Robustness for regression models with asymmetric error distribution

Daria Pupashenko

Datum der Disputation: 16 März 2014

Vom Fachbereich Mathematik der Technischen Universität Kaiserslautern zur Verleihung des akademischen Grades Doktor der Naturwissenschaften (Doctor rerum naturalium, Dr. rer. nat.) genehmigte Dissertation.

1. Gutachter: Priv.-Doz. Dr. Peter Ruckdeschel
2. Gutachter: Prof. Dr. Helmut Rieder

D 386

Abstract

In this work we focus on the regression models with asymmetrical error distribution, more precisely, with extreme value error distributions. This thesis arises in the framework of the project "Robust Risk Estimation". Starting from July 2011, this project won three years funding by the Volkswagen foundation in the call "Extreme Events: Modelling, Analysis, and Prediction" within the initiative "New Conceptual Approaches to Modelling and Simulation of Complex Systems". The project involves applications in Financial Mathematics (Operational and Liquidity Risk), Medicine (length of stay and cost), and Hydrology (river discharge data). These applications are bridged by the common use of robustness and extreme value statistics.

Within the project, in each of these applications arise issues, which can be dealt with by means of Extreme Value Theory adding extra information in the form of the regression models. The particular challenge in this context concerns asymmetric error distributions, which significantly complicate the computations and make desired robustification extremely difficult. To this end, this thesis makes a contribution.

This work consists of three main parts. The first part is focused on the basic notions and it gives an overview of the existing results in the Robust Statistics and Extreme Value Theory. We also provide some diagnostics, which is an important achievement of our project work. The second part of the thesis presents deeper analysis of the basic models and tools, used to achieve the main results of the research.

The second part is the most important part of the thesis, which contains our personal contributions. First, in Chapter 5, we develop robust procedures for the risk management of complex systems in the presence of extreme events. Mentioned applications use time structure (e.g. hydrology), therefore we provide extreme value theory methods with time dynamics. To this end, in the framework of the project we considered two strategies. In the first one, we capture dynamic with the state-space model and apply extreme value theory to the residuals, and in the second one, we integrate the dynamics by means of autoregressive models, where the regressors are described by generalized linear models.

More precisely, since the classical procedures are not appropriate to the case of outlier presence, for the first strategy we rework classical Kalman smoother and extended Kalman procedures in a robust way for different types of outliers and illustrate the performance of the new procedures in a GPS application and a stylized outlier situation.

To apply approach to shrinking neighborhoods we need some smoothness, therefore for the second strategy, we derive smoothness of the generalized linear model in terms of L_2 differentiability and create sufficient conditions for it in the cases of stochastic and deterministic regressors. Moreover, we set the time dependence in these models by linking the distribution parameters to the own past observations. The advantage of our approach is its applicability to the error distributions with the higher dimensional parameter and case of regressors of possibly different length for each parameter. Further, we apply our results to the models with generalized Pareto and generalized extreme value error distributions.

Finally, we create the exemplary implementation of the fixed point iteration algorithm for the computation of the optimally robust influence curve in `R`. Here we do not aim to provide the most flexible implementation, but rather sketch how it should be done and retain points of particular importance. In the third part of the thesis we discuss three applications, operational risk, hospitalization times and hydrological river discharge data, and apply our code to the real data set taken from Jena university hospital ICU and provide reader with the various illustrations and detailed conclusions.

Acknowledgements

In the first place, I want to thank my supervisor Dr. habil. Peter Ruckdeschel for his constant support, important advices and constructive criticism in the writing of this thesis. In three years of our cooperation I took part in publication of two papers and participated in various prestigious conferences, so I am very grateful for this experience.

Further, I want to express my great appreciation to Prof. Dr. Matthias Kohl for supporting my work as my second supervisor and providing me with numerous advices and valuable remarks.

My appreciation also goes to all my colleagues in the project "Robust Risk Estimation" for the fruitful cooperation, team work experience and gained knowledges in various fields, i.e. Dr. Bernhard Spangl, Dr. Gerald Kroisandt, Dr. Sascha Desmettre and M. Sc. Mykhailo Pupashenko. I am also very thankful to Volkswagen foundation for their financial support.

I want to express special thanks to the Financial Mathematics Group of Fraunhofer ITWM, where I have spent all three years of my PhD research, for the comfortable working atmosphere. It was a pleasure to be part of this friendly company.

Further, I would like to express my appreciation to Dr. habil. Christoph Lossen and Dr. Falk Triebsch for their helpful advices during my studies. A special thank goes to Jessica Borsche for her continued support and tenderness.

I deeply wish to thank the International School for Graduate Studies (ISGS), especially Dipl.-Math. Arthur Harutyunyan, for helping me integrate into a foreign country.

I would also like to express my deepest gratitude to my parents for their love, patience, support and believing in me, my beloved brother, Mykhailo Pupashenko, and his wife, Olga Pupashenko, for always being there for me.

Contents

List of Figures

List of Tables

Abbreviations

a.e.	almost everywhere
ALE	**A**symptotically **L**inear **E**stimator
AO	**a**dditive **o**utliers
BP	**B**reakdown **P**oint
c.d.f.	**c**umulative **d**istribution **f**unction
CF	**C**ovariance **F**ilter algorithm
Ch.	Chapter
CLT	**C**entral **L**imit **T**heorem
COPRA	**C**omputer **O**rganized **P**atient **R**eport **A**ssistant
Def.	Definition
dd plot	**d**istance-**d**istance plot
dp plot	**d**istance-**p**rojection plot
DRG	**D**iagnosis **R**elated **G**roups
e.g.	for example
EM-algorithm	**E**xpectation-**M**aximization algorithm
et al.	and others
etc.	and the others
EVT	**E**xtreme **V**alue **T**heory
FI	**F**isher **I**nformation
FSBP	**F**inite **S**ample **B**reakdown **P**oint
GARCH	**G**eneralized **A**uto**r**egressive **C**onditional **H**eteroskedasticity Model
GES	**G**ross **E**rror **S**ensitivity
GEVD	**G**eneralized **E**xtreme **V**alue **D**istribution
GLM	**G**eneralized **L**inear **M**odel
GMM	**G**eneralized **M**ethod of **M**oments Estimator

GPD	**G**eneralized **P**areto **D**istribution
ICU	**I**ntensive **C**are **U**nit
i.e.	that is to say
IF	**I**nfluence **F**unction
iff	if and only of
i.i.d.	independent identically distributed
IO	**i**nnovation **o**utliers
KF	**K**alman **F**ilter
kMedMAD	**M**edian-**kMAD**
LDA	**L**oss **D**istribution **A**pproach
LDE	**L**ocation-**D**ispersion **E**stimator
LLN	**L**aw of **L**arge **N**umbers
LOS	**L**ength **o**f **S**tay
MAD	**M**edian of **A**bsolute **D**eviation
maxbias	maximal asymptotic bias
MBRE	**M**ost **B**ias-robust **E**stimator
MCD	**M**inimum **C**ovariance **D**eterminant
MDA	**M**aximum **D**omain of **A**ttraction
MDE	**M**inimum **D**istance **E**stimator
MLE	**M**aximum **L**ikelihood **E**stimator
MME	**M**ethod of **M**oments **E**stimator
MSE	**M**ean **S**quared **E**rror
OBRE	**O**ptimal **B**-**R**obust **E**stimator
OMSE	**O**ptimal **M**ean **S**quared **E**rror Estimator
OR	**O**perational **R**isk
POT	**P**eak-**O**ver-**T**hreshold
pp plot	**p**rojection-**p**rojection plot
QQ plot	**Q**uantile-**Q**uantile plot
RD	**R**obust **D**istances
RMXE	**R**adius **M**inima**x** **E**stimator
r.v.	**r**andom **v**ariable
Sec.	Section
SO	**s**ubstitutive **o**utliers

SOFA	**S**epsis-related **O**rgan **F**ailure **A**ssessment
SSM	**S**tate-**S**pace **M**odels
SRCF	**S**quare **R**oot **C**ovariance **F**ilter
s.t.	such that
Thm.	Theorem
w.r.t.	with respect to

Notations

Sets and Functions

$\mathbb{N}$	set of natural numbers
$\mathbb{R}$	set of real numbers
$\mathbb{R}^k$	set of real vectors
$\mathbb{R}^+$	set of positive real numbers
$\times$	Cartesian product of sets
$\mathbf{1}$	indicator function
Θ	open parameter domain
Γ	Gamma function
Λ	L_2 derivative
I	Fisher information

Measures and Norms

$\lvert \cdot \rvert$	Euclidean norm on $\mathbb{R}^k$ ($\mathbb{R}$) space
ν	counting or dominating measure
$\lVert \cdot \rVert_{L_2^k}$	norm on the respective $L_2^k(\nu)$ space
$(\Omega, \mathcal{A})$	measurable space
$\mathcal{M}_1(\mathcal{A})$	the set of probabilities on the σ-algebra $\mathcal{A}$
$\otimes$	product of measures

Random Variables

$X = \{X_1, ..., X_n\}$	sample of random variables

$\mathbf{P}$	probability
$\mathbf{E}X$	expectation of the random variable X
$\mathrm{Cov}(X) = \mathbf{E}(X^2)$	covariance of the random variable or vector X
$\bar{X}$	sample mean
$\mathrm{Med}(X)$	sample median
$\sigma(X)$	σ-algebra generated by X
$\hat{\boldsymbol{\vartheta}}_n$	estimate of the parameter ϑ
$\hat{\boldsymbol{\vartheta}}_\infty$	asymptotic value of the estimate $\hat{\boldsymbol{\vartheta}}_n$
X^{id}	random variable generated by the ideal distribution
X^{di}	random variable generated by the distorting distribution
X^{re}	random variable generated by the realistic distribution

Distributions

$\mathfrak{L}(X)$	distribution of the random variable X
$X \sim Q$	random variable X is Q distributed
$X_t \sim^{indep.} Q$	$X_t, t \in \mathbb{N}$ independent identically Q distributed
$\overline{Q}(x) = \mathbf{P}(X > x)$	tail distribution function for X
$\mathrm{Bin}(n, p)$	Binomial distribution
$\mathcal{N}(\mu, \sigma)$	Normal distribution
$\mathcal{CN}_2(r, \mu, R, \mu_c, R_c)$	contaminated bivariate normal distribution

Mathematical Symbols

$\dot{f} = \frac{\partial}{\partial x}(f(x))$	derivative of the function f w.r.t. x
a^{T}	transpose of vector a
$\mathbb{I}$	unit matrix
$\det(A)$	determinant of matrix A
A^{-1}	inverse of matrix A
$\mathrm{tr}A$	trace of the matrix A
$\mathrm{diag}(a_1, ...a_k)$	diagonal matrix
$\approx$	approximately equal
$\rightarrow^p$	convergence in probability

$\downarrow$	limit from the right
$o(n)$	little-o (Landau) notation
δ_x	Dirac measure at x

This thesis is dedicated to Mykola Pupashenko, my oldest brother and my good friend, who has been an example and mainstay for me all my life.

Chapter 1

Introduction

1.1 Motivation

From the title of the thesis it becomes clear, that here we focus on the the regression models with asymmetrical error distribution, more precisely, with extreme value error distributions. These regression models are applied in a variety of different application domains, e.g. hydrology, finance and public health. In all these settings classical estimation and inference is enhanced by the common use of robust statistics. The particular challenge in this context comes from the asymmetric error distributions, which significantly complicate the computations and make desired robustification extremely difficult.

While for i.i.d. observations from extreme value distributions there already is a sizable amount of robustification available, and in particular the approach underlying this thesis has been covered by Horbenko (2011). So far these approaches do not make use of potentially available additional information,in form of predictors and regressors, to enhance predictable power. As a consequence its scale and shape parameters vary from observation to observation.

Main focus of the research belongs to two types of the regression models. The first type covers dynamical regression models, more specifically, state-space models, i.e. typical observation driven models (compare Cox (1981)). Our interest in this model is caused by the fact, that it can be treated as dynamical system for the measuring of some sort of the signal. Later, discussing applications of our research, we motivate the choice of this model by the hydrological application, concerning river discharge data.

The second type of regression models covered here are generalized linear models, which are typical example of the parameter driven time dependency. Moreover, in contrast to

the usual definition, here we focus on the generalized linear models with error distributions not necessarily belonging to an exponential family. In this thesis we apply these models to public health, focusing on the length of stay and costs prediction.

Working with the real data we usually suspect that it can be contaminated by some proportion of outliers, therefore the focus of this research is also aimed to compute the robust versions for the methods we apply to the mentioned models.

More precisely, it is known that the classical Kalman filter does not perform well in the presence of outliers. Hence, in this work we use robust versions of the Kalman filter and rework classical Kalman smoother and extended Kalman filter in a robust way for different types of outliers. To assess the performance of our constructed procedures, we apply it at real data and stylized outlier situation. Moreover, we compare efficiency of our procedures to other existing approaches.

The connection to extreme value theory is given by a separation approach where we try to extract the dynamics fitting an state-space model to the data and delegate extreme value analysis to the respective residuals, which are then treated as i.i.d.

Talking about robustness for generalized linear models, we are not only interested in consistency results for specific estimators, but rather in local asymptotic normality in the sense of Hájek (1972) and LeCam (1970). Hence, following Rieder (1994), we derive smoothness of the model in terms of L_2 differentiability and aim to create its sufficient conditions for the generalized linear models, covering both cases of stochastic and deterministic regressors. Moreover, we set the time dependence in these models by linking the distribution parameters to the own past observations.

We check suitability of the introduced L_2 differentiability conditions on the models with discrete error distribution (Binomial or Poisson) and then pass over to the generalized Pareto or generalized extreme value continuous error distributions.

Besides, we review robustness properties of some estimators for the generalized Pareto model, proven before, and obtain similar robustness results for the generalized extreme value distribution.

Last, but not least, purpose of this thesis is to give to the reader an idea of the fixed point iteration algorithm for the computation of the optimally robust influence curve. Comparing to other similar algorithms, our version of it uses another techniques to get some intermediate values. We not only discuss it step by step and point out its week stages, but also implement it in `R`. Later, we apply our implementation to the real data set taken from the Jena university hospital ICU.

1.2 Project "Robust risk estimation"

This PhD-thesis is written in the framework of the project "Robust Risk Estimation", based on the cooperation of four different institutions: Fraunhofer ITWM (Institut für Techno- und Wirtschaftsmathematik), Technical University of Kaiserslautern (TU KL), Furtwangen University (HFU) and University of Natural Resources and Life Sciences in Vienna (BOKU).

This project was funded by Volkswagen Foundation within the call "Extreme events: Modeling, Analyses, and Prediction" in the years 2011-2014. The principal researchers of this project were Dr. habil. Peter Ruckdeschel (coordinator), Prof. Dr. Ralf Korn (TU KL and Fraunhofer ITWM), Prof. Dr. Matthias Kohl (HFU) and Dr. Bernhard Spangl (BOKU). They were supported by the post-docs Dr. Nataliya Horbenko (Fraunhofer ITWM and TU KL), and after here leaving to KPMG, Frankfurt, Dr. Gerald Kroisandt (Fraunhofer ITWM) and Dr. Sascha Desmettre (TU KL), as well as by PhD students M.Sc. Daria Pupashenko (HFU and TU KL) and M.Sc. Mykhailo Pupashenko (TU KL).

The main goal of the project was to develop robust procedures for risk management of complex systems in the presence of extreme events, i.e. apply Robust Statistics to Extreme Value Theory.

Project members were divided into three teams regarding to three reference application examples, i.e. Financial Mathematics (financial risks, in particular operational and liquidity risk of a bank), Medicine (unit length of stay and cost in intensive care of a university clinic), and Hydrology (river discharge data of Austrian rivers). In order to cover all these applications, in the meantime we discovered some specific problems in the general approach to be solved. As a benefit of this broad range of applications we could transfer domain-specific methodologies from one pillar to the other one, and provide a common infrastructure for all of them in form of a unified robustness approach and a common `R` infrastructure. The applications themselves are discussed in details below in the Chapter 7.

For each example and its parametric model we aimed to determine optimally-robust estimators minimizing the maximal asymptotic risk on neighborhoods about the ideal model. The main achievement of the project is the invention of the specific robustness approach for this estimation and development of the diagnostic tools to quantify and visualize the influence and outlyingness of data, see Section 2.6.

1.3 Outline

This PhD thesis consists of three main parts. The first part is focused on the basic notions and gives an overview of the existing results on the Robust Statistics and Extreme Value Theory.

As every Chapter of this thesis, Chapter 2 starts from the short description of the previous treatment in literature concerning robustness. Then, we introduce the parametric model, which is the basis for the further research, and give definition for L_2 differentiability notion with the overview of some existing results and simple examples. Further, we discuss tools which capture local and global robustness and introduce most common in use classical and optimally robust estimators. We close Chapter 2 by the model diagnostics discussion, which is the important achievement of our project work. We also give the overview of the software infrastructure so far available in `R`, including our invented package `RobExtremes`, and conclude.

The second Chapter of the first part starts with the general discussion of the existing sources concerning Extreme Value Theory. Then, we present the basic concepts, i.e. extreme value distributions, limit theorems and approaches of fitting extreme value distributions. Next, we discuss smoothness conditions for the models with two types of distributions, generalized extreme value and generalized Pareto distributions. Further, we review global and local robustness properties of some estimators for the parameters of generalized Pareto model. The new result, which we prove in the Appendix of the thesis, concerns the obtaining similar robustness results for the generalized extreme value distribution. It is presented in the end of the Chapter together with the software overview and conclusions.

The second part of the thesis starts with the deeper analysis of the basic models and tools, used to achieve our results. Chapter 4 shows the relation between two main subjects of the research and describes the procedures of interest in the classical form.

Here we pass over to the main part of the thesis which contains our personal results. Chapter 5 presents robust Kalman filter and our robust versions of the Kalman smoother and extended Kalman procedures specialized on the different types of outliers. We also test all procedures in different outlier situations to conclude about their performance in the situations they were created for. Moreover, we compare introduced procedures to one chosen non-parametric filtering method. In the last Section of the Chapter 5 we give overview of the existing `R`-software infrastructure for the classical and robust Kalman procedures.

Another important Chapter of this research, Chapter 6, is focused on the new approach for the L_2 differentiability of the generalized linear models. Here we extend already existing approach on L_2 differentiability for linear regression models to the case of non-exponential scale-shape families, e.g. generalized extreme value and generalized Pareto distributions. Analogically to the previous authors, we treat cases of stochastic and deterministic regressors separately and compute corresponding L_2 differentiability conditions for them. The advantage of our new approach is its applicability to the error distributions with the higher dimensional parameter and case of regressors of possibly different length for each parameter. Later, we also test our methods on the linear regression, Binomial and Poisson generalised linear models and generalized extreme value and generalized Pareto joint shape-scale models. Important step for two last models is to choose the appropriate componentwise link function, what we discuss in details.

The last Section of this Chapter describes fixed point iteration algorithm for the computation of the optimally robust influence curve. We write the algorithm itself and present our exemplary implementation of it in `R`, under the name `FixPglm`. Here we do not aim to provide the most flexible implementation, but rather sketch how it should be done and retain everything that is necessary. We also test our function `FixPglm` on the `R`-data "carrots" for the generalized linear model with Binomial error distribution.

Last part of the thesis is focused on the application examples, already mentioned in the previous Section. Chapter 7 contains hydrological, medical and financial applications concerning river discharge data, unit length of stay and costs and operational and liquidity risk of a bank respectively. Moreover, in the Chapter 8 we apply our `FixPglm` algorithm to the real medical data taken from the Jena university hospital ICU. We aim to conclude about the performance and the speed of our algorithm.

Part I

Foundations

Chapter 2

Robustness

Why are robust methods needed? When we want to describe a set of observations in some statistical modeling problem, we often get information about the data which can be formalized in a number of assumptions. It often happens in practice that this assumptions hold approximately, describing the majority of observations. But some observations follow a different pattern or no pattern at all. Such atypical data are called *outliers*.

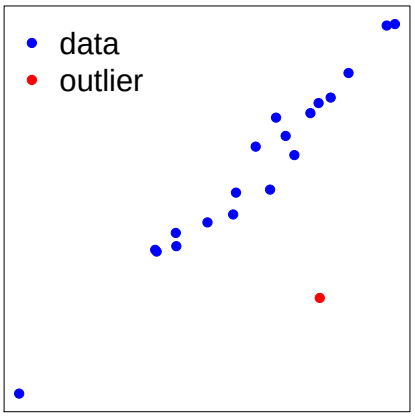

FIGURE 2.1: Data with outlier

Even a single outlier can have a large distorting influence on a classical statistical method. Outliers may also be correct, but they should always be checked for transcription errors. The robust approach to statistical modeling and data analysis aims at deriving methods that produce reliable results not only when the data follow a given assumption exactly, but also approximately, in the presence of outliers. So, if the data contains no outliers, the robust method gives results, which are very close to the results of the classical method. However, if data carries small proportion of outliers, robust method gives approximately same results as the classical method applied to the "typical" data.

Somewhat larger deviations from the model should not cause a catastrophe in the robust procedures.

We note that outliers are not the only deviation from model assumptions against which robustness provides a remedy. More precisely, if done properly it protects against any small deviation in suitable distances of probabilities like Hellinger, total variation, or, ideally, even Prokhorov. This comprises in particular small shifts of the whole distribution. On the other hand outliers are also one focus of diagnostics. We will come to this aspect later on in Chapter 2.6.

The idea of robustness can be traced back at least to the end of the 20th century. Primarily, robustness is associated with the names Tukey J., Huber P. and Hampel F.; see Hampel (1968, 1971, 1974), Huber (1964, 1965, 1981), Tukey (1960, 1962). The first theoretic approach to robust statistics is introduced by Huber (1964). He was working with *neighborhoods* of a stochastic model, earlier introduced by Tukey, and he found estimator that behaves optimally over the whole neighborhood. Later, his basic idea was extended by other approaches. Large contribution to the development of robust methods is made by the fundamental work of Hampel et al. (1986), Maronna et al. (2006), Rieder (1994). For all notions we mention further in this Chapter we refer to these books.

We essentially limit ourselves to the presentation of the theory as far as we will need it in the subsequent sections rather than full overview of the subject. We also refer to the corresponding monographs giving a broader view on the respective notions. For a detailed introduction to robust statistics we mainly refer to Hampel et al. (1986, Ch. 1).

2.1 Model assumptions

2.1.1 Parametric model

A family of distributions, which can be described using a finite number of parameters, is called a *parametric model* or a *parametric family*. More precisely, first we define the measurable space $(\Omega, \mathcal{A})$ with $\mathcal{M}_1(\mathcal{A})$, the set of probabilities on the σ- algebra $\mathcal{A}$. For each parameter ϑ from the open domain $\Theta \subset \mathbb{R}^k$ we denote corresponding distribution as Q_ϑ. Then the following family of distributions

$$\mathcal{Q} = \{Q_\vartheta | \vartheta \in \Theta\} \subset \mathcal{M}_1(\mathcal{A}) \tag{2.1}$$

is a *parametric model* with open parameter domain $\Theta \subset \mathbb{R}^k$.

Next, we need the notion of the absolute continuity, which is the weakest generalisation of the fundamental theory of calculus. Later we use it in order to link the concepts of the derivative and of the integral of a function.

Definition 2.1 (Absolute continuity)**.** Let $I \subset \mathbb{R}$ be some interval. A function $f : I \to \mathbb{R}$ is said to be *absolutely continuous*, if for every $\epsilon > 0$ there exist $\delta > 0$ s.t.

$$\sum_{i=1}^{n} |f(b_i) - f(a_i)| \leq \epsilon,$$

for every finite number of non overlapping intervals (a_i, b_i), $i = 1, ..., n$, with $[a_i, b_i] \subset I$ and

$$\sum_{k=i}^{n} (b_i - a_i) \leq \delta.$$

Remark 2.2. *The notion of absolute continuity can be extended to higher dimensions in the following way. Let $I \subset \mathbb{R}^k$ and the function $f : I \to \mathbb{R}$. We call f* absolutely continuous *function, if for all $a, b \in \mathbb{R}^k$ the function $G\colon [0,1] \to \mathbb{R}$, s.t. $s \mapsto G(s) = f(a + s(b - a))$ is absolutely continuous in the sense of Def. 2.1.*

When the model consists of absolutely continuous distributions, the corresponding density functions can be defined. Following Rieder (1994) and LeCam (1970) we write dQ_ϑ to denote the densities q_ϑ w.r.t. some counting or dominating measure ν on the sigma-algebra $\mathcal{A}$, i.e. $dQ_\vartheta = q_\vartheta d\nu$. Then the parametric model can be alternatively specified in terms of density functions

$$\mathcal{Q} = \{q_\vartheta | \vartheta \in \Theta\}$$

In order to give examples of the parametric model we consider the discrete and continuous cases separately.

Example 2.3 (Binomial parametric model)**.** As an example of a family of discrete probabilities we introduce the Binomial family of distributions for fixed known parameter $n \in \mathbb{N}$. Then, this model is parametrized by success probability $p \in [0, 1]$. The probability mass function, i.e. the density w.r.t. some counting measure, of the Binomial distribution is

$$q_p(y) = \binom{n}{x} p^x (1 - p)^{n-x}, \tag{2.2}$$

for $x \in \{0, \ldots, n\}$. Then, for some open parameter domain $\Theta \subset [0, 1]$ the *Binomial parametric family* is defined as $\mathcal{Q} = \{q_p | p \in \Theta\}$.

Example 2.4 (Generalized extreme value parametric model)**.** As an example of the a continuous distribution family we consider the family of generalized extreme value distributions (GEVD), specified by three parameters $\mu \in \mathbb{R}$, $\sigma > 0$ and $\xi \in \mathbb{R}$. Parameter

μ can be treated as parameter of interest, nuisance parameter or as fixed and known. Then, the model is parametrized by $\vartheta = (\sigma, \xi)$. In this example, for simplicity, we restrict ourselves to the case $\xi \neq 0$ so that the cumulative distribution function of GEVD is defined for $1 + \xi\frac{y-\mu}{\sigma} > 0$ as

$$Q_{\sigma,\xi}(y) = \exp\big(-(1+\xi\frac{y-\mu}{\sigma})^{-\frac{1}{\xi}}\big), \tag{2.3}$$

and the density function (simply the derivative of the distribution function $Q_{\sigma,\xi}(y)$, when it exists) is the following:

$$q_{\sigma,\xi}(y) = \frac{1}{\sigma}(1+\xi\frac{y-\mu}{\sigma})^{-\frac{1}{\xi}-1}\exp\big(-(1+\xi\frac{y-\mu}{\sigma})^{-\frac{1}{\xi}}\big). \tag{2.4}$$

For the further use we introduce here the quantile function (the inverse of the cumulative distribution function $Q_{\sigma,\xi}(y)$) of GEV distribution:

$$F_{\sigma,\xi}(y) = Q^{-1}_{\sigma,\xi}(y) = \mu - \frac{\sigma}{\xi}(1-(-\log y)^{-\xi}). \tag{2.5}$$

In principle, knowing one of these functions gives us two others, so the GEVD itself is defined by any one of the functions (2.3), (2.4) and (2.5).

Then, for some suitable open parameter domain $\Theta \subset \mathbb{R}^+ \times \mathbb{R}\backslash\{0\}$, the corresponding *Generalized extreme value parametric model* is defined as $\mathcal{Q} = \{Q_{\sigma,\xi} | (\sigma, \xi) \in \Theta\}$.

We intentionally chose this distribution as an example of a parametric model, since it will often appear later in this thesis. We discuss this distribution in more detail, with other options for the parameters μ and ξ and some properties, in Section 3.1.1.

Remark 2.5. *We should mention that all our theory only holds for interior points, i.e. we don't allow our parameter sitting on the margin of the parameter space. Otherwise we can face some problems, which are discussed in Pollard et al. (1997, pp.305-314).*

2.1.2 Smoothness

Under *smoothness* we usually understand existence and continuity of the function derivatives up to some desired order over some domain. The smoothness of the parametric model $\mathcal{Q} = \{q_\vartheta | \vartheta \in \Theta\}$ implies the linearization of the density function, i.e.

$$q_{\vartheta+h_n}(y) = q_\vartheta(y) + h_n\frac{\partial}{\partial\vartheta}q_\vartheta(y) + r(y, \vartheta, h_n), \tag{2.6}$$

for $|h_n| \neq 0$, $|h_n| \leq h \in \mathbb{R}$ and the remainder $r(y, \vartheta, h_n)$ s.t.

$$\lim_{n\to\infty} \frac{r(y, \vartheta, h_n)}{|h_n|} = 0, \tag{2.7}$$

as $h_n \to 0, n \to \infty$ in $\mathbb{R}^k$.

Usual "pointwise" approach requires convergence condition (2.7) to hold for all $y \in \mathbb{R}$, which is too much to claim, since the density function $q_\vartheta(y)$ is only $\nu(dy)$-a.e. defined. Therefore density is not even defined for all y. Moreover, we need to be able to interchange derivative and expectation operations, i.e. differentiation and integration, which is not self evident if we use a pointwise approach.

Besides, from the graduate course in analysis we know that pointwise convergence is not enough to conclude on the convergence of the integrals and we need additional information like, e.g. dominated or monotone convergence. Therefore, even if we assume (2.7) to be fulfilled for all $y \in \mathbb{R}$, it is not sufficient to conclude the square integrability of the remaining term, i.e.

$$\frac{\int r^2(y, \vartheta, h_n)\nu(dy)}{|h_n|^2} \to 0, \quad \text{as} \quad h_n \to 0, n \to \infty, \tag{2.8}$$

which is necessary in order to have a Hilbert space structure.

If we express the remainder in terms of the density function from the linearization (2.6) and plug it in the integral convergence (2.8), we get the following:

$$\frac{\int (q_{\vartheta+h_n}(y) - q_\vartheta(y) - h_n \frac{\partial}{\partial\vartheta} q_\vartheta(y))^2 \nu(dy)}{|h_n|^2} \to 0, \quad \text{as} \quad h_n \to 0, n \to \infty.$$

Therefore, to require square integrability of the remaining term (2.8), we need the density function also to be square integrable, and this is too strong condition to claim.

Nevertheless, there is a way to avoid the demand of the density square integrability, presented by LeCam (1970) and Hájek (1972). Since the density function is always positive, idea is to take the square roots of it (see also Rieder (1994, Ch. 2)), i.e.

$$\frac{\int (\sqrt{q_{\vartheta+h_n}(y)} - \sqrt{q_\vartheta(y)} - h_n \frac{\partial}{\partial\vartheta} \sqrt{q_\vartheta(y)})^2 \nu(dy)}{|h_n|^2} \to 0, \quad \text{as} \quad h_n \to 0, n \to \infty.$$

In this way we get to the notion of L_2 differentiability, which we define following Rieder (1994, Def. 2.3.6.). Fix $\vartheta \in \Theta$.

Definition 2.6 (L_2 differentiability)**.** The model $\mathcal{Q}$ is called L_2 *differentiable* at ϑ if there exists a function $\Lambda_\vartheta^{\mathcal{Q}} \in L_2^k(Q_\vartheta)$ such that

$$\left\| \sqrt{dQ_{\vartheta+h}} - \sqrt{dQ_\vartheta}(1 + \frac{1}{2}(\Lambda_\vartheta^{\mathcal{Q}})^{\mathrm{T}} h) \right\|_{L_2^k} = o(|h|), \tag{2.9}$$

as $h \to 0$ and

$$I_\vartheta^{\mathcal{Q}} = \mathbf{E}_\vartheta \Lambda_\vartheta^{\mathcal{Q}} (\Lambda_\vartheta^{\mathcal{Q}})^{\mathrm{T}} > 0. \tag{2.10}$$

Then the function $\Lambda_\vartheta^{\mathcal{Q}}$ is the L_2 *derivative* and the $k \times k$ matrix $I_\vartheta^{\mathcal{Q}}$ is the *Fisher information* of the parametric model $\mathcal{Q}$ at ϑ.

We say that the model $\mathcal{Q}$ is *continuously* L_2 *differentiable* at ϑ if, for any $h \to 0 \in \mathbb{R}^k$,

$$\sup_{t \in \mathbb{R}^k : |t| \le 1} \left\| \sqrt{dQ_{\vartheta+h}} (\Lambda_{\vartheta+h}^{\mathcal{Q}})^{\mathrm{T}} t - \sqrt{dQ_\vartheta} (\Lambda_\vartheta^{\mathcal{Q}})^{\mathrm{T}} t \right\|_{L_2^k} = o(1). \tag{2.11}$$

Remark 2.7. *Condition (2.10) in Rieder (1994, Def. 2.3.6) is required for the local identifiability and we can drop it when we are only interested in smoothness. We claim Fisher information to be a symmetric and positive-semidefinite matrix.*

Remark 2.8. *In Rieder (1994, Thm. 2.3.7) it is proven that if the model $\mathcal{Q}$ is L_2 differentiable at ϑ, then its L_2 derivative $\Lambda_\vartheta^{\mathcal{Q}}$ is uniquely determined in $L_2^k(Q_\vartheta)$, moreover $\boldsymbol{E}_\vartheta \Lambda_\vartheta^{\mathcal{Q}} = 0$.*

As the main criteria for the L_2 differentiability of the parametric model we would require the Hájek (1972, Lem. A.1–A.3) conditions to be fulfilled.

Proposition 2.9 (Hájek)**.** *Assume that in some $\vartheta_0 \in \Theta$ surrounded by some open neighborhood U, model $\mathcal{Q}$ satisfies*

(H.1) The densities $dQ_\vartheta(y)$ are absolutely continuous in each $\vartheta \in U$ for Q_{ϑ_0}-a.e. y.

(H.2) The derivative $\frac{\partial}{\partial \vartheta} dQ_\vartheta(y) = \Lambda_\vartheta(y)\, dQ_\vartheta(y)$ exists in each $\vartheta \in U$ for Q_{ϑ_0}-a.e. y.

(H.3) The Fisher information $I_\vartheta = \int \Lambda_\vartheta(y) \Lambda_\vartheta^T(y) Q_\vartheta(dy)$ exists, (i.e., the integral is finite) and is continuous in ϑ on U.

Then $\mathcal{Q}$ is continuously L_2 differentiable in ϑ_0 with derivative Λ_{ϑ_0} and Fisher information I_{ϑ_0}.

Remark 2.10. *The first condition (H.1) gives us the pointwise smoothness of the distribution and the second (H.2), uniform square integrability of the scores, meaning that the variance is finite. Apparently, these two conditions are implied by the continuous differentiability of the densities $dQ_\vartheta(y)$ w.r.t. ϑ. Condition (H.3) provides continuity*

of the Fisher information of the distribution, so that the variance for close parameter values stays close.

We can note that both parametric models from Examples 2.3 and 2.4 satisfy Hájek conditions (H.1)-(H.3), therefore they are L_2 differentiable. Let us take a closer look at the L_2 derivatives and the Fisher information matrices for these examples.

Example 2.11 (Binomial parametric model)**.** In the Binomial model $\mathcal{Q} = \{q_p | p \in \Theta\}$ let for simplicity $n = 1$. Then the first Hájek condition (H.1) is fulfilled by the probability mass function (2.2), since it is absolutely continuous for each p taken from the open domain $\Theta \subset [0, 1]$.

By condition (H.2), the L_2 derivative is defined as $\Lambda^{\mathcal{Q}}_{\vartheta}(y) = \frac{\partial}{\partial \vartheta}(q_\vartheta(y))/q_\vartheta(y)$, which exists for the Binomial parametric family and for $p \in \Theta$:

$$\Lambda^{\mathcal{Q}}_{p}(y) = \frac{y - p}{p(1 - p)}$$

The last Hájek condition (H.3) requires existence and continuity of the Fisher information, which is also fulfilled by

$$I^{\mathcal{Q}}_{p} = \mathbf{E}_p(\Lambda^{\mathcal{Q}}_{p})^2 = \frac{1}{p(1 - p)} > 0.$$

Example 2.12 (GEV parametric model)**.** For parametric model $\mathcal{Q} = \{Q_{\sigma,\xi} | (\sigma, \xi) \in \Theta\}$ with GEV cumulative distribution function (2.3) and domain $\Theta \subset \mathbb{R}^+ \times \mathbb{R}$, the the L_2 derivative takes up the structure of the parameter, so it consists of two coordinates

$$\Lambda^{\mathcal{Q}}_{\sigma,\xi}(y) = (\Lambda^{\mathcal{Q}}_{\sigma}(y), \Lambda^{\mathcal{Q}}_{\xi}(y))^{\mathrm{T}},$$

where $\Lambda^{\mathcal{Q}}_{\sigma}$ is the L_2 derivative for parameter σ and $\Lambda^{\mathcal{Q}}_{\xi}$ for parameter ξ.

The Fisher information for this model is a 2×2 symmetric matrix of the form:

$$I^{\mathcal{Q}}_{\sigma,\xi} = \mathbf{E}_{\sigma,\xi}\Lambda^{\mathcal{Q}}_{\sigma,\xi}(y)(\Lambda^{\mathcal{Q}}_{\sigma,\xi}(y))^{\mathrm{T}} = \begin{pmatrix} I_{\sigma\sigma} & I_{\sigma\xi} \\ I_{\sigma\xi} & I_{\xi\xi} \end{pmatrix},$$

where $I_{\sigma\sigma} = \mathbf{E}_{\sigma,\xi}(\Lambda^{\mathcal{Q}}_{\sigma}(y))^2$ is information only about parameter σ, $I_{\xi\xi} = \mathbf{E}_{\sigma,\xi}(\Lambda^{\mathcal{Q}}_{\xi}(y))^2$ is the information about parameter ξ and $I_{\sigma\xi} = \mathbf{E}_{\sigma,\xi}(\Lambda^{\mathcal{Q}}_{\sigma}(y)\Lambda^{\mathcal{Q}}_{\xi}(y))$ contains all mixed information. Additionally, from Remark 2.8 one can see that $I_{\sigma\sigma} = \mathrm{Cov}(\Lambda^{\mathcal{Q}}_{\sigma}(y))$, $I_{\xi\xi} = \mathrm{Cov}(\Lambda^{\mathcal{Q}}_{\xi}(y))$ and $I_{\sigma\xi} = \mathrm{Cov}(\Lambda^{\mathcal{Q}}_{\sigma}(y)\Lambda^{\mathcal{Q}}_{\xi}(y))$.

