

5.1 PCA of the telematics data

We have established in section 3.7 that the Principal Component Analysis can be comfortably performed using the Singular Value Decomposition (SVD) and we introduced implementation ideas for the SVD. There are optimized SVD routines available in different libraries.

The SVD routine for the model matrices of telematics data was taken out of the C++ library Eigen, which is a high-level template library and its routine JacobiSVD performs similarly to the approach we discussed in Section 3.7 by pivoted QR-preconditioning and Jacobi-rotations.

As mentioned above, we considered the SVDs for model matrices with different property sets in this project, in order to get more insights about characteristic driving styles and suitable regression models based on them.
For better clarity, we will restrict ourselves to two different property sets in the presentation. The first one consists of only six properties which are supposed to be risk relevant due to prestudies and which are not too strongly correlated. In that setting the principal components and the resulting regression models are nicely interpretable.
The second property set consists of 26 selected properties. By means of that set we show exemplarily how the SVD can be used to reduce the dimensionality of the respective telematics regression model.

The analysis of the first property set serves as a initial step to get insights about key properties and characteristical linear combinations of them, which can in turn help to interpret bigger models later. As mentioned, key properties are ones that turned out to be probably significantly risk relevant in prestudies.
However, due to the large amount and high resolution of the given data, there is also a large number of possible properties, so much more important properties like e.g. about overtaking behavior will probably be found in further telematics research.

Remark 5.1. One has to keep in mind that the SVD only recognizes linear relationships. So in the process of extracting properties out of the big data matrix, it might be helpful to also consider non-linear functionals. For example, for the key property to measure the acceleration behavior, we took the squared g-forces, which led to seemingly more reasonable results compared to the linear values.

The resulting principal components and singular values for the first property set are shown in figure 5.1.
Based on this results we will build a small PC-regression model, which will be relatively easy to interpret. Before that, we will also consider a standard regression model based on the key properties itself to get some preliminary insights about the isolated impact of single properties.

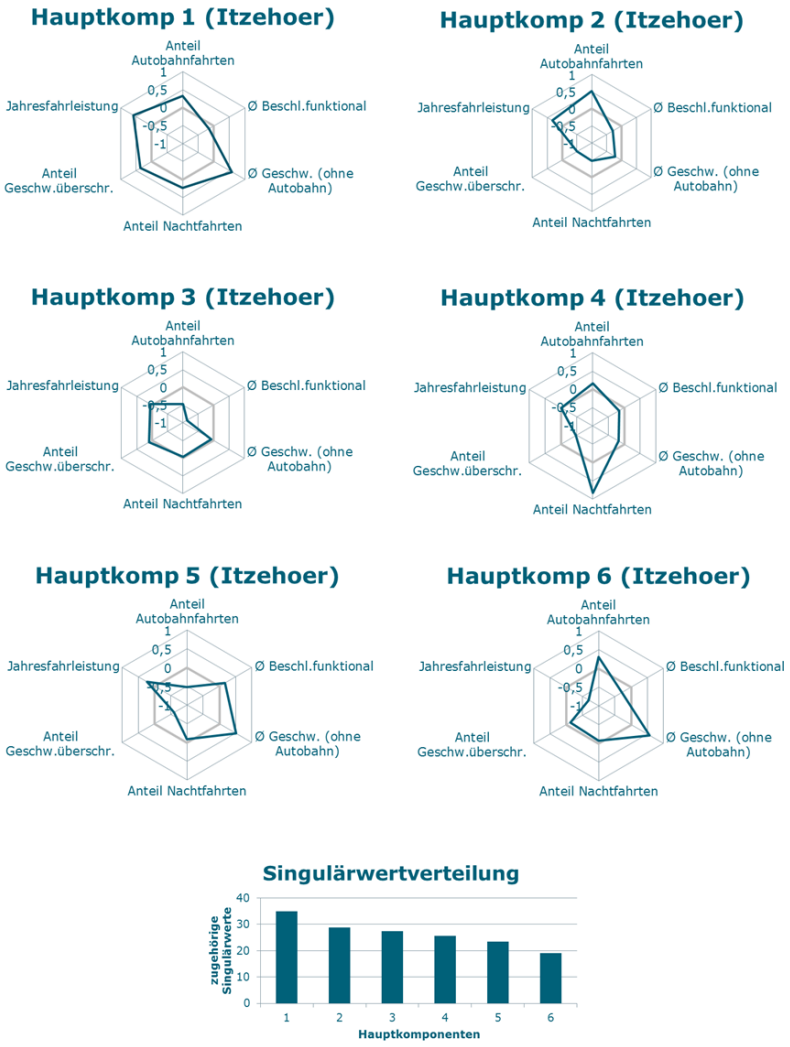

Figure 5.1: SVD of the model matrix with just 6 properties: the 6-dimensional normalized principal component vectors, i.e. the columns of the matrix V in the SVD, are shown in a graphical representation