The Fisher information is a positive-semidefinite matrix, therefore by Sylvester's criterion, $I_{\sigma\sigma} \geq 0$ and the determinant $\det(I^{\mathcal{Q}}_{\sigma,\xi}) \geq 0$.

Remark 2.13. *The computation of the Fisher information terms here is not easy, because it involves integrating w.r.t. large support and slow decay density, that is why we use the fact that the expectation of a random variable is the integral of its quantile, i.e.*

$$I_{\sigma\sigma} = \boldsymbol{E}_{\sigma,\xi}(\Lambda_\sigma^{\mathcal{Q}})^2 = \int (\Lambda_\sigma^{\mathcal{Q}}(y))^2 dQ_{\sigma,\xi}(y) = \int_0^1 (\Lambda_\sigma^{\mathcal{Q}}(F_{\sigma,\xi}(y)))^2 dy,$$

for quantile function of GEV distribution $F_{\sigma,\xi}$ from (2.5).

2.2 Neighborhoods

One way of quantifying the distance between measures in mathematics is by metrics. For further use we introduce here four of them.

The *Hellinger distance*, defined in terms of the Hellinger integral (see Nikulin (2001)), can be defined by the expression:

$$d_h^2(Q, Q_\vartheta) = \frac{1}{2} \int \left| \sqrt{dQ} - \sqrt{dQ_\vartheta} \right|^2. \tag{2.12}$$

The *total variation distance* is the largest possible difference between the probability distribution functions assigned to the same event, i.e.

$$d_v(Q, Q_\vartheta) = \frac{1}{2} \int |dQ - dQ_\vartheta| = \sup_{A \in \mathcal{A}} |Q(A) - Q_\vartheta(A)|. \tag{2.13}$$

The *Kolmogorov distance* is the supremum of the absolute difference between the distribution functions:

$$d_k(Q, Q_\vartheta) = \sup_{y \in \mathbb{R}^k} |Q(y) - Q_\vartheta(y)|. \tag{2.14}$$

The square of the *Cramér-von-Mises distance* is given as the integral of the squared difference between the distribution functions, i.e.

$$d_m^2(Q, Q_\vartheta) = \int (Q(y) - Q_\vartheta(y))^2 \nu(dy). \tag{2.15}$$

The aim of robust methods is, roughly speaking, to develop estimates which have a "good" behavior in a "neighborhood" of the ideal distribution. We define

$$U_*(\vartheta, \epsilon) = \{Q | d_*(Q_\vartheta, Q) \leq \epsilon, \epsilon \in [0, \infty)\} \subset \mathcal{M}_1(\mathcal{A})$$

as the *neighborhoods* about distribution Q_ϑ of radius ϵ generated by some distance measure d_*. In robust statistics the basic types of such neighborhoods are: *Hellinger*

($* = h$), *total variation* ($* = v$), *Kolmogorov* ($* = k$) and *Cramér-von-Mises* ($* = m$), which are based on the metrics (2.12), (2.13), (2.14) and (2.15) correspondingly.

From other side, assume that a proportion $1 - \epsilon$ of the observations is generated by the (true, ideal) distribution Q_ϑ, while a proportion ϵ is generated by an unknown mechanism. Such real data set can be modeled by the well-known *Gross Error Model* (convex contamination) of the form:

$$Q = (1 - \epsilon)Q_\vartheta + \epsilon G.$$

Here the radius $\epsilon \in [0, 1)$ is the amount of gross errors (contamination) and G is an unknown, uncontrollable, unpredictable outlier generating distribution.

The Gross Error Model defines one of the most practicable types of neighborhoods, which is able to capture deviations of distributions or outlier phenomena. This neighborhood is called *contamination* ($* = c$) neighborhood and it consists of convex combinations, i.e.

$$U_c(\vartheta, \epsilon) = \{Q | Q = (1 - \epsilon)Q_\vartheta + \epsilon G, G \in \mathcal{G}\},$$

where $\mathcal{G}$ is a suitable set of distributions (often the set of all distributions).

Remark 2.14. *To balance variance and bias in the situation of n observations it turns out useful to scale the radius by the sample size n i.e.*

$$\epsilon = \frac{r}{\sqrt{n}},$$

for some radius $r \in [0, \infty)$ (compare Rieder (1994)). This shrinking can also be motivated by detectability of outliers, compare Rieder (2006).

In one-dimensional case, for esch neighborhood defined above, i.e. $U_*(\vartheta, \epsilon)$ where $* = h, v, k, \mu$, there is an explicit expression for the bias term (see Rieder (1994, Prop. 5.3.3)), but there is no explicit solution in multivariate case, whereas for the contamination neighborhood $U_c(\vartheta, \epsilon)$ we get explicit expressions for the bias term for all dimensions. Therefore further in this thesis we focus only on contamination neighborhoods.

Huber (1981) proposed two interpretations for the contamination of the sample. Either we let large changes in a few observations or small changes in all of them. Similar to Ruckdeschel et al. (2014b), here and further, we denote the *ideal* model assumptions by the suffix "id", the *distorting* (contaminating) situation by "di" and the suffix "re" indicates the *realistic* contaminated situation. Then, we either replace few observations of the sample, i.e.

$$X^{\text{re}} = (1 - \epsilon)X^{\text{id}} + \epsilon X^{\text{di}}, \tag{2.16}$$

so that X^{di} is generated by the contaminated distribution $Q \in U_c(\vartheta, \epsilon)$, or we let all observations X^{re} be a bit distorted, i.e. they all have distribution $Q \in U_c(\vartheta, \epsilon)$. In this thesis we focus on the first interpretation, since it is more convenient for our purposes.

2.3 Measuring Robustness

In robust statistics one is interested in estimators which have certain stability w.r.t. the contamination of the ideal model. We distinguish between global and local robustness of an estimator. *Local robustness* asks how small deviations, in extreme cases a single outlier, influence the value of the estimator. *Global robustness* of the estimator describes the behavior of the estimator under massive distortions.

In this Section we follow the notations of Maronna et al. (2006, Ch. 3) and Hampel et al. (1986, Ch. 2). For the parametric model $\{Q_\vartheta | \vartheta \in \Theta\}$, with open parameter domain Θ, consider the estimator $\hat{\boldsymbol{\vartheta}}_n(Q_\vartheta)$ of the parameter ϑ, depending on a sample $X = \{X_1, ..., X_n\}$ of i.i.d. random variables with distribution Q_ϑ. In order to formalize next notions we study the behavior of the estimates when the sample size tends to infinity ("asymptotic behavior"). We define the *asymptotic value* of the estimate as $\hat{\boldsymbol{\vartheta}}_\infty(Q_\vartheta)$, s.t. $\hat{\boldsymbol{\vartheta}}_n(Q_\vartheta) \to^p \hat{\boldsymbol{\vartheta}}_\infty(Q_\vartheta)$ as $n \to \infty$. Typical examples are the mean of the distribution $\hat{\boldsymbol{\vartheta}}_\infty(Q_\vartheta) = \mathbf{E}_{Q_\vartheta} X$ for the sample mean $\hat{\boldsymbol{\vartheta}}_n = \bar{X}$ or the distribution median $\hat{\boldsymbol{\vartheta}}_\infty(Q_\vartheta) = Q_\vartheta^{-1}(0.5)$ for the sample median $\hat{\boldsymbol{\vartheta}}_n = \text{Med}(X)$.

Further we are interested in the behavior of the asymptotic estimate $\hat{\boldsymbol{\vartheta}}_\infty(Q)$ when Q ranges over the contamination neighborhood $U_c(\vartheta, \epsilon)$.

2.3.1 Local robustness. Influence function

The local robustness of an estimator may be captured by the *influence function* (IF). By Hampel (1974), IF is an approximation to the behavior of $\hat{\boldsymbol{\vartheta}}_\infty(Q_\vartheta)$ when the sample contains a small fraction ϵ of identical outliers. It measures the dependency of the estimator on the value of one of the points in the sample. An estimator is considered locally robust if its IF is bounded, because then we are able to ensure that small deviations from the model distribution do not cause large changes in the estimate. The classical definition of IF is taken from Huber (1981).

Definition 2.15 (Influence function)**.** The *influence function* is the functional derivative of the estimator with respect to the distribution. It is defined as the Gâteaux derivative

in the direction of Dirac measure (point-mass) δ_x in x:

$$\psi_\vartheta(x) := IF(x; \hat{\boldsymbol{\vartheta}}_n, Q_\vartheta) = \lim_{\epsilon \downarrow 0} \frac{(\hat{\boldsymbol{\vartheta}}_\infty((1-\epsilon)Q_\vartheta + \epsilon\delta_x) - \hat{\boldsymbol{\vartheta}}_\infty(Q_\vartheta))}{\epsilon},$$

provided the limit exists. "$\downarrow$" stands for "limit from the right".

Remark 2.16. *If we have a higher dimensional parameter ϑ, then $\hat{\boldsymbol{\vartheta}}_\infty(Q_\vartheta)$ is a p-dimensional vector and so is its IF.*

Definition 2.17 (Asymptotically linear estimator)**.** Assume that the estimator $\hat{\boldsymbol{\vartheta}}_n$ has an expansion in the sample $X = \{X_1, ..., X_n\}$, i.e.

$$\hat{\boldsymbol{\vartheta}}_n = \vartheta_n^0 + \frac{1}{n}\sum_{i=1}^{n} \psi_\vartheta(X_i) + R_n,$$

where $\sqrt{n}\,|R_n| \to \infty$, as $n \to \infty$, for the starting estimator ϑ_n^0 and ψ_ϑ being IF of $\hat{\boldsymbol{\vartheta}}_n$, for which we require the following to hold (see Rieder (1994, Lemma 4.2.18)):

$$\mathbf{E}\psi_\vartheta = 0, \quad \mathbf{E}\psi_\vartheta \Lambda_\vartheta^{\mathrm{T}} = \mathbb{I}_k.$$

Then, such an estimator $\hat{\boldsymbol{\vartheta}}_n$ is called *asymptotically linear* (ALE).

The *asymptotic (co-)variance matrix* of the asymptotically linear estimator $\hat{\boldsymbol{\vartheta}}_n$ is then determined as the following matrix (see Rieder (1994, Rem. 4.2.17)):

$$\mathrm{asVar}(\hat{\boldsymbol{\vartheta}}_n) = \int \psi_\vartheta \psi_\vartheta^{\mathrm{T}} dQ_\vartheta$$

The *gross error sensitivity* (GES) for asymptotically linear estimator $\hat{\boldsymbol{\vartheta}}_n$ is defined in Hampel et al. (1986, Ch. 2.1) as

$$\gamma(\hat{\boldsymbol{\vartheta}}_n) := \sup_x |\psi_\vartheta(x)|\,.$$

Then, the estimator is locally robust iff its GES is finite.

2.3.2 Global robustness

Maximal asymptotic bias

Here we introduce a global robustness measure called *maximal asymptotic bias* (maxbias), which is the most complete and accurate measure of robustness for a point estimate.

The maxbias is originally defined by Huber (1964) and later developed and applied to other statistical models.

An estimator of the parameter of a parametric family F_ϑ is called *consistent* if its asymptotic value fits the true value of the parameter, i.e. $\hat{\boldsymbol{\vartheta}}_\infty(Q_\vartheta) = \vartheta$.

For most distributions Q in the ϵ-neighborhood $U_c(\vartheta, \epsilon)$ of the parametric distribution Q_ϑ the Fisher consistency of the model parameters $\vartheta \in \Theta$ does not hold, i.e. $\hat{\boldsymbol{\vartheta}}_\infty(Q) \neq \vartheta$ and this parameter estimate is subject to the *asymptotic bias*

$$b_{\hat{\boldsymbol{\vartheta}}_n}(Q, \vartheta) = \hat{\boldsymbol{\vartheta}}_\infty(Q) - \vartheta.$$

Definition 2.18 (Maximal asymptotic bias)**.** The overall bias performance in the neighborhood $U_c(\vartheta, \epsilon)$ can then be measured by the *maximal asymptotic bias* defined as

$$B_{\hat{\boldsymbol{\vartheta}}_n}(\epsilon, \vartheta) = \max\left\{|b_{\hat{\boldsymbol{\vartheta}}_n}(Q, \vartheta)| : Q \in U_c(\vartheta, \epsilon)\right\}.$$

Remark 2.19. *In the shrinking neighborhood setup (see Rieder (1994, Lemma 5.3.3)), the $\sqrt{n}$-standardized, maximal asymptotic bias of an asymptotically linear estimator $\hat{\boldsymbol{\vartheta}}_n$ in the gross error model with the radius given by $\epsilon = r/\sqrt{n}$, is the following*

$$B_{\hat{\boldsymbol{\vartheta}}_n}(r, \vartheta) = r\gamma(\hat{\boldsymbol{\vartheta}}_n).$$

Breakdown point

The global robustness of an estimator may be quantified by the *Breakdown Point* (BP). The BP is the largest amount of contamination that the data may contain, s.t. the estimator $\hat{\boldsymbol{\vartheta}}_n$ still gives some information about ϑ.

Definition 2.20 (Breakdown Point)**.** The *breakdown point* (BP) of the estimate $\hat{\boldsymbol{\vartheta}}_n$ at Q is the maximal radius ϵ^* s.t. the maxbias is finite, i.e.

$$\epsilon^* = \sup\left\{\epsilon : B_{\hat{\boldsymbol{\vartheta}}_n}(\epsilon, \vartheta) < \infty\right\}.$$

It is obvious that for "reasonable" estimates there must be more "typical" points than outliers, so $\epsilon^* \leq 1/2$.

Remark 2.21. *If $B_{\hat{\boldsymbol{\vartheta}}_n}(\epsilon, \vartheta)$ is differentiable at $\epsilon = 0$, the corresponding derivative is the gross error sensitivity of the estimator $\hat{\boldsymbol{\vartheta}}_n$, i.e.*

$$\gamma(\hat{\boldsymbol{\vartheta}}_n) = \frac{\partial}{\partial\epsilon}(B_{\hat{\boldsymbol{\vartheta}}_n}(\epsilon, \vartheta))_{\epsilon=0}.$$

2.4 Estimators

In this Section we present the most common classical estimators and discuss their main properties. Well-known behavioural properties of the estimators are (weak, strong) consistency, asymptotic normality with the corresponding asymptotic variance, efficiency, i.e. estimators with smallest possible asymptotic variance in the class of all asymptotically normal estimators, and robustness. Here we focus more on the robust properties of the estimators, which can be described by the breakdown point or the influence function as we mentioned in the Section 2.3. Later in Sections 3.3 and 3.4 we compute some of these estimates for some certain distributions and discuss in detail their robust properties.

2.4.1 Maximum likelihood estimator

One of the most useful methods of estimating parameters of a statistical model is the maximum-likelihood estimation. To give its formal definition, suppose we have i.i.d. random sample $\{X_1, ..., X_n\}$ with an unknown parametric distribution function Q_ϑ and density function $q_\vartheta(x)$ associated with an unknown parameter value ϑ. We define *average log-likelihood* by considering observed values $X_1, ..., X_n$ to be fixed, i.e.

$$l_n(\vartheta|X_1, ..., X_n) = \frac{1}{n}\sum_{i=1}^{n} \log q_\vartheta(X_i)$$

Then the *maximum likelihood estimate* (MLE) is an estimate of the parameter obtained by maximizing the average log-likelihood function, i.e.

$$\hat{\boldsymbol{\vartheta}}^{\mathrm{MLE}} = \arg\max\{l_n(\vartheta|X_1, ..., X_n), \vartheta \in \Theta\}.$$

Remark 2.22. *The maximum likelihood estimator is often used in the classical setup due to its properties - consistency, asymptotic normality and efficiency, i.e. it is an estimator with the minimal asymptotic variance. However, for robust statistics this estimator is too sensitive to contaminations.*

2.4.2 M-estimators

A generalization of the maximum likelihood estimator is given by so-called maximum likelihood type estimator or *M-estimator* (see Huber (1964, 1981)). These estimators use some function ρ instead of the likelihood function for optimization. The M-estimate

of the unknown parameter ϑ is defined as the solution of the following minimization problem:

$$\hat{\boldsymbol{\vartheta}}^{\mathrm{ME}} = \vartheta : \quad \sum_{i=1}^{n} \rho(\vartheta, x_i) = \min!, \tag{2.17}$$

where ρ is an arbitrary function.

Note, that the choice $\rho(\vartheta, x) = -\log q_\vartheta(x)$ gives the ordinary MLE.

Remark 2.23. *This is one of the two possible approaches for these estimators. If the function ρ is differentiable, by denoting $\psi(\vartheta, x) = \partial/\partial\vartheta\rho(\vartheta, x)$, the problem (2.17) can be described by the implicit equation*

$$\sum_{i=1}^{n} \psi(\vartheta, x_i) = 0. \tag{2.18}$$

The second approach is to search for the zero, as in equation (2.18). Therefore, if the maximizing value of equation (2.17) is obtained as a zero, the resulting estimator is called Z-estimators *(see van der Vaart (1998)).*

Remark 2.24. *M-estimators may be more computationally efficient and more robust (resistant to deviations from the assumptions) than MLE. Moreover, due to their generality, the high breakdown point, and efficiency, M-estimators now appear to dominate among other approaches to robust estimation. More details on this can be found in Huber (1981) and van der Vaart (1998).*

2.4.3 Minimum distance estimators

The *minimum distance estimation* (MDE) is the method to obtain an estimator of a parameter by minimizing some distance between the empirical distribution function $\hat{\mathbf{Q}}_n$, defined for the sample $\{X_1, ..., X_n\}$ as $\hat{\mathbf{Q}}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(X_i,\infty)}$ and the theoretical parametric distribution function Q_ϑ. The MD-estimator is defined as

$$\hat{\boldsymbol{\vartheta}}^{\mathrm{MDE}} = \arg\min\{d_*(\hat{\mathbf{Q}}_n, Q_\vartheta), \vartheta \in \Theta\},$$

for Hellinger (2.12), total variation (2.13), Kolmogorov (2.14) and Cramér-von-Mises (2.15) distance measures, so $* = h, v, k, m$ correspondingly.

Remark 2.25. *In the paper of Wolfowitz (1953-1954) it is shown on several problems of different levels of difficulty that the MDE method compared to the MLE obtains consistent estimators rather cheaply. Moreover, for some distances, e.g. Cramér-von-Mises, the minimum distance estimator is asymptotically normal and efficient for a variety of*

models. This class of estimators is also sensitive to outliers. Robust properties of the minimum distance estimations are discussed in details in Rieder (1994, Ch. 6).

Remark 2.26. *The maximum likelihood estimator, the M-estimator and the minimum distance estimator are each minimizing a certain criterion. In a multiparameter setting this relies on finding the global minimum which computationally is not a trivial task in general. In extant software, most algorithms to this end require a decent starting / initial value for the parameter to start their (local) search. This starting estimator of course then affects the found minimizer as well in general. This is a problem for all three estimators introduced above.*

Looking at the code in the package `ismev` *(based on Coles (2001)) we notice, that what they do is even more dangerous. For MLE they start with a fixed value* $\xi = 0.1$*. It means that, because it is a local algorithm, it risks to get stuck somewhere and if* ξ *is far from* 0.1 *it risks to never get there.*

A step forward in this context is not just to consider one fixed value, but consider a sort of representative grid. So at every starting point we risk to get stuck somewhere, but if we do it in the whole range, then we know at least that we looked already over all of them.

2.4.4 k-step estimators

One of the alternatives to the M-estimator is so-called *k-step estimator*. For the finite IF ψ_ϑ and for some suitably chosen *starting estimator* ϑ^0 the k-step estimation for $k = 1, 2, 3, ...$ is defined recursively as

$$\vartheta^k = \vartheta^{k-1} + \frac{1}{n}\sum_{i=1}^{n} \psi_{\vartheta^{k-1}}(x_i). \tag{2.19}$$

We fix value of k and iterate this expression k times.

Remark 2.27. *Iterating (2.19) until convergence amounts to a fixed point iteration, which converges to some point quite quickly, and when it converges we get, that equation (2.19) turns to*

$$\frac{1}{n}\sum_{i=1}^{n} \psi_{\vartheta^{k-1}}(x_i) = 0,$$

starting from some natural number k, and then computing k-step estimator transforms to the zero problem, similarly to the Z-estimator (2.18).

Remark 2.28. *One should note, that for* $k = 1$ *we already achieve the desired expansion. Moreover, for values* $k \leq 2$ *explicit terms for the higher order asymptotic correction*

terms, for the maximal MSE on the neighborhoods, can be obtained in special case, see Ruckdeschel (2010a), Ruckdeschel (2010b).

However, there are previous researches, which all deal with the situation of the ideal model. Ibragimov and Linnik (1971) improve the asymptotics underlying CLT, then Pfanzagl (1985) tries to optimize higher order behavior and Field and Ronchetti (1990) uses saddle point approximation. Ibragimov and Linnik (1971) stops on the third iteration for convenience, i.e. terms get too complicated for $k \geq 2$. Pfanzagl (1985) and Field and Ronchetti (1990) stop at the second order with the fundamental reason, i.e due to the ill-posedness of the optimal problem for the second order. Taniguchi and Kakizawa (2002) cover third order a bit more generally.

For $k = 2$ we can specify the asymptotics up to $o_{P(n^{-3/2})}$. Fixing k ex ante is crucial for the preservation of the breakdown point (it can degenerate in an unbounded number of correction steps). Due to the lack of equivariance, each step requires a new computation of the Lagrange multipliers and hence is computationally demanding. Therefore, we restrict ourselves to $k = 1, 2$.

Remark 2.29. *The k-step (one-step) estimator inherits the breakdown property of the starting estimator and yields a high efficiency.*

2.4.5 Moment based estimators

The *method of moments estimation* is based solely on the law of large numbers (LLN). The idea of the method is to match the sample moments with the corresponding distribution moments. In some situations the solution can be found explicitly, even if in the same situation the MLE requires numerical solvers; in these situations this is then a benefit of the procedure.

For the i.i.d. random sample $X = \{X_1, ..., X_n\}$ with the distribution associated to an unknown parameter value $\vartheta \in \mathbb{R}^p$, the j–th moment about 0 is a function of ϑ, denoted by $\mu_j(\vartheta) = \mathbf{E}_\vartheta X^j$, for $j = 1, ..., p$. The j–th sample moment about 0 is defined as $m_j(X) = \frac{1}{n}\sum_{i=1}^n X_i^j$. Then the method of moments estimator (MME) is solution to the system of equations

$$\mu_j(\hat{\boldsymbol{\vartheta}}^{\mathrm{MME}}) = m_j(X).$$

We often need only the first two moments for this method, i.e. the theoretical and the empirical means and variances.

Remark 2.30. *Under reasonable conditions the moment based estimators in the ideal model are asymptotically normal. More details about these estimators one can find in van der Vaart (1998).*

The generalization of MME is called the *generalized method of moments* estimation (GMM) and it focuses on the solving specific minimization problems based on the "moment conditions". Here we omit the detailed definition of the method, which can be found in Hansen (1982).

2.4.6 Quantile based estimators

Instead of moments matching, as it is for the method of moments estimation, we can also match empirical quantiles of the considered distribution. It leads to solving the system of equations and obtaining parameter estimates in terms of the quantiles, what makes the method easy to use. These are the so-called *quantile based estimators* and one well known example of such an estimator is the *Pickands estimator*, first proposed by Pickands (1975), which is based on the empirical 50% and 75% quantiles.

Remark 2.31. *Pickands estimator is consistent and asymptotically normal as one can see from Embrechts et al. (1997).*

2.4.7 Examples in R

Here we give some simple examples for all introduced estimators. For the maximum Likelihood (MLE) and minimum distance (MDE) estimators we upload R-data "carrots" from the package `robustbase` as follows:

```
> require(robustbase)
> data(carrots)
> data0 <- as.vector(do.call(rbind, carrots))
```

The functions for these estimators are implemented in the R-package `distrMod`, so we require this package

```
> require(distrMod)
```

and having needed functions available, we estimate the success probability of the Binomial distribution with the number of trials equal to the data size. After the estimation is done, by calling the result we get the following output:

```
> MLE <- MLEstimator(data0, BinomFamily(size = 96))
> MLE
Evaluations of Maximum likelihood estimate:
```

```
----------------------------------------
Object of class "Estimate"
generated by call
  MLEstimator(x = data, ParamFamily = BinomFamily(size = 96))
samplesize:   96
estimate:

  0.12657986
 (0.00346356)
fixed part of the parameter:
size
  96
asymptotic (co)variance (multiplied with samplesize):
[1] 0.00115164
Criterion:
negative log-likelihood
                    Inf
```

Next we apply Cramér-von-Mises minimum distance estimator to the same data set and get:

```
> MDE <- MDEstimator(data0, BinomFamily(size = 96), distance = CvMDist)
> MDE
Evaluations of Minimum CvM distance estimate:
----------------------------------------
Object of class "Estimate"
generated by call
  MDEstimator(x = data, ParamFamily = BinomFamily(size = 96), distance = CvMDist)
samplesize:   96
estimate:
      prob
0.03713059
fixed part of the parameter:
size
  96
Criterion:
CvM distance
   0.1541651
```

For the k-step estimator we require the `R`-package `RobAStBase` and compute the classical optimal influence function

```
> require(RobAStBase)
> IC <- optIC(model=BinomFamily(size = 96, prob = 0.5),risk=asCov())
```

Then, using the function `kStepEstimator` we get the following output:

```
> kStep <- kStepEstimator(data0, IC, start = 0.5)
> kStep
Evaluations of 1-step estimate:
-------------------------------
Object of class "Estimate"
generated by call
  kStepEstimator(x = data, IC = IC, start = 0.5)
samplesize:   96
estimate:
      prob
  0.126579861
 (0.005208333)
fixed part of the parameter:
size
  96
asymptotic (co)variance (multiplied with samplesize):
[1] 0.002604167
Infos:
     method
[1,] "kStepEstimator"
[2,] "kStepEstimator"
     message
[1,] "1-step estimate for Binomial family"
[2,] "computation of IC, trafo, asvar and asbias via useLast = FALSE"
asymptotic bias:
NULL
(partial) influence curve:
An object of class "IC"
### name:  Classical optimal influence curve for Binomial family
### L2-differentiable parametric family:  Binomial family

### 'Curve': An object of class "EuclRandVarList"
Domain: Real Space with dimension 1
[[1]]
length of Map:  1
Range: Real Space with dimension 1

### Infos:
     method  message
[1,] "optIC" "optimal IC in sense of Cramer-Rao bound"
steps:
[1] 1
```

For the moment-base estimators we can use the functions from the R-package gmm, where the generalized method of moments is implemented. As an example we estimate a simple linear model in the following way:

```
> require(gmm)
> N <- 1000
> u <- rnorm(N)
> x <- 1 + rnorm(N)
> y <- 1 + x + u
> GMME <- gmm(y ~ x, x)
> GMME
Method
 twoStep

Objective function value:  8.255603e-29

(Intercept)            x
    0.99239      1.01498
```

The last estimator to show is the Pickands estimator from the R-package RobExtremes. By construction it can be applied only to extreme value distributions, which is introduced in detail in Section 3.1.1. Here we take the generalized Pareto distribution with some chosen location, scale and shape parameters:

```
> N <- 1000 # total sample size
> alpha <- 0.05 # percentage of outliers in the data
> N1 <- floor((1-alpha) * N)
> N2 <- N - N1
>
> GP <- GPareto(loc=100,scale=1000,shape=0.4) #ideal distr
> GPfam <-  GParetoFamily(loc=100,scale=1000,shape=0.4)
>
> GP1 <- GPareto(loc=1000,scale=10000,shape=1.4) #contamination distr
>
> GP3 <- r(GP)
> GP4 <- r(GP1)
> data1 <- GP3(N1)
> data2 <- GP4(N2)
> data0 <- c(data1, data2)
```

Then, by applying the function, we obtain the corresponding estimator:

```
> PickandsE <- PickandsEstimator(data0, ParamFamily = GPfam)
> PickandsE
Evaluations of PickandsEstimator:
---------------------------------
Object of class "Estimate"
generated by call
  PickandsEstimator(x = data0, ParamFamily = GPfam)
samplesize:   1000
estimate:
      scale          shape
  1139.6189885      0.3636657
 ( 105.8763514) (   0.1172541)
asymptotic (co)variance (multiplied with samplesize):
            scale        shape
scale 11209801.79 -10537.13674
shape   -10537.14     13.74853
Infos:
     method              message
[1,] "PickandsEstimator" ""
```

2.5 Optimally robust estimators

We discussed a few classical estimation methods, which are sensitive to outliers. In addition, we would like to give a brief overview of the optimally robust estimators.

In Section 2.3.1 we mentioned that desirable robustness property for an estimator is that it has a bounded IF. By Hampel et al. (1986) such an estimator is also called *B-robust* (bias-robust). In the paper Dupuis and Field (1998) authors construct the *Optimal B-robust Estimator* (OBRE), which is originally named *Most Bias-robust Estimator* (MBRE) and also defined in Hampel et al. (1986).

The *Most Bias-robust Estimator* minimizes the maximal bias on convex contamination neighborhoods $U_c(\vartheta, \epsilon)$ of the underlying distribution Q_ϑ. Unlike Dupuis and Field (1998) we note, that MBRE and OBRE are not the same. In the case when the law of the scores is continuous, MBRE can also be obtained as a limit within the class of OBRE estimators, provided that the bound of the bias converges to the minimum (minimax bias).

The estimator with the lowest mean square error (MSE) in the asymptotic distribution neighborhoods is called the *Optimal Mean Squared Error Estimator* (OMSE). It is very similar to the previous optimally robust estimators. Moreover, following Rieder (1994),

or Ruckdeschel and Horbenko (2013) in the context of GPD, it is a special case of the OBRE.

If the radius r, which expresses the proportion of the outliers (see Remark 2.14), is not (precisely) known, it can be computed by the minimax-principle introduced by Rieder et al. (2008). Proposed method is called *Radius minimax* (RMX) and it consists of the following steps. First, for starting radius r_0, possibly miss-specified, we determine the relative efficiency of the optimal solution with radius r_0 to the solution for the true radius s. Then, varying, we calculate the "least favourable" radius s_0 for r_0, which corresponds to the minimal relative efficiency. Next, varying the radius r_0 we choose the value, where the efficiency is maximal in r_0 and minimal in s_0. It turns out, that the estimator, which corresponds to this chosen radius is again optimally robust and it is called *Radius minimax estimator* (RMXE).

All these estimators are defined through their optimal influence functions; we realize these estimators as k-step estimators, defined in Section 2.4.4. More about optimality of these estimators can be found in the PhD thesis of Horbenko (2011, Ch. 6.6).

2.6 Diagnostic

After fitting the model to some data it is important to determine whether all the necessary model assumptions are valid. If there are any violations, one can make the wrong conclusions. Therefore, it is crucial to perform appropriate model diagnostics.

As we have mentioned before, the aim of robustness is protection against outliers. It is easy to conclude the presence of outliers if by applying some procedure, we get a complete break down. But we face problems when the procedure is affected by some proportion of small deviations, which cannot be detected surely, and it is still works. One approach to deal with such outliers is to use diagnostics. While detecting of such small deviations is a main purpose of diagnostics, robust statistics offers a more differentiated view of the data, hence by means of robustness we also can enhance diagnostics. The research described further can be found later in the paper of Ruckdeschel et al. (2014a). All functions for diagnostic plots are implemented in the framework of the `R`-package `RobExtremes`.

2.6.1 General Principles

The concept for diagnostic plots in the `distr` and `RobASt` families of `R`-packages follows some basic principles. The first one to mention is the *flexibility,* i.e., user should be

given the full possibility to customize the respective plots.

Other plot functions in the `R`-package offer a very high level of flexibility. There are good facilities for setting colors, line types and widths, plot characters, fonts, sizes and other features. Further capabilities for plotting large amount of data are offered, e.g. alpha transparency. These additional features significantly improve corresponding plots.

The main principle of setting up the graphical diagnostics, is to pass-through all these features in the user interfaces so that the user can profit from this built in flexibility. This distinguishes our approach from several others, where only default settings are available.

A very important feature is *rescaling*. While working with data it is possible that very large "input" can have a great impact on the procedure. In particular, for the GEVD and GPD, unboundedness of the IF for the MLE is visible only for very large values. Then, plotting influence curves we want to display not only some specific observations, but the whole curve $x \mapsto \psi_\vartheta(x)$. Therefore, in order to make these vulnerabilities visible, by a suitable rescaling of the axes, we can plot unlimited observation and parameter regions. Moreover, user is given guidance how to rescale the axis using his own transformations.

When it comes to regularly repeated diagnostics, it is often helpful to have some text comments. Therefore a couple of *automatic text* templates are also offered to the user. It is very useful to have the information about the creation date, name of the author, maybe some model details, the respective class of the input and other comments on the plot, to be able to distinguish it from all attempts, analyze and compare it with others.

What is the most important from our opinion, is to provide user with easy-to-use *wrapper functions* with a restricted flexibility, which is the call to the full-fledged functions in order to make the first working experience for the user much easier. Wrapper functions take most of arguments of the original function by default. Only the key parameters, necessary for producing the plots, have to be given. However, the user has the opportunity to enter the necessary additional parameters. Beside that, wrapper functions improve the original functions in handling the parameters. Example of the wrapper function can be found later in Section 2.6.4.

2.6.2 Diagnostic plots

To show how diagnostics work, we choose a parametric model, which is required to be smooth, because IF requires smoothness. We take the scale-shape generalized Pareto (GPD) parametric model, i.e. the location parameter μ is given, whereas scale and shape parameters σ and ξ are unknown. Since we use this model only for the diagnostics

example, we prefer to omit the detailed description of the generalized Pareto parametric model here, but we give its full overview and properties in Section 3.1.

In order to avoid legal problems with original data used in the paper of Ruckdeschel et al. (2014a), we illustrate all diagnostic procedures on synthetic data. As the ideal model we take GPD$(100, 1000, 0.4)$. Then, we generate data with 1000 observations and take around 5% of it from the GPD with all three parameters significantly increased, i.e. GPD$(1000, 10000, 1.4)$. Then, we apply Maximum likelihood (MLE) and robust Radius-minimax (RMXE) estimators, defined in Sections 2.4.1 and 2.5 correspondingly. We construct diagnostic plots for both estimators and compare their results, concluding about the performence of both estimators.

Influence Curve Plot

We start with the diagnostic plots for the influence curves (see Section 2.3.1 for definition). For this plot we use the mapping $x \mapsto \psi_\vartheta(x)$. Recalling here the definition of the asymptotically linear estimator (Def. 2.17),

$$\hat{\boldsymbol{\vartheta}}_n = \vartheta_n^0 + \frac{1}{n}\sum_{i=1}^{n} \psi_\vartheta(X_i) + R_n,$$

it becomes clear that, by plotting influence curves, we can check the local influence of the data on the estimated parameters of the model. Moreover, one can conclude which way (up- or downwards) and how much each observation can shift the respective parameter. For example the value of 2 in the ordinate in the first plot on Figure 2.2 indicates that the respective parameter in a first order approximation is shifted upwards by $2/n$ by a corresponding observation made at the respective abscissa.

From these graphs we obtain two types of information. First, the actual plot line, the theoretical influence curve, helps us to identify some future potential vulnerabilities. Here it is particularly important to be able to use the rescaling of the axes. As in our example for GPD, unboundedness of influence function of MLE shows up only on a logarithmic scale of the x-axis.

The second type of information can be obtained from the circles on top of the influence curve, which represent the influence of the actual data in the sample. Observations with larger overall influence have larger radii.

In the chosen example we work in a setting with a multivariate (two dimensional) parameter with coordinates shape and scale. So we can plot the influence curve for each of the parameter coordinates. But if the parameter dimension is much higher it is possible

to focus only on the most important 3-5 coordinates. It can happen, that some observation has strong influence if we consider all coordinates simultaneously. But, showing only these 3-5 coordinates, it may seem to be innocent, having a little impact. Therefore, by the means of additional plotting the individual coordinate-wise influence of all observations by the corresponding circle of specific size, we still can see the influence of the coordinates that are not shown. It is very helpful to identify suspicious observations in the data set.

To construct the corresponding graphs one can use the command `plot(IC, data,...)` or the wrapper function `PlotIC(IC, data,...)`.

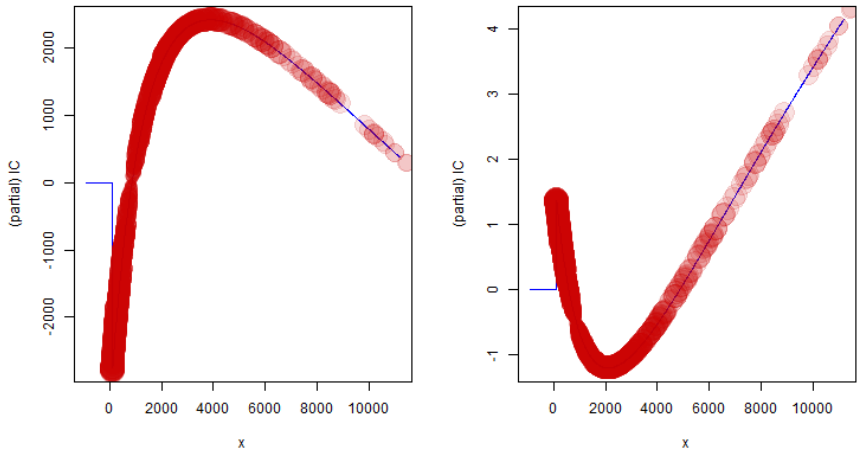

FIGURE 2.2: "Scale" and "shape" components for classical optimal influence curve for Generalized Pareto family

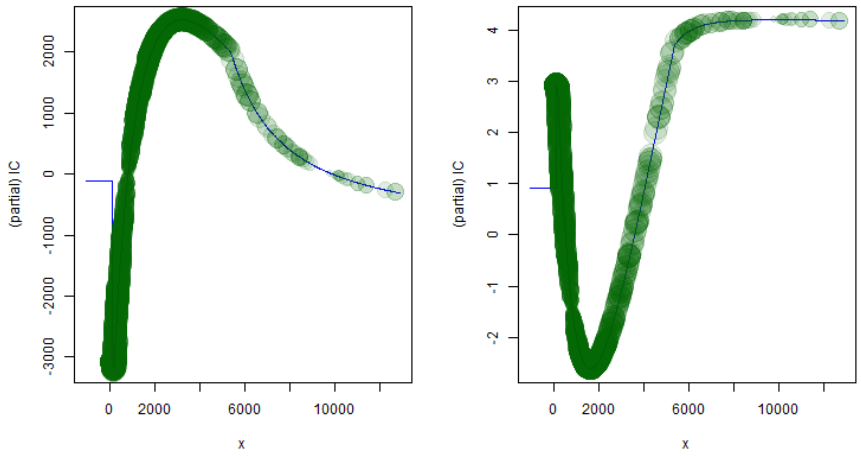

FIGURE 2.3: "Scale" and "shape" components of the influence curve of contamination type for Generalized Pareto family

Figure 2.2 represents the influence curves for the scale and shape parameters of GPD (left and right panel correspondingly) for the MLE and the second Figure 2.3 - for RMXE. As we already mentioned, each circle on the plot represents the observation and the radius of the circle reflects the influence of this observation on the corresponding parameter estimation. Besides, for this graph we have used the alpha-transparency feature. It describes the concentration of the observation in a specific point, i.e. more transparent circles have few of them, while the solid ones contain a lot.

On the left panel of the first graph (Figure 2.2) one can see that the influence function of MLE for the scale is bounded, whereas the one on right panel is unbounded for the shape. This underlines the well-known fact that the crucial difficulties for the estimation of the GPD family occur during the estimating of the shape parameter (see Section 3.1). It is important that, applying the RMXE procedure, the influence function becomes bounded for both parameters. This shows that the MLE of the shape is much more vulnerable to outlieres within the data sample than the RMXE and, thus, the RMXE can cope better with contaminated data.

ComparePlot

There is a function in the `R`-package `RobAStBase`, called `comparePlot`, which plots from 2 to 4 influence curves for the same model. This makes the comparison of the influence functions of different estimating procedures easier and clearer. This function is called by the command `comparePlot(obj1,obj2,...,data,...)` with the arguments `obj1`, `obj2`, which denote the corresponding influence curves, which we compare. Optional argument `data` for plotting observations into the plot is followed by other general arguments. The wrapper function is created for an easier use of these general arguments is `ComparePlot`.

On the graphs of Figure 2.4 we compare previously seen MLE and RMXE influence functions, left plot - for scale and right - for shape parameters of GPD family. It can be easily seen, that the MLE and the RMXE do not differ very much for the scale parameter in contrast to the shape parameter. As we mentioned before, this is caused by the difficulties arising while estimating the shape parameter. On the left-hand side we can see, that both estimators for the scale coincide for small values, then differ moderately for medium size values and converge against each other for large values. In contrast to that, looking at the right-hand side for the shape parameter, we see that the MLE and the RMXE are close to each other for very small values, differ already remarkably for values of medium size, with even higher influence for the robust estimator. Moreover, as

can be seen on previous plots, the ML-estimator is unbounded for large values, whereas the RMXE is bounded.

Information Plots

Another important diagnostic plots are information plots. They can quantify two types of information, i.e. so called *absolute* or *total* and *relative* information.

Here we note, that highly influential data points are those, which have a large Euclidean norm of the value of the influence function, which can also be multivariate. In the ideal situation such points contain much information about the parameter. In the robust context, however, it might be dangerous to induce correspondingly large bias values if outliers shift mass to this x value.

From the definition of the ALEs (Def. 2.17) it is easy to see, that the trace of the covariance of any ALE is just the expected squared Euclidean norm of the corresponding influence function. From the other side, the absolute information is simply the squared Euclidean norm of the influence function evaluated at the trace of the empirical covariance matrix. Hence, the values of the total information also represent the contribution of each observation to the corresponding trace of the covariance.

Here we use a squared norm scale. The reason is that we can either start with the aggregation over the parameter dimensions, i.e. to sum up all the corresponding squares,

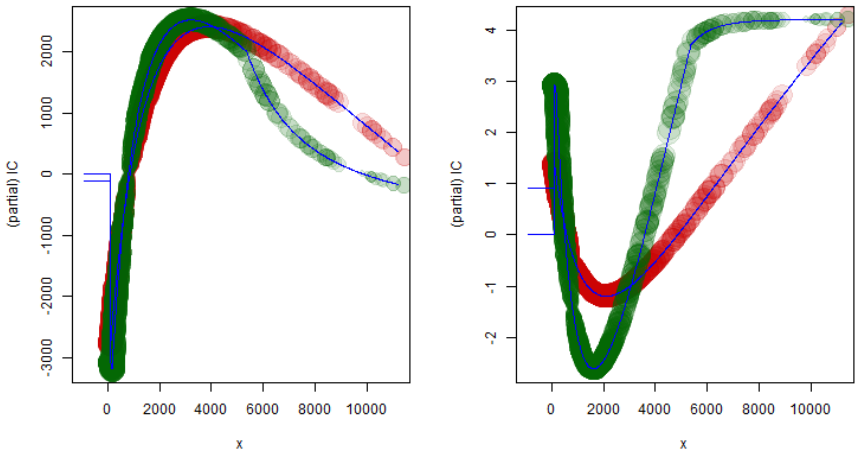

FIGURE 2.4: "Scale" and "shape" components of (partial) influence curve for Generalized Pareto family

or just aggregate over the observations. This double linearity in both parameter dimensions and observations also makes it easier to quantify the respective contribution of each parameter coordinate. This is what we call the relative information.

Hence, relative information plots can check percentage of the information, used per observation for each parameter coordinate. For example, when some specific observation assigns much less of its information to the scale than to the shape, it means, that it contains more information about the tail behavior than about the overall scaling.

For this plot we also can use rescaling of the axes, alpha-transparency, and plot all observations as circles of some specific sizes, which visualize additionally the total information on the top of the relative information curves. As we mentioned before, this is useful especially for the truly multivariate parameter setting.

To construct this plot we use the command `infoplot(IC, data,...)`, with the wrapper function `InfoPlot(IC, data,...)` for easier use.

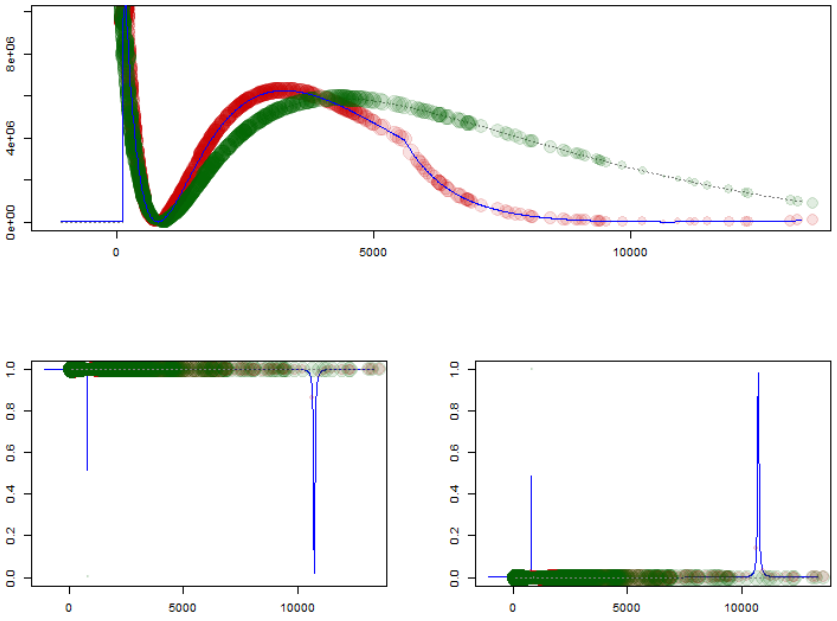

FIGURE 2.5: Absolute information and relative information of "scale" and "shape" components of (partial) influence curve for Generalized Pareto family

Using this diagnostic in our GPD context we get Figure 2.5 with three graphs. The first one for the absolute information and other two for the relative information for each parameter coordinate, scale and shape correspondingly. From the plots one can see that most observations are used for the scale parameter of GPD, whereas the shape is determined by few observations. Therefore we conclude that most observations contain

more information about the overall scaling, whereas only few of them are used for the tail behavior.

QQ Plot

The next plot we consider is the QQ plot. It checks the goodness of fit of the statistical model and describes how well the model fits a set of observations. QQ plots can be used to compare two probability distributions, either empirical distributions of two data samples or an empirical distribution with a theoretical one, by plotting their quantiles against each other.

For QQ plots we use log-rescaling for both axes. Moreover, on the graph, the true log-quantities on the x-axis are plotted against estimated log-quantities on the y-axis. In the literature QQ plots are shown in the inverse way, i.e. with interchanged x and y axes. Here we also have an option to flip the graph, but by default we leave it as it is.

Additionally, assuming that we are aware of 5% outliers in the data, we produce two types of 95%-confidence intervals: the outlier-adjusted pointwise and simultaneous confidence bands, which are based on the Central Limit Theorem and the Kolmogorov-Smirnov test statistics correspondingly. Alternatively, for parametric models one can use the profile based confidence intervals, as it is done in Coles (2001). They are even narrower than the pointwise or simultaneous bands, but in that case we restrict ourselves to the parametric models only. In the robust setting it is basically better to have nonparametric models. Another point here is that these intervals can only distinguish different parameter values within the GPD model, but we always assume that the model is GPD model. Therefore, we deside for the outlier-adjusted pointwise and simultaneous confidence intervals.

The plot is called in general by the command `qqplot(data,model, ...)`.

After applying QQ plot to our GPD model one can see from the Figure 2.6, that the curve has quite linear behavior. Only after the value of 10 on the abscissa, the curve starts to deviate a bit from the linearity. Moreover, as expected, the pointwise confidence interval, which is labeled by the green dotted lines, is narrower than the simultaneous one, bounded by the red lines. One can see, that most of the observations more or less fit the simultaneous interval, whereas observations with abscissa value larger than 10, where the curve deviates from the linearity, also exceed the bounds of the pointwise confidence interval. Here, as before, each observation is plotted as a circle with the radius according to the influence on the parameter estimation.

Outlyingness Plot

Outlyingness plot can be used for the identification of outliers. In the article of Hubert et al. (2005) distance-distance (dd), distance-projection (dp) and projection-projection (pp) plots are distinguished. These plots can be constructed by the function `dd.plot` plugging various distances in the function.

The first candidate for the distance is the *Mahalanobis distance*, which is very sensitive to the presence of outliers. For the observation x and the group of observations with the mean μ and covariance matrix Σ, the Mahalanobis distance is defined by the expression:

$$d_m(x) = \sqrt{(x-\mu)^{\mathrm{T}}\Sigma(x-\mu)}. \tag{2.20}$$

It might be unclear from the first look, that Mahalanobis distance can be heavily affected by single extreme observation, or group of outliers. The reason is the sensitivity of the arithmetic mean μ and the sample covariance matrix Σ to outliers. Moreover, in the presence of outliers, applying classical methods can lead to masking effects, i.e. large outliers can hide some group of small ouliers, so they can no longer be identified. The benefit of robustness here is that we can protect ourselves against this masking effect by using robust *minimum covariance determinant* (MCD) estimator, see Todorov (2009), to estimate the covariance for the Mahalanobis distance.

Therefore, we apply robust estimation to the location and scale in the formula for the Mahalanobis distance and get the so-called robust distances (RDs). Then, we construct the distance-distance graph (`dd.plot`) by plotting the classical Mahalanobis distance

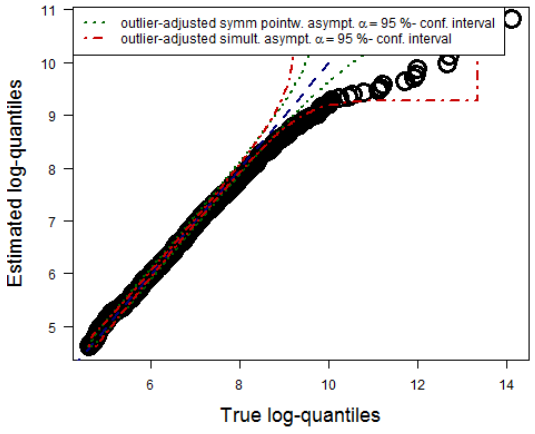

FIGURE 2.6: QQ plot with outlier-adjusted symmetric pointwise and simultaneous $\alpha = 95\%$-confidence intervals for Generalized Pareto family

based on the ordinary sample covariance matrix of the data against the robust distance obtained using MCD estimator.

A similar technique can be used for the distance-projection plot for the GPD, where the Mahalanobis distance of the score function is plotted against the log-transformed theoretical GPD quantiles.

The function `outlyingPlotIC(data, IC.class, IC.rob, ...)` constructs the corresponding outlyingness plot.

Applying the outlyingness plot we can identify outliers according to both size and influence. The influence in this case is measured at the classical, non-robust scale. Here it is crucial to determine parameter by robustness. However, to distinguish observations, it is important for the plot to use non-robust criteria. For x-axis it is usually more convenient to use rescaling according to log.

To construct the plot we use small set of the data (100 points) to get better look of the plot and to be able to recognize particular observations.

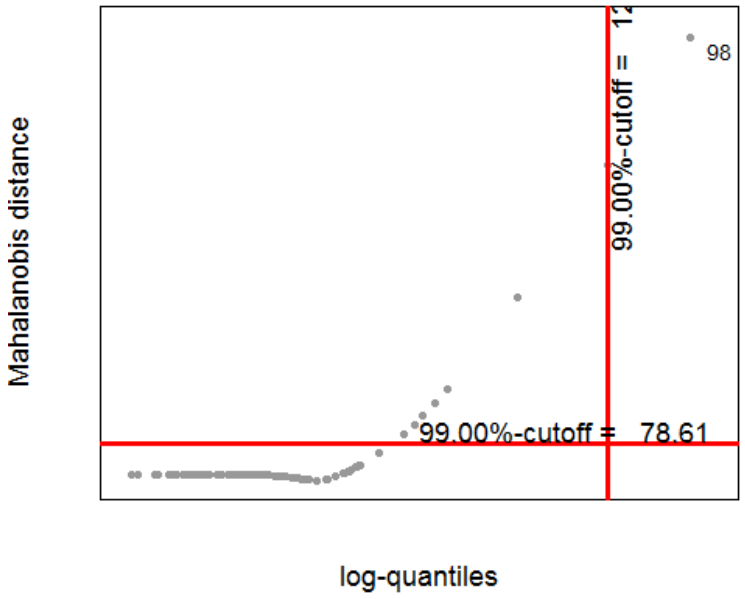

FIGURE 2.7: Outlyingness according to the size (vertical line) and the influence (horizontal line) of the outliers for Generalized Pareto family

On Figure 2.7 one can see two red cutoffs. The vertical line is made for the size and the horizontal is drawn for the influence of outliers. Both cutoffs are chosen according to robust procedures. Then, we get the critical area inside right upper quadrant and claim that all observations which fall into it are suspicious to be outliers. The data taken to construct the plot contains one outlier, which falls into the quadrant of interest.

On this plot we use labeling for suspicious observations. This feature is very important, because it gives us the opportunity to look behind the label and identify particular observation. Then, user can analyze this data point and decide if it is an outlier. As one can see in our example, observation in the critical area is labeled by number 98.

Model outlyingness was, to some extent, discovered before, but not in an automated setting for usage with extreme value distributions as is now achieved in package RobExtremes, which is one of our contributions in this PhD thesis. Therefore, one of our main contributions is the application of Model Outlyingness in the extreme value context.

Cniper Plot

Using robust estimation we analyze data only on the outliers themselves, but it is also very important to check their influence on the underlying estimator. Considering contamination of the sample as in (2.16), we have to decide how small X^{di} can be so that we still get considerable bias of the estimator and the candidate robust procedure becomes profitable, i.e. it beats the classic one. To do so, we do not admit arbitrary outlier generating distributions, but only contaminations by Dirac measures at some well-chosen gross error points. The notion of the cniper contamination and its interpretation can be found in the PhD thesis of Kohl (2005).

The main purpose of the cniper plot is to analyze the effect of an extra outliers on the estimator. By means of this plot one can distinguish cniper points, i.e. points s.t. under contamination with Dirac measures at these points the minimax risk for the optimally robust estimator is smaller than for the classical optimal estimator.

So on the cniper graph we plot the dirac points against the asymptotic risk difference for classical and robust estimation procedures and pay attention to the observations above the x-axis. It helps us to find points which cause the optimally robust estimator to perform better than the classical optimal estimator.

To compute this cniper plot we use the function `cniperPointPlot(L2Fam, data, ...)`.

Analyzing Figure 2.8 we conclude that, as far as the asymptotic risk difference for the classical and robust estimators is negative, classical procedure is preferable and we do not suspect outliers presence. However, for all points above the x-axis, robust estimation is better than the classical one, so these observations become suspicious and, similar to the previous plot, they can be identified using their labels.

2.6.3 Main features

There are some general plotting features, available in any plot function in R. These are the titles for the whole plot, or for each axis, the coloring, line attributes, e.g the width, the type of the lines e.g. dashed or dotted, character symbols and others. Of course, it is not the full list of them, but there are some special features, which we count as our contribution to the diagnostics. Although each feature is shown only on some or even none of the plots, they are all available for each diagnostic plots.

On the first plots (2.2, 2.3 and 2.4) we use the *alpha-transparency*, which describes the concentration of the observation in one point of coordinate system.

Most of the plots (see e.g. Figures 2.2, 2.3, 2.4, 2.5 and 2.6) plot each observation as a circle with *the radius* according to the total influence on the parameter estimation. We already mentioned before, that it might be very useful for the higher dimensional parametric models, when it is not possible to display all parameter coordinates. Using circles one can see the influence of the coordinates which are not displayed on the parameter estimation.

In Section 2.6.1 we introduced the idea and the importance of *rescaling*. On the Figures 2.2, 2.3, 2.6 and 2.7 we apply the log-rescaling for some or both axes. Rescaling helps to draw some very large observations, that have some important impact on the procedure, but can be omitted due to the limited amount of data.

Here we took the influence curves for the MLE of GPD scale and shape parameters and rescaled both axes. One can see that for the x-value "infinity" one can make sure that the influence function of the MLE is unbounded for the shape.

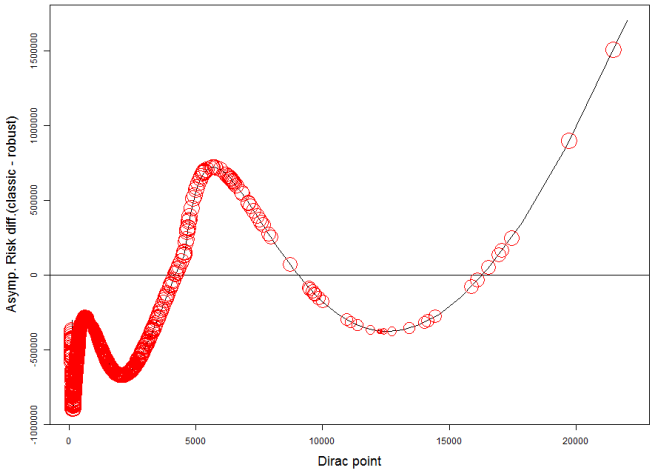

FIGURE 2.8: Cniper points for Generalized Pareto family

In Figure 2.7 we also used *labeling*, which let us determine observations behind the labels in order to be able to analyze them.

For the influence curve plot there is the additional option to draw the *rug plot*. This is the compact way of illustrating the marginal distributions of the variable along the axis. The positions of the data points are denoted by tick marks or circles. Formally, the rug plot is the series of short lines along the axis, positioned at each value of the data variation.

2.6.4 Wrapper function example

To give an example of the wrapper function we consider the diagnostic we used to plot Figure 2.2. The wrapper function used there for the influence curve plot is computed for the ML-estimation (see Section 2.6.2) is `PlotIC`. It has the following input parameters:

`IC` - object of class IC - influence curve;

`data` - optional data argument for plotting observations in the plot;

... - additional parameters of the wrapper function;

`alpha.trsp` - alpha-transparency argument, made for better view of plots with high number of points. Any number from 0 to 100 can be given as an input, meaning a percentage of the transparency of the points. As a default value we use an automatically adjusted transparency argument which depends on the number of points to be plotted;

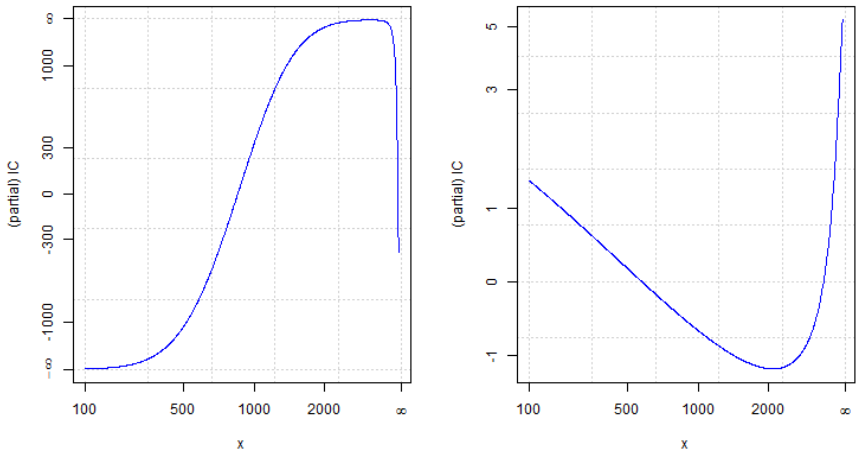

FIGURE 2.9: Rescaled "scale" and "shape" components for classical optimal influence curve for Generalized Pareto family

`with.legend` - indicator for showing the legend of the plot, which by default is set to be `TRUE` and the legend is drown;

`rescale (scaleX, scaleY)` - the flag for rescaling the axes for a better view of the plot. Rescaling is done automatically for the user if the corresponding argument `rescale` of the wrapper functions is set to be `TRUE` (then both axes are rescaled). If the user wants to rescale one axis, he may set respective argument, `scaleX` or `scaleY`, to be `TRUE`. By default, these arguments, however, are set to be `FALSE`.

`withCall` - the flag for the call output. Since the wrapper functions are thought to give some easy first look at the corresponding diagnostic plot, it can be useful for user to see, which parameters the wrapper have given to the diagnostic plot by default. For this purpose, the argument `withCall` is introduced, which has the default value `TRUE`.

The wrapper has the following usage:

```
PlotIC(IC, data, ..., alpha.trsp, with.legend, rescale ,withCall)
```

The simplest way to use the wrapper function `PlotIC`, is to take all possible parameters values by default. Additionally, to see these default values, we set the parameter `withCall` to be `TRUE`, i.e. we use the function:

```
PlotIC(ICmle, data0, with.call=TRUE)}
```

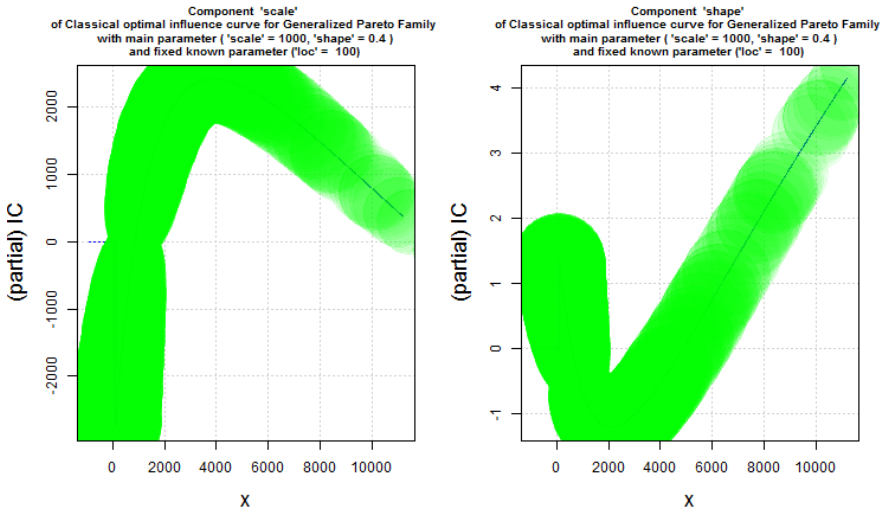

FIGURE 2.10: Example of the wrapper function usage when all parameters are taken by default

Beside drawing this plot, which is the same as in Figure 2.2, except color adjustment and other style options, this wrapper function gives as an output the true call of the original diagnostic plot `PlotIC`:

```
plot(adj = 0.5, alpha.trsp = 50, bmar = par("mar")[1], bty = "o",
         cex = 1.5, cex.inner = 0.8, cex.lab = 1.5, cex.main = 1.5, cex.pts = 0.3,
         col = "blue", col.inner = par("col.main"), col.lab = "black", col.main =
         "black", col.MBR = par("col"), col.pts = addAlphTrsp2col(rgb(0, 255, 0,
         maxColorValue = 255), substitute(50)), inner = TRUE, jitter.fac = 1,
         legend.bg = "white", legend.cex = 0.8, legend.location = "bottomright",
         lty.MBR = "dashed", lwd.MBR = 0.8, main = FALSE, MBR.fac = 2, MBRB = NA,
         mfColRow = TRUE, panel.first = grid(), pch.pts = 19, return.Order =
         FALSE, scaleN = 9, scaleX = FALSE, scaleY = FALSE, sub = FALSE,
         tmar = par("mar")[3], with.call = TRUE, with.lab = FALSE, with.legend =
         FALSE, withMBR = FALSE, withSweave = getdistrOption("withSweave"),
         x = ICmle, y = data0)
```

We mentioned before in Section 2.6.2, that the wrapper functions already implemented for diagnostics are follows:

- `CniperPointPlot` is the wrapper for `cniperPointPlot`,
- `InfoPlot` is the wrapper for `infoPlot`,
- `ComparePlot` is the wrapper for `comparePlot`,
- `PlotIC` is the wrapper for `plot`.

More about these functions can be found later in the paper of Ruckdeschel et al. (2014a).

2.7 Software infrastructure

`RobASt` (Robust Asymptotic Statistics) is the family of `R`-packages, which consists of `distr`-family packages created mostly by M. Kohl and P. Ruckdeschel. Version 2.6 of these packages is used in this thesis. More precisely, these are the following packages:

- The `R`-package `distr` provides classes for distributions, including discrete distributions e.g. Binomial, Poisson etc. and absolute continuous distributions, e.g. Normal, Exponential, Uniform etc. There are corresponding calls for the random number generator, density, cumulative distribution and quantile functions, which are identical for both types of distributions. The methods, available in the package, cover simple arithmetic operations with distributions as well as more complicated once. The vignette of this package is written by Ruckdeschel et al. (2005), and a detailed description of it can be found in Ruckdeschel et al. (2006).
- The `R`-package `distrEx` extends the package `distr`, introducing the expectation operator (see Kohl and Ruckdeschel (2005) and Ruckdeschel et al. (2006)). The

main benefit of this package is that this expectation operator can be used for extreme value distributions.

- The R-package `distrMod` widely uses distribution classes from the package `distr`, as well as functions and methods from the package `distrEx`. In addition, it includes functions and methods to compute maximum likelihood and the minimum distance estimators (see Ruckdeschel and Kohl (2008) and Kohl and Ruckdeschel (2010)). Moreover, it represents the most flexible implementation of the minimum criterion estimators for the univariate distributions available in R so far.

- The R-package `distrSim` is created for the standardized treatment of simulations, also under contaminations (see Ruckdeschel et al. (2006)).

- The R-package `distrTEst` contains classes and methods for evaluations of statistical procedures on such simulations and can also be found in the article of Ruckdeschel et al. (2006).

- The R-package `distrTeach` creates illustrations for basic statistics courses using all distribution classes (see more in Ruckdeschel et al. (2008a) and Ruckdeschel et al. (2008b)).

Except for the `distr`-family packages, the `RobASt`-family contains other robust paskages (for each package the version 1.0 is used), i.e.

- The R-package `RandVar` provides classes for the random variables or vectors, which extends and requires the packages `distr` and `distrEx`. It applies all arithmetic (and matrix) operations, possible in `R` with numeric variables (vectors), to the random variables and some further methods which can be found in Kohl and Ruckdeschel (2013b).

- The R-package `RobAStBase` (see Kohl and Ruckdeschel (2013e)) includes some necessary S4 class infrastructure like neighborhoods, influence curves and robust models.

- The R-package `ROptEst` provides classes for optimally-robust estimation in infinitesimal robustness setup (see Kohl and Ruckdeschel (2013c)). Using this package we are able to construct asymptotically linear estimators, one-step-estimators etc. and apply various methods to them. Optimally-robust estimators can be constructed for different neighborhood types, risks, bias-types and norms.

- The R-package `RobLox` includes functions for the computation of many well known influence curves (e.g., Huber-, Hampel-, Tukey- etc.) for normal location and scale in the framework of our asymptotic setup (see Kohl and Ruckdeschel (2013a)).

- The R-package RobRex, which can be found in Kohl (2013), provides computation of the optimally robust ICs for regression and scale parameters in regression-type models.

- The R-package ROptRegTS (see Kohl and Ruckdeschel (2013d)) contains functions and methods for optimally robust estimation for regression and time series models.

There are another R-packages on robustness submitted to CRAN, e.g package robustbase (see Todorov and Filzmoser (2009) and Rousseeuw et al. (2012)), based on the book of Maronna et al. (2006), and package robust which can be found in Wang et al. (2014). Both are aimed to provide maximum tools for analyzing data with robust methods. Corresponding versions 0.91.1 and 0.4.16 of those packages are used here.

2.8 R-package RobExtremes

Here we present new member of the RobASt-family of R-packages, package RobExtremes (see Ruckdeschel et al. (2013)). This package is based on and extends the framework of most packages mentioned above - distr and RobASt families of R-packages available on CRAN.

Package RobExtremes provides infrastructure for optimally robust estimation in scale-shape models, covering Gamma, Weibull, and in particular generalized Pareto distribution and generalized extreme value distribution models.

As starting estimators for the Generalized Pareto and Generalized Extreme Value Distribution models, RobExtremes implements general Location-Dispersion (LD) estimators including the high-breakdown point estimators medSn, medQn, and medkMAD discussed in Ruckdeschel and Horbenko (2012).

To speed-up computation of the optimally-robust estimators and overcome problems caused by limited equivariance structure of the scale-shape models, package RobExtremes applies interpolation technique. Moreover, all diagnostics discussed in Section 2.6 belong to the package RobExtremes. Version of this package used for computing the plots and for other examples is 1.0.

In the paper Ruckdeschel et al. (2014a) one can find four reference examples from extreme value statistics on how to use this package on the real data sets covering hospital length of stay, liquidity risk, operational risk of a bank, and hydrology.

2.9 Conclusions

Main goal of this Chapter is to give some understanding of robustness and get to know some notions and methods from robust statistics. At the beginning of this Chapter we have briefly discussed the concept of robustness itself and made some literature overview concerning this topic.

Next, we introduced parametric model, which is the basis of our further research, and gave some examples of it. Here we defined notions of the absolute continuity, smoothness and L_2 differentiability. Then, we presented important for our research result of Hájek, which contains conditions for L_2 differentiability of the parametric model and for better understanding checked this conditions for Binomial and GEV parametric models.

Further, we introduced different types of neighborhoods and then focused on the contamination neighborhoods based on the Gross Error Model.

Then we devoted one Section to the different notions which capture local or global robustness, including influence function and breakdown point.

In the next two Sections we introduced the most common classical estimators and gave some simple examples of computing these estimators in software programming language `R`. Next we defined optimally robust estimators.

At the end of the Chapter we paid attention to the model diagnostics already implemented in new `R`-package `RobExtremes`, which draws diagnostic plots to see different aspects of the taken model. We also gave the overview of the software infrastructure including our new package `RobExtremes` in the last Section.

Chapter 3

Extreme value statistic

Extreme value theory (EVT) has been developed in parallel with the central limit theory, but instead of the partial sums, EVT is concerned with the stochastic behavior of the sample extremes, i.e. the minimum or the maximum of i.i.d. random variables. It plays the same fundamental role in the extremes of the random variables as the central limit theorem (CLT) in their sum. More precisely, we take the sample of the i.i.d. random variables and let its size tend to the infinity. Then we get some limiting distributions for the extreme values of the sample. These distributions are defined as the *extreme value distributions*. They are widely used in finance, insurance, economics, telecommunications and many other industries dealing with extreme events. Extreme value distributions are also often used in hydrology to model some natural phenomena, e.g. sea levels, river heights, stream flows, and rainfall, in order to obtain the distribution of the annual maxima.

Extreme value theory studies these kinds of the distributions and their properties. It is the theory of modeling events which occur with very small probability, so-called rare events. Per definition, risky events happen with low probability, therefore, EVT is very useful in the risk modeling. The statistical analysis of extreme data is important for the various disciplines, including not only hydrology, insurance, finance, but also engineering and environmental sciences.

Origins of the extreme value theory go back to the research of Firsher and Tippett (1928) and Gnedenko (1943), but the first allusion on it appears even earlier, in the articles of Bortkiewicz (1922), where the *distribution of the largest value* is introduced for the first time; of Tippett (1925), in which the exact cumulative distribution function and moments of the largest order statistic are studied; and Fréchet (1927), where one possible limit distribution for the largest order statistic is defined. Next year, authors of the article Firsher and Tippett (1928) showed, that extreme limit distribution can only be

one of three types. Along with Tippett and Fisher, well-known German mathematician Gumbel plays an important role in the development of the EVT, especially his work Gumbel (1958).

Due to the large variety of the applications, analysis of the extremes became popular area for the research and a lot of books and articles on the extreme value theory have appeared. H. Harter wrote an authoritative bibliography of EVT, see Harter (1978); authors of the book Leadbetter et al. (1983) worked with the extremes of the stationary processes; books and articles of Kotz and Nadarajah (2000), Coles (2001), de Haan and Ferreira (2006), Reiss and Thomas (2007) and Falk et al. (2011) provide a self-contained introduction to the analysis of extremes and their applications in the different fields. Publications of Castillo (1988) are well-known by applications of EVT in engineering and science. Beirlant et al. (1996) and Embrechts et al. (1997) provide a practical analysis of extreme values with emphasis on finance and insurance applications.

Similarly to Chapter 2, here we restrict ourselves to giving only some part of the EVT, which is used in the next sections of the thesis. In each Section we refer to the relevant monographs, where detailed overview of one or the other concept can be found. Theoretical background of this Chapter is mainly taken from the books of Embrechts et al. (1997), Kotz and Nadarajah (2000) and Falk et al. (2011).

3.1 Basic concepts

3.1.1 Extreme value distributions

As was already mentioned, *extreme value distributions* are the limiting distributions for the extreme values of the i.i.d. sample. Two main members of this family are *generalized extreme value* distribution and *generalized Pareto* distribution.

Generalized extreme value distribution

In probability theory and statistics, the *generalized extreme value distribution* (GEVD), derived by Firsher and Tippett (1928), is a family of the continuous probability distributions, which combines into a single form three possible types of the limiting distribution for extreme values. In some applications GEVD is also known as Fisher-Tippett distribution.

In the Example 2.4 we introduced simplified form of the generalized extreme value distribution, i.e. by the cumulative distribution function (2.3). Here we give the general definition of GEVD.

Definition 3.1 (GEVD)**.** The *generalized extreme value distribution* $\mathrm{GEVD}(\mu, \sigma, \xi)$ is defined by the c.d.f. of the following form :

$$Q^{\mathrm{GEVD}}_{\mu,\sigma,\xi}(x) = \begin{cases} \exp\big(-(1+\xi\frac{x-\mu}{\sigma})^{-\frac{1}{\xi}}\big), & \xi \neq 0 \\ \exp\big(-\exp(-\frac{x-\mu}{\sigma})\big), & \xi = 0, \end{cases} \tag{3.1}$$

for the input domain $1 + \xi\frac{x-\mu}{\sigma} > 0$. Support of $\mathrm{GEVD}(\mu, \sigma, \xi)$ corresponds to

$$x \in \begin{cases} (\mu - \frac{\sigma}{\xi}, \infty), & \xi > 0 \\ (-\infty, \mu - \frac{\sigma}{\xi}), & \xi < 0 \\ (-\infty, \infty), & \xi = 0. \end{cases}$$

GEVD is specified by three parameters: location parameter $\mu \in \mathbb{R}$, positive scale parameter $\sigma > 0$ and the shape $\xi \in \mathbb{R}$, which governs the tail behavior of the distribution.

GEVD can be also described by the density and quantile functions, given in Chapter 2 by the equations (2.4) and (2.5) correspondingly. Just to remind, these are the following functions:

$$q^{\mathrm{GEVD}}_{\mu,\sigma,\xi}(x) = \frac{1}{\sigma}(1+\xi\frac{x-\mu}{\sigma})^{-\frac{1}{\xi}-1}\exp\big(-(1+\xi\frac{x-\mu}{\sigma})^{-\frac{1}{\xi}}\big),$$

$$F^{\mathrm{GEVD}}_{\mu,\sigma,\xi}(y) = \mu - \frac{\sigma}{\xi}(1-(-\log y)^{-\xi}).$$

Definition 3.2. Family of the probability distributions $\mathcal{Q}$ is called *location-scale (affine) invariant*, if for all constants $a \in \mathbb{R}$ and $b \in \mathbb{R}^+$, and for all $X \sim Q \in \mathcal{Q}$, affine transformation $a + bX$ has is Q'-distributed, i.e. $a + bX \sim Q'$, and $Q' \in \mathcal{Q}$.