For further insights to possible rewards for specific driving styles, one can try to associate types of drivers to the principal components.
For example, the second PC could be associated with a commuter driving style, since it contains over-average mileage and rate of freeway drives, but under-average rates of night drives and speeding. Furthermore, the 4th PC could be associated with a new driver because of the very low rate of speeding and strongly over-average rate of night drives which indicates a younger age. In contrast, PC 6 could be associated with a holiday driver due to the low mileage and over-average rate of freeway drives.

As the second property set for the presentation, we use a selection of 26 properties to show, how this data mining approach can be utilized for dimensionality reduction. This approach can then also be applied to much larger property sets.
Selected principal components and the singular values can be seen in figure 5.2.

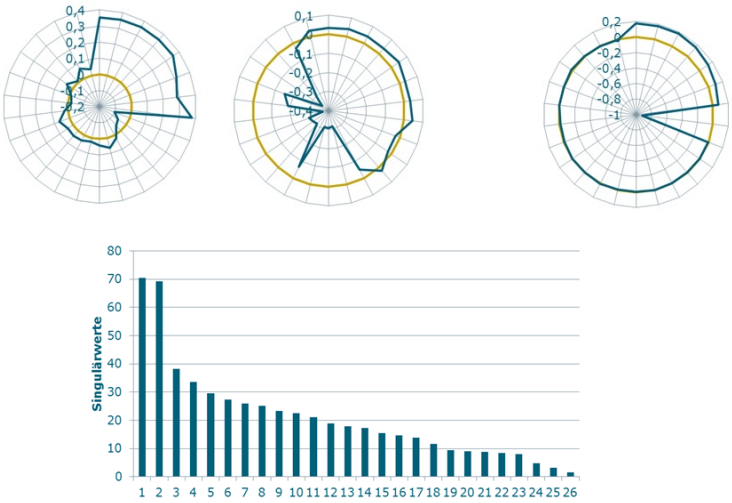

Figure 5.2: SVD of the model matrix with 26 properties, shown are from left to right the first, second and 26. principal component in a graphical representation and the singular values

In figure 5.2 we see that for the selection of 26 properties we have a much more inhomogeneous distribution of the singular values. Two singular values are quite dominant and for example the last principal component does have many entries very close to zero. This indicates that there are strongly correlating properties in the selection. Therefore we will perform a very radical dimensionality reduction and evaluate a regression model based on only the first two PC scores.

Remark 5.2. It is a natural approach and the most popular one to choose the PC scores with the largest variance for building the regression model. However, it should be mentioned that Jolliffe found examples in [21] in which this approach is not optimal for the regression results. For further studies on that topic one should be mindful of slightly different definitions of principal components across the literature.

Before we get to the dimensional reduction, we first return to the small property set and build regression models based on that properties.

5.2 Constructing a GLM based on selected driving properties

Now, as a first preliminary step to get insights about the influence of individual properties on the expected claims expenditures, we construct a GLM based on the six properties in the first property set directly.

After the aggregation to a driver level and data mining we are able to build a reasonable GLM and therefore chose this regression method here instead of one without distribution assumptions for unstructured data. As underlying distribution the Poisson distribution was taken and as link function the canonical log-link. These assumptions were chosen because they turned out to be suitable in many insurance applications and they give a multiplicative pricing structure which is very nice to interpret.
As response variable we take the corresponding premiums of the drivers.

Remark 5.3. The claims expenditures were not fully available yet, so the corresponding premiums of the policyholders were taken as a replacement response variable, which implicitly contain the expected claims expenditures of a driver.

For example for the property 'rate of night drives' we see in figure 5.3 that a high rate leads to distinctly higher premiums. So punishing night drives in a telematics tariff could lead to smaller claims expenditures.

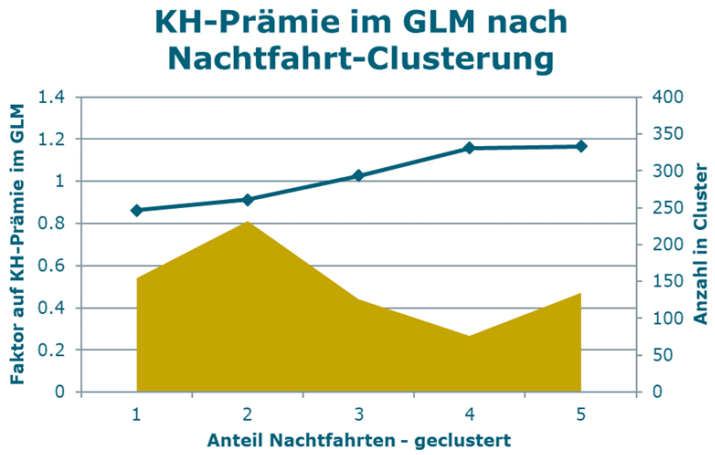

Figure 5.3: property "rate of night drives" in Poisson GLM based on 6 properties

Remark 5.4. There are different approaches to measure the quality of a regression model. The statistical approach is often by applying hypothesis tests. Another popular measure for the goodness of fit is the R^2-value, which calculates the decrease of errors in comparison with a null model, that just takes the average as constant fit for all values.
The approach in the Machine Learning framework is generally to validate the model with a

test data set. The frameworks coincide though and using test data is also a known technique in the field of statistics.

In the presentation we will use a method, that can be reasonably applied to any model, the cross-validation.
Cross-Validation is a good technique to validate the goodness of predictions of a model, which we are particularly interested in.
As already mentioned in remark 3.47, it is done by splitting the model matrix into parts of equal or similar size, iteratively taking one part as test data and building the regression model based on the remaining model matrix.

More precisely, we randomly split the driving profiles in our model matrix into 4 parts. To do that, we assign a random number to each driving profile using a linear congruential generator, as described in section 4.2, order them and build 4 partitions of similar size. Following that we take the first part as test data and build a GLM based on the coefficients a_1 and a_2 of the remaining driving profiles. Then we take the second part as test data and so on. This procedure is called 4-fold cross validation.

The results can be seen in figure 5.4. It shows the distribution of the relative differences between the current premiums and the ones predicted by the GLM which is based on six selected properties.

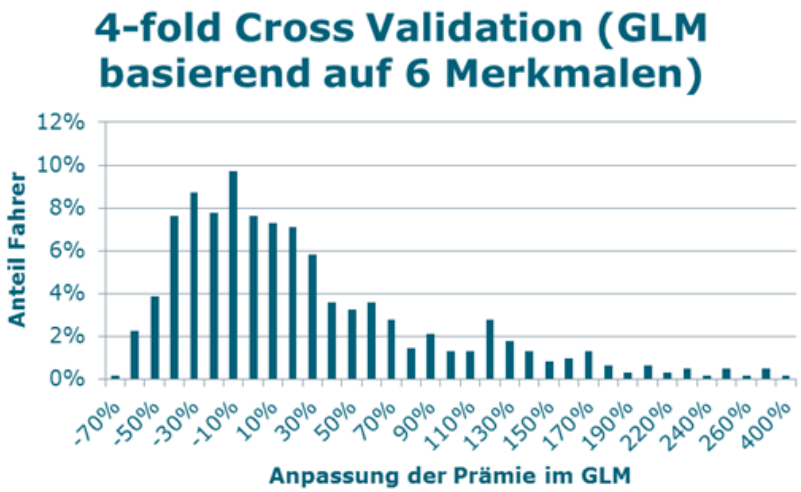

Figure 5.4: Cross Validation of the Poisson GLM based on 6 selected properties

5.3 Constructing a GLM based on the PCA

Our main idea is to estimate the claims expenditures based on the principal component scores in the SVD, i.e. on the similarity to characteristic driving styles.
Recall, that since the first r columns of V build an orthonormal basis for the columns of the rank r matrix X, i.e. for the driving profiles, we can represent any driving style v in terms of the orthonormal columns V_i, i.e. $v = \sum a_i V_i$ with coefficients or PC scores $a_i = \langle v, V_i \rangle$.

With this approach we do not analyze the impact of specific properties, but the contribution of characteristic driving profiles. As regression method we choose a Poisson GLM again for the same reasons. Firstly, for the model matrix with just 6 properties one of the results can be seen in figure 5.5.