Remark 3.3. *Note, that GEVD family is location-scale invariant. In other words, using affine transformations of the data we do not leave the model class.*

Definition 3.4. Function $F : X \mapsto Y$ is called *location-scale (affine) equivariant*, if it is not affected by the affine transformations, i.e. for all constants $a \in \mathbb{R}$ and $b \in \mathbb{R}^+$, and for all $x \in X$, holds that $F(a + bx) = a + bF(x)$.

Remark 3.5. *If we consider transformation acting on the location and the scale parameters, $T : (0, 1) \mapsto (\mu, \sigma)$, then one can immediately see from the structure of the quantile function, that GEVD is equivariant w.r.t. location μ and scale σ, i.e. $F^{GEVD}_{\mu,\sigma,\xi}(y) = F^{GEVD}_{T(0,1),\xi}(y) = \mu + \sigma F^{GEVD}_{0,1,\xi}(y)$.*

For the completeness we define *standard generalized extreme value distribution* $\mathrm{GEVD}(\xi)$ (as in Embrechts et al. (1997, Def. 3.4.4)), specified only by the shape parameter ξ,

taking the expression $(x-\mu)/\sigma$ as an argument for the distribution function, i.e.

$$Q_{\xi}^{\mathrm{GEVD}}(x) = \begin{cases} \exp\big(-(1+\xi x)^{-\frac{1}{\xi}}\big), & \xi \neq 0 \\ \exp\big(-\exp(-x)\big), & \xi = 0. \end{cases}$$

The *tail distribution function* for any distribution with c.d.f. $Q(x)$ is defined as:

$$\overline{Q}(x) = \mathbf{P}(X > x),$$

then one can define the *long (right) tail distribution* as the one, with the tail distribution satisfying:

$$\lim_{x\to\infty} \mathbf{P}(X > x + t | X > x) = \lim_{x\to\infty} \frac{\overline{Q}(x+t)}{\overline{Q}(x)} = 1.$$

Working with the statistical distributions, one also can distinguish the *thick* or *heavy* tails, meaning that they converge to zero slowly in the extremes. More precisely, distribution with c.d.f. $Q(x)$ is called *heavy-tailed*, if its tail distribution decays slower than exponentially, i.e.

$$\lim_{x\to\infty} e^{\lambda x}\overline{Q}(x) = \infty \quad \text{for all} \quad \lambda > 0.$$

Any long-tailed distribution is heavy-tailed, but not vise versa. It is possible to construct heavy-tailed distributions that are not long-tailed.

The way to measure the "thickness" of the tail for heavy-tailed distribution is to use the *tail-index*.

Definition 3.6 (Tail-index)**.** For the distribution with c.d.f. $Q(x)$ and quantile function $F(x)$, the *tail-index* is defined as follows:

$$\alpha = \frac{F(0.99) - F(0.5)}{F(0.75) - F(0.5)} \Big/ \frac{\Phi(0.99) - \Phi(0.5)}{\Phi(0.75) - \Phi(0.5)},$$

where $\Phi(x)$ is the standard normal quantile function. Obviously, for the Normal distribution tail-index is equal to 1.

Remark 3.7. *Another important index, which can be used to measure the degree of clustering is called extremal index. The definition and the further explanations of it one can find in the book of Leadbetter et al. (1983).*

Remark 3.8. *GEVD family depends on the tail index, moreover, the tail-index for it can be expressed in the terms of the shape parameter, i.e.* $\alpha = 1/\xi$*. Often, for the estimation of the shape, tail-index estimation is used. There are some well-studied estimators for it, e.g. Hill estimator introduced by Hill (1975).*

Due to this relation between the shape and the tail-index, one can distinguish three sub-families of the GEVD w.r.t. the shape parameter, which have different types of the tails.

More precisely, for the zero shape it is called the *Gumbel distribution* (see Gumbel (1958)). In Figure 3.1 one can see, that this distribution has light (medium) upper tail and it is positively skewed. Whereas if the shape is positive, we obtain the *Fréchet distribution* (see Fréchet (1927)) with heavy upper tail. For the negative shape it is called the *Weibull distribution* (see Weibull (1939, 1951)), the distribution with bounded (short) upper tail.

The Weibull distribution has a larger index of tail weight than an exponential when the shape parameter is small, and the index decreases as the shape increases.

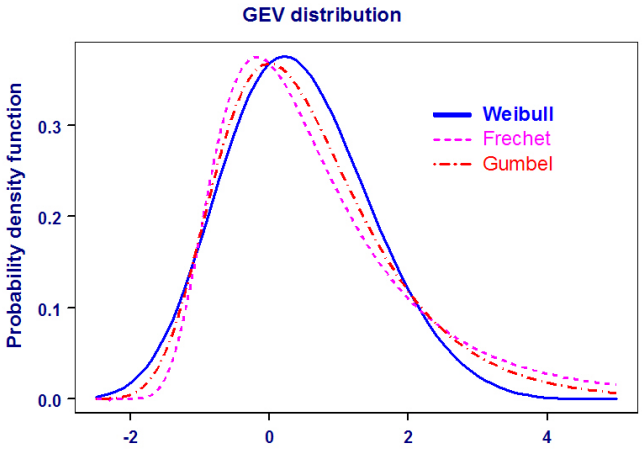

FIGURE 3.1: Generalized extreme value distributions: Gumbel, Fréchet, Weibull

Cumulative distribution functions for these sub-families of the GEVD are:

- Gumbel or type I extreme value distribution ($\xi = 0$):

$$\text{Gumbel} = Q^{\text{G}}_{\mu,\sigma}(x) = e^{-e^{\frac{x-\mu}{\sigma}}}, \quad x \in \mathbb{R} \tag{3.2}$$

- Fréchet or type II extreme value distribution ($\xi = \alpha^{-1} > 0$):

$$\text{Fréchet} = Q^{\text{F}}_{\mu,\sigma,\xi}(x) = \begin{cases} 0, & x \leq \mu \\ e^{-(\frac{x-\mu}{\sigma})^{-\alpha}}, & x > \mu. \end{cases} \tag{3.3}$$

- Weibull or type III extreme value distribution ($\xi = \alpha^{-1} < 0$):

$$\text{Weibull} = Q^{\text{W}}_{\mu,\sigma,\xi}(x) = \begin{cases} e^{-(\frac{x-\mu}{\sigma})^{-\alpha}}, & x < \mu \\ 1, & x \geq \mu. \end{cases} \tag{3.4}$$

Generalized Pareto distribution

Generalized Pareto distribution (GPD) is another extreme value distribution, specified by the same set of parameters, i.e. the location $\mu \in \mathbb{R}$, the scale $\sigma > 0$ and the shape $\xi \in \mathbb{R}$.

Definition 3.9 (GPD)**.** The cumulative distribution function of the *generalized Pareto distribution* $\mathrm{GPD}(\mu, \sigma, \xi)$ is of the form:

$$Q^{\mathrm{GPD}}_{\mu,\sigma,\xi}(x) = \begin{cases} 1 - (1 + \xi \frac{x-\mu}{\sigma})^{-\frac{1}{\xi}}, & \xi \neq 0 \\ 1 - e^{-\frac{x-\mu}{\sigma}}, & \xi = 0. \end{cases} \tag{3.5}$$

The support of the $\mathrm{GPD}(\mu, \sigma, \xi)$ corresponds to

$$x \in \begin{cases} (\mu, \infty), & \xi \geq 0 \\ [0, \mu - \sigma/\xi), & \xi < 0. \end{cases}$$

GPD can be described by the density function of the form

$$q^{\mathrm{GPD}}_{\mu,\sigma,\xi}(x) = \begin{cases} \frac{1}{\sigma}(1 + \xi \frac{x-\mu}{\sigma})^{-\frac{1}{\xi}-1}, & \xi \neq 0 \\ \frac{1}{\sigma} e^{-\frac{x-\mu}{\sigma}}, & \xi = 0. \end{cases} \tag{3.6}$$

Density function $q^{\mathrm{GPD}}_{0,1,0.7}(x)$ is plotted in Figure 3.2.

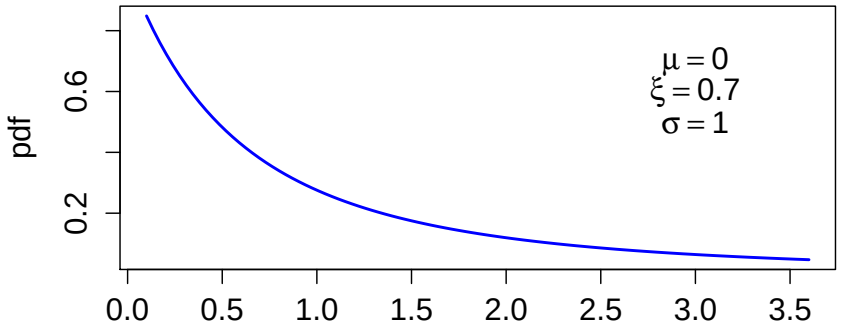

FIGURE 3.2: Generalized Pareto distribution

Quantile function of the GPD is the following

$$F^{\mathrm{GPD}}_{\mu,\sigma,\xi}(y) = \mu - \frac{\sigma}{\xi}(1 - (1 - y)^{-\xi}). \tag{3.7}$$

Remark 3.10. *Similarly to the GEVD, the GPD family is location-scale invariant and equivariant. Moreover, moving the threshold affects the scale but not the shape, i.e. if we use some transformation to the location parameter $\mu \mapsto \mu'$, then $\sigma \mapsto \sigma'$, but $\xi \mapsto \xi$.*

Remark 3.11. *The tail-index (see Def. 3.6) of the GPD can be also expressed in terms of the shape parameter ξ, i.e. $\alpha = 1/\xi$.*

We define *standard generalized Pareto distribution* GPD(ξ) (as in Embrechts et al. (1997, Def. 3.4.9)), specified by the shape ξ, taking the expression $(x-\mu)/\sigma$ as an argument x for the distribution function, i.e.

$$Q_{\xi}^{\mathrm{GPD}}(x) = \begin{cases} 1-(1+\xi x)^{-\frac{1}{\xi}}, & \xi \neq 0 \\ 1-e^{-x}, & \xi = 0. \end{cases} \tag{3.8}$$

The special cases of the GPD w.r.t. some specific parameter values are the following

- Exponential distribution ($\mu = 0$, $\sigma = \lambda^{-1}$ and $\xi = 0$):

$$Q_{\lambda}^{\mathrm{E}}(x) = 1 - e^{-\lambda x}, \quad x \geq 0. \tag{3.9}$$

- Uniform distribution ($\mu = a$,$\sigma = b-a$ and $\xi = -1$,):

$$Q_{a,b}^{\mathrm{U}}(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & x \in [a,b) \\ 1, & x \geq b. \end{cases} \tag{3.10}$$

- Pareto distribution ($\mu = \sigma/\xi = x_m$, $\sigma > 0$ and $\xi = \alpha^{-1} > 0$):

$$Q_{x_m,\alpha}^{\mathrm{P}}(x) = \begin{cases} 1-(\frac{x_m}{x})^{\alpha}, & x \geq x_m \\ 0, & x < x_m. \end{cases} \tag{3.11}$$

3.1.2 Extreme value theorems

There are two well-known general results in the extreme value theory regarding asymptotic distribution of extreme order statistics. These extreme value theorems are called *Fisher–Tippett–Gnedenko theorem* (see Firsher and Tippett (1928), Gnedenko (1943)) and *Pickands–Balkema–de Haan theorem* (see Balkema and de Haan (1974), Pickands (1975)). Next we give the statements of these theorems.

Fisher–Tippett–Gnedenko theorem

Fisher–Tippett–Gnedenko theorem, also called the first theorem in the extreme value theory, shows that the GEVD is the only possible limit distribution for the properly normalized maximum of the sequence of i.i.d. random variables. More precisely, following the formulation taken from the book of Falk et al. (2011, Thm. 2.1.1), this theorem states:

Theorem 3.12 (Fisher–Tippett–Gnedenko theorem)**.** *Let $\{X_1, ..., X_n\}$ be the sample of i.i.d. random variables. Denote the maximum of the sample as $M_n = \max\{X_1, ..., X_n\}$. If for some constants $a_n > 0$ and $b_n \in \mathbb{R}$ holds*

$$\lim_{n\to\infty} P\Big(\frac{M_n - b_n}{a_n} \leq x\Big) = Q(x), \tag{3.12}$$

where Q is the non-degenerate distribution function, then the limit distribution Q is GEVD, i.e. it belongs to one of three sub-families: the Gumbel, the Frećhet or the Weibull family.

Remark 3.13. *Theorem does not state that a limit distribution exists: this additionally requires regularity conditions on the tail of the distribution.*

Pickands–Balkema–de Haan Theorem

The Pickands–Balkema–de Haan theorem is also called the second theorem in the extreme value theory. It states that the best approximation of any tail distribution for the data above the threshold is GPD.

To give the statement of the following theorem, first we need to define so-called *conditional excess distribution function* (see Embrechts et al. (1997, Def. 3.4.6)). It is conditional distribution over a certain threshold u (in practice threshold is sufficiently large), i.e.

$$\overline{Q}_u(x) = \mathbf{P}(X - u \leq x | X > u) = \frac{Q(u+x) - Q(u)}{1 - Q(u)},$$

for the values $0 \leq x \leq x_Q - u$, where x_Q is the *right endpoint* of the distribution Q.

Then Balkema and de Haan (1974) and Pickands (1975) posed the following theorem.

Theorem 3.14 (Pickands–Balkema–de Haan Theorem)**.** *Let $\{X_1, ..., X_n\}$ be the sample of i.i.d. random variables with distribution function Q. Then hold that*

$$\lim_{u\to\infty} \overline{Q}_u(x) = Q^{GPD}_{\mu,\sigma,\xi}(x),$$

for the GPD $Q^{GPD}_{\mu,\sigma,\xi}(x)$ with the c.d.f. (3.6).

3.1.3 Relation between GEVD and GPD

One can see from the expressions (3.1) and (3.5), that c.d.f. of the GPD can be formulated in terms of the GEVD c.d.f. in the following way

$$Q^{\mathrm{GPD}}_{\mu,\sigma,\xi}(x) = 1 + \log(Q^{\mathrm{GEVD}}_{\mu,\sigma,\xi}(x)).$$

From this expression one concludes, that whenever we have convergence in one distribution, we have it in the other one as well. Moreover, this expression is the key to another relations between GEVD and GPD, which we present further.

Maximum domain of attraction

We start with the definition of the maximum domain of attraction taken from the book of Embrechts et al. (1997, Def. 3.3.1).

Definition 3.15 (Maximum domain of attraction)**.** We say that random variable X, as well as its distribution function Q, belongs to the *maximum domain of attraction* of the GEVD, denoting it as $X \in \mathrm{MDA}(Q^{\mathrm{GEVD}}_{\mu,\sigma,\xi})$, if convergence (3.12) holds for some constants $a_n > 0$ and $b_n \in \mathbb{R}$.

Relation between maximum domains of attraction of the GEVD and the GPD can be described by the property presented in Embrechts et al. (1997, Th. 3.4.13). It states that for every $\xi \in \mathbb{R}$ distribution function Q belongs to the MDA of the standard GEVD, i.e. $Q \in \mathrm{MDA}(Q^{\mathrm{GEVD}}_{\xi})$, iff

$$\lim_{u \uparrow x_Q} \sup_{0<x<x_Q-u} |\overline{Q}_u(x) - Q^{\mathrm{GPD}}_{0,\sigma(u),\xi}(x)| = 0,$$

for some positive function σ. This characterisation of the MDA of GEVD immediately leads to the definition of the GPD (see also Embrechts et al. (1997, Th. 3.4.5)).

Remark 3.16. *It is important to note, that the scale σ is the function of the threshold u, whereas the shape is constant. Therefore, changing the threshold has affects on the scale but not on the shape.*

Block maxima and peak-over-threshold

Here we present two approaches of analyzing the extreme values or fitting extreme value distributions, called *block maxima* approach and *peak-over-threshold* (POT) method, see Embrechts et al. (1997, Ch. 6), Coles (2001) and Ferreira and de Haan (2013). The

POT method is developed by Pickands (1975), so the block maxima approach is the older one (see e.g. Gumbel (1958)).

The idea of the **block maxima** approach is to divide the data into non-overlapping blocks of equal length and restrict attention to the maximum observation in each block, e.g. annual maxima of daily precipitation amounts. Right choice of the block size is very important, because too small size can lead to the bias and with too big size, we can generate too few block maxima and get large estimation variance (see Coles (2001, Ch. 3)). Then, by the Fisher-Tippett-Gnedenko theorem, new observations created by this approach approximately follow GEVD. Therefore, the block maxima approach is closely associated with the use of the GEVD family.

By the **peak-over-threshold** method, suggested by hydrologists, we chose some certain high threshold value and select those of the initial observations, which exceed this threshold. As one can conclude from the Pickands-Balkema-de Haan theorem 3.14, probability distribution of these selected observations, under extreme value conditions, is approximately the GPD. Again, a bias may appear, since GPD is not the exact distribution of the selected observations.

Remark 3.17. *These two approaches are closely related. Moreover, we get convergence in the block maxima to the GEVD iff we have convergence of the POT against the GPD. Besides, the limiting shapes for both approaches coincide.*

On the one hand, the POT seems to make better use, since it picks up all relevant high observations and it is justified under exact well-known conditions (see Ferreira and de Haan (2006)). However, the block maxima method is preferable when the observations are not exactly i.i.d., e.g. seasonal periodicity in the case of the yearly maxima. Moreover, the block maxima method may be easier to apply, since the block periods appear naturally in many real situations.

3.2 Model smoothness

In this Section we discuss the Hájek conditions (H.1)-(H.3) from Proposition 2.9 for the introduced extreme value distributions.

3.2.1 Smoothness of GEVD

We start with the generalized extreme value parametric model from the Example 2.4. We let location parameter μ be unknown and consider three dimensional general parameter $\theta = (\mu, \sigma, \xi) \in \Theta \subset \mathbb{R} \times \mathbb{R}^+ \times \mathbb{R}$.

The density function q_θ^{GEVD}, given by (2.4), is absolutely continuous in each $\theta \in \Theta$, except singularity point $\xi = 0$, therefore, Hájek condition (H1) is fulfilled for all $\xi \neq 0$.

In order to check the second condition, we compute the derivative $\frac{\partial}{\partial\theta} q_\theta^{\mathrm{GEVD}}(x) = \Lambda_\theta^{\mathrm{GEVD}}(x)\, q_\theta^{\mathrm{GEVD}}(x)$. As we mentioned in the Example 2.12, the L_2 derivative for the GEVD takes up the structure of the parameter, therefore, it consists of three coordinates, one for the location, one for the scale and one for the shape, i.e.

$$\Lambda_\theta^{\mathrm{GEVD}} = (\Lambda_\mu^{\mathrm{GEVD}}, \Lambda_\sigma^{\mathrm{GEVD}}, \Lambda_\xi^{\mathrm{GEVD}})^{\mathrm{T}}$$

Next, we compute these L_2 derivative coordinates, and using notations $z := \frac{x-\mu}{\sigma}$ and $s := 1 + \xi z$, we get the following

$$\Lambda_\mu^{\mathrm{GEVD}} = \frac{(\xi+1) - s^{-\frac{1}{\xi}}}{s\sigma},$$

$$\Lambda_\sigma^{\mathrm{GEVD}} = \frac{z(1 - s^{-\frac{1}{\xi}}) - 1}{s\sigma},$$

$$\Lambda_\xi^{\mathrm{GEVD}} = \frac{1}{\xi}\big(1 - s^{-\frac{1}{\xi}}\big)\Big(\frac{1}{\xi}\log s - \frac{s}{z}\Big) - \frac{s}{z}.$$

Therefore, the L_2 derivative for the GEVD exists for all $\xi \neq 0$ and Hájek condition (H2) is fulfilled.

Another singularity point for the GEVD is $\xi = -1/2$. This is reflected by the the Fisher information matrix $\mathcal{I}_\theta^{\mathrm{GEVD}} = \mathbf{E}_\theta \Lambda_\theta^{\mathrm{GEVD}} (\Lambda_\theta^{\mathrm{GEVD}})^{\mathrm{T}}$, which in the case of three parameters if 3×3 symmetric matrix of the form:

$$\mathcal{I}_\theta^{\mathrm{GEVD}} = \mathrm{diag}(\sigma^{-1}, \sigma^{-1}, \xi^{-1}) \begin{pmatrix} I_{\mu\mu} & I_{\mu\sigma} & I_{\mu\xi} \\ I_{\mu\sigma} & \xi^{-2} I_{\sigma\sigma} & \xi^{-2} I_{\sigma\xi} \\ I_{\mu\xi} & \xi^{-2} I_{\sigma\xi} & \xi^{-2} I_{\xi\xi} \end{pmatrix} \mathrm{diag}(\sigma^{-1}, \sigma^{-1}, \xi^{-1}).$$

We calculate all components of this Fisher information matrix an get the following

$$I_{\mu\mu} = (\xi+1)^2\Gamma(2\xi+1), \quad I_{\mu\sigma} = (\Gamma(\xi) - 2(\xi+1)\Gamma(2\xi))(\xi+1),$$

$$I_{\mu\xi} = (2(\xi+1)\Gamma(2\xi) - (\xi+2)\Gamma(\xi) - \xi\Gamma'(\xi))(\xi+1),$$

$$I_{\sigma\sigma} = (\xi+1)^2\Gamma(2\xi+1) - 2\Gamma(\xi+2) + 1,$$

$$I_{\sigma\xi} = -(\xi+1)^2\Gamma(2\xi+1) + (\xi+3)\Gamma(\xi+2) + (\xi+1)(\xi^2\Gamma'(\xi) - 1) - \xi\Gamma'(1),$$

$$I_{\xi\xi} = (\xi+1)^2\Gamma(2\xi+1) - 2\Gamma(\xi+3) - 2\xi^2(\xi+1)\Gamma'(\xi) +$$

$$+2\xi(\xi+1)\Gamma'(1) + \xi^2(\Gamma''(1) + (\Gamma'(1))^2) + (\xi+1)^2.$$

Therefore, the third Hájek condition (H3) is fulfilled, as long as $\xi \in (-1/2, 0)$ or $\xi > 0$. Then, by Proposition 2.9, generalized extreme value parametric model is continuously L_2 differentiable with the L_2 derivative $\Lambda_\theta^{\mathrm{GEVD}}$ and the Fisher information matrix I_θ^{GEVD}.

3.2.2 Smoothness of GPD

For the generalized Pareto parametric model, only the case of μ known is studied until now, therefore, we restrict ourselves to this case. We consider parameter $\theta = (\sigma, \xi) \in \Theta \subset \mathbb{R}^+ \times \mathbb{R}$.

The density function q_θ^{GPD}, given by equation (3.6), is absolutely continuous in each $\theta \in \Theta$, except the singularity point $\xi = 0$, therefore, the first Hájek condition is fulfilled for $\xi \neq 0$.

In order to check the condition (H2), first, we compute the derivative $\frac{\partial}{\partial\theta} q_\theta^{\mathrm{GPD}}(x) = \Lambda_\theta^{\mathrm{GPD}}(x)\, q_\theta^{\mathrm{GPD}}(x)$. Similarly to the GEVD case, the L_2 derivative for the GPD model takes up the structure of the parameter, i.e. it consists of two coordinates, one for the scale and one for the shape. Therefore, with the same notations as before, we obtain L_2 derivative for the GPD model of the form

$$\Lambda_\theta^{\mathrm{GPD}} = (\Lambda_\sigma^{\mathrm{GPD}}, \Lambda_\xi^{\mathrm{GPD}})^{\mathrm{T}}, \tag{3.13}$$

where

$$\Lambda_\sigma^{\mathrm{GPD}} = -\frac{1}{\sigma} + \frac{(1+\xi)z}{\sigma s}, \quad \Lambda_\xi^{\mathrm{GPD}} = \frac{1}{\xi^2}\log(s) - (\frac{1}{\xi} + 1)\frac{z}{s}. \tag{3.14}$$

Hence, for $\xi \neq 0$ the L_2 derivative exists and the condition (H2) is fulfilled.

For the last Hájek condition we find the Fisher information matrix for the GPD model. We obtain the following

$$I_\theta^{\mathrm{GPD}} = \mathbf{E}_\theta \Lambda_\theta^{\mathrm{GPD}} (\Lambda_\theta^{\mathrm{GPD}})^{\mathrm{T}} = \frac{1}{1+2\xi} \begin{pmatrix} \frac{1}{\sigma^2} & \frac{1}{\sigma(\xi+1)} \\ \frac{1}{\sigma(\xi+1)} & \frac{2}{(\xi+1)} \end{pmatrix}. \tag{3.15}$$

From this matrix we get the second singularity point $\xi = -1/2$. Then, the third Hájek condition (H3) is fulfilled as long as $\xi \in (-1/2, 0)$ or $\xi > 0$ and, by the Proposition 2.9, generalized Pareto parametric model is continuously L_2 differentiable with the L_2 derivative $\Lambda_\theta^{\mathrm{GPD}}$ and the Fisher information matrix I_θ^{GPD}.

3.3 Robustness properties of the GPD estimator

In this Section we make brief overview of the existing results concerning the different estimators for the scale and the shape parameters of the generalized Pareto model and discuss their global and local robustness properties. We focus mainly on the results presented in the papers of Ruckdeschel and Horbenko (2010) and Ruckdeschel and Horbenko (2012), and in the PhD-thesis of Horbenko (2011). In particular, these authors consider maximum likelihood and skipped maximum likelihood estimators, moment-based and Cramér-von-Mises minimum distance estimators.

In the paper of Ruckdeschel and Horbenko (2012), authors discussed two options for the highly-robust, easy-to-compute initial estimators,i.e. Pickands-type and Location-Dispersion-type estimators, e.g. kMedMAD (see Horbenko (2011, Ch. 6.5)). PhD-thesis of Horbenko (2011) also contains all optimally robust estimators which were presented in Section 2.5. In the listed researches the estimators are computed together with their finite sample breakdown points (FSBP), influence functions and statistical accuracy measured by asymptotic bias, variance, and mean squared error.

3.3.1 Likelihood based estimators

For the parameters of the GPD model there is no explicit solution of the MLE. Influence function of the MLE is of the form:

$$IF(x, \mathrm{MLE}, Q_\theta^{\mathrm{GPD}}) = (I_\theta^{\mathrm{GPD}})^{-1}\Lambda_\theta^{\mathrm{GPD}}(x) = (\psi_\xi(x), \psi_\sigma(x))^{\mathrm{T}},$$

for the score function $\Lambda_\theta^{\mathrm{GPD}}$ as in (3.13) and the Fisher information I_θ^{GPD} from (3.15).

According to asymptotic minimax theorem (see Rieder (1994, Thm. 3.3.8)), the MLE attains the smallest asymptotic variance among all asymptotically linear estimators. For the shortness we introduce additional notations $z := \frac{x-\mu}{\sigma}$ and $u := (1+\xi z)^{-\frac{1}{\xi}}$, then influence function of the MLE consists of the following terms

$$\psi_\xi(u) = \frac{\xi+1}{\xi^2}(-(\xi^2+\xi)\log u + (2\xi^2+3\xi+1)u^\xi - (\xi^2+3\xi+1)),$$

$$\psi_\sigma(u) = \frac{\xi+1}{\xi^2}(\xi\log u - (2\xi^2+3\xi+1)u^\xi + 3\xi + 1).$$

In the paper of Ruckdeschel and Horbenko (2010, Sec. 2.2.) influence function for the skipped maximum likelihood estimators is also computed. As well as MLE, the SMLE enjoys the same asymptotic equivariance.

3.3.2 Cramér-von-Mises minimum distance estimators

Following Horbenko (2011, Sec. 6.2), the influence function for the Cramér-von-Mises minimum distance estimator for the GPD model is of the form

$$IF(x, \mathrm{MDE}, Q_\theta^{\mathrm{GPD}}) = (\mathcal{J}_\theta)^{-1}(\varphi_\xi(x), \varphi_\sigma(x))^{\mathrm{T}},$$

where the Cramér-von-Mises information matrix is given as

$$\mathcal{J}_\theta = 3(\xi+3)^2 \begin{pmatrix} \frac{18(\xi+3)}{2\xi+9} & -3\sigma \\ -3\sigma & 2\sigma^2 \end{pmatrix}.$$

The explicit terms of this influence function are the following

$$\varphi_\xi(x) = \frac{19+5\xi}{36(3+\xi)(2+\xi)} - \frac{1}{2\xi^2}(1+\frac{\xi}{\sigma}x)^{-2/\xi}\log(1+\frac{\xi}{\sigma}x) + \frac{2-\xi}{4\xi^2}(1+\frac{\xi}{\sigma}x)^{-2/\xi} -$$

$$-\frac{1}{\xi^2(2+\xi)}(1+\frac{\xi}{\sigma}x)^{-2/\xi-1},$$

$$\varphi_\sigma(x) = \frac{5+\xi}{6(3+\xi)(2+\xi)\sigma} - \frac{1}{2\xi\sigma}(1+\frac{\xi}{\sigma}x)^{-2/\xi} + \frac{1}{\xi\sigma(2+\xi)}(1+\frac{\xi}{\sigma}x)^{-2/\xi-1}.$$

Apparently, the same (asymptotic) in-/equivariance as for the MLE and SMLE holds for the Cramér-von-Mises minimum distance estimator as well.

3.3.3 Method of moments estimators

Method of moments estimators for the shape and the scale parameters of the GPD are following

$$\hat{\boldsymbol{\xi}}^{\mathrm{MME}} = \frac{m_2 - 2m_1^2}{m_2 - m_1^2}, \quad \hat{\boldsymbol{\sigma}}^{\mathrm{MME}} = \frac{m_1 m_2}{2(m_2 - m_1^2)},$$

with

$$m_1 = \frac{\sigma}{1-\xi}, \quad m_2 = \frac{2\sigma^2}{(1-\xi)(1-2\xi)}.$$

The influence function of the method of moments estimator for the GPD is computed in the PhD-thesis of Horbenko (2011, Sec. 6.3) and has the following form

$$IF(x, \mathrm{MME}, Q_{\sigma,\xi}^{\mathrm{GEVD}}) = D(x - m_1, x^2 - m_2)^{\mathrm{T}},$$

where matrix D is also calculated, i.e.

$$D = \begin{pmatrix} \frac{2(\xi-1)^2(2\xi-1)}{\sigma} & \frac{(2\xi-1)^2(\xi-1)^2}{2\sigma^2} \\ (4\xi-3)(\xi-1) & \frac{(2\xi-1)^2(\xi-1)}{2\sigma} \end{pmatrix}.$$

3.3.4 Starting estimators

In the PhD-thesis of Horbenko (2011), author estimates the GPD parameters in a robust way, taking as the starting estimator Pickands (-type) estimator or applying newly developed Median-kMAD (kMedMAD) method. Definition of the Pickands estimator is introduced in the Section 2.4.6, hence, here we only present the estimators and the influence function for the GPD parameters. As for the kMedMAD estimator, we give its definition and link to Horbenko (2011, Ch. 6.5) and the paper of Ruckdeschel and Horbenko (2010) for the detailed computation.

Quantile-based estimators

For this Section we let location parameter of GPD be known, i.e. $\mu = 0$ and denote empirical 50% and 75% GPD quantiles as F_2 and F_3 correspondingly. Then, Pickands estimators for the GPD scale and shape parameters are

$$\hat{\boldsymbol{\xi}}^{\mathrm{PE}} = \frac{1}{\log 2} \log\Big(\frac{F_3 - F_2}{F_2}\Big), \quad \hat{\boldsymbol{\sigma}}^{\mathrm{PE}} = \hat{\boldsymbol{\xi}} \frac{F_2^2}{F_3 - 2F_2}.$$

Influence function for the Pickands estimator can be separated to two coordinates, one for each parameter, i.e.

$$IF(x, \mathrm{PE}, Q^{\mathrm{GPD}}_{\sigma,\xi}) = (IF_\xi(x, \mathrm{PE}, Q^{\mathrm{GEVD}}_{\sigma,\xi}), IF_\sigma(x, \mathrm{PE}, Q^{\mathrm{GEVD}}_{\sigma,\xi}))$$

Following Rieder (1994, Ch. 1.5) one can also calculate influence function for each parameter by the following expression

$$IF_k(x, \mathrm{PE}, Q^{\mathrm{GPD}}_{\sigma,\xi}) = h_{k1} \frac{0.75 - \mathbb{1}(x \le F(0.75))}{1/\sigma(0.25)^{1+\xi}} + h_{k2} \frac{0.5 - \mathbb{1}(x \le F(0.5))}{1/\sigma(0.5)^{1+\xi}},$$

where k distinguishes shape and scale parameter IF, function F is the quantile function of the GPD and $h_{ki}, i = 1, 2$ are the weights, i.e. $h_{ki} = \partial\hat{\mathbf{k}}/\partial F_{i+1}$. For the Pickands estimator we also have (asymptotic) equivariance.

Location-dispersion estimators

For the computation of the estimates for some specific parametric family of probability measures with the scale and shape parameters, one can use location-dispersion estimator. The idea of this estimator is to match location and dispersion functionals against empirical counterparts. There is the `R`-function `LDEstimator` in the package `RobExtremes`, which provides a general way to do that.

In particular, in our research we focus on the scale and shape estimators, presented in the paper of Ruckdeschel and Horbenko (2010). They are based on the matching empirical median, denoted by $\hat{\mathbf{m}}_n$, and the median of absolute deviations (MAD) $\hat{\mathbf{M}}_n$ against their population counterparts m and M within the GPD model. For $k > 0$ we define

$$\text{kMAD} := \inf\left\{t > 0 | Q^{\text{GPD}}_{\sigma,\xi}(m + kt) - Q^{\text{GPD}}_{\sigma,\xi}(m - t) \geq 1/2\right\}$$

where $k = 1$ reproduces MAD. Corresponding estimator for ξ and σ is called *kMedMAD* and consists of two estimating equations, one for the median and one for the respective kMAD. The first equation, using the quantile function of the GPD, $F^{\text{GPD}}_{\sigma,\xi}$, converts to the following

$$m = m(\xi, \sigma) = F^{\text{GPD}}_{\sigma,\xi}(0.5) = \frac{\sigma(2^{\xi} - 1)}{\xi}.$$

The second equation has to be solved numerically, searching for the unique root M of the function

$$f_{m,\xi,\sigma;k}(M) = -\nu_+ + \nu_- - 0.5,$$

where

$$\nu_+ := (1 + \xi\frac{kM + m}{\sigma})^{-\frac{1}{\xi}}, \quad \nu_- := (1 + \xi\frac{m - M}{\sigma})^{-\frac{1}{\xi}}.$$

Influence function for such estimator is computed in the paper of Ruckdeschel and Horbenko (2010). Moreover, it is shown that the reasonable choice of k is the value $k = 10$. This estimator is also implemented in the `R`-package `RobExtremes` and the function for it is named `medkMAD`.

Hybrid estimator

There is the essential drawback of the kMedMAD estimator. Solving corresponding equations for it, with the value $k = 10$, can fail even for the small sample size ($n = 40$). To be safe from such fails hybrid estimator *Hybr* can be used. By default this estimator returns kMedMAD for $k = 10$, but if procedure fails, it tries another values of k. Hybrid estimator takes starting value $k = 3.23$, then, each value of k which results in a failure, multiplied by factor 3. It stops either when success has been achieved and returns the first estimator which did not fail, or when after 20 attempts with the different values of k have been made. `medkMADhybr` is the `R`-function for this type of the estimator. It is available in the package `RobExtremes`. For more details about hybrid estimator see Horbenko (2011, Ch. 6.5).

3.4 Robustness properties of the GEVD estimators

Analogically to the case of GPD from the previous Section, here we construct some estimators, presented from Section 2.4, for the GEVD, what has not been done yet. Further, we analyze their robustness properties and compute influence function for each estimator. We start with the classical moment based estimator and then pass over to the Cramér-von-Mises minimum distance estimator. All results described in this Section are new and belong to my own results.

3.4.1 Method of moments estimators

As we mentioned in Section 2.4.5, method of moments estimator can be computed by matching the sample moments with the corresponding distribution moments. In the case of GEVD(μ, σ, ξ), with known location parameter μ, the first and the second empirical moments are enough to estimate scale and shape. These first two theoretical moments are respectively:

$$m_1 = \frac{\sigma(g_1 - 1)}{\xi}, \quad m_2 = \sigma^2 \frac{g_2 - 2g_1 + 1}{\xi^2} \tag{3.16}$$

where we used notations g_1 and g_2 for the corresponding Gamma functions, i.e. $g_k := \Gamma(1-k\xi)$, $k = 1, 2$. here we restrict ourselves to $\xi < 0.5$, so that second moment is finite.

in order to construct method of moments estimator, we have to solve system of equations (3.16) w.r.t. the unknown parameters ξ and σ. We express the scale from the first equation in terms of the first moment and the shape, i.e.

$$\sigma = \frac{\xi m_1}{g_1 - 1}$$

and plug it in the second equation, so we get

$$m_2 = \frac{m_1^2(g_2 - 2g_1 + 1)}{(g_1 - 1)^2}.$$

For the shape we do not get the explicit solution. Estimator $\hat{\boldsymbol{\xi}}^{\mathrm{MME}}$ is the value of ξ which satisfies the following equality

$$m_1^2 - m_2 = -m_1^2\Gamma(1 - 2\xi) + m_2\Gamma^2(1 - \xi) + 2(m_1^2 - m_2)\Gamma(1 - k\xi),$$

and for parameter σ we get corresponding estimator

$$\hat{\boldsymbol{\sigma}}^{\mathrm{MME}} = \frac{\hat{\boldsymbol{\xi}}^{\mathrm{MME}} m_1}{\Gamma(1 - \hat{\boldsymbol{\xi}}) - 1}.$$

Theorem 3.18. *The influence function of the method of moments estimator for the GEVD has the following form*

$$IF(x, MME, Q^{GEVD}_{\sigma,\xi}) = D(x - m_1, x^2 - m_2)^T,$$

where matrix D consists of the terms

$$d_{11} = \frac{2\xi^2(g_2 - 2g_1 + 1)}{\sigma(2(g_1'\xi - 1)(g_2 - 2g_1 + 1) + \xi(1 - g_1)(g_2' - 2g_1'))},$$

$$d_{12} = \frac{-\xi((g_2' - 2g_1')\xi - 2(g_2 - 2g_1 + 1)}{\sigma(2(g_1'\xi - 1)(g_2 - 2g_1 + 1) + \xi(1 - g_1)(g_2' - 2g_1'))},$$

$$d_{21} = \frac{-\xi^3(g_1 - 1)}{\sigma(2(g_1'\xi - 1)(g_2 - 2g_1 + 1) + \xi(1 - g_1)(g_2' - 2g_1'))},$$

$$d_{22} = \frac{\xi^2(g_1'\xi - g_1)}{\sigma(2(g_1'\xi - 1)(g_2 - 2g_1 + 1) + \xi(1 - g_1)(g_2 - 2g_1'))},$$

for the Gamma function $g_k = \Gamma(1 - k\xi)$ *and* $\xi < 0.5$.