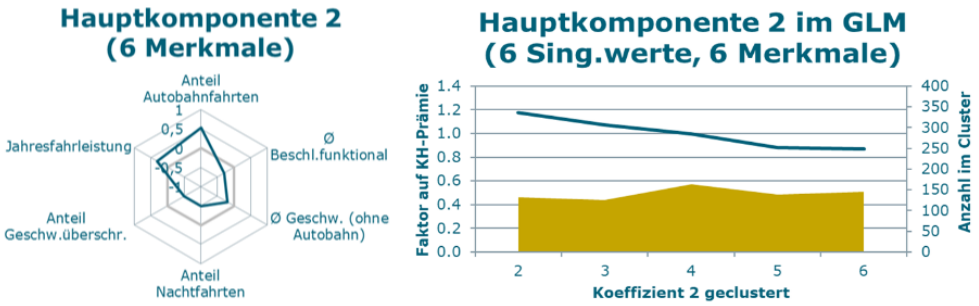

Figure 5.5: SVD based GLM for model matrix with just 6 properties

As already mentioned, the second principal component for the model matrix with just 6 properties can be associated with a commuter driving style.
The GLM indicates that this driving style is a safe driving style, since a larger corresponding coefficient leads to lower premiums for a driver.
Since we considered the entire decomposition of the model matrix and the same property set as in the preceding model, the distribution of the errors is similar here.

Remark 5.5. Note, that the sign of a principal component V_i is arbitrary since we have $U\Sigma V = (-U)\Sigma(-V)$ in a decomposition. However, in the GLM this is adjusted, since the sign of the coefficients or PC scores $(-U)\Sigma$ also change when we consider $-V$ instead of V. So we would get that the opposite characteristic driving style $(-V_i)$ is a risky one instead of the original one being a safe driving style.

Using the SVD of the model matrix with 26 properties, we will show how we we can reduce the dimensionality of the regression model. Since figure 5.2 shows that the first two singular values are quite dominant, we cut off after these two and build a rough but simple Poisson GLM on the coefficients of the first two principal components.
The analysis for the second principal component can be seen in figure 5.6.

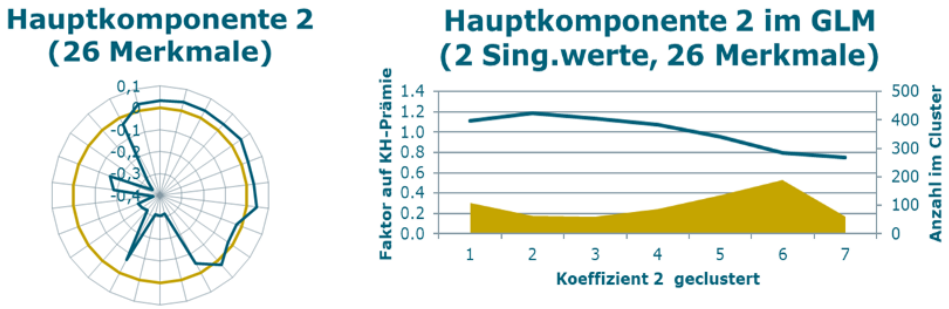

Figure 5.6: SVD based GLM for model matrix with 26 properties

In this model matrix with 26 properties the second principal component comprises, among others, slightly over-average rates of routine drives and under-average rates of night drives. The GLM indicates that this driving style is a safe one. It is quite remarkable that we get a differentiation of about 50% higher premiums for the coefficients in the lowest cluster compared to the highest cluster.

The results of the cross-validation can be seen in figure 5.7, where the distribution of the relative differences between the original premiums and the ones predicted by the GLM based on only two coefficients or PC scores is shown.

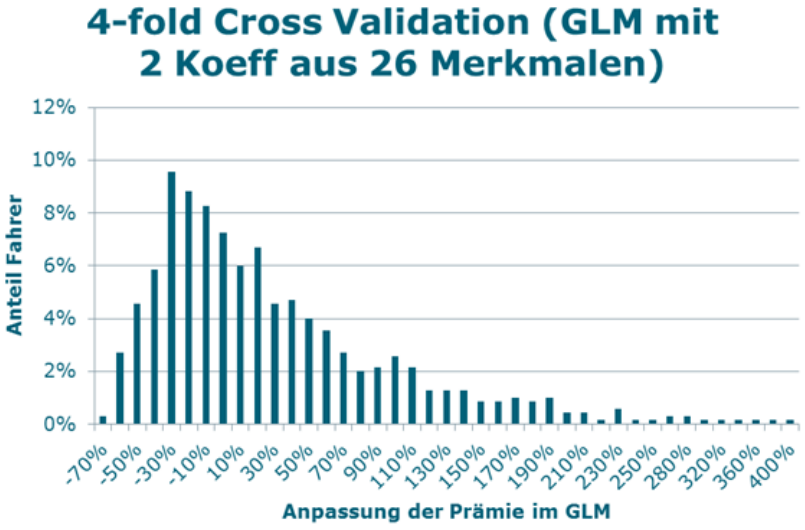

Figure 5.7: Cross Validation of the Poisson GLM based on the first 2 PC scores in the model matrix with 26 properties

Considering we only use two PC scores to build the GLM, the predictions are quite good. They are already comparable to the predictions of the 6-dimensional GLM in figure 5.4.

Overall we were able to predict, to a limited extent, the premiums of project participants based on their driving behavior already with very low-dimensional models. The premiums were only a replacement response variable though and since there is supposed to be new important information gained from driving styles, the matching probably cannot be completely accurate with any model. However, when there is significant data about the corresponding claims expenditures available, these studies can serve as a basis for bigger models with the actual claims expenditures as response variables.

5.4 Outlook: Polynomial Chaos Expansion for broader model fitting

The empirical distribution function of driving styles is bounded by the maximum and minimum values that appear in the dataset. In bigger driver populations there will likely appear more extreme driving styles though.

Therefore an idea is to smooth the empirical distribution function by matching a polynomial function with the same first moments to the empirical distribution function, which is a form of polynomial chaos expansion. Then new driving styles can be simulated as described in Chapter 4.

The corresponding algorithm could not be programmed stably yet, but it is an interesting approach for further studies.

Bibliography

[1] Alan Agresti, *Foundations of Linear and Generalized Linear Models*, Wiley Series in Probability and Statistics, 2015

[2] Mehryar Mohri, Afshin Rostamizadeh and Ameet Talwalkar, *Foundations of Machine Learning*, Massachusetts Institute of Technology Press, 2012

[3] Anthony O'Hagan, *Polynomial Chaos: A Tutorial and Critique from a Statistician's Perspective*, University of Sheffield, 2013

[4] Ronald Christensen, *Plane Answers to Complex Questions: The Theory of Linear Models*, Springer Texts in Statistics, 2011

[5] Viorel Barbu and Teodor Precupanu, *Convexity and Optimization in Banach Spaces*, Fourth Edition, Springer Monographs in Mathematics, 2012

[6] J.C. De los Reyes, *Numerical PDE-Constrained Optimization*, Springer Briefs in Optimization, Springer International Publishing, 2015

[7] Zlatko Drmač and Krešimir Veselić, *New Fast and Accurate Jacobi SVD Algorithm I*, SIAM Journal on Matrix Analysis and Applications Vol. 29(4), pp. 1322-1342, 2008

[8] Zlatko Drmač and Krešimir Veselić, *New Fast and Accurate Jacobi SVD Algorithm II*, SIAM Journal on Matrix Analysis and Applications Vol. 29(4), pp. 1343-1362, 2008

[9] Zlatko Drmač, Matjaž Omladič and Krešimir Veselić, *On the perturbation of the Cholesky factorization*, SIAM Journal on Matrix Analysis and Applications Vol. 15(4), pp. 1319-1332, 1994

[10] Damien Lamberton and Bernard Lapeyre, *Introduction to Stochastic Calculus Applied to Finance*, Chapman & Hall, 1996

[11] Irina Shevtsova, *On the absolute constants in the Berry–Esseen type inequalities for identically distributed summands*, Moscow State University, 2011

[12] John E. Kolassa, *Series Approximation Methods in Statistics*, Springer Lecture Notes in Statistics, 2006

[13] W. Feller, *An Introduction to Probability Theory and its Applications*, New York: Wiley, 1971

[14] T. Gard, *Introduction to stochastic differential equations*, Marcel Dekker, 1988

[15] Ioannis Karatzas and Steven E. Shreve, *Brownian Motion and Stochastic Calculus*, Springer New York, 1988

[16] Dongbin Xiu, *Numerical Methods for Stochastic Computations: A Spectral Method Approach*, Princeton University Press, 2010

[17] Alex J. Smola and Bernhard Schölkopf, *A Tutorial on Support Vector Regression*, Kluwer Academic Publishers, Statistics and Computing 14(3), pp. 199-222, 2004

[18] Stephen Boyd and Lieven Vandenberghe, *Convex Optimization*, Cambridge University Press, 2009

[19] Alecos Papadopoulos, *The Information Matrix Equality: proof, misspecification, and the quasi-maximum likelihood estimator*, Athens University, 2014

[20] Gene H. Golub and Charles F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Third Edition, 1996

[21] I.T. Jolliffe, *A Note on the Use of Principal Components in Regression*, Journal of the Royal Statistical Society, Series C (Applied Statistics), Vol. 31(3), pp. 300-303, 1982

[22] I.T. Jolliffe, *Principal Component Analysis*, Springer Series in Statistics, 2002

Selbstständigkeitserklärung

Ich habe die Arbeit selbständig verfasst, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und bisher keiner anderen Prüfungsbehörde vorgelegt.

Ort, Datum	Unterschrift