Proof of this Theorem can be found in Appendix A.1.

3.4.2 Cramér-von-Mises minimum distance estimators

Theorem 3.19. *The influence function of the Cramér-von-Mises minimum distance estimator is of the from*

$$IF(x, MDE, Q^{GEVD}_{\sigma,\xi}) = \mathcal{J}_\theta^{-1}\Big(\varphi_\xi(x), \varphi_\sigma(x)\Big)^T,$$

where Cramér-von-Mises information matrix contains the following terms

$$J_{11} = \frac{1}{27\xi^2\sigma}\Big(- \Gamma''(3) + 2\ln 3\Gamma'(3) - 2(\ln(3))^2 + \frac{2}{\xi 3^\xi}(\Gamma'(\xi + 3) - \ln 3\Gamma(\xi + 3)) -$$

$$-\frac{2}{\xi}(\Gamma'(3) - 2\ln 3) - \frac{1}{\xi^2 3^{2\xi}}\Gamma(2\xi + 3) + \frac{2}{\xi^2 3^\xi}\Gamma(\xi + 3) - \frac{2}{\xi^2}\Big),$$

$$J_{12} = \frac{1}{27\xi^2\sigma^2}\Big(\frac{1}{3^\xi}(\Gamma'(\xi + 3) - \ln 3\Gamma(\xi + 3)) - \Gamma'(3) + 2\ln 3 + \frac{1}{\xi 3^{2\xi}}\Gamma(2\xi + 3) -$$

$$-\frac{2}{\xi 3^\xi}\Gamma(\xi + 3) + \frac{2}{\xi}\Big), \quad J_{22} = \frac{1}{27\xi^2\sigma^3}\left(-\frac{1}{3^{2\xi}}\Gamma(2\xi + 3) + \frac{2}{3^\xi}\Gamma(\xi + 3) - 2\right).$$

Functions $\varphi_\xi(x)$ *and* $\varphi_\sigma(x)$ *for the influence function are computed and of the form*

$$\varphi_\xi(x) = \frac{1}{\xi\sigma}\Big(\frac{1}{2^{\xi+2}\xi}\Gamma\Big(\xi + 2, (1 + \xi\frac{x - \mu}{\sigma})^{-\frac{1}{\xi}}\Big) - \frac{1}{4}\Big((1/\xi - \log 2)\Gamma(2, u) +$$

$$+\Gamma'\Big(2,(1+\xi\frac{x-\mu}{\sigma})^{-\frac{1}{\xi}}\Big)\Big)+\Big(\frac{1}{3^{\xi+2}}-\frac{1}{2^{\xi+2}}\Big)\frac{1}{\xi}\Gamma(\xi+2)+\frac{5}{36\xi}+\frac{5}{36}\Gamma'(2)+\frac{\log 3}{9}-\frac{\log 2}{4}\Big),$$

$$\varphi_\sigma(x)=\frac{1}{\xi\sigma^2}\Big(\frac{1}{4}\Gamma\Big(2,(1+\xi\frac{x-\mu}{\sigma})^{-\frac{1}{\xi}}\Big)-\frac{1}{2^{\xi+2}}\Gamma\Big(\xi+2,(1+\xi\frac{x-\mu}{\sigma})^{-\frac{1}{\xi}}\Big)+$$

$$+\Big(\frac{1}{2^{\xi+2}}-\frac{1}{3^{\xi+2}}\Big)\Gamma(\xi+2)-\frac{5}{36}\Big).$$

Proof of this Theorem one can be found in Appendix A.2.

3.4.3 Starting estimator for GEVD and GPD

As we mentioned in Section 3.3.4, author of the PhD-thesis Horbenko (2011) tried the kMedMAD method and the Pickands (-type) estimators as the starting estimators for the GPD parameters. Although kMedMAD worked decently well for a wide range of shape parameters, still, it failed from time to time. In the case of the GEVD for the kMedMAD and, similarly, for the Pickands estimator, this was much worse and the starting estimator failed in many occasions. Therefore, there was need for some improvement.

The first promising idea was to use the Cramér-von-Mises MDE. But in the case of the multidimensional parameter this involves a call to the `R`-function `optim`, hence, needs a starting estimator again. But this is not as bad as it appears on the first glance.

For the one-dimensional parameter Cramér-von-Mises minimum distance estimator uses the `R`-function `optimize` as a line search, which only needs a reasonable search interval. Therefore, the idea is to fix some value of the shape parameter ξ and robustly determine scale σ by Cramér-von-Mises MDE. In this way we compute an admissible starting estimator for the joint-estimation of both parameters σ and ξ, i.e. using notations from the Chapter 2, estimation of the parameter $\theta=(\sigma,\xi)\in\Theta\subset\mathbb{R}^+\times\mathbb{R}$.

The only drawback of this idea is the deterministically chosen starting value for the shape ξ. Moreover, we are not the only ones to fix ξ deterministically (see `R`-packages `evir`, `isevm` and others).

In order to improve this situation, we decide to use a deterministic grid of the shape values to start with. For each start ξ we get the corresponding Cramér-von-Mises MDE scale σ and in the second step, we get the joint minimum distance estimate of both parameters, together with the corresponding Cramér-von-Mises distance value. By means of the latter, we can order the obtained (σ,ξ)-pairs so that the "optimal" θ then is optimal for the set of all starting ξ-s. This strategy is cumbersome in the sense, that it involves multiple starting values, but this also adds an insurance not to miss the best (σ,ξ)-pair due to the falsely chosen suboptimal ξ in the beginning.

Summarizing, we use the following algorithm for the GPD parameter estimation:

St.1. Try out the hybrid estimator using R-function `medkMADhybr`,

if `medkMADhybr` does not fail and no errors appear, we get estimate $\theta_0 = (\sigma_0, \xi_0)$:

St.2. Evaluate the Cramér-von-Mises MDE, with the R-function `MDEstimator`, for the pair (σ, ξ) with the starting value obtained from the hybrid estimation $\theta_0 = (\sigma_0, \xi_0)$, to get new value of the parameter and distance value for it.

St.3. Check whether this parameter estimate is admissible and set current best value of distance to d_0.

if `medkMADhybr` fails:

St.2. Run through the prescribed grid for the shape parameter $\{\xi_i^0\}$. For each fixed value of the shape from the grid ξ_i^0 determine the respective univariate Cramér-von-Mises MDE σ_i^0.

St.3. Use the pair (σ_i^0, ξ_i^0) as the start for the Cramér-von-Mises MDE of the parameter θ to compute estimate $\theta_i = (\sigma_i, \xi_i)$.

St.4. Afterwards, check the admissibility of this estimate, i.e. if the condition $1 + \sigma/\xi(x - \mu) > 0$ is satisfied.

St.5. If it is admissible, check whether θ_i generates new optimal distance d_i.

St.6. If so, store current optimal distance d_i and respective θ_i and return the optimal admissible pair (σ_i, ξ_i).

For the GEVD algorithm is similar. The only difference is that instead of the hybrid estimator `medkMADhybr`, we start with the Pickands estimator (R-function `PickandsEstimator`) taken by default.

3.5 Software infrastructure

The order of the packages in this Section is caused by their appearance.

The R-package `evd` is created by Stephenson (2002) and focused on the distributions which often arised in the analysis of the extreme values. Moreover, it contains functions for the simulation and calculation of the distribution, density and quantile functions, for the various univariate and multivariate parametric extreme value distributions. The package provides fitting functions which calculate the maximum likelihood estimates

for the univariate and bivariate models, and for the univariate and bivariate threshold models. The current version of this package is version 2.3.0.

Later, the R-package `evir` was submitted to `CRAN` by Pfaff and McNeil (2012). This package is primarily designed for applying extreme value methodology to the financial data. It implements standard stationary univariate extreme value modeling, including maximum likelihood fitting of the GPD and GEVD. The package provides functions for the calculating expected shortfalls and quantiles, for the extracting records and declustering, and for the estimating the extremal index. The version 1.7.3 of the package is available in `CRAN`.

One of the main R-packages on the extreme value statistics is the package `ismev`, created by Heffernan and Stephenson. (2012). It is based on the book of Coles (2001), which provides an introduction to the topic at a relatively simple statistical level. The functions of the package cover estimation of the distributions for the block maxima and threshold model approaches. The package includes functions for diagnosing the quality of the fitted distributions e.g., probability and qq-plots, histograms, as well as the functions useful for the selection of the appropriate threshold for the threshold models. The current version of this package is 1.39.

The R-package `extRemes` is essentially a graphical user interface to the package `ismev`, created by Gilleland and Katz (2011) the same year. Nevertheless, it includes some additional functionality. In particular, for the GEVD and GPD it allows L-moments estimation for the stationary case and has some capability for the extremal index and the number of clusters calculation. The last version of the package available in `CRAN` is 2.0.1.

The R-package `fExtremes` was built by Wuïtz in 2009 as an open source solution for teaching financial market analysis (see `Rmetrics` software collection). It provides explicit calculation of the financial measure known as value-at-risk. The package is developed using codes from other R-packages, e.g. `evd` and `evir`. Functions for the univariate simulation and distribution functions are available, as well as the estimation of the stationary models for the GEVD and GPD using maximum likelihood and probability weighted moments. Current version of the package 3010.81 can be found in Wuertz (2013).

Other R-packages on the extreme value theory are `POT` discussed in Ribatet (2007) and focused only on the modeling of exceedances over a threshold; `SpatialExtremes` (see Ribatet and Singleton (2013)) devoted to the modeling of spatial extremes and others. More detailed list of the packages and deeper description of them, one can find in the article of Gilleland et al. (2013).

The R-package RobExtremes, already mentioned in the Section 2.8, also covers scale-shape models with Gamma, Weibull, and extreme value distributions, i.e. GPD and GEVD models. As it was mentioned, it contains infrastructure for the LD estimators and optimally-robust estimation with speed-up by interpolation technique.

3.6 Conclusions

In the beginning of this Chapter we described the main ideas of the Extreme Value Theory and gave some overview of the sources concerning this topic. In the first Section we introduced some basic concepts, used further in this thesis, including two main types of the extreme value distributions: generalized extreme value and generalized Pareto distributions. The importance of these distributions is caused by two main theorems of extreme value theory regarding asymptotic distribution of extreme order statistics, which were also presented in this Chapter.

Next, we devoted one Section to the discussion of the smoothness of the generalized extreme value and generalized Pareto parametric models, checking Hájek conditions, introduced in the previous Chapter. We obtained the explicit form of L_2 derivative and Fisher information matrix for each model.

Then, we discussed robustness properties of some estimators for the generalized Pareto parametric model, giving the brief overview of the published results. Afterwards, we proved similar results for the classical moment based and Cramér-von-Mises minimum distance estimators for the GEVD scale and shape parameters.

At the end of the Chapter, we presented software infrastructure concerning extreme value theory available in R and stressed the functionality of our package RobExtremes.

Part II

Interplay of foundations

Chapter 4

Structured models

4.1 Regression models

So far GEVD and GPD context potentially ignored additional information available for each observation. This information could make our statements more precise in the sense that parameters of extreme value distributions could now vary from the observation to observation. That is how we come to the idea of regression in our approach.

From the other side, one could also get more precise statements about the time sequences of the observations, i.e. if we have the dynamic model it would show how observation today depends on the observation yesterday. Therefore, we set up time-series models for GEVD and GPD context and state-space models provide very flexible setup to do that.

As we already mentioned in the Section 1.1, time series models with the time-varying parameters categorized onto two classes: observation driven models and parameter driven models.

Observation driven time dependency leads to the time variation of the parameters by making them dependent on their own lagged values, past observations, and exogenous variables, or even some specific functions of them. Although the parameters are stochastic, they are usually predictable given the past information. Hence, in this context we introduce state-space models and Kalman filtering procedure as the estimation procedure for the state in the presence of noise.

The alternative to the observation driven models are *parameter driven models*. Here parameters are stochastic processes, which are subject to their own source of error. Therefore, the parameters are not perfectly predictable given the past and the current observations. Typical example of the parameter driven time dependency is the generalized linear model, which is further introduced.

In this Chapter we give the overview of the theory and notions needed for the next two Chapters of the thesis, where our main achievements are presented. Therefore, here we cover two main topics, i.e. Kalman filtering for the state-space models and L_2 differentiability of the generalized linear models.

One remark to be done here is related to the title of the Section. We claim that both topics can be considered in the regression context. It is obvious for the generalized linear models, whereas for the SSM with Kalman filtering one needs some more explanations of this statement.

The idea of connection between the Kalman procedures and the regression theory can be obtained from the lemma, which is proved in the articles of Dunkan and Horn (1972) and Cipra and Romera (1991), and later, is taken over in the PhD thesis Ruckdeschel (2001, Ch. 3). This lemma states, that classical Kalman filter can be considered as a weighted least-squares regression estimator. Proof of this statement one can find in Ruckdeschel (2001, Lemma 3.1.1). Moreover, in the next Section we show that any state-space model has regression representation.

4.2 Dynamics

As we already mentioned, here we are going to introduce the state-space models, which build a flexible but still manageable class of the dynamic models. These models are useful for a wide range of the applications. As an example, master thesis of Pupashenko (2011) is focused on the engineering application in the context of the GPS problem with linear and non-linear state-space formulations.

Kalman filter is first described and partially developed in the technical papers of Kalman (1960) and Kalman and Bucy (1961). Together with the Kalman filtering procedures and their extensions, state-space models become even more useful. Nevertheless, for the full use of the Kalman procedures we lack robustness. We discuss this drawback and our ways out of it later, in Chapter 5.

4.2.1 State-space models

The mathematical notion for a fixed rule which describes the time dependence of a point in the geometrical space is called *dynamical system*. At any specific time, dynamical system has a *state*, given in the form of the set of real numbers or the vector, that can be represented by the point in an appropriate state (geometrical) space. To describe these dynamical systems one can use *state-space representation*.

State-space models (SSM) are originally developed by control engineers for some navigation applications. They are also very useful in many types of the time-series problems, e.g. forecasting problem. To introduce state-space models we focus on the books of Chatfield (1996) and Brockwell and Davis (2002) and link to them for the deeper study.

It is typical that when we measure any sort of a signal, we get it contaminated by some noise, so that the actual observation is given as some combination of the signal and noise. As in any dynamical system, the signal in the state-space model can be expressed in terms of so-called state variables, which constitute the state vector. This vector describes the state of the whole system at some specific moment of time. The state vector cannot be observed directly, hence we use the observations to make inference about the state vector.

General state-space model is composed from two equations. We consider SSM consisting of an unobservable p-dimensional state X_t and the time series of q-dimensional observations Y_t . The state-space models are based on the Markov property, hence the state vector summarizes all information from the past that is necessary to predict the future. Therefore, the first equation of the state-space model, so-called *state equation*, is the following

$$\text{State equation:} \quad X_t = f_t(X_{t-1}, u_t, v_t), \tag{4.1}$$

with $p-$dimensional random vectors, called innovations v_t, some user defined control u_t and sequence of the smooth known state update functions f_t.

The second equation of the state-space model is called *observation equation*. It is an expression of the q-dimensional observations in terms of the states, involving some additional error ε_t, user defined control w_t and corresponding sequence of the smooth known output functions z_t,

$$\text{Observation equation:} \quad Y_t = z_t(X_t, w_t, \varepsilon_t) \tag{4.2}$$

In the ideal setup we work in a Gaussian context, i.e. we assume

$$v_t \sim^{indep.} \mathcal{N}_p(0, Q_t), \qquad \varepsilon_t \sim^{indep.} \mathcal{N}_q(0, V_t), \qquad X_0 \sim \mathcal{N}_p(a_0, Q_0), \tag{4.3}$$

and $\{X_0, v_s, \varepsilon_t; s, t \in \mathbf{N}\}$ stochastically independent.

If there exists a state-space model (4.1) and (4.2) for the time series $\{Y_t\}$ we say that this time series has the *state-space representation*.

In general, functions f_t and z_t are arbitrary. Here, as the special case of the general SSM, we also consider its linearization, *linear SSM*. The state and observation equations

of the linear SSM are autoregressive processes of the first order and the system can be written in the following matrix form:

$$\text{State equation:} \quad X_t = F_t X_{t-1} + v_t, \tag{4.4}$$

$$\text{Observation equation:} \quad Y_t = Z_t X_t + \varepsilon_t, \tag{4.5}$$

for the corresponding transition matrices $F_t \in \mathbf{R}^{p\times p}$ and observation matrices $Z_t \in \mathbf{R}^{q\times p}$. The convenience of the linear state-space representation lies in the simple structure of the state equation 4.4, which makes analysis of the state process relatively simple.

For our research, we assume all hyper–parameters of the SSM, i.e. F_t, Z_t, Q_t, V_t, a_0, to be known.

4.2.2 Kalman filter

The most important problem in the state-space modeling is the estimation of the signal in the presence of the noise. In other words, we are interested in the "best estimator" of the unobservable states X_t by means of the observations Y_t.

Following abbreviation of the paper of Ruckdeschel et al. (2014b), we denote the series of the observations as $Y_{1:t} = (Y_1, \ldots, Y_t)$, $Y_{1:0} := \emptyset$ and σ-algebra generated by this series as $\sigma(Y_{1:t})$. By the "best estimator" we mean that it is the minimum mean square error estimator, i.e. the solution to the following equation

$$\mathbf{E}\big|X_t - f_t\big|^2 = \min_{f_t}, \qquad f_t \text{ measurable w.r.t. } \sigma(Y_{1:s}) \tag{4.6}$$

The general solution of the problem (4.6) is the conditional expectation $\mathbf{E}[X_t|Y_{1:s}]$, which is usually rather expensive to compute. Therefore, Kalman (1960) introduced another way to obtain the "best estimator" for the state vectors when the next observation becomes available, well-known as *Kalman filter* (KF). Moreover, if all observations are given in advance, we can further improve estimation procedure by using so-called Kalman smoother, which computes the estimate of the state vector based on all observation data. Next, we present both procedures in the classical setup.

4.2.2.1 Classical Kalman procedures

Classical Kalman filter

Here we present the classical Kalman filter for the linear state-space model (4.4) and (4.5). Kalman filter is an recursive procedure which has three stages called initialization, prediction and correction.

The first step, initialization, defines the base of the recursions. Since in the assumptions (4.3) the initial state vector has multivariate Gaussian distribution, i.e. $X_0 \sim N_p(a_0, Q_0)$, we take these distribution parameters as the initial values.

On the prediction step of the filter we compute the *best one-step predictor* $X_{t|t-1}$, which is the random vector whose components are the best linear mean square predictors in terms of all components of the observations $Y_1, ..., Y_{t-1}$. Afterwards, we pass over to the correction step and obtain the *best estimator* $X_{t|t}$, based on the observations $Y_1, ..., Y_t$.

More precisely, we get the following recursive scheme to compute the optimal linear filter:

$$\text{Initialization:} \quad X_{0|0} = a_0, \qquad \Sigma_{0|0} = Q_0; \tag{4.7}$$

$$\text{Prediction:} \quad X_{t|t-1} = F_t X_{t-1|t-1}, \qquad \Sigma_{t|t-1} = F_t \Sigma_{t-1|t-1} F_t^{\mathrm{T}} + Q_t; \tag{4.8}$$

$$\text{Correction:} \quad X_{t|t} = X_{t|t-1} + K_t \Delta Y_t, \quad \Delta Y_t = Y_t - Z_t x_{t|t-1} \tag{4.9}$$

$$K_t = \Sigma_{t|t-1} Z_t^{\mathrm{T}} C_t^{-1} \qquad \Sigma_{t|t} = (\mathbf{I}_p - K_t Z_t)\Sigma_{t|t-1}, \tag{4.10}$$

$$C_t = Z_t \Sigma_{t|t-1} Z_t^{\mathrm{T}} + V_t \tag{4.11}$$

with *Kalman gain* K_t and covariance matrices $\Sigma_{t|t} = \mathrm{Cov}(X_t - X_{t|t})$ and $\Sigma_{t|t-1} = \mathrm{Cov}(X_t - X_{t|t-1})$.

One can notice that all steps of the filtering procedure inherit the linearity of the model, what makes KF very easy to use.

Classical Kalman smoother

So far we considered the best estimator for the state vector X_t in terms of the observations up to time t, i.e. taking in account only the "past" information. As we have mentioned, there is a way to improve the estimator considering also the "future" observations related to the state vector. This method is called *Kalman smoother*.

In this thesis, for simplicity, we assume that all hyper-parameters of the state-space model are given. In the other case, i.e. when hyper-parameters have to be estimated, one

can use the *Expectation-Maximization-algorithm* (EM-algorithm), which can be found in the article of Shumway and Stoffer (1982). EM-algorithm is an efficient iterative procedure to compute the MLE in the presence of missing or hidden data. In the master-thesis of Pupashenko (2011) EM-algorithm was applied explicitly to the linear and quadratic state-space models.

In many situations, in particular for the estimation of the hyper-parameters applying EM-algorithm, it is common to use filtered values in retrospective, accounting for the information (observations) available in the meantime, i.e. use Kalman smoother.

Kalman smoother is the backward recursion, which takes the filtered estimate as the initial condition. For the observations set $\{Y_1, ..., Y_T\}$ this procedure can be described by the following scheme (see Anderson and Moore (1990, Sec.7.4, (4.5))):

$$X_{t|T} = X_{t|t} + J_t(X_{t+1|T} - X_{t+1|t}), \qquad J_t = \Sigma_{t|t} F_t^{\mathrm{T}} \Sigma_{t+1|t}^{-1} \tag{4.12}$$

with smoothing covariance:

$$\Sigma_{t|T} = \Sigma_{t|t} + J_t(\Sigma_{t+1|T} - \Sigma_{t+1|t}) J_t^{\mathrm{T}}. \tag{4.13}$$

This recursive procedure is also easy and especially useful for online-purposes.

4.2.2.2 Extended Kalman procedures

Extended Kalman filter

If we consider the general (nonlinear) state-space model (4.1) and (4.2), we can use the *extended Kalman filter* (see Wan and van der Merwe (2002)). The main idea of this approach is to approximate the nonlinear system with the linear, using first-order Taylor approximation. In this way one gets the following recursive scheme:

$$\begin{array}{lcl}
\text{Initialization:} & X_{0|0} = a_0, & \Sigma_{0|0} = Q_0; \\
\text{Prediction:} & X_{t|t-1} = f_t(X_{t-1|t-1}, u_t, \bar{v}_t), & \Sigma_{t|t-1} = F_t \Sigma_{t-1|t-1} F_t^{\mathrm{T}} + B_t Q_t B_t^{\mathrm{T}}; \\
\text{for} & F_t = \frac{\partial}{\partial x} f_t(x, u_t, v_t)|_{x_{t-1|t-1}}, & B_t = \dfrac{\partial}{\partial v} f_t(X_{t-1|t-1}, u_t, v)|_{v_t}, \\
\text{Correction:} & X_{t|t} = X_{t|t-1} + K_t \Delta Y_t, & \Delta Y_t = Y_t - z_t(x_{t|t-1} w_t, \bar{\varepsilon}_t) \\
& K_t = \Sigma_{t|t-1} Z_t^{\mathrm{T}} C_t^{-1} & \Sigma_{t|t} = (\mathbb{I}_p - K_t Z_t) \Sigma_{t|t-1}, \\
& C_t = Z_t \Sigma_{t|t-1} Z_t^{\mathrm{T}} + V_t & \\
\text{for} & Z_t = \frac{\partial}{\partial x} z_t(x, w_t, \varepsilon_t)|_{x_{t-1|t-1}}, & D_t = \dfrac{\partial}{\partial \varepsilon} f_t(X_{t-1|t-1}, w_t, \varepsilon)|_{\varepsilon_t}.
\end{array}$$

Extended Kalman smoother

Similarly, using linearization, one can write corresponding *extended Kalman smoother* recursive equation in the form:

$$X_{t|T} = X_{t|t} + J_t(X_{t+1|T} - X_{t+1|t}), \qquad J_t = \Sigma_{t|t} F_t^{\mathrm{T}} \Sigma_{t+1|t}^{-1},$$

for Jacobian matrices $F_t = \frac{\partial}{\partial x} f_t(x, u_t, v_t)|_{x_{t-1|t-1}}$.

4.3 Regression case

It is clear, that for the scale-shape models, e.g. GEVD and GPD models, the parameter domain is not the whole set of real numbers $\mathbb{R}$, but one can link scale and shape parameters to parameters $\beta_i \in \mathbb{R}$ with the use of the link function. This is exactly the concept of generalized linear models.

Generalized linear models were first introduced by Nedler and Wedderburn (1972) for exponential families. There is a large amount of literature on these models and we cannot refer all of them, hence we give only partial literature overview for the generalized linear models.

For the basic information about the models we mainly refer to McCullagh and Nedler (1989) and Fahrmeir and Tutz (2001). When it comes to the regularity assumptions, we use literature where focus falls on the models from exponential families, e.g. articles of Fahrmeir (1990) and Fahrmeir and Kaufmann (1985), although in some situations, exponential families are a too narrow class.

In this thesis we are mainly interested in the asymptotic results and robustness for the generalized linear models, more precisely, the local asymptotic normality in the sense of Hájek (1972) and LeCam (1970). Therefore, our goal is to obtain smoothness of the generalized linear models in terms of L_2 differentiability. For the exponential families, this has already been achieved by Schlather (1994).

In this Section we mostly focus on the book of Rieder (1994) and present the results on L_2 differentiability for linear regression models. Later, in Chapter 5, we generalize them, covering higher dimensional error distributions and case of regressors of possibly different length for each parameter.

4.3.1 Generalized Linear Models

Definition 4.1 (Generalized linear model)**.** *Generalized linear model* consists of three elements.

(1) The first component, also called *random component* of the model, specifies the conditional distribution of the response vector given the values of the explanatory variables in the model. Usually, this probability distribution is taken from the exponential family, denoted as Q_ϑ, with one or more dimensional parameter ϑ.

(2) If we regress this distribution and retain linearity, we get so-called *linear predictor*, i.e. linear combination of the regressors and regression parameters

$$\theta = X\beta = X_1\beta_1 + ... + X_p\beta_p. \tag{4.14}$$

Here regressors are prespecified functions of the explanatory variables.

(3) The last component of the model is smooth and invertible linearizing *link function*, usually denoted by l, via which linear model is related to the regressors, i.e. $\vartheta = l(\theta) = l(X\beta)$.

Remark 4.2. *One of the advantages of GLMs is that the structure of the linear predictor is the familiar structure of a linear model. Moreover, linear predictor θ may take over arbitrary real values, whereas ϑ usually is restricted as parameter of some distribution.*

Further, we closely follow Rieder (1994, Ch. 2). As an example of a structured model author considered the linear model with real-valued errors. The error distribution F is required to have finite Fisher information of the location, i.e. F is dominated by the Lebesgue measure λ and has absolutely continuous density f with derivative $\dot{f}$, s.t.

$$dF = f d\lambda, \quad I_f = \mathbf{E}(\Lambda_f)^2 < \infty, \quad \Lambda_f = -\frac{\dot{f}}{f}.$$

For the regression we can observe random and deterministic carriers, treating regressors as random variables or using some given array of regressors correspondingly. Random carriers can be typically applied to the time series models as (stochastic) past values of the observation series, e.g. in hydrology. This example is discussed among other applications, in Section 7.1. As for the deterministic carriers, one of the examples where we need them is the planned treatment in the hospital context, where each patient gets some medicine, which is not random and affects length of stay of the patient. More about medical application one can find in Section 7.2.

4.3.2 Random Carriers

The linear model may be brought back to the i.i.d. case by handling the regressors as random variables, i.e. $x_1, ..., x_n$ are i.i.d. realizations of the regressor $x \sim K$ for some probability distribution K. Then, we create n i.i.d. observations of the form

$$y_i = x_i^{\mathrm{T}}\vartheta + u_i.$$

Here $u_1, ..., u_n$ are i.i.d. copies of the error $u \sim F$, regressor x and error u are stochastically independent and parameter $\vartheta \in \Theta$.

The corresponding parametric model $\mathcal{Q} = \{Q_\vartheta | \vartheta \in \Theta\}$ can be written in the form

$$Q_\vartheta(dx, dy) = F(dy - x^{\mathrm{T}}\vartheta)K(dx) = f(y - x^{\mathrm{T}}\vartheta)\lambda(dy)K(dx). \tag{4.15}$$

It is proved in the Rieder (1994, Theorem 2.4.7), that with some additional assumptions model (4.15) is L_2 differentiable at every $\vartheta \in \Theta$.

4.3.3 Deterministic Carriers

Case of the deterministic carriers for the linear regression is also proved in Rieder (1994, Theorem. 2.4.2). In this setup we work with a given array of regressors $x_{n,i}$ for $n \geq 1$ and $1 \leq i \leq i_n$. Then, for unknown regression parameter $\vartheta \in \Theta$, we compute real-valued observations in the following way

$$y_{n,i} = x_{n,i}^{\mathrm{T}}\vartheta + u_{n,i},$$

where the errors $u_{n,1}, ..., u_{n,i_n}$ are i.i.d. with distribution F.

The corresponding probabilities of the parametric model $\mathcal{Q}_{n,i} = \{Q_{n,i,\vartheta} | \vartheta \in \Theta\}$ can be written in the form

$$Q_{n,i,\vartheta}(dy) = f(y - x_{n,i}^{\mathrm{T}}\vartheta)\lambda(dy), \tag{4.16}$$

for Lebesgue measure λ.

Next, we reformulate definition of the L_2 differentiability (Def. 2.6) in more general setting, i.e. for the stochastically independent variables which are not identically distributed (see Rieder (1994, Def. 2.3.8)).

Let $(\Omega_{n,i}, \mathcal{A}_{n,i})$ be general sample spaces, $\mathcal{M}_1(\mathcal{A}_{n,i})$ is the set of all probabilities on $\mathcal{A}_{n,i}$ for $n \in \mathbf{N}$ and $i = 1, ..., i_n$. Consider array of parametric families of probability measures $\mathcal{Q}_{n,i} = \{Q_{n,i,\vartheta} | \vartheta \in \Theta\} \subset \mathcal{M}_1(\mathcal{A}_{n,i})$, with open parameter set of finite dimension $\Theta \subset \mathbb{R}^p$.

Definition 4.3. Parametric model $\mathcal{Q} = (\otimes_{i=1}^{i_n} \mathcal{Q}_{n,i})$ is called L_2 differentiable at fixed $\vartheta \in \Theta$, if there exist an array of functions $\Lambda^Q_{n,i,\vartheta} \in L_2^k(Q_{n,i,\vartheta})$, s.t. for all $i = 1, ..., i_n$ and $n \geq 1$ holds

$$\mathbf{E}_{n,i,\vartheta} \Lambda^Q_{n,i,\vartheta} = 0, \tag{4.17}$$

and, for all $\varepsilon \in (0, \infty)$ and all $t \in \mathbb{R}^p$

$$\lim_{n\to\infty} \sum_{i=1}^{i_n} \int_{\left\{\left|t^{\mathrm{T}}(I^Q_{n,\vartheta})^{-\frac{1}{2}} \Lambda^Q_{n,i,\vartheta}\right| > \varepsilon\right\}} \left|t^{\mathrm{T}}(I^Q_{n,\vartheta})^{-\frac{1}{2}} \Lambda^Q_{n,i,\vartheta}\right|^2 dQ_{n,i,\vartheta} = 0, \tag{4.18}$$

and for all $b \in (0, \infty)$

$$\lim_{n\to\infty} \sup_{|t|\leq b} \sum_{i=1}^{i_n} \left\| \sqrt{dQ_{n,i,\vartheta+t}} - \sqrt{dQ_{n,i,\vartheta}}(1 + \frac{1}{2} t^{\mathrm{T}}(I^Q_{n,\vartheta})^{-\frac{1}{2}} \Lambda^Q_{n,i,\vartheta}) \right\|^2_{L_2^k} = 0. \tag{4.19}$$

Then, array $(\Lambda^Q_{n,i,\vartheta})$ is the L_2 derivative and $p \times p$ matrix $I^Q_{n,\vartheta} = \sum_{i=1}^{i_n} \mathbf{E}_{n,i,\vartheta} \Lambda^Q_{n,i,\vartheta} (\Lambda^Q_{n,i,\vartheta})^{\mathrm{T}}$ is the Fisher information of the parametric model $\mathcal{Q}_{n,i}$ at ϑ and time n.

Remark 4.4. *Comparing to the Rieder (1994, Def. 2.3.8), we drop the local identifiability condition $I^Q_{n,\vartheta} > 0$ with the same reasons as in Remark 2.7.*

Remark 4.5. *Parametric model $\mathcal{Q}$ is continuously L_2 differentiable at fixed $\vartheta \in \Theta$, if it is L_2 differentiable and for each sequence $h_n \to 0 \in \mathbb{R}^p$ holds*

$$\lim_{n\to\infty} \sup_{|t|\leq b} \sum_{i=1}^{i_n} \left\| \sqrt{dQ_{n,i,\vartheta+h_n}} U_{n,i,\vartheta+h_n}(t) - \sqrt{dQ_{n,i,\vartheta}} U_{n,i,\vartheta}(t) \right\|^2_{L_2^k} = 0, \tag{4.20}$$

where, for simplicity, we use additional notation for the following expression $U_{n,i,\vartheta}(t) = t^T (I^Q_{n,\vartheta})^{-\frac{1}{2}} \Lambda^Q_{n,i,\vartheta}$.

In Rieder (1994, Theorem 2.4.7) author shows, that for the deterministic carriers conditions (4.18) and (4.19) follow from the (uniform) smallness of so-called *hat matrix*, which is of the following form:

$$H_n = H_{n;i,j} = x^{\mathrm{T}}_{n,i} (\sum_{k=1}^{i_n} x_{n,k} x^{\mathrm{T}}_{n,k}) x_{n,j}. \tag{4.21}$$

This matrix should satisfy *Feller type condition*, i.e. it should get uniformly small along with its diagonal,

$$\lim_{n\to\infty} \max_{i=1,...,i_n} H_{n;i,i} = 0, \tag{4.22}$$

Remark 4.6. *Condition (4.18) is also known under the name Lindeberg condition. It is easy to check that Feller condition (4.22) follows from the Lindeberg condition, but there are simple examples which prove that vice versa statement does not hold.*

4.4 Conclusions

Main goal of this Chapter was to show the relation between two main subjects of the research, i.e. state-space models with the Kalman procedures and generalized linear models with the concept of the L_2 differentiability. First, we have explained why we treat both models as the special models of the regression analysis. Moreover, we distinguished two classes of the time series models with time-varying parameters and attached each model to the class.

Then, we gave the overview of notions and methods used for our research, i.e. definition of the state-space model, classical Kalman filter and smoother and extended Kalman procedures for nonlinear state-space model. Next, we passed over to the second subject of the research and introduced generalized linear models. We described the difference between the random and the deterministic carriers for the regression model. Then, we made deeper analysis, comparing to Section 2.1.2, of L_2 differentiability notion and presented existing results, concerning L_2 differentiability, for the linear regression.

Chapter 5

Kalman filter

As we mentioned in the previous Chapter, classical Kalman filter does not perform very well in the presence of outliers. However, there is the way to rewrite Kalman procedures in the robust way. Here the input parameters are the model distributions, so robustness should be understood in a distributional sense. We define suitable neighborhoods about the ideal model and allow for the deviations in the respective assumptions, which capture various types of outliers. Our research is mostly based on the approach of Ruckdeschel (2001, 2010c) for distributional-robust Kalman filtering.

In this Chapter we summarize main findings of the paper Ruckdeschel et al. (2014b). This paper presents new robust Kalman filters and smoothers as well as specialized versions for non-propagating outliers. This is illustrated in the GPS application and at the stylized outlier situation. Finally, the efficiency of our new procedures in comparison to competitors is discussed.

As mentioned in the cited references, there is a huge amount of the existing literature on robustifications of the Kalman procedures. Since these methods are not the main focus of the thesis, here we only refer the reader to review few articles as Ershov and Lipster (1978), Kassam and Poor (1985), Stockinger and Dutter (1987), Schick and Mitter (1994), Künsch (2001), Ruckdeschel (2001), Spangl (2008). For the full literature overview on this topic we guide reader to the introduction of the paper Ruckdeschel et al. (2014b).

5.1 Deviations from the ideal model

In Section 2.2, describing contamination of the sample, we used some special notations borrowed from the paper of Ruckdeschel et al. (2014b). Since we are going to use them

further, we repeat them briefly here. We denote the *ideal* model assumptions by suffix "id", *distorting* (contaminating) situation by "di" and suffix "re" indicates the *realistic* contaminated situation.

First, we define different types of outliers. In time series it is common to distinguish between system-endogenous outliers, which propagate, or -exogenous, non-propagating outliers. For the notions of the types of outliers we use the terminology of Fox (1972), but in a some more general sense. Fox distinguishes *innovation outliers* (IOs), which affect the state and hence propagate and *additive outliers* (AOs), which only affect single observations and do not propagate. Originally, for the linear state-space model AOs and IOs are defined as follows:

Definition 5.1 (Innovation and additive outliers for linear SSM)**.** The innovation and additive outliers which affect the innovations and observation errors of the state-space model (4.4) and (4.5) correspondingly defined as:

$$\text{IO}: \quad v_t^{\text{re}} \sim (1 - r_{\text{IO}})\mathfrak{L}(v_t^{\text{id}}) + r_{\text{IO}}\mathfrak{L}(v_t^{\text{di}}), \tag{5.1}$$

$$\text{AO}: \quad \varepsilon_t^{\text{re}} \sim (1 - r_{\text{AO}})\mathfrak{L}(\varepsilon_t^{\text{id}}) + r_{\text{AO}}\mathfrak{L}(\varepsilon_t^{\text{di}}), \tag{5.2}$$

where $\mathfrak{L}(v_t^{\text{di}})$ and $\mathfrak{L}(\varepsilon_t^{\text{di}})$ are arbitrary, unknown and uncontrollable distributions and $0 \leq r_{\text{IO}} \leq 1$, $0 \leq r_{\text{AO}} \leq 1$ are the IO- and AO-contamination radii, which specify the sizes of the corresponding neighborhoods.

We use these notions of Fox in a wider sense:

- IOs denote endogenous outliers affecting the state equation in general, also covering level shifts or linear trends;
- AOs denote general exogenous outliers which do not propagate. This also covers substitutive outliers.

It turns out, that in order to obtain explicit solution it payoff to replace the outlier model (5.2) by the following substitutive outlier (SO) model:

$$Y^{\text{re}} = (1 - U)Y^{\text{id}} + UY^{\text{di}}, \qquad U \sim \text{Bin}(1, r) \tag{5.3}$$

for SO-contamination radius $0 \leq r \leq 1$, which specifies size of the corresponding neighborhood. Here U is assumed to be independent of (X, Y^{id}) and (X, Y^{id}), as well as observations Y^{di} are independent of X. As usual, the contaminating distribution $\mathcal{L}(Y^{\text{id}})$ is arbitrary, unknown and uncontrollable.

As for the IOs, they assume that the state equation of the SSM is divided into two steps. For the linear state-space model (4.4) and (4.5) this model is written in the form

$$\tilde{X}_t = F_t X_{t-1}^{\mathrm{re}} + v_t^{\mathrm{id}}, \qquad X_t^{\mathrm{re}} = (1-\tilde{U}_t)\tilde{X}_t + \tilde{U}_t X_t^{\mathrm{di}}, \qquad Y_t^{\mathrm{re}} = Z_t X_t^{\mathrm{re}} + (1-\tilde{U}_t)\varepsilon_t^{\mathrm{id}}, \quad (5.4)$$

where $\tilde{U}_t$ and X_t^{di} are defined in analogy to U_t and Y^{di}, i.e. with independence from all ideal distributions and the past.

Due to the different nature of these outliers, we differently react to the presence of IOs and AOs. AOs are usually downweighted as far as possible, since they are exogenous, whereas we always try to detect IOs as fast as possible, because they can make structural changes in the whole system. The situation when we face both types of outliers is more difficult, since we cannot distinguish IO from AO type immediately after a suspicious observation, but in reality both types of outliers are usually presented in the data.

IOs and AOs in our sense still do not cover all the possible types of outliers, but in the framework of this PhD thesis we restrict ourselves to these two types only.

5.2 Robustification of the least squares solution

The idea of the new robust procedures, presented below, is based on the filter, introduced for the additive outliers by Ruckdeschel (2001), more precisely, so-called *r*obustifying *r*ecursive *L*east *S*quares: `rLS`. Here, we extend this `rLS` filter for the AOs and denote it as `rLS.AO` and construct the IO-robust version of this filter, named `rLS.IO`. As we mentioned, our procedures `rLS.AO` and `rLS.IO` assume the outlier models (5.3) and (5.4) correspondingly. We prefer to start with the `rLS.AO` filter, since it turns out to be easier.

5.2.1 `rLS.AO` filter

By the definition, AOs enter only observations of the model. In the classical Kalman filtering scheme (4.7), one can notice, that observations appear only in the correction step. Therefore, we do not make any changes in the initialization and prediction for AO-robustification.

As for the correction step, we use the method introduced in Ruckdeschel (2000). The idea is to replace the term $K\Delta Y$ by its *Huberization* $H_b(K\Delta Y)$, where vector function $H_b(x)$ is defined as

$$H_b(x) = x \min\{1, b/|x|\},$$

for some suitably chosen clipping height b. Natural candidates for the norm, to be used in the Huber function, are Euclidean and Mahalanobis norms. There are other options for the robustification of the Kalman filter, but we prefer this one since it has some optimality properties (see Ruckdeschel et al. (2014b, Appendix A.2)).

In the master-thesis Pupashenko (2011, Sec, 3.1.1) calculations for the error covariance matrix in the correction step were made, but it turns out that we do not gain too much from this change, therefore, in this thesis we leave covariance matrices unchanged. Hence, the only modification we do to AO-robustufy KF is in the correction step, i.e.

$$X_{t|t} = X_{t|t-1} + H_b(K_t \Delta Y_t). \tag{5.5}$$

Another benefit of our choice in favor of this robustification is that, doing only one substitution in the correction step, we keep Kalman filter simple and non iterative, hence especially useful for online-purposes.

Choice of the clipping height

To complete `rLS.AO` filtering scheme we should choose corresponding clipping heights. This issue was studied by Ruckdeschel (2010c) in detail. Author distinguished two approaches. Both are based on one additional simplifying assumption, which turns out to be only approximately correct. Nevertheless, denoting expectation w.r.t. the ideal distribution as $\mathbf{E}_{\text{id}}$, we let $\mathbf{E}_{\text{id}}[\Delta X|\Delta Y]$ be linear.

The first way of choosing the clipping height is to select $b = b(\delta)$ according to an Anscombe (1960) criterion, i.e.

$$\mathbf{E}_{\text{id}}\big|\Delta X - H_b(K\Delta Y)\big|^2 \stackrel{!}{=} (1+\delta)\mathbf{E}_{\text{id}}\big|\Delta X - K\Delta Y\big|^2, \tag{5.6}$$

where δ is also called "insurance premium" to be paid in terms of the efficiency and usually is given as $\delta = 0.05$. For computational reasons, equation (5.6) is transformed to expression involving the covariances (e.g. see Pupashenko (2011, Sec, 3.1.1)).

The second possible way to choose the clipping height uses the radius of the SO-contamination neighborhood, $r \in [0,1]$, and computes $b = b(r)$, s.t.

$$(1-r)\mathbf{E}_{id}(|K\Delta Y| - b)_+ \stackrel{!}{=} rb$$

This approach can be extended to the case when we do not know the radius itself, but only the interval it lies in, see Rieder et al. (2008) and Ruckdeschel (2010c).

5.2.2 `rLS.IO` filter

In this Section we mainly present the results of the papers of Ruckdeschel (2010c) and Ruckdeschel et al. (2014b). So far the presented approach does not cover IO's, although this is an important problem as, e.g. classical Kalman filter in the situation of data with IOs behaves much better than in the AOs presence. Nevertheless, the classical filter is too inert and there is a way to improve this procedure.

First let us consider simplest one-dimensional linear state-space model with the observation coefficient equal to 1, i.e.

$$Y = X + \varepsilon. \tag{5.7}$$

In the correction step of the classical Kalman filter, based on observation residual, we want to improve innovation residual. In the case of our simplified model, equation (5.7) shows useful symmetry of X and ε, moreover, we get the following relation:

$$\mathbf{E}[X|Y] = Y - \mathbf{E}[\varepsilon|Y].$$

Hence, to obtain corresponding IO-reconstruction, we reconstruct ε in the AO-robust way, using already studied `rLS.AO`-filter, and plug new observation error in the last relation. In this case if `rLS.AO`(ε) gets damped, `rLS.IO` value of $\mathbf{E}[X|Y]$ gets closer to ΔY, hence follows the signal more closely than the classical Kalman filter. We should note, that in this structure we rely on identically distributed errors ε.

Returning to the general structure, note that in the ideal setting for state-space model (4.4) and (4.5) we get that

$$\mathbf{E}[\varepsilon_t|\Delta Y_t] = (\mathbb{I}_q - Z_t K_t)\Delta Y_t,$$

therefore, the correction step for the ideal model can be rewritten as follows:

$$X_{t|t} = X_{t|t-1} + Z_t^{\Sigma}[\Delta Y_t - \mathbf{E}[\varepsilon_t|\Delta Y_t]],$$

where $Z_t^{\Sigma} := \Sigma_{t|t-1} Z_t^{\mathrm{T}} (Z_t^{\mathrm{T}} \Sigma_{t|t-1} Z_t)^{-1}$ is suitably generalized inverse for Z_t. This inverse is necessary for higher dimension case when rank of Z is less than p. If the observation matrix is invertable, then the matrix Z_t^{Σ} is simply the inverse for it.

Therefore, in the presence of IOs we construct the `rLS.IO`-filter remaining the initialization and prediction steps as in the classical Kalman filter and replacing correction step

by the following:

$$X_{t|t} = X_{t|t-1} + Z_t^{\Sigma}[\Delta Y_t - H_b((\mathbb{I}_q - Z_t K_t)\Delta Y_t)] \tag{5.8}$$

Remark 5.2. *Arguments for the choice of the norm and the clipping height here are the same as for the* `rLS.AO`*. For the optimality properties of the* `rLS.IO` *see Ruckdeschel et al. (2014b, Appendix A.3).*

5.2.3 Robust smoother

Often, to improve filtered values we apply Kalman smoother (4.12)-(4.13), considering information available in the meantime. Moreover, it is very important for the further use of the estimators, e.g. for estimation of the hyper-parameters of the state-space model applying EM-algorithm. Therefore, robustness of the Kalman smoother is also very important issue and here we describe new results on it, presented in the paper of Ruckdeschel et al. (2014b).

To conclude what if the Kalman smoother is robust, we rewrite backward recursive equation of it (4.12) in the following form:

$$X_{t|T} - X_{t|t} = J_t[(X_{t+1|T} - X_{t+1|t+1}) + (X_{t+1|t+1} - X_{t+1|t})]$$

One can see that the first summand in the brackets of the right hand side $X_{t+1|T} - X_{t+1|t+1}$ is just the previous iteration of the left hand side in the recursion $X_{t|T} - X_{t|t}$. As for the second summand $X_{t+1|t+1} - X_{t+1|t}$, it is already robustified, as an increment of the correction step (5.5) in the robust Kalman filter.

Therefore, we conclude that for outlier models with IO and AO contamination (5.3) and (5.4), modification has to be done only in the second summand $X_{t+1|t+1} - X_{t+1|t}$, treating it as the one from the robust Kalman filter. There is no further need for robustification in the Kalman smoother.

5.2.4 Robust versions of extended Kalman procedures

In this Section we reproduce findings of the paper Ruckdeschel et al. (2014b). In Section 4.2.2.2 we introduced extended Kalman filter and smoother for the general (non-linear) state-space models. These procedures can also be robustified for both types of outliers, AOs and IOs.

The idea of the reconstruction in the extended procedures is the same as in the classical ones. In the filter the only changes to be done concern correction step, where we simply replace the term $K_t\Delta Y_t$ by $H_b(K_t\Delta Y_t)$ in the AO-case and by $Z_t^{\Sigma}(\Delta Y_t - H_b((\mathbb{I} - Z_tK_t)\Delta Y_t)$ for IOs. As for the smoother, with the same arguments as in the Section 5.2.3, there is no need for robustification in the extended Kalman smoother except the treating second summand of the backward recursion as the one from robust extended Kalman filter.

5.3 Behavior of the filters at stylized outlier situations

In this Section, we study the behavior of introduced filters in the ideal case as well as in the stylized outlier situations, for which they are not necessarily designed. We take some specifically chosen models and display performance of our procedures on the corresponding plots, also comparing them to another existing methods.

5.3.1 The ideal situation, AO- and IO-contamination

Here we reproduce an illustration from the papaer of Ruckdeschel et al. (2014b). We analyze the behavior of classical Kalman and `rLS` filters for three different types of generated data. First, we aplly these methods to the ideal situation, then we contaminate data with IOs and AOs correspondingly. As the hyper-parameters for the state-space model we take the following:

$$a_0 = \begin{pmatrix} 20 \\ 0 \end{pmatrix}, \quad Q_0 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},$$
$$F_t = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}, \quad Z_t = \begin{pmatrix} 0.3 & 1 \\ -0.3 & 1 \end{pmatrix}, \quad Q_t = \begin{pmatrix} 0 & 0 \\ 0 & 9 \end{pmatrix}, \quad V_t = \begin{pmatrix} 9 & 0 \\ 0 & 9 \end{pmatrix}.$$

From the given state matrix F_t one can see that our state process consists of two coordinates, where the first is a random walk, therefore non-stationary, and the second coordinate is white noise.

We simulate the innovations v_t and the observation errors ε_t from the contaminated bivariate normal distribution of the form:

$$\mathcal{CN}_2(r, 0, R, \mu_c, R_c) = (1 - r)\mathcal{N}_2(0, R) + r\mathcal{N}_2(\mu_c, R_c) \ , \tag{5.9}$$

where amount of contamination is specified by the radius $r = 10\%$. Notation R can be replaced by the matrix Q_t for the innovations or matrix V_t in the case of the observation

errors. Moments of this bivariate normal distribution we set as $\mu_c^{\mathrm{T}} = (25, 30)$ and $R_c = \mathrm{diag}(0.9, 0.9)$.

After applying classical Kalman, `rLS.IO`, and `rLS.AO` filters we plot the first coordinate of the state vector in three different contamination situations.

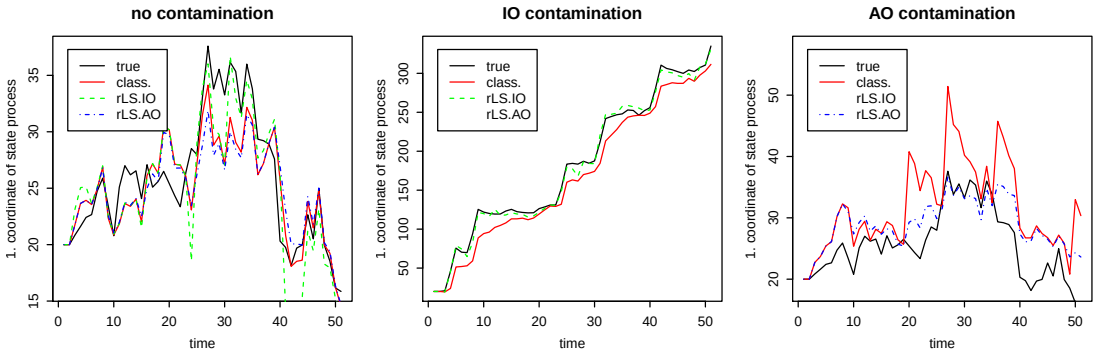

FIGURE 5.1: Results of three filters (classical KF, `rLSIO` and `rLSAO`) for different contamination situations

In Figure 5.1 the true state process is plotted by the thick black line, while classical Kalman, `rLS.IO`, and `rLS.AO` filters are plotted by the light red, dotted green and dot-dashed blue lines respectively.

From the first plot of Figure 5.1 we conclude that in the ideal situation all three filters perform very similar. Only `rLS.IO` does not work perfectly well showing some sudden jumps. The reason for that is the higher dimensional state-space model with the observation matrix which is not of full rank.

At IO contamination situation, drawed on the middle plot of Figure 5.1, the `rLS.IO` filter almost immediately follows the true state. Classical Kalman filter performs also well in this case, but it is only able to track the true state with a certain delay, what is worse comparing to the `rLS.IO` filter.

What is important to mention in the case of AO contamination, is that by definition AOs affect only the observation equation, therefore their impact cannot be seen directly in Figure 5.1. Nevertheless, effect of additive outliers is indirectly observable in the spikes for the filter estimate of the classical Kalman filter.

5.3.2 Changes in oscillation patterns and level shifts

Here we consider behavior of the classical and robustified versions of the Kalman filter comparing to the non-parametric filtering method `ADORE` in three situations. First we consider IO- and AO-contamination, and then additionally study the case when the part

of the state is replaced by a completely artificial signal. Chosen non-parametric filter `ADORE` is introduced by Schettlinger (2009) and it uses automatic selection of the window width.

For the state here we take an autoregressive process of order 2. Observations are one dimensional and hyper parameters are the following:

$$a_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad Q_0 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},$$
$$F_t = \begin{pmatrix} 1 & -0.9 \\ 1 & 0 \end{pmatrix}, \quad Z_t = \begin{pmatrix} 1 & 0 \end{pmatrix}, \quad Q_t = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad V_t = \begin{pmatrix} 1 \end{pmatrix}.$$

To complete state-space model we compute the innovations v_t in the IO situation from the contaminated bivariate normal distribution (5.9) with the following moments

$$\mu_c^{\mathrm{T}} = (30, 0), \quad R_c = \begin{pmatrix} 0.1 & 0 \\ 0 & 0 \end{pmatrix}.$$

As for the observation errors ε_t, their contaminating distribution is chosen to be $\mathcal{N}(10, 0.1)$.

For the case of endogenous contamination, we replace whole parts of the state process by so called *block signal* (see Donoho and Johnstone (1994)), which consists of pieces of the random length and amplitude.

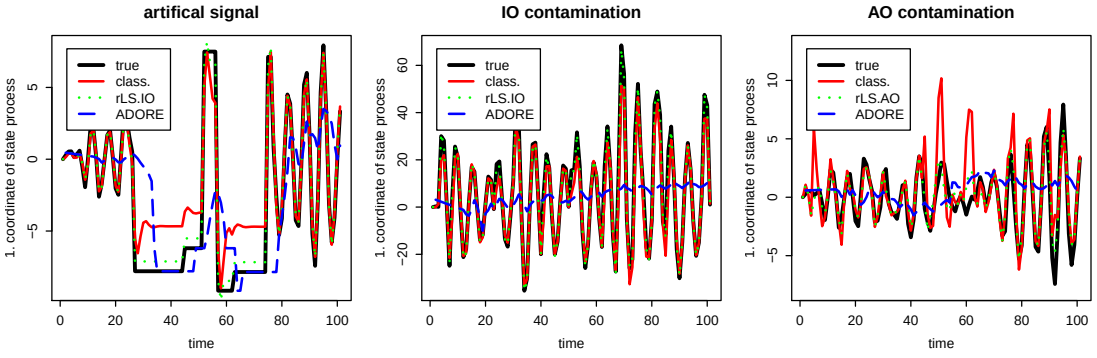

FIGURE 5.2: Results of the simulated state-space model for different contamination situations

In Figure 5.2 the black line again represents the true state process, the red line is drwan for the classical Kalman filter, whereas rLS and ADORE filters are plotted by the dashed green and dot-dashed blue lines respectively.

In the situation of the artificial signal `rLS.IO` filter follows the true state very close and performs better than other filters. Similarly to the previous example, here classical

Kalman filter does not track the level shifts. As for the non-parametric filter `ADORE`, it displays curve very similar to the true process, but does it with some time delay.

On the middle plot of Figure 5.2 we observe that `rLS.IO` filter performs very well in the IO-contamination situation, whereas the classical Kalman filter fails to track the spikes of the state signal. The `ADORE` filter in this case fits, but it extremely smoothes underlying process.

The last plot is drawn for the AO-contamination case. Here one can see that `rLS.AO` filter is not affected by the spiky outliers, but the classical Kalman filter is prone to them. Non-parametric filter `ADORE` shows similar behavior as in the IO-contamination, estimating only an overall trend of the true process.

5.3.3 Coping with non observed aspects

In this Section we study behavior of all filters in the special case of some non-observed aspects, i.e. for the following setup:

$$T = 50\ , \quad F = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix}\ , \quad Z = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}\ , \quad Q = \mathrm{diag}(0, 0, 0.001)\ ,$$
$$V = \mathrm{diag}(0.1, 0.001)\ , \quad a_0 = (0, 0, 0)^{\mathrm{T}}\ , \qquad Q_0 = \mathrm{diag}(1, 0.1, 0.001)\ .$$

As one can notice observation matrix Z_t here has a non-trivial null space. State signals which are falling to this null space are not visible at filtering time. In this case smoothing can improve results of the filtering with a certain time delay. The reason is that transitions F_s move the invisible states and at some later stage they become visible to Z_s.

In the contamination models for AO's and IOs, i.e. (5.3) and (5.4), we take equal radii $r_{\mathrm{IO}} = r_{\mathrm{AO}} = 0.1$. We choose multivariate Cauchy contaminating distribution for the states $X_t^{\mathrm{di}} \sim$ multiv.Cauchy$(0, Q)$ and special form of Cauchy contaminating distribution for the observations $Y_t^{\mathrm{di}} \sim$ Cauchy/1000 (one can easily compute these distributions using `R`-packages `mvtnorm` and `MASS`).

Figure 5.3 displays how our filters and smoother can cope with the introduced non-observed aspects situation. The left plot shows behavior of the classical Kalman filter and smoother, the middle one is drawn for the IO-robust filter and the right plot reflects AO-robust filter. Here we plot only the second coordinate of the state process, which

lies in $\ker Z_t$. The black line represents the true state process, red line draws the IO-contaminated state process, i.e., the real situation. `rLS` filter and smoother are drawn by the dashed green and dot-dashed blue lines correspondingly.

From all plots of Figure 5.3 we conclude that none of the proposed filters can cope with this situation.

5.3.4 Application

The application of our procedures used above is described in detail in the paper of Ruckdeschel et al. (2014b). It is based on the real data, captured from the vehicle moving on some track. Data consists of four data channels, including time, speed, altitude and pitch angle speed. The object of interest is the slope, i.e. change of the altitude over distance. Since the original data is obviously distorted, there is a need to use robustified methods. Therefore, for three state-space models of the different levels of complexity, we applied `rLS.AO` and `rLS.IO` filters comparing their performance to the classical one. From this real-data application we conclude not only about the importance of our filters, but also about the quality of their performance and enough level of complexity for the model in this case.

5.4 Software infrastructure

Probably due to the simple form of the Kalman recursions there is no great abundance of the packages implementing them in `R`. We give the overview of the existing ones though.

The package `dse` P.Gilbert (2011) counts as the first one offering Kalman filtering. First version of the package was submitted to `CRAN` in year 2000. This package is focused

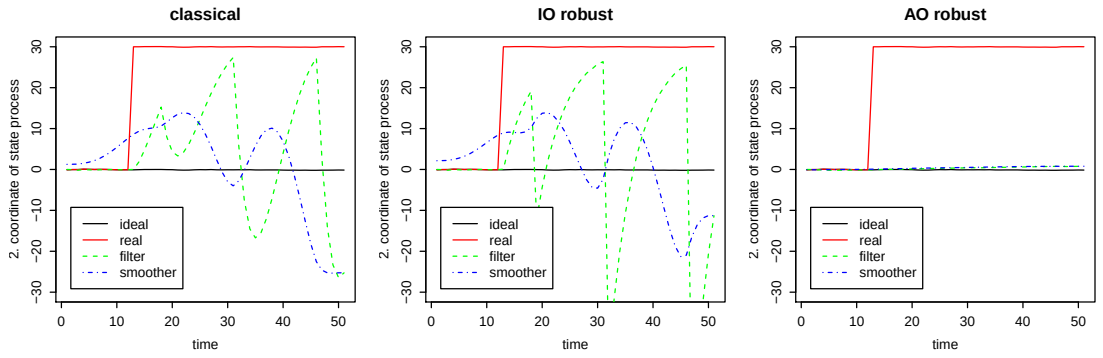

FIGURE 5.3: Filter estimates of the simulated state-space model using different filters and smoothers

on multivariate time series. It covers state-space representations, and methods for converting between them, including estimation techniques and forecasting models. Kalman filter and smoother estimates can be obtained with the functions of the `dse` package. The state-space model reduction techniques are also implemented in this package.

Another R-package available in CRAN is `sspir`, created by Dethlefsen et al. (2009). This package covers state-space models, offering the function `ssm`, which is based on the familiar formula notation of the functions like `lm`, `glm`, etc. It includes functions for Kalman filtering and smoothing, returning a new object with the filtered (or smoothed) estimates of the state, and their covariance matrices. Moreover, the package `sspir` contains implementations for the expectation-Maximization algorithm, used for the case of unknown SSM hyper-parameters.

The R-package `dlm` first appeared in CRAN in August 2006 and the actual version of it can be found in Petris (2010). This package contains set of the R-functions, which help us with the specification of state-space models. Maximum likelihood estimation and Kalman filtering and smoothing for the linear version of the state-space models are implemented in the package. It also includes some specific form of square root filter, that is more robust and general than the standard square root filters. In addition, `dlm` package contains the "outer sum" operator, which combines models for different time series into a joint model. This eases the usage of the models with components of different dynamics.

The R-package `FKF` with the most actual version presented in Luethi et al. (2010), was submitted to CRAN in year 2009. As we understanfd from the name of the package, i.e. Fast Kalman Filter, it is mostly focused on the speed of the Kalman procedures. It also covers maximum likelihood estimation, Kalman filtering and smoothing and Expectation-Maximization algorithm, which does much faster switching between E and M steps than before, due to faster computation of the filtered estimates.

One of the most recent packages in R, which contains Kalman procedures, is the package `KFAS`, introduced by Helske (2010). It includes functions for Kalman filtering, smoothing, simulation smoothing and disturbance smoothing for multivariate exponential family state space models. The package also covers the case of the models with unknown distributions of some or all elements of the initial state vector.

Features and speed comparison

Here we make quick comparison of the features of all presented R-packages, putting them together in the Table 5.1.

	dse	sspir	dlm	FKF	KFAS
Coded in	R+Fortan	R	R+C	R+C	R+Fortan
Class model	S3	S3	S3	S3	
Algorithm	CF	CF	SRCF	CF	CF
Sequential processing					*
Exact diffuse initialization					*
Missing values allowed			*	*	
Time varying matrices		*	*	*	*
Simulator	*		*		
Smoother	*	*	*		*
Simulation smoother		*	*		*
Disturbance smoother					*
MLE routine	*	*	*		
Non-Gaussian models		*			*

TABLE 5.1: Quick packages comparison

Abbreviation CF is made for the covariance filter algorithm and SRCF is for the square root covariance filter. We conclude from the table that the `sspir` and `dlm` packages are quite useful in different situations, i.e. they deal with time-varying state space models and both have implemented smoother. As we mentioned, the package `FKF` is focused on the KF simulation speed, so it is not a surprise that it does not cover most of the illustrated features. And the package `KFAS` seems to be the most general, since most of the procedures are implemented in it.

More detailed overview of the Kalman procedures implemented in `R`, with examples of using introduced packages, one can find in the article of Fernando (2011).

5.5 Package `robkalman`

In the framework of this PhD-thesis, together with P. Ruckdeschel and B. Spangl, we developed new package based on the Kalman procedures, named `robKalman`. Last developer version available in `R-Forge` is 0.3. The goal of the package is not only to provide routines for robust Kalman filtering, introduced earlier in this Chapter, but also cover most of the possible situations.

Function `classEKF` covers classical extended Kalman filter routines, which include functions to return list of parameters for all three steps of the filter, with error covariance matrices and Kalman gain.

The **R**-package `robKalman` implements `rLS`-filter in the function `rLSeKF` including both types of outliers, AOs and IOs. It also provides function for the (extended) Kalman filter to create the state and observation matrices and covariance matrices of innovations

and observation errors. Functions `recSmoother` and `calibrateRLS` compute extended Kalman smoother and clipping height correspondingly.

5.6 Conclusions

In this Chapter we presented the robust versions of the Kalman filter and smoother which are specialized on the spiky outliers, AOs and IOs. Here is important to note that we were first to compute general IO-robust filter and introduce new idea for the smoother.

We have tested our procedures in different stylized outlier situations, i.e. in the presence of AOs or IOs, where we conclude that our procedures perform very well in the situations they were created for. Moreover, our procedures also can cover wider variety of outlier situations.

We also compared `rLS.AO` and `rLS.IO` to one non-parametric filtering method and obtained that our filters beat it in all contamination situations.

All our procedures are recursive, therefore they are quite fast and convenient for online using. They can be used not only for the robustification of the classical filter, but also for the extended Kalman filter applied to the non-linear state-space models.

Still, there are some open issues in our procedures, which are topics for further research, e.g. IO-robust smoother have to be essentially improved. Besides, after checking filters in the case of some non observed aspects, i.e. when the observation matrix of the model is non invertible, we conclude that all filters cannot cope with this situation.

It is also very important to mention that in reality both types of outliers are usually presented in the data, therefore using one of two introduced robust filters does not bring to much. In general we would need some hybrid filter and smoother, which will combine `rLS.IO` and `rLS.AO` procedures in one, used for these mixed situations. First attempts were made in the Ruckdeschel (2010c, Ch. 5).

In the last Section of this Chapter we also gave overview of the existing software infrastructure in `R` for the Kalman procedures and introduced our `R`-package `robKalman`, which beside classical filter and smoother covers our robust procedures.

Chapter 6

Generalized linear models

6.1 L_2 Differentiability of Generalized Linear Models

In this Chapter we extend already existing theory on the L_2 differentiability of the parametric models to the generalized linear models. This has been studied already for the GLMs, which are exponential families, by Schlather (1994). We consider non-exponential scale-shape families, e.g. the generalized extreme value and generalized Pareto distributions.

Here we generalize result of Rieder (1994, Sec. 2.4) on L_2 differentiability for the linear regression models. We also cover higher dimensional error distributions and the case of regressors of possibly different length for each parameter. We separately treat cases of stochastic regressors, which is of particular interest for incorporating (space-)time dependence, and deterministic regressors as occurring in planned experiments.

The results of this Section have been submitted as separate contribution in the paper Pupashenko et al. (2014).

6.1.1 General settings

Earlier, in Section 4, we introduced the idea of the L_2 differentiability for the model $\mathcal{Q} = \{Q_\vartheta | \vartheta \in \Theta\} \subset \mathcal{M}_1(\mathcal{A})$ parameterized by ϑ from the open parameter domain $\Theta \subset \mathbb{R}^k$. Remind, that the notion for the densities of the distributions from the model are $dQ_\vartheta = q_\vartheta$.

Here we turn model $\mathcal{Q}$ into a regression model $\mathcal{P}$ parametrized by regression parameter β. We do it using continuously differentiable link function $l : \mathbb{R}^k \to \Theta$, with derivative denoted as $\dot{l}$.

First, we introduce the partition $\pi = (p_h)_{h=1,\dots,k}$, which groups p coordinates of the regressor into k blocks of dimensions p_h, where $\sum_h p_h = p$. With the help of this function each $x \in \mathbb{R}^p$ can be represented in the form $x = (x_{h,j})_{\substack{h=1,\dots,k \\ j=1,\dots,p_h}}$.

For the later use, we also need some additional operators based on the function π, i.e.

$$T_\pi : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^k, \quad (a,b) \mapsto T_\pi(a,b) =: a^{\mathrm{T}_\pi} b = (\sum_{j=1}^{p_h} a_{h,j} b_{h,j})_{h=1,\dots,k}; \tag{6.1}$$

$$\rho_\pi : \mathbb{R}^k \times \mathbb{R}^p \to \mathbb{R}^p, \quad (c,a) \mapsto \rho_\pi(c,a) =: c \cdot_\pi a = (c_h a_{h,j})_{\substack{h=1,\dots,k \\ j=1,\dots,p_h}}; \tag{6.2}$$

$$M_\pi : \mathbb{R}^{k\times k} \times \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^{p\times p}, \quad (C,a,b) \mapsto M_\pi(C,a,b) = (C_{h_1 h_2} a_{h_1,j_1} b_{h_2,j_2})_{\substack{h_1,h_2=1,\dots,k \\ j_1,j_2=1,\dots,p_h}}. \tag{6.3}$$

Later, we also apply operator ρ_π to the $k \times m$ matrix C, denoting it as $C \cdot_\pi a$ and meaning that we obtain corresponding $p \times m$ matrix $(c_{h,l} a_{h,j})_{\substack{h=1,\dots,k \\ j=1,\dots,p_h}; l=1,\dots,m}$.

Then, for the regressor X and regression parameter β we obtain a regression as $\vartheta = l(\theta)$, where $\theta = X^{\mathrm{T}_\pi}\beta$ and the corresponding GLM induced by this link function is given by

$$\mathcal{P} = \{P_\beta(dx,dy) = Q_{l(x^{\mathrm{T}_\pi}\beta)}(dy|x)K(dx) | \beta \in \mathbb{R}^p, Q_\vartheta \in \mathcal{Q}\}. \tag{6.4}$$

We have already mentioned before, that for the linear regression case Rieder (1994, Theorem 2.4.7) obtained L_2 differentiability with some additional assumptions. We prove some similar result for the introduced GLM (6.4), distinguishing two cases of stochastic and deterministic regressors.

6.1.2 Random Carriers

In Section 4.3.2 we already introduced parametric model for the random carriers in the linear setting. Here we treat regressors x as stochastic variables with distribution K, but the pairs (x,y) are modeled as i.i.d. observations. We suppose, that the model $\mathcal{Q}$ is L_2 differentiable, with corresponding L_2 derivative $\Lambda_\vartheta^{\mathcal{Q}}$ and the Fisher information matrix $I_\vartheta^{\mathcal{Q}}$. Then, we state the following (see also Ruckdeschel et al. (2014b, Thm. 2.5)).

Theorem 6.1. *Let $\beta_0 \in \mathbb{R}^p$. For the link function $l : \mathbb{R}^k \to \Theta$ holds $\vartheta_t = l(\theta_t)$ for $\theta_t = x^{T_\pi}(\beta_0 + t)$, s.t. $\dot{l}_t = \dot{l}(\theta_t)$. Denote the Frobenius matrix norm as $|\cdot|$, i.e. $|A|^2 = trA^2$.*

If the following conditions (i)-(iii) hold:

(i) Model $\mathcal{Q}$ satisfies Hájek conditions (H.1)-(H.3) from (2.9), replacing "Q_{ϑ_0}-a.e. x" by expression "P_{β_0}-a.e. (x, y)",

(ii)

$$\int |I^{\mathcal{P}}_{\vartheta_0}(x)| K(dx) < \infty, \tag{6.5}$$

(iii) for all $b > 0$ holds

$$\lim_{s\to 0} \sup_{|t|\le b} \int \Big| |I^{\mathcal{P}}_{\vartheta_{st}}(x)| - |I^{\mathcal{P}}_{\vartheta_0}(x)| \Big| K(dx) = 0, \tag{6.6}$$

then generalized linear model $\mathcal{P}$ from (6.4) is L_2 differentiable in β_0 with the L_2 derivative of the form:

$$\Lambda^{\mathcal{P}}_{\beta_0}(x, y) = \dot{l}_0^T \Lambda^{\mathcal{Q}}_{\vartheta_0} \cdot_\pi x \tag{6.7}$$

and the Fisher information matrix

$$I^{\mathcal{P}}_{\beta_0} = \boldsymbol{E}_{\beta_0} \Lambda^{\mathcal{P}}_{\beta_0} (\Lambda^{\mathcal{P}}_{\beta_0})^T = \int I^{\mathcal{Q}}_{\vartheta_0}(x) K(dx) \tag{6.8}$$

For the proof of this theorem we need some additional lemma, which we formulate here. Both proofs, of the lemma and of theorem itself, can be found in Appendix B.1 and Appendix B.2 correspondingly.

Remark 6.2. *Conditions (6.5) and (6.6) can be also strengthened to the following form:*

$$\int |I^{\mathcal{Q}}_{\vartheta_0}| |\dot{l}_0|^2 |x|^2 K(dx) < \infty,$$

and for all $b > 0$

$$\lim_{s\to 0} \sup_{|t|\le b} \int \Big| |I^{\mathcal{Q}}_{\vartheta_{st}}| |\dot{l}_{st}|^2 - |I^{\mathcal{Q}}_{\vartheta_0}| |\dot{l}_0|^2 \Big| |x|^2 K(dx) = 0.$$

Lemma 6.3 (Chain rule)**.** *Assume, that parametric model $\mathcal{Q} = \{Q_\vartheta | \vartheta \in \Theta\}$ with open parameter domain $\Theta \subset \mathbb{R}^k$ is L_2 differentiable in $\vartheta_0 \in \Theta$, with derivative $\Lambda^{\mathcal{Q}}_{\vartheta_0}$ and the Fisher information $I^{\mathcal{Q}}_{\vartheta_0}$.*

Let link function $l : \Theta' \to \Theta$, $\Theta' \subset \mathbb{R}^k$ be differentiable in some θ_0, s.t. $\vartheta_0 = l(\theta_0)$ and its derivative is $\dot{l}_0 = \dot{l}(\theta_0)$.

Then $\tilde{\mathcal{Q}} = \{\tilde{Q}_\vartheta = Q_{l(\theta)} | \theta \in \Theta'\}$ is L_2 differentiable in $\theta_0 \in \Theta'$ with derivative $\Lambda^{\tilde{\mathcal{Q}}}_\theta = (\dot{l}(\theta_0))^T \Lambda^{\mathcal{Q}}_{\vartheta_0}$ and the Fisher information $I^{\tilde{\mathcal{Q}}}_\theta = (\dot{l}(\theta_0))^T I^{\mathcal{Q}}_{\vartheta_0} \dot{l}(\theta_0)$. Moreover, if model $\mathcal{Q}$ is continuously L_2 differentiable in $\vartheta_0 \in \Theta$, then $\tilde{\mathcal{Q}}$ is continuously L_2 differentiable in θ_0.

Remark 6.4. *Chain rule 6.3 holds for both, deterministic and random cases.*

6.1.3 Deterministic Carriers

Here we aim to get the analogous result to Rieder (1994, Theorem. 2.4.2) for GLMs. Hence, we make $i_n \geq 1$ real valued observations $y_{n,i}$, with given array of the regressors $x_{n,i} \in \mathbb{R}^p$.

For these deterministic regressors we define corresponding GLM in the following way:

$$\mathcal{P} = (\otimes_{i=1}^{i_n} \mathcal{P}_{n,i}), \tag{6.9}$$

$$\mathcal{P}_{n,i} = \{P_{n,i,\beta_0}(dy) = Q_{\vartheta_{n,i}}(dy) | \beta_0 \in \mathbb{R}^p, \vartheta_{n,i} = l(\theta_{n,i}), \theta_{n,i} = x_{n,i}^{\mathrm{T}_\pi}\beta_0, Q_{\vartheta_{n,i}} \in \mathcal{Q}_{n,i}\}. \tag{6.10}$$

As we have mentioned in Section 4.3.3, idea of the proof of the deterministic carriers conditions (4.18) and (4.19), described in Rieder (1994, Theorem 2.4.7), is based on the smallness of hat matrix (4.21). For our general framework we can still define analogical hat matrix in the following way:

$$H_n = H_{n;i,j;\beta_0} = L_{n;i;\beta_0}^{\mathrm{T}} (I_{n,\beta_0}^{\mathcal{P}})^{-1} L_{n;i;\beta_0}, \quad L_{n;i;\beta_0} = \dot{l}(\theta_{n,i})^{\mathrm{T}} (I_{n,i,\beta_0}^{\mathcal{P}})^{\frac{1}{2}} \cdot_\pi x_{n,i}.$$

If we perform analogically to the proof of the linear regression case, as in Rieder (1994, Theorem 2.4.7), first we have the change in the fitted parameter $\vartheta_{n,i}$ of the form:

$$\vartheta'_{n,i} = \vartheta_{n,i} + \sum_{j=1}^{i_n} (I_{n,\beta_0}^{\mathcal{P}})^{-\frac{1}{2}} H_{n;i,j} (I_{n,\beta_0}^{\mathcal{P}})^{-\frac{1}{2}} \Lambda_{\vartheta_{n,j}}^{\mathcal{Q}}(y_{n,j}).$$

One should note, that in the linear case distribution of the standardized scores, i.e. $(I_{n,\beta_0}^{\mathcal{P}})^{-\frac{1}{2}} \Lambda_{\vartheta_{n,j}}^{\mathcal{Q}}(y_{n,j})$, is invariant in β_0, whereas in our setting it does not hold anymore. Since this property is used at some stage of the proof of Rieder (1994, Theorem 2.4.7), we fail at this point.

Therefore, in the general case we have to strengthen hat matrix condition (4.22). We propose the following theorem:

Theorem 6.5. *Suppose that the following conditions (i)-(iii) hold:*

(i) Model $\mathcal{Q}$ fulfills Hájek conditions (H.1)-(H.3) from (2.9),

(ii) Lindeberg condition (4.18) from the Definition 4.3 holds,

(iii) Let $\beta_0 \in \mathbb{R}^p$. For the link function $l : \mathbb{R}^k \to \Theta$ holds $\vartheta_{n,i,t} = l(\theta_{n,i,t})$ for $\theta_{n,i,t} = x_{n,i}^{T_\pi}(\beta_0 + (I_{n,\beta_0}^{\mathcal{P}})^{-\frac{1}{2}} t)$, s.t. $\dot{l}_{n,i,t} = \dot{l}(\theta_{n,i,t})$. For simplicity, we use additional notations as $I_{n,i,t}^{\mathcal{Q}} = I_{\vartheta_{n,i,t}}^{\mathcal{Q}}$ and $I_{n,i,t}^{\mathcal{P}} = M_\pi(\dot{l}_{n,i,t}^T I_{\vartheta_{n,i,t}}^{\mathcal{Q}} \dot{l}_{n,i,t}, x_{n,i}, x_{n,i})$. Then,

for all $b > 0$ holds

$$\lim_{n\to\infty} \sup_{|t|\le b} \sum_{i=1}^{i_n} t_n^T (I^{\mathcal{P}}_{n,i,t} - I^{\mathcal{P}}_{n,i,0}) t_n = 0. \tag{6.11}$$

Then generalized linear model $\mathcal{P}$ from (6.9) is continuously L_2 differentiable in β_0 with the L_2 derivative $\Lambda^{\mathcal{P}}_{n,i,\beta_0}(x,y) = \Lambda^{\mathcal{P}}_{\beta_0}(x_{n,i}, y)$, where $\Lambda^{\mathcal{P}}_{\beta_0}$ is obtained by the chain rule, i.e. $\Lambda^{\mathcal{P}}_{\beta_0} = \dot{l}(\theta)^T \Lambda^{\mathcal{Q}}_{\vartheta}(y) \cdot_\pi x$ and the Fisher information matrix given in the Definition 4.3.

Proof of this Theorem one can find in Appendix B.3.

Remark 6.6. *Similarly to the random carriers, here we also can obtain analogues to the conditions in Remark 6.2, which are*

$$\limsup_{n\to\infty} \sup_{|t|\le b} |t_n|^2 \sum_{i=1}^{i_n} |I^{\mathcal{Q}}_{n,i,0}| |\dot{l}(n,i,0)|^2 |x_{n,i}|^2 < \infty,$$

and for all $b > 0$

$$\lim_{n\to\infty} \sup_{|t|\le b} |t_n|^2 \sum_{i=1}^{i_n} \left| |I^{\mathcal{Q}}_{n,i,t}| |\dot{l}_{n,i,t}|^2 - |I^{\mathcal{Q}}_{n,i,0}| |\dot{l}_{n,i,0}|^2 \right| |x_{n,i}|^2 = 0.$$

6.2 Examples

In this Section we give some easy as well as more complicated examples, of applying presented theorems to show L_2 differentiability of the next models.

Example 6.7 (Linear regression)**.** Obviously, Theorem 6.1 can be applied to the linear regression model $\mathcal{P}$ about one dimensional location model $\mathcal{Q}$, i.e. for $\mathcal{Q} = \{Q_\vartheta(dy) = F(dy - \vartheta)\}$ we have the following model

$$\mathcal{P} = \{P_\beta(dx, dy) = F(dy - x^{\mathrm{T}}\beta) K(dx)\}, \tag{6.12}$$

for some probability F on $(\mathbb{R}, \mathbf{B})$ with finite Fisher information of the location (see Huber (1981, Def. 4.1/Thm. 4.2)). Then, condition (i) of Theorem 6.1 follows from the finiteness of the Fisher information of location and condition (ii) boils down to $\int |x|^2 \, K(dx) < \infty$. Condition (iii) here is void.

Example 6.8 (Binomial GLM with logit link and Poisson GLM with log link)**.** Here we consider Binomial model $\mathrm{Binom}(m, p)$ for known size $m \in \mathbb{N}$, e.g. $m = 1$, and unknown success probability $p \in (0, 1)$. Error distribution of such model has counting density $q_p(y) = \binom{m}{y} p^y (1-p)^{m-y}$ for $y \in \{0, ..., m\}$.

Condition (i) of Theorem 6.1 is obviously fulfilled with the Fisher information $I_p^{\mathcal{Q}} = m(p(1-p))^{-1}$.

For the link function in this case we take logit link, i.e., $l(\theta) = e^\theta/(1+e^\theta)$. Then, we compute the term $I_p^{\mathcal{Q}}\dot{l}(\theta)^2 = mp(1-p)$ and the conditions (ii) and (iii) of Theorem 6.1 turn to

$$\text{(ii)} \qquad \int \frac{e^{x^{\mathrm{T}}\beta}}{(1+e^{x^{\mathrm{T}}\beta})^2}|x|^2 K(dx) < \infty,$$

$$\text{(iii)} \qquad \int e^{x^{\mathrm{T}}\beta}\frac{(e^{x^{\mathrm{T}}s}-1)(1-e^{x^{\mathrm{T}}(2\beta+s)})}{(1+e^{x^{\mathrm{T}}(\beta+s)})^2(1+e^{x^{\mathrm{T}}\beta})^2}\,|x|^2\,K(dx) \to 0, \quad s \to 0.$$

One can see that in both expressions integrands are bounded pointwise in x, hence, if $|x|^2$ is integrable w.r.t. K, the Binomial GLM with logit link function is continuously L_2 differentiable.

Next, we consider the Poisson model Pois(λ) for parameter ($\lambda \in (0,\infty)$). This parametric model has error distribution with counting density $q_\lambda(y) = \frac{e^{-\lambda}\lambda^y}{y!}$ for $y \in \mathbb{N}$.

Condition (i) of Theorem 6.1 is obviously fulfilled with the Fisher information $I_\lambda^{\mathcal{Q}} = \lambda^{-1}$.

Here we take log link for the link function, i.e., $l(\theta) = e^\theta$, so that $I_\lambda^{\mathcal{Q}}\dot{l}(\theta)^2 = \lambda$. Then, conditions (ii) and (iii) of Theorem 6.1 turn to

$$\text{(ii)} \quad \int e^{x^{\mathrm{T}}\beta}|x|^2 K(dx) < \infty, \qquad \text{(iii)} \quad \int e^{x^{\mathrm{T}}\beta}(e^{x^{\mathrm{T}}s}-1)|x|^2 K(dx) \to 0, \quad s \to 0.$$

Hence, if $e^{|x|(|\beta|+\delta)}|x|^2$ is integrable w.r.t. K, then the Poisson GLM with log-link function is continuously L_2 differentiable.

Conditions additionally required for the L_2 differentiability of these models, i.e. $|x| \in L_2(K)$ for Binomial logit and $e^{|x|(|\beta|+\delta)}|x|^2 \in L_1(K)$ for the Poisson GLM with log-link, recover the conditions mentioned in Fahrmeir and Tutz (2001, p.47).

Example 6.9 (GEVD and GPD joint shape-scale models with componentwise log link)**.** Here we check the L_2 differentiability of the generalized extreme value distribution GEVD(μ,σ,ξ) from Definition 3.1 and the generalized Pareto distribution GPD(μ,σ,ξ) from Definition 3.9.

For the GEVD, three dimensional model is L_2 differentiable for the shape values $\xi \in (-1/2, 0)$ and $\xi \in (0,\infty)$. Unfortunately, our theory for L_2 differentiable error models does not cover model including the threshold parameter in the GPD case. This problem is based on the fact, that observations which are close to the endpoint of the support in the GPD model, carry extremely much information on the threshold. To avoid such

problems, we assume μ to be known in both models and, for simplicity, let $\mu = 0$. Then we work with the two dimensional parameter, which consists of scale and shape $\vartheta = (\sigma, \xi)$.

If we write the scores $\Lambda^{\mathcal{Q}}_{\vartheta}$ on the quantile scale for both models, i.e., $\Lambda_{\vartheta}(F_{\vartheta}(u))$ for $F_{\vartheta}(u)$ the respective quantile function (2.5) or (3.7), we see that both scores include terms of order $(1-u)^{\xi}$. Therefore, to fulfill condition (i) we assume that $\xi > -1/2$. This is the most general restriction for the shape and in other cases of interest it is natural to assume $\xi > 0$, e.g. for the case of Fréchet distributions or $\xi \geq 0$ for the GPD.

Here we take continuously differentiable componentwise link function $l : \mathbb{R}^2 \to \Theta$, where $l(\theta) = (l_\sigma(x_\sigma^{\mathrm{T}}\beta_\sigma), l_\xi(x_\xi^{\mathrm{T}}\beta_\xi))$. We also partition the p-dimensional regressor x and regression parameter β according to the parameter $\vartheta = (\sigma, \xi)$, i.e. $x = (x_\sigma, x_\xi)$ and $\beta = (\beta_\sigma, \beta_\xi)$. Moreover, we get that $\theta = x^{\mathrm{T}_\pi}\beta = (x_\sigma^{\mathrm{T}}\beta_\sigma, x_\xi^{\mathrm{T}}\beta_\xi)$.

The Fisher information for these both models is 2×2 symmetric matrix of the form:

$$\mathcal{I}^{\mathcal{Q}}_{\sigma,\xi} = \mathbf{E}_{\sigma,\xi}\Lambda^{\mathcal{Q}}_{\sigma,\xi}(y)(\Lambda^{\mathcal{Q}}_{\sigma,\xi}(y))^{\mathrm{T}} = \begin{pmatrix} I_{\sigma\sigma} & I_{\sigma\xi} \\ I_{\sigma\xi} & I_{\xi\xi} \end{pmatrix},$$

therefore we obtain

$$\dot{l}^{\mathrm{T}}\mathcal{I}^{\mathcal{Q}}_{\sigma,\xi}\dot{l} = \begin{pmatrix} \dot{l}_\sigma^2 I_{\sigma\sigma} & \dot{l}_\sigma\dot{l}_\xi I_{\sigma\xi} \\ \dot{l}_\sigma\dot{l}_\xi I_{\sigma\xi} & \dot{l}_\xi^2 I_{\xi\xi} \end{pmatrix}.$$

Plugging last expressions to the conditions (ii) and (iii) of Theorem 6.1 we rewrite conditions in the following from

$$\text{(ii)} \qquad \int \dot{l}_\sigma^2(I_{\sigma\sigma} + I_{\sigma\xi})|x_\sigma|^2 K(dx) + \int \dot{l}_\xi^2(I_{\xi\xi} + I_{\sigma\xi})|x_\xi|^2 K(dx) < \infty,$$

$$\begin{aligned}\text{(iii)} \qquad & \int ((\dot{l}^2_{\sigma+s}I_{\sigma+s\sigma+s} - \dot{l}^2_\sigma I_{\sigma\sigma})|x_\sigma|^2 + 2(\dot{l}_{\sigma+s}\dot{l}_{\xi+s}I_{\sigma+s\xi+s} - \dot{l}_\sigma\dot{l}_\xi I_{\sigma\xi})|x_\sigma||x_\xi| + \\ & +(\dot{l}^2_{\xi+s}I_{\xi+s\xi+s} - \dot{l}^2_\xi I_{\xi\xi})|x_\xi|^2)K(dx) \leq \int (\dot{l}^2_{\sigma+s}(I_{\sigma+s\sigma+s} + I_{\sigma+s\xi+s}) - \\ & -\dot{l}^2_\sigma(I_{\sigma\sigma} + I_{\sigma\xi}))|x_\sigma|^2 K(dx) + \int (\dot{l}^2_{\xi+s}(I_{\xi+s\xi+s} + I_{\sigma+s\xi+s}) - \\ & -\dot{l}^2_\xi(I_{\xi\xi} + I_{\sigma\xi}))|x_\xi|^2 K(dx) \to 0, \quad s \to 0.\end{aligned}$$

Next, we closer consider case of each model.

GEVD model: We start with the scale-shape model $\mathrm{GEVD}(0, \sigma, \xi)$ which has error distribution $Q_\vartheta(y) = \exp\big(-(1+\xi\frac{y}{\sigma})^{-\frac{1}{\xi}}\big)$.

The Fisher information matrix of this model can be written explicitely in the following way

$$\mathcal{I}^{\mathcal{Q}}_{\sigma,\xi} = \xi^{-2} D \begin{pmatrix} I_{\sigma\sigma} & I_{\sigma\xi} \\ I_{\sigma\xi} & I_{\xi\xi} \end{pmatrix} D, \quad \text{where} \quad D^{-1} = \operatorname{diag}(\sigma, \xi),$$

$$\begin{aligned} I_{\sigma\sigma} &= (\xi+1)^2\Gamma(2\xi+1) - 2(\xi+1)\Gamma(\xi+1) + 1, \\ I_{\sigma\xi} &= -(\xi+1)^2\Gamma(2\xi+1) + (\xi^2+4\xi+3)\Gamma(\xi+1) + (\xi^2+\xi)\Gamma'(\xi)\Gamma(\xi+1) - \xi\Gamma'(1) - \xi - 1, \\ I_{\xi\xi} &= (\xi+1)^2\Gamma(2\xi+1) - 2\Gamma(\xi+3) - 2\xi\Gamma'(\xi)\Gamma(\xi+2) + 2\xi(\xi+1)\Gamma'(1) + \\ &\quad + \xi^2(\Gamma''(1) + (\Gamma'(1))^2) + (\xi+1)^2. \end{aligned}$$

One can see, that the Fisher information matrix has singularities in the values $\xi = 0$ and $\xi = -1/2$, what confirms our conclusion, mentioned above, that condition (i) of Theorem 6.1 are fulfilled only as long as $\xi \in (-1/2, 0)$ or $\xi > 0$.

GPD model: The GPD$(0, \sigma, \xi)$ scale-shape model has error distribution function $Q_\vartheta(y) = 1 - (1 + \xi\frac{y}{\sigma})^{-\frac{1}{\xi}}$ with the Fisher information matrix:

$$\mathcal{I}^{\mathcal{Q}}_{\sigma,\xi} = \frac{1}{1+2\xi} D \begin{pmatrix} 1, & 1 \\ 1, & 2(\xi+1) \end{pmatrix} D, \qquad D^{-1} = \operatorname{diag}(\sigma, \xi+1).$$

Again condition (i) is fulfilled for $\sigma > 0$ and $\xi > -\frac{1}{2}$, what is reflected by a singularity at $\xi = -1/2$ of the Fisher information.

Link function: As for the componentwise link function, trivial choice for the scale for both models is log link, i.e. $l_\sigma(x_\sigma^{\mathrm{T}}\beta_\sigma) = \exp(x_\sigma^{\mathrm{T}}\beta_\sigma)$.

Due to a lack of equivariance in the shape, it is harder to choose the link function for it and none of the canonical link functions fits in this case. The admissible link function for the shape should be smooth and strictly increasing (for identifiability). Empirical information (non-regression-based), received from our GEVD and GPD applications, restricts the shape ξ to be in the interval $(0, 2)$. Therefore, good link function should not exclude values which are out of this interval $\xi \notin (0, 2)$, but make them hard to reach.

Moreover, main challenge while modeling parameter driven time dependencies in the terminology of Cox (1981), with the usage of our GLMs with generalized extreme error distributions in the time series context, is to design link functions, which let regressors follow GEVD or GPD distribution themselves. The problem here is that then we have to integrate against very heavy tails. In particular, we aim to construct autoregressive-type

time series for the scale and shape of the form

$$X_t \sim \text{GEVD}(l(X^{\mathrm{T}}_{(t-1):(t-p_1)}\beta_\sigma, X^{\mathrm{T}}_{(t-1):(t-p_2)}\beta_\xi)) \quad \text{for} \quad X_{(t-1):(t-p)} = (X_{t-1}, ..., X_{t-p}) \tag{6.13}$$

Here all negative values of β_ξ dampen clustering of extremes, as then usually the large value obtained from the large positive shape is followed by an observation with low, or even negative, shape parameter (hence with much lighter tails), thus, in general a smaller value; correspondingly β_ξ positive will foster clustering of extremes.

Therefore, the first idea is to use the log link function for the shape parameters as well as for the scale. But using this link for GEVD or GLM time series we get that the integrability condition (ii), equation (6.5), is not be satisfied in this case. From this fact we conclude that the admissible link function for the shape should also grow very slowly.

To design such link function, we also note, that in the case of GEVD errors all terms of the Fisher information matrix are dominated by the term $\Gamma(2\xi+1)$, hence conditions (ii) and (iii) of Theorem 6.1 are fulfilled if for large positive values θ_ξ, the link function grows so slowly to ∞ that $\Gamma(2l_\xi(\theta_\xi)) \approx \log(\theta_\xi)$, which for large x behaves like the iterated logarithm $\log(\log(x))$.

Applying similar technique for the case of GPD errors, we obtain link function for the shape parameter behaving like the logarithm, i.e. $l_\xi(\theta_\xi) \approx \log(\theta_\xi)$.

After we collected all required properties for the shape link, we suggest the candidates of it for GLM with GEVD and GPD error distributions. For simplicity, let $p = 1$, the link for the GEVD would be of the form

$$l_\xi(\theta_\xi) = \log(f(\log(x_\xi)^{\mathrm{T}}\beta_\xi)),$$

where function $f(x)$ behaves like quadratic function, e.g. $x^2/2 + x + 1$ for $x > 0$, and for $x < 0$ it is like the function $a_1/(\log(a_2 - x))^2 + a_3$ with $a_1, a_2, a_3 > 0$ such that f is continuously differentiable in 0 and $f(x) > e^{-1/2}$ for all x.

In Appendix B.4 we check if our choice of the link function for the GEVD shape-scale model fulfills conditions (ii) and (iii) of Theorem 6.1 and calculate approximate values for the constants a_1, a_2, a_3.

Remark 6.10. *Obviously, the next question would concern (asymptotic) stationarity of the time series (6.13) for $t \geq 0$ and for given starting values $x_{-1}, \ldots, x_{-\max(p_1,p_2)}$, using proposed link function. We leave this question open for the further research.*

6.3 Fixed-point algorithm

In this Section we focus on the robust optimality problem described in detail in Rieder (1994, Ch. 5). Solving this standardized MSE problem for implicitly defined Lagrange multipliers is a difficult question. This issue was looked already by Hampel (1968). Later, in the book of Hampel et al. (1986) fixed point iteration algorithm for the computation of the optimally robust influence curve (also called Hampel-type IC) was proposed. The notion of the optimal influence function was studied in detail by Rieder (1994, Thm. 5.5.1 and 5.5.7 (b)) and regression optimal influence function was discussed in Rieder (1994, Ch. 7). Moreover, in our infinitesimal setting described above, there is corresponding algorithm sketched in Rieder (1994, Rem. 5.5.2). More on the optimal influence function for regression model one can find it in Kohl (2005).

Here we present general algorithm of computing the Hampel-type optimal influence curve, which, comparing to other similar algorithms, uses another techniques to get some intermediate values. We have implemented this algorithm in `R` in the function named `FixPglm`. To keep it simple, we wrote exemplary versions of the code for two specific cases, Binomial model and Generalized Pareto shape model.

For both cases, `FixPglm` is a function which requires next input parameters: the matrix of regressors X; fixed parameters of the corresponding distributions, i.e. number of trials for Binomial case and location and scale parameters for the GPD; matrix A_0 which is simply the inverse of the Fisher information matrix; the link function l; the regression parameter β; the clipping hight b, which can be computed solving Anscombe criteria; and ϵ for the stopping criteria of the algorithm. In `R` this function can be called by the following command:

```
FixPglm(X, param, A0, link, beta, b, eps)(x,y)
```

Here, in order to make clear the real structure of the algorithm, we prefer not to explicate the stages, where we might need to insure against dividing by zero, or check suitability of some intermediate elements. Later, in Chapter 8 we discuss some challenges of the algorithm for the case of the GPD model.

`FixPglm` algorithm

(-1). We start our algorithm with the preparation step inside the function, where we determine the link function l and its derivative $\dot{l}$ (as slots of the input link function), calculate value of the term $X^{\mathrm{T}}\beta$ and plug it in the both functions in oder to get

the values of $\vartheta = l(X^{\mathrm{T}}\beta)$ and its derivative $\dot{l}(X^{\mathrm{T}}\beta)$. Here we also compute the L_2 derivative function correspondingly to the fixed parameters.

(0). After we defined everything needed for the further use, we pass to the introductory part of the algorithm. First, we rename the matrix A_0, assigning it notion A. We define additional function z of x, and on this stage let it be zero. Then, we denote the difference between the L_2 derivative function and function z as a new function v, i.e.

$$z_0(x) = 0, \quad , v_0(x, y) := \Lambda^{\mathcal{Q}}_{\vartheta}(y) - z_0(x).$$

(1). Here starts the main block of our algorithm, where we define several additional functions to make the computation easier to follow. We introduce the iteration symbol i, which provides iteration of the corresponding procedure from $i = 1$, until some stopping criteria breaks it. We start with the function

$$c_i(x) := \frac{b}{\left| A_i X \dot{l}(X^{\mathrm{T}}\beta) \right|}. \tag{6.14}$$

Then, we compute function of two variables x and y, where we divide function c by the norm of function derived in the zero step

$$w_i(x, y) := \min(1, \frac{c_i}{|v_i(x, y)|}). \tag{6.15}$$

Next, we redetermine function z_0 to be the division of two expectation w.r.t. the parameter ϑ, i.e.

$$z_{i+1}(x) := \frac{\mathbf{E}_{\vartheta}(\Lambda^{\mathcal{Q}}_{\vartheta}(y) w_{i+1}(x, y))}{\mathbf{E}_{\vartheta}(w_{i+1}(x, y))}. \tag{6.16}$$

We also reassign function v_0, similarly to its previous form, but now with the updated version of the function z, i.e.

$$v_{i+1}(x, y) := \Lambda^{\mathcal{Q}}_{\vartheta}(y) - z_{i+1}(x) \tag{6.17}$$

Another intermediate function is the expectation taken w.r.t. to the variable y

$$t_{i+1}(x) := \mathbf{E}_{\vartheta}(v^2_{i+1}(x, y) w_{i+1}(x, y)). \tag{6.18}$$

At last, we rewrite matrix A as the following expectation w.r.t. regressors distribution

$$A_{i+1} := (\mathbf{E}(X X^{\mathrm{T}} (\dot{l}(X^{\mathrm{T}}\beta))^2 t_{i+1}(x)))^{-1} \tag{6.19}$$

For simplicity, we suppose that $\mathbf{P}(X = X_j) = 1/n$ for $j = 1, ..., n$, therefore, in the code we treat this expectation as arithmetic mean of the underlying vectors.

Note, that in (6.19) we already get not necessarily optimal, but valid influence function.

(2). We iterate block (1)., i.e. equations (6.14)-(6.19), until one of two stopping criteria breaks the iteration. Both criteria are based on the relative difference between current and previous iterations, first for the function z and second for the matrix A. If such relative difference becomes smaller than the chosen value of ϵ, iteration stops. Usually, the first criteria breaks the loop.

(3). After iteration stopped, we compute the optimal IF as a function of pair (x, y), multiplying last iteration values of the corresponding functions, i.e.

$$IF(x, y) := AX\dot{l}(X^{\mathrm{T}}\beta)v(x, y)w(x, y).$$

Note, that after each iteration in (1) after (6.19) we obtain a valid regression influence function $\psi(x, y) = A_{i+1}(\Lambda(y) - z_{i+1}(x)) \min(1, c_i(X)/|v_i(x, y)|)$, i.e. $\mathbf{E}[\psi(x, y)|x] = 0$ and $\mathbf{E}[\psi(x, y)\Lambda(y)^{\mathrm{T}}] = \mathbb{I}_p$.

First what we would like to have, working with any iterative procedure, is the proof of convergence of the algorithm, but as far as we know, there is no such proof for our algorithm up to now. What we can check, is the accuracy of the algorithm and the optimality at each iteration step. Here we can claim, that ones we reached the stationary point, where function z or matrix A do not change much from the iteration to the iteration, we know that the influence curve we computed is optimal. Moreover, this algorithm provides some continuity in the sense, that if we even do not reach the limiting point, but we are close to it, we are also close to the optimal solution (see Kohl (2005)).

Main difficulties implementing this algorithm we faced in the block (0).The reason is that, for the computation of the expectations in the block (1)., we need the L_2 family distribution with different values of the parameters on each iteration. Therefore, already in the (0). block, we compute list of such L_2 family distributions and plug its elements in the expectation operator one by one.

Binomial example of `FixPglm`

For this example we took the `R`-data "carrots" from the package `robustbase`, already mentioned in the Section 2.4.7. To load these data one can use the following commands

```
> require(robustbase)
> data(carrots)
```

Then, we compute corresponding matrix of the regressors X

```
success<-carrots$success
total<-carrots$total
logdose<-carrots$logdose
block<-carrots$block
blockb1<-c(rep(1,8),rep(0,16))
blockb2<-c(rep(0,8),rep(1,8),rep(0,8))
blockb3<-c(rep(0,16),rep(1,8))
X<-cbind(success,total,logdose,blockb1,blockb2,blockb3)
n <- dim(X)[1] #number of the regressors
k <- dim(X)[2] #dimension of each regressor
```

As usually we take the logit link function and extract derivative function from it. Then, we define the fixed parameter of the Binomial distribution

```
link <- make.link("logit")
linkfct <- link$linkinv
linkder <- link$mu.eta
fixedparam<-total
```

For the computation of the regression parameter vector β we use the following GLM model:

```
Cfit1 <- glm(cbind(success, total-success)~ X-1, data=carrots, family=binomial)
beta=coef(Cfit1)
```

Next we calculate matrix A_0 as inverse of the Fisher information matrix and again for simplicity we treat the expectation as arithmetic mean (using function `rowMeans`) of the underlying vectors, i.e. in our case

```
vartheta <- X%*%beta
param <- sapply(vartheta,linkfct)
paramder <- sapply(vartheta, linkder)
X1 <- X*paramder*sqrt(fixedparam)/sqrt(param*(1-param))
A <- rowMeans(apply(X1,1,function(x)x%*%t(x)))
dim(A) <- c(k,k)
A <- ginv(A)
A0<-A
```

and choose clipping hight b and parameter of the stopping criteria ϵ

```
b<-800
eps<-0.05
```

To get the value of oprimal IF we need to set pair of (x, y), so

```
x<-X[1,]
y<-carrots$success[1]
```

Results and the speed of the `FixPglm` performance for the introduced data are the following

```
> FixPglm(X=X, fixedparam=fixedparam, A0=A0, link = "logit", beta=beta,
+ b=b, eps=eps)(x,y)
          [,1]      [,2]      [,3]     [,4]     [,5]    [,6]
[1,] -66.72038 -90.60094 -2901.978 10776.38 9734.015 8943.32
> system.time(FixPglm(X=X, fixedparam=fixedparam, A0=A0, link = "logit",
+ beta=beta, b=b, eps=eps)(x,y))
   user  system elapsed
   2.26    0.00    2.28
```

Since we did not face some obvious problems and got reasonable results in this example, we conclude that our algorithm performs well for the Binomial model. Of course there might be the way to improve the algorithm and increase the speed of the computation, but since the main goal was to implement this algorithm in the easiest way, we are more than satisfied with the result.

Note, that we apply this algorithm for the GPD shape model to the real hospital data in Chapter 8, where we also point out main challenges or drawbacks of the algorithm.

6.4 Conclusions

In this Chapter we presented our approach of L_2 differentiability for the generalized linear models. The main idea was to generalize theory of Rieder (1994) on L_2 differentiability for linear regression models to the required case. We focused on the non-exponential scale-shape families, i.e. generalized extreme value and generalized Pareto distributions.

Our important achievement is that our approach also covers higher dimensional error distributions and case of regressors of possibly different length for each parameter, what is new. Similarly to Rieder (1994), we separately considered cases of stochastic and

deterministic regressors and computed corresponding theorems with L_2 differentiability conditions, L_2 derivatives and the Fisher information matrices for each case.

We also tested our approach on the linear regression, Binomial and Poisson GLMs with respective link functions and, finally, GEVD and GPD joint shape-scale models. Main challenge for two last models was to obtain the appropriate componentwise link function, what have been done and discussed in details.

The last Section of this Chapter was focused on the algorithm `FixPglm`, which computes Hampel-type optimal IC's. We presented the iteration scheme and discussed its properties. We tested function `FixPglm` for the Binomial case, with the `R`-data "carrots", and checked the time of its performance. As we could conclude, for the exemplary implementation of the algorithm, `FixPglm` showed quite good results.

Part III

Applications

Chapter 7

Applications

We have mentioned in Chapter 1, that this thesis is written in the framework of the project "Robust Risk Estimation", based on the cooperation of four different institutions. We focus on three research areas: hydrological application for modeling discharge data, clinical application for the estimation of length of stay at an intensive care unit and operational risk of a banks. Members of the project, who are responsible for each application, are Bernhard Spangl from University of Natural Resources and Life Sciences, Matthias Kohl from Furtwangen University and Peter Ruckdeschel from Fraunhofer ITWM, Kaiseslautern respectively. All other members, i.e. Gerald Kroisandt, Sascha Desmettre and Mykhailo Pupashenko, focused partially on the problems arising in each of three applications.

For this dissertation we give the overview on the problematic and reached results of all three applications, but hydrology and operational risk are not prescribed further in this thesis, whereas length of stay is considered more explicitly in the real data example.

Each of these applications has its own specific problems, but the ways to solve these problems are based on the methods described in the previous Chapters, what combines them together in one project.

7.1 Hydrology

We start with the hydrological application and first, discuss the research questions concerning the river discharge data, done under the guidance of Bernhard Spangl. It is obvious, that based on environmental reasons , e.g. irregular climate pattern, non-stationarity (neither geographically, nor temporal), river discharge data are full of extremes and spikes. Hence, the main challenge in this application is related to the analysis

and modeling of extreme events in discharge data, especially their frequency and magnitude. Our primary research question is to create specific approaches to solve this problem.

One of the challenges in this application is to distinguish anthropogenic impact and natural fluctuations. For that we would follow Tukey's idea of "borrowing strength", where one can use information from other datasets for the estimation at one specific dataset. The hydrological view on this method was proposed by Hosking and Wallis (1997), where authors raised the issue of regionalisation and seasonality.

7.1.1 Available data

In this application we worked with data, collected in various sites in Austria over the last 35 years, or in some cases even longer. These data consist of hourly and daily average discharge time series of rivers from various Austrian regions.

Single time series of the data were measured in the alpine and high-alpine areas. Some of data were taken from the pre-alps and the Bohemian Massif, as well as from some large rivers, like the Danube or the Salzach. Each category includes not only the pristine rivers, but also rivers with an anthropogenic impact caused by water transfers or storages.

Data contain some additional geographical variables, e.g., longitude, latitude, sea-height or catchment area.

We also had access to data collected in Saxony, where annual maxima based on daily average discharge data were measured, and two daily average discharge series from Bavaria.

7.1.2 Main approach

We considered two approaches for this application. The first one is based on the idea of filtering the average discharge time series in order to get rid of systematic trend or seasonality in the data and then applying robust extreme value theory. Under "filtering" in this method we mean the use of the robust filtering combined with the robust signal extraction, like in Fried et al. (2007).

Doing so, we extract the dynamic structure from the series and the remainder data (innovations) do not show any dynamics, that is why we can apply EVT. We get the estimated resulting innovations which still contain extremes and therefore, keep the ability to detect extremes and spikes in those data.

For this approach we apply Kalman filtering procedures, but not the classical ones, since they will not perform particularly well in this situation. Instead, we use our new robust recursive filters and smoothers, discussed in Chapter 5. More precisely, to daily average discharge data we apply methods implemented in the `R`-package `robKalman`, which is described in Section 5.5. To remind, they include classical and robust, extended and unscented Kalman filters and corresponding smoothers.

Second approach uses generalized linear models for the GPD errors in the way introduced in Chapter 6. We choose some suitable link function and link parameters of the GPD to the corresponding regressors. This approach more explicitly models time dependence for the extremes or exceedances directly.

7.2 Medicine

Length of stay (LOS) is a term to describe the duration of single patient hospitalization. It is one of the most important notions used for the various purposes in the medical application. LOS can be treated as an indicator of the hospital activity and applied for the management of the hospital care, quality control, appropriateness of hospital use.

The second application of our results we considered, guided by Matthias Kohl, is related to the hospital LOS at an intensive care unit (ICU) and respective costs spent on each patient. From year 2004, so-called Diagnosis Related Groups (DRG) started to classify patients of German hospitals, so that the hospitals are paid by cases. Therefore, the comparison of the length of stay and costs between different departments, clinics or classification schemes is very important for the health care system and for each hospital particularly.

LOS is stochastic, therefore it has some natural fluctuations, this is why predictions for expected LOS have to be complemented by some assessments of fluctuations (some risk measure).

One should note, that in any hospital we can have some atypical patients, which have longer LOS than we expected, hence require higher costs. Such extreme cases can be caused by lots of reasons and can be treated as outliers from the captured data. Although, for the correct prediction these extreme cases should be modeled as well. Moreover, since LOS often has very skewed distribution, impact of such atypical patients with longer length of stay can be dramatic.

Therefore, our main goal is to make robust prediction of the length of stay and the costs constructing corresponding regression models, i.e. we apply extreme value theory

in combination with regression-type models. Moreover, we face the problem of model selection and validation.

7.2.1 Available data

Data used for this application is taken from the Jena university hospital ICU. The department for anesthesiology and intensive care medicine in the 1990s adopted the electronic patient-data-management-system named COPRA (Computer Organized Patient Report Assistant). COPRA was developed in clinical practice at the University of Leipzig to get rid of the need in manual or hand-written documentation. COPRA enables the calculation of expenses for the complex intensive care treatments and additional costs for medicine and blood.

The current version of the data, "COPRA V", includes all relevant vital parameters for each patient, e.g. diagnoses, laboratory results, medications, LOS etc. It has been successfully used for the last years and became one of the most complete databases for critical ill patients. "COPRA V" includes more than 52000 cases with more than 210000 patient days.

7.2.2 Main approach

The main challenge we faced in this application was a large number of covariates and in addition a inhomogeneous population. It means, that we had to work with non i.i.d. EVT, where every patient has his own extreme value distribution parameters. Our solution to this problem was our GLM approach which could capture that.

First, working with the data described above, we decided that it is too large to use, as e.g. version of the data we use in Chapter 8 contains 209 variables and over 21000 observations. The way out of the problem is to apply classical and robust variable selection techniques to the data, in order to reconstruct it and reduce its dimension. In this way one can select the most informative variables for LOS and costs. Within the project we spent some time to formalize variable selection techniques.

As we have mentioned, we aim to construct regression models for the robust prediction of the LOS and costs. Since typical LOS distributions are skewed and contain outliers, main model candidates here would be Weibull, Gamma, GPD and GEV distributions. All these distribution are implemented in the R-package `RobExtremes`, which was discussed in Section 2.8.

For linear and GLM context we speed up our algorithms computing algorithm, which, with usage of the two-dimensional interpolation, quickly calculates and saves Lagrange multipliers arising in the optimally-robust procedures on a grid of parameter values offline. Partly it is implemented in the function `FixPglm`, studied in Section 6.3, which obtains optimally-robust estimators.

In order to improve the obtained models for robust prediction of LOS and costs, we developed a concept which we called Bed-at-Risk. It is a high upper quantile of the LOS distribution and may be used to control average length of stay. The usage of this concept in the planned surgeries was also sketched in the framework of the project, paper in preparation.

7.3 Operational risk

Here we would like to present another application of our research, to the financial mathematics domain, which was mainly guided by Peter Ruckdeschel. We focus on the notion of the operational risk (OR). We cite definition from the second of the Basel Accords, which contain recommendations on banking laws and regulations, Basel II. Operational risk is "the risk of direct or indirect loss resulting from inadequate or failed internal processes, people and systems or from external events".

Basel II also states, that all banks have to maintain regulatory capital, so that unexpected losses caused by these risks will not lead to the bankruptcy. There are various approaches for this purpose, suggested in Basel II. We focus on so-called Loss Distribution Approach (LDA). Idea of this method is to group operational risk data into cells, representing the bank's business lines and risk events. We fit each cell to historical data and model separately severity and frequency of losses and determine total loss from the respective compound distribution. In our research we basically focused on the robustness issues of the approach.

For this application we also suggested to distinguish "expected" and "unexpected" losses as body and tail of the corresponding distribution. Our propose was to take log-normal or Weibull body distribution and GPD for the tail.

Main goal of this application was to quantify the regulatory capital, which can be computed as some risk measure evaluated at the distribution, resulting from aggregation of the cell-wise fitted model distributions. In a realistic modeling, taking into account possible model deviations, one cannot tell (without error) whether these events are singular outliers or reproducible and, hence, contribute valuable evidence for future losses.

Therefore, we aim to create methods of fitting cell-individual severity distributions to the data, which remain stable under model deviations, hence robust.

7.3.1 Available data

First, one should note that obviously data collected from one bank is rather short. Hence, in order to increase amount of the given information, banks usually pool their data in consortia, e.g. the most important data provider in this field is ORX association (www.orx.org). Since then it is not clear if the used data is appropriate for each bank, amount of the robustness problems in the approach increases.

Another possible challenge using external data is so-called censoring problem, since data usually is only reported beyond a certain threshold, whereas very large operational losses are observed rarely.

The data we worked with is operational loss data collection `Algo OpData` of Algorithmics Inc. This database has been collected within last 40 years and the majority of the losses were observed during last 20 years. By July 2010, `Algo OpData` contained more than 12,000 operational risk losses from all industry sectors. Moreover, these data provides detailed information about operational loss events over 1 million USD from 2431 financial institutions according to Basel II business line and event type definition.

As we just mentioned, usually data collected from public sources is censored, therefore the severity of losses is likely to be (heavy-tailed). This makes these data different from other external operational loss data as e.g. ORX database. Here we consider the "unexpected" losses and use them to model the extreme tails of severity distributions.

7.3.2 Main approach

The main challenges in this application were the aggregation, outliers presence and intertemporal stability, meaning that the risk figure today should be close to the one yesterday and two last problems were captured by the robustness.

More precisely, the first step of the research was related to the robust scaling problem of the data. As we have mentioned, the external operational loss data is collected from different banks, so in order to estimate operational risk, a scaling step is necessary to scale the data to make losses comparable among banks. N. Horbenko and P. Ruckdeschel worked on quantile regression for this scale. Another more direct approach directly builds up on our GLM-results for scale-shape regression with GPD or GEVD errors.In

particular the optimally robust estimators sketched in Section 2.5 directly contribute to a more stable assessment of this scaling.

Using quantile regression on various factors, we tried to assess the bank-individual severity of operational losses. Here we tried two different approaches, the first based on the standard quantiles and the second, modeling scale of a GPD with a GLM.

Since both methods are not robust against leverage points in response variables, we have to use an optimal robust estimation procedure for regression, and the challenge here is to compute a globally robust starting estimator. After this problem is solved, we can obtain optimally robust estimator by the one-step reweighting estimation, starting with a globally robust estimate and using the influence function of the MSE-optimally robust estimator.

Next we would need simultaneous regression for shape and scale, where we will use multivariate regression mentioned before in other applications.

7.3.3 Conclusions

In this Chapter we presented the applications, which were studied in the framework of the project "Robust Risk Estimation". For each of them, we discussed main issues, presented the available data and pointed the specific problems, which can be solved by using our results.

Chapter 8

Examples

8.1 Example on the real data

8.1.1 Hospital data

In this Section we apply the algorithm, presented in Section 6.3, to the real data taken from the Jena university hospital ICU. These data contain 309 vital parameters for each patient, e.g. diagnoses, laboratory results, medications, length of stay etc.

For our example we focus on a couple subjectively selected variables, taking 5 parameters from the hospital data as the regressors. The main purpose of this variable selection is to get the relevant parameters using some expert opinion, in our case received from Dr. Gordon Otto (Clinic for Anesthesiology and Intensive Care department of the Jena University Hospital). The second important selection criterion is to work with both types of variables, i.e. boolean and numerical.

More precisely, three of the chosen parameters are boolean, i.e. they have only one of two values for each patient, i.e. `TRUE` and `FALSE`. These parameters indicate whether each patient has some specific disease, e.g. `Cancer` or `Sepsis`, and if this patient receives special treatments as e.g. `Dialyse` (used primarily for people with renal failure).

Two remaining of chosen parameters belong to the SOFA (Sepsis-related Organ Failure Assessment) scoring system, which determines the extent of the person's organ function or the rate of failure. The score is based on six different scores for the respiratory, the cardiovascular, the hepatic, the coagulation, the renal and the neurological systems. In our example we focus on the parameters `Gesamtscore.SOFA` and `SOFA.max`.

```
X<-cbind(Cancer, Sepsis, Dialyse, Gesamtscore.SOFA, SOFA.max)
```

To complete the data, for y we take the parameter `ITS.Tage`, i.e. number of days, that were spent in the ICU.

The given data set contains information about 21757 patients, and the same is the dimension of the observation. First, to be able to apply our algorithm to the data, we through away all regressors which contain value `NA`. More precisely, parameters `Gesamtscore.SOFA` and `SOFA.max` contain a large number of the not available values. After cleaning regressors, we achieve reduced dimension equal to 18623.

For the further analysis, to have better idea about the possible values of these parameters, we draw the histograms for their values.

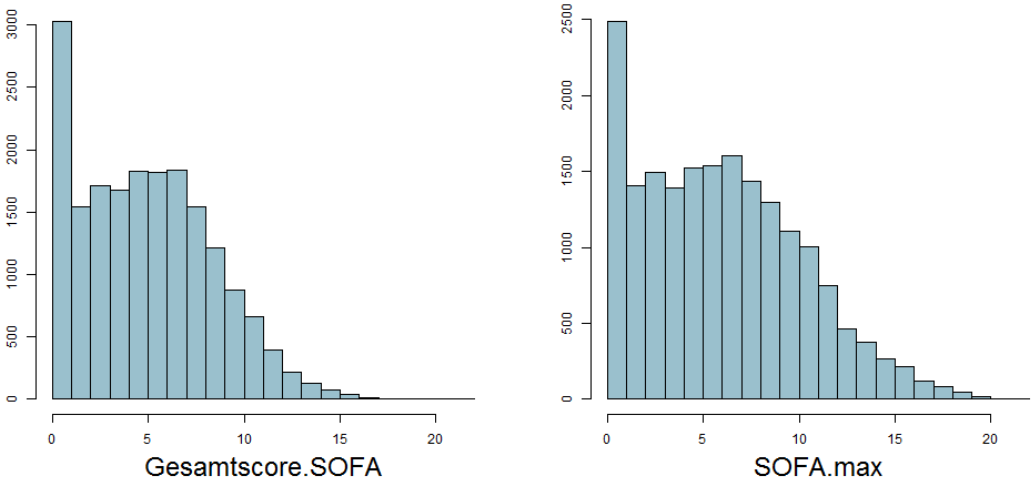

FIGURE 8.1: Gesamtscore.SOFA and SOFA.max histogram for 18623 patients

As for the boolean parameters, we calculate amount of the `TRUE` of 18623 values for each of them, and get:

```
> trueCancer
[1] 4258
> trueSepsis
[1] 982
> trueDialyse
[1] 1095
```

8.1.2 Model description

In this Chapter, we fit described data with the GLM with generalized Pareto error distribution. We link shape parameter ξ of the GPD to the linear predictor $X\beta$ via

corresponding link function and, later, estimating regression parameter β, compute unknown shape parameter. In this Section we discuss how we select all elements of the corresponding GLM, i.e. nuisance parameters of the GPD and link function.

Threshold selection remains a delicate questions and has not been covered by this thesis. Obviously one needs a compromise between a good approximation quality in the Pickands–Balkema–de Haan theorem 3.14, which encourages high thresholds, and a decent number of remaining observations for inference beyond this threshold which would speak for the lower threshold. In fact the threshold could be chosen by cross-validation techniques trying to minimize the MSE.

In our thesis we have selected the threshold by the requirement that $n = 1000$ observations are beyond this threshold, which amounts to taking the threshold at the upper 5.4% quantile of the data. Fact that all observations are strictly larger than the threshold is very important for our function, since otherwise, when we compute L_2 derivative for the shape parameter (3.14), we might get problems obtaining the value $\log 0$.

Parallel to the choice of the threshold, which turns to be equal to 19, we reduce the dimension of the data to 1000.

Next, we do the scale σ selection for our GPD. Since we estimate scale together with the β estimation up to some point, and compute MLE and robust estimate (using skipping technique) for the scale, it will be discussed in the next Section in details.

As for the choice of the link function, first, we follow our arguments from Section 6.2. There we obtained, that for the GPD errors, the link function for the shape parameter behaves like the logarithm, i.e. $l_\xi(\theta_\xi) \approx \log(\theta_\xi)$.

After we applied this link function to our real data, we observed several drawbacks in the design of the function, therefore, we adjust it to avoid various errors.

The first error we faced is in the argument of the link function θ_ξ. This expression can get zero value, especially for the small size of the regressors. The problem is that our logarithm link function is not defined in zero. Therefore, we aim to secure ourselves from such error.

Next possible error can come from the fact, that observations with the Generalized Pareto distribution have to lie in the following support:

$$y \in \begin{cases} (\mu, \infty), & \xi \geq 0 \\ [0, \mu - \sigma/\xi), & \xi < 0. \end{cases}$$

If we reformulate it to create restrictions for the shape parameter of the GPD, we get the following:

$$\xi \geq -\frac{\sigma}{y-\mu}, \quad \text{when} \quad \xi < 0.$$

therefore, this also has to be satisfied by the link function.

Next what we should keep in mind is the L_2 derivative for the GPD is not defined for $\xi = 0$, see Section 3.2.2, moreover converging against zero for positive shapes will explode the Fisher Information, i.e. $\lim_{\xi \to 0} I_\xi = \infty$. Therefore, we have to consider two separate parameter areas, $\xi \in (-1/2, 0)$ and $\xi > 0$. Note, that we do not exclude zero value from our link, but we make it very hard to achieve. For both areas we should compute appropriate link functions. Here we compute the link function for $\xi > 0$ and for $\xi \in (-1/2, 0)$ any binomial link l_0, with values in $(0, 1)$, after transformation $l(x) = l_0(x)/2 - 1/2$ will fit.

Adjusting link to the data we work with, we choose the following function

$$l_\xi(x) = \frac{\exp(\frac{x}{20})}{(\exp(\frac{x}{20}) + 1)} h_1(x), \tag{8.1}$$

with

$$h_1(x) = \begin{cases} 2/5, & x < 0 \\ 1/5 + \log(1 + x), & x > 0. \end{cases}$$

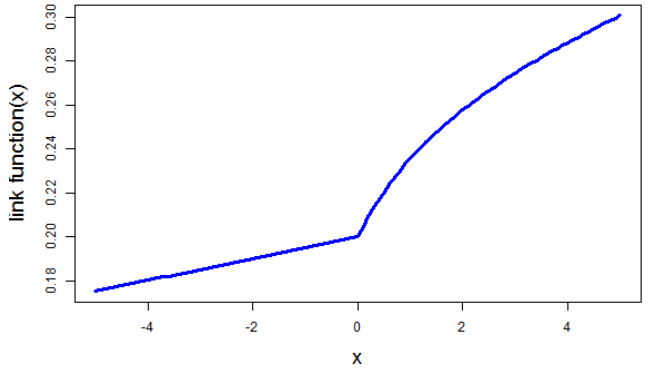

FIGURE 8.2: Link function for the shape parameter of GPD

In Figure 8.2 one can see the plot of the final choice of the link function for the shape parameter of the GPD. Derivative of this link is the following:

$$l_\xi(x) = \frac{\exp(\frac{x}{20})}{200(\exp(\frac{x}{20}) + 1)} h_2(x), \tag{8.2}$$

with

$$h_2(x) = \begin{cases} 4/(exp(x/20)+1), & x < 0 \\ \log(1+x)/(\exp(x/20)+1) + 20/(1+x), & x > 0. \end{cases}$$

8.1.3 Regression parameters selection

In this thesis we do not work out the interpolation step in `FixPglm` - from theory, the influence function $\psi(x, y)$ maps arbitrary combinations (x, y) to the (tangent space of) the parameter space. In particular, in the optimal influence functions, the centering z, i.e. (6.16), must be a function of x. For the sake of this thesis we limit ourselves to computation of the IF at the actual data, at the set $(X[i,], y[i])$, which makes computation considerably simpler. While this has clear drawbacks for diagnostics (e.g. whatif analysis gets much harder) it pays off in terms of computational time. With these reasons, we make our function `FixPglm` be the function of the index i, i.e.

```
FixPglm(X, locat, scale, A0, link, beta, b, eps)(i)
```

Here we describe the selection of the regression parameter β and, as we have mentioned above, scale parameter of the GPD. We aim to obtain the robust starting estimator for both parameters.

First, we apply link function to the linear predictor, i.e. to the product of regression matrix and regression parameter, $X\beta$. Then, we compute likelihood function, as the function of two parameters, β and σ, and maximize it (i.e. multiply it with -1 and minimize) using the `R`-function `optim`. In this way we obtain some kind of the MLE for both parameters. For 1000 observations of dimension 5 we get `thetaMLE`, where first coordinate is the MLE of the logarithm of the scale, i.e. $\log(\sigma)$ and remaining are MLE of β.

```
> thetaMLE
[1]  2.480620 -7.616404 19.819251 53.171399  1.098835 -4.354091
> scaleMLE
[1] 11.94867
```

For this estimation we require some robustness, which is obviously not covered by the ML estimation. Lacking a better robust starting estimator, we use skipping technique to robustify the classical MLE. First, we compute the vector which contains squared euclidean norms of the inverse of the Fisher information matrix multiplied with L_2 derivative for each observation, more precisely:

$$N_i = |I_\beta^{-1} \Lambda_\beta(y_i, x_i)|^2.$$

Then, we drop out of the data 5% of regressors and observations with the largest N_i, i.e. ones with the largest impact. With the reduced sample we go back to the optimization, redo it and get some robust estimate of the parameters β and σ. This estimation gives the following results:

```
> thetaRob
[1]   2.409787  -7.682626  19.562721  52.655627  -7.812973 -18.443984
> scaleRob
[1] 11.13159
```

As an effect we have immunized our estimator against the effect of at least the 5% most influential data. Here we stop the estimation procedure for the scale and further, we use the obtained values as the nuisance parameters of the GPD.

Getting hand on a (globally) robust, consistent starting estimator in this context is all but trivial. The usual technique to use a minimum distance estimator to a distance based on the distribution function (Kolmogorov, Cramér-von-Mises) in our case suffers from the need to compute the multivariate (i.e., here 6-dimensional) density. Failing to find a better solution we instead propose the following procedure.

We drop the 5% most influential observations (in terms of the value $|I^{-1}\Lambda|$) and compute the MLE on the remaining data. Obviously, this will already lead to a bias in the ideal model in general; Dupuis and Morgenthaler (2002) to this end have sketched a general procedure to tackle this problem, but for the sake of this example we skip this step for the moment, hoping that the subsequent k steps in the k step iteration will then already reduce the bias sufficiently again. As a control we compare the resulting parameter estimates with the ones of the MLE (on the whole data set). When the difference in the real data set is small, we can expect our hope to be justified.

Applying the link function to the regression matrix multiplied with the corresponding MLE or robust estimate of the parameter β, we receive the shape parameter estimations. To compare them, we calculate the mean, minimum and maximum of all coordinates of the shape estimators, i.e.

	mean	max	min
MLE	0.138	0.564	0.0092
Robust	4.019e-05	0.0103	1.487e-11

TABLE 8.1: Mean, minimum and maximum of the shape estimators for 1000 observations

One can see that estimations of the parameter β significantly differ in the two last coordinates, which are the parameters associated with the parameters `Gesamtscore.SOFA`

and `SOFA.max`. This difference causes the visible difference between shape parameter estimations. The only possible reason for such difference can be the robust issue itself. We can conclude, that 5% of data with the largest impact significantly influence on the shape, so when we remove it, we get much lighter tail of the GPD. To be sure that we have got some reasonable results here, we approximately calculated confidence intervals for both estimations, and we conclude that they fit them quite well.

Before we start with the k-step estimation, first we decide about the choice of the clipping hight b. If we want to select it instead of computing, we can face few problems. First, it is hard to tell how large is the range between the most robust estimator MBRE and the MLE, which is not robust at all. This range can be extremely small and if our guess of b does not get into it, we can have problems with the convergence of the fixed point iteration.

As we have mentioned in Section 6.3, the best way to compute the clipping hight, is to solve Anscombe criteria, but it is hard and we do not use it here. Another way is to use one of four explicit equations for the selection of b in presented in Ruckdeschel (2014, Sec. 4.4).

The second problem of the clipping hight selection is that, usually, it is not easy to make good prediction of its value before taking look at the whole model. Four mentioned equations from Ruckdeschel (2014) give more model-independent criteria for this selection.

What we do here is, using the following approximation in the classical setting equality $\mathbf{E}(|I^{-1}\Lambda|^2) \approx \mathrm{tr}I^{-1}$, we compute the value of b in the following way

$$b = c * \sqrt{\mathrm{tr}I^{-1}}. \tag{8.3}$$

Here, choosing appropriate constant c, we search for the first b, for which the whole algorithm converges after 3-4 iteration. If it does not work for the chosen c, we increase it. In our real data example with the dimension 200, choice of $b = 25 * \sqrt{\mathrm{tr}I^{-1}}$ led to the convergence of the algorithm after 3 iterations, whereas for 1000 observations we took $b = 70 * \sqrt{\mathrm{tr}I^{-1}}$ to get the influence after 3 iterations.

Next, we compute k-step estimator of the parameter β, and, since k-step procedure is very durable, we apply it only to the amount of 200 observations. We obtain location parameter for this dimension to be equal to 38. Then, we compute the MLE and the robust estimator for this reduced data and obtain:

```
> thetaMLE
[1]  2.843544 -142.497857 -17.954474  157.657757 -17.762091 1.343739
```

```
> thetaRob
[1]  2.711753e+00 -1.443249e+02 -2.690242e+01 1.570307e+02 -1.857499e+01
[6] -1.118916e-04
```

Hence, values of the estimated scale parameter are:

```
> scaleMLE
[1] 17.17653
> scaleRob
[1] 15.05564
```

Next, starting with the robust estimates of the scale and β parameters, we compute k-step estimator for $k = 1$ and $k = 2$, and on each step we get the following results:

```
> betaKstep1
[1] -705.419849 -277.536523 274.182498 -8.389895 5.043223
> betaKstep2
[1] -690.699222 -278.729817 276.208028 -8.277089 3.937345
```

Mean, minimum and maximum of all coordinates of the shape estimators are illustrated in the Table 8.2.

	mean	max	min
MLE	0.1522	0.7178	1.231e-10
Robust	0.0896	0.706	1.613e-11
k=1	0.2581	0.785	5.175e-23
k=2	0.2147	0.781	4.752e-23

TABLE 8.2: Mean, minimum and maximum of the shape estimators for 200 observations

For the comparison of the results of the MLE, robust and k-step estimation procedures, we draw corresponding pp-plot. First we generate random sample of the size 200 for the GPD with the chosen location parameter and estimated scale and shape. Plotting these samples against sequence $1/n, ..., 1$, we get the plot from the Figure 8.3.

The approximation quality increases with the n getting larger for all 4 estimators. One should note, that strange behavior of all estimators for the small values is not a surprise. It is familiar phenomenon for extreme value plots, in particular for GPD, that for the lower quantiles we do not get good approximation.

Red points on Figure 8.3 represent results of the maximum likelihood estimation, whereas the blue color is chosen for the robust estimators of the scale and shape. Comparing results of these two estimations, we cannot easily conclude from the plot, which of them

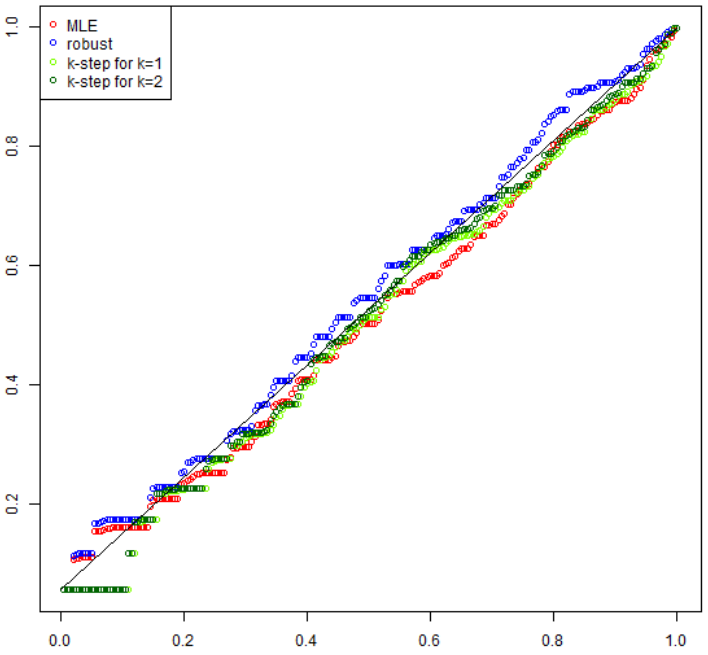

FIGURE 8.3: PP-plot for the MLE, robust, k=1- and k=2-step estimation of the shape parameter for GPD

performs better. Red points do not perfectly follow the line $y = x$, since MLE is not robust, i.e. it is affected by the outliers in the data. The line constructed by the blue points should perform better in this aspect, but one can see that it is a bit shifted from the black line due to the bias caused by the construction of the robust estimator.

From the pp-plot we observe that both lines, of light and dark green points, closely follow line $y = x$. Light green color here represents the 1-step estimation, and the dark green in chosen for the 2-step estimator of the shape. As was expected, second step slightly improves the 1-step estimation results.

Next, we aim to draw analogs of the diagnostic plots from Section 2.6.2 for our results. First, we plot the influence curves for each of 5 coordinates of the k-step estimation of the parameter β, comparing them to the MLE influence, obtained from the expression $I^{-1}\Lambda_\beta^{\mathcal{P}}$.

On Figure 8.4, Figure 8.5 and Figure 8.6 coordinates of the vector β are plotted. Since green color represents influence calculated with the 1-step estimator, it is clear that it is close to zero for all 5 parameters. Analyzing all plots one can notice, that for all coordinate, except second, MLE influence, which is drawn by red color, does not differ too much from the robust estimation influence, which consists of the blue points. Difference for the second component of the parameter β, which corresponds to the

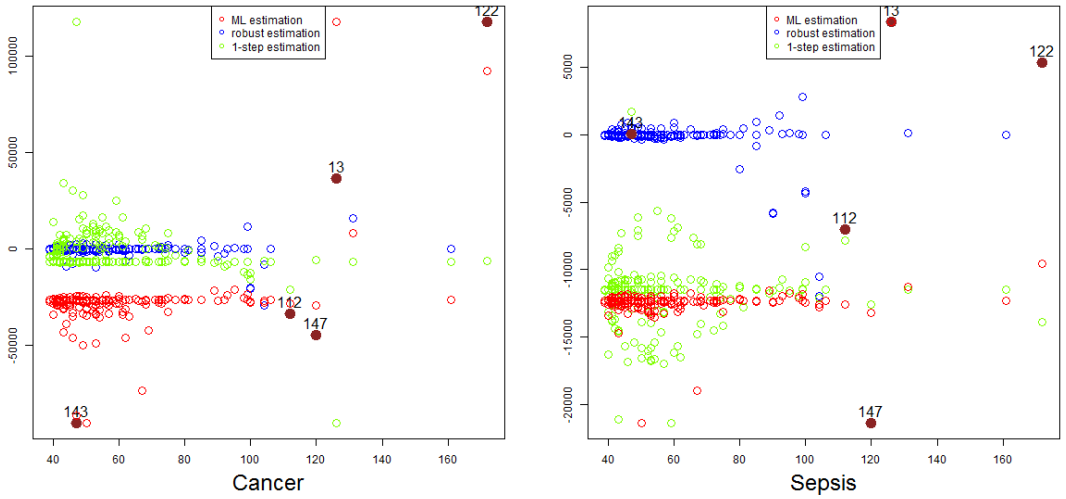

FIGURE 8.4: The first two components of the optimal influence curves for GPD

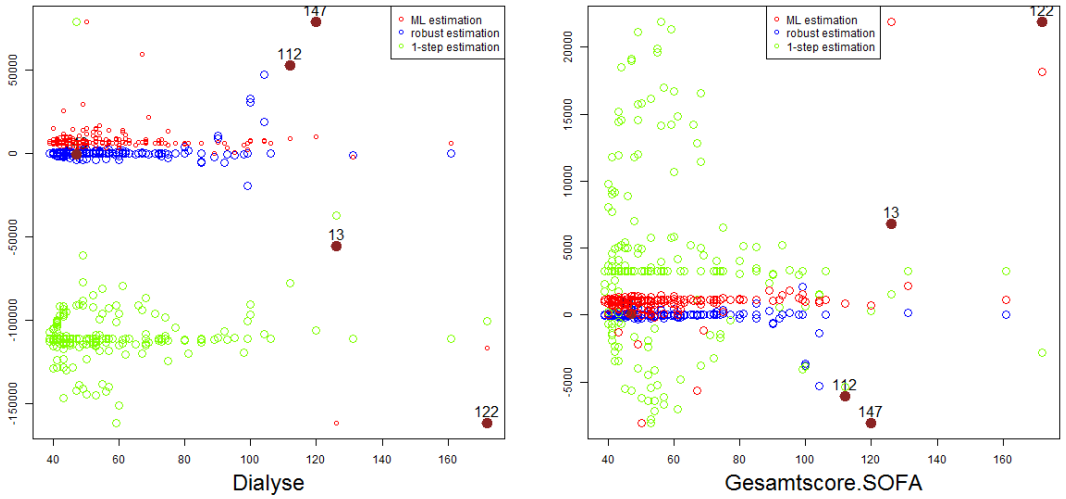

FIGURE 8.5: The second two components of the optimal influence curves for GPD

`Sepsis` parameter, means that removed 5% of the most influential observations had the only impact on the shape parameter, because after removing them, the general influence of this parameter on the shape becomes zero.

On this stage by the maximal norms of the influence we detect the most influential observations. Vector `top` represents the numbers of the observations with the norms of the influence `infnorms[top]`. Then we display top 5 observations themselves and corresponding regressors. `FixPglm[,top]` contains the influences of the top 5 observations as the columns.

```
> top
[1] 122 147 143  13 112
```

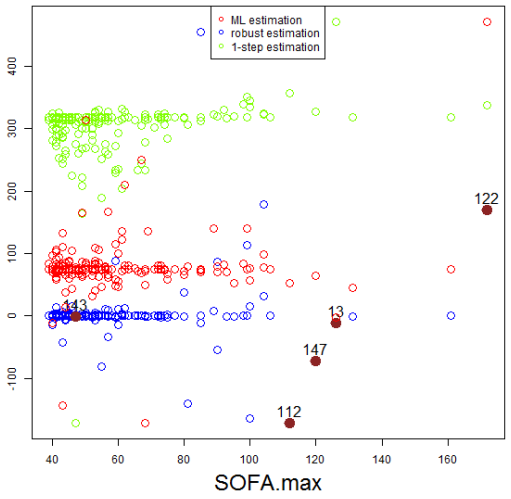

FIGURE 8.6: The last component of the optimal influence curves for GPD

```
>  FixPglm[,top]
                        [,122]      [,147]       [,143]       [,13]      [,112]
Cancer              117969.949 -44872.1938 -90233.56998   36740.527 -33637.482
Sepsis                5375.379 -21377.8497     76.54543   8410.0869  -7013.071
Dialyse            -161287.980  78920.2473   -631.27602 -55436.7673  52632.565
Gesamtscore.SOFA     21898.232  -8061.2293      80.5967   6777.7163  -6075.129
SOFA.max               169.700     -72.126     -0.68553     -10.987   -171.101
> infnorms[top]
[1] 201094.95  93615.81  90235.85  67377.81  63148.92
> y[top]
[1] 172 120  47 126 112
> X[top,]
       Cancer Sepsis Dialyse Gesamtscore.SOFA SOFA.max
[122,]      0      0       1               10        18
[147,]      0      0       1                7        15
[143,]      1      1       1                1        15
[13,]       0      1       1                9        16
[112,]      0      1       1                6        13
```

These top 5 observations are also plotted on Figure 8.4, Figure 8.5 and Figure 8.6 as dark brown points and labeled by their numbers. E.g. observation number 122 has the biggest of all observations influence, what can be seen from all 5 plots, as it is far from the area of the points concentration. Interesting here is that the observation 143 has low influence on all coordinates except first one, but this one influence is huge enough to bring it to the top 5 influential observation.

From this result we can conclude that the most influential parameter coordinates are `Dialyse`, as it is `TRUE` in each regressor, and `SOFA.max`, which, as one can see from the Histogram 8.1, has quite big values in the chosen regressors.

To make more accurate conclusion about the percentage of the information, used per observation for each parameter coordinate, we plot analog to the information plots from Section 2.6.2.

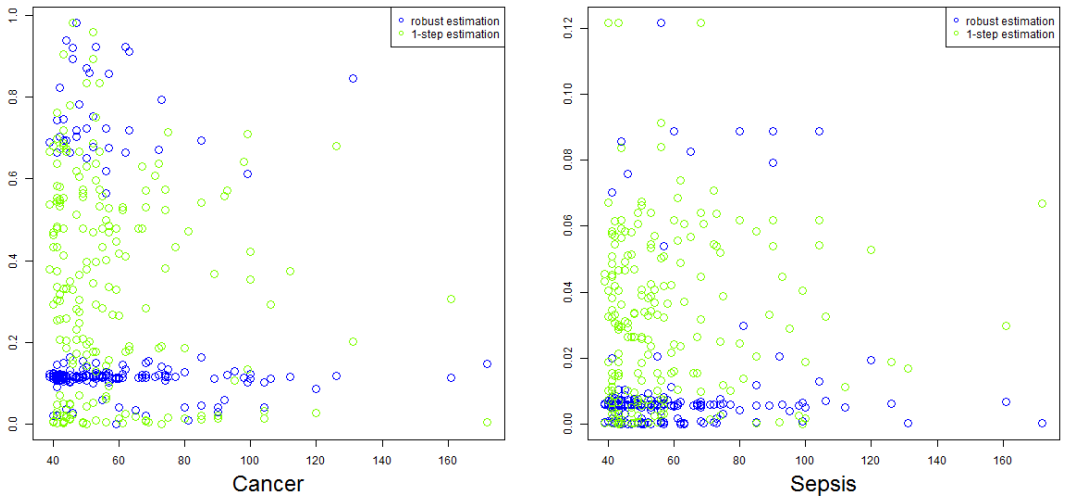

FIGURE 8.7: Relative information of first two components of (partial) influence curve for GPD

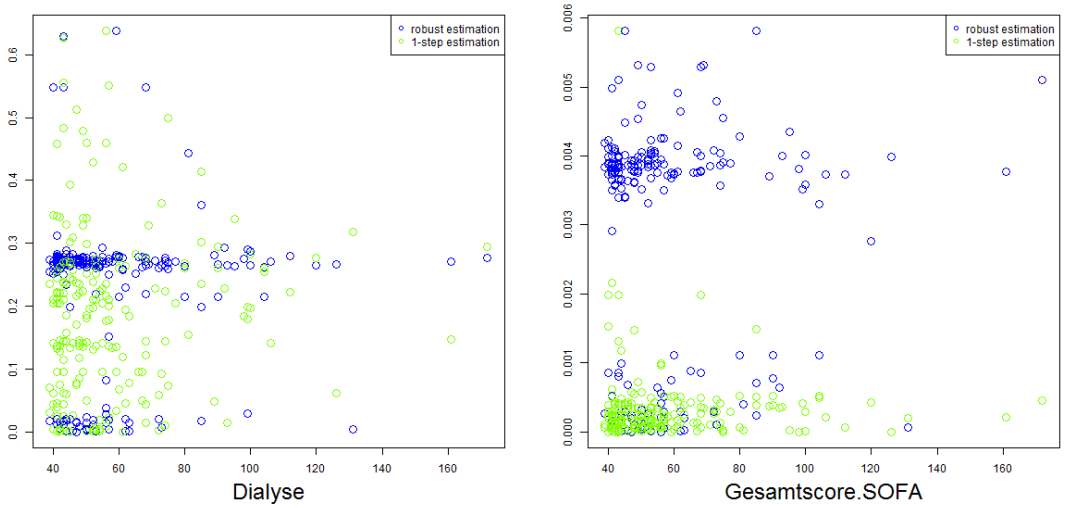

FIGURE 8.8: Relative information of second two components of (partial) influence curve for GPD

From these plots we confirm the assumption about the big influence of the third coordinate `Dialyse`, see Figure 8.8. From Figure 8.9 one can see that fifth coordinate `SOFA.max` does not have much influence, so this assumption was premature.

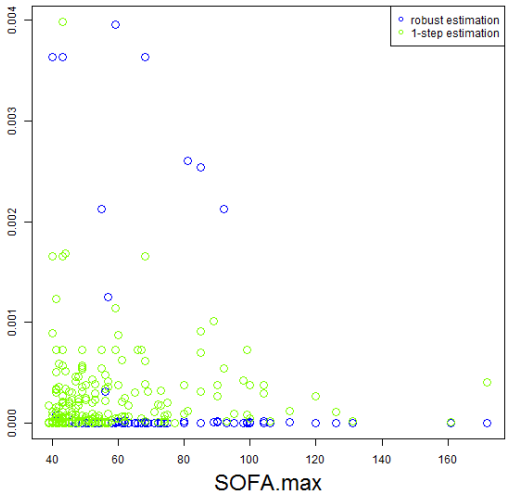

FIGURE 8.9: Relative information of the last component of (partial) influence curve for GPD

8.1.4 Speed of the algorithm

In this Section we conclude about the speed of our function `FixPglm`. First, we try this function on 1000 observations and 5-dimensional regressors. We measured performance speed of the algorithm using the `R` function `proc.time()` in the following way

```
begin <- proc.time()

Influence1 <- FixPglm(X=X, A0=A0, locat=myloc, scale=scaleRob, linkfct=linkfct,
      beta=betaRob, b=b, eps=eps)(1)

elaspedTime <- proc.time() - begin
```

In the case of 1000 observations algorithm stops after 3d iteration for the values $\epsilon = 0.5$ and clipping hight $b = 70 * \sqrt{\mathrm{tr} I^{-1}}$. Time of the algorithm performance in this case is

```
> elaspedTime
   user  system elapsed
4836.52    1.35 4924.08
```

Next, we apply our function to the smaller number of the observations, e.g. 200, and 5-dimensional regressors. In this case we also choose clipping hight to get convergence after 3 iterations were used and the results were the following:

```
> elaspedTime
   user  system elapsed
 219.90    0.16  224.13
```

Another comparison we do is reducing amount of the regressor dimension. First, we try it for the 2-dimensions, taking the `Cancer` and the `Gesamtscore.SOFA` parameters only. The problem we face in this case is that we get some two-dimensional zero regressors. Later, we plug them in the link function and obtain zero shape parameter. This makes the computation of the L_2 derivative function more complicated.

One possible way out of this problem is to use interception. We can replace one of the chosen regressors or simply add one parameter, which is always equal to 1. In the regression matrix it is additional column of ones and regressors never become zero vectors.

Nevertheless, we decide to keep 3 parameters, `Cancer`, `Sepsis` and `Gesamtscore.SOFA`. Here we are safe from the described problem, hence, we do not use interception. Dimension of the regressors is 1000 as in the first example of this Section. Here we get the following speed of 3 iterations used:

```
> elaspedTime
   user  system elapsed
2899.66    0.49 2938.38
```

Here we conclude about the speed of the algorithm performance. With the computations described above we track the dependence of our function on the amount of the observations and the number of the taken parameters. First, we observe conspicuous time reduction of the algorithm performance of 5-times reduced sample. We expected that the speed of our algorithm is linearly dependent on the observation dimension, but here we observe 20-times longer performance of the 1000-dimensional data, than for 200-dimensional. One also should note, that the most durable stages of our algorithm are the ones with computation of the expectations, i.e. (6.16) and (6.18). The speed of the used function for the calculation of these expectations, `E` from the `R`-package `distrEx`, is also not linear w.r.t. the observation dimension.

However, we also predicted the linear dependence of the function performance time on the number of the chosen data parameters, what seems to be true. The speed of the algorithm applied to 3-parametric data turns to be 1,66 faster than the one for 5-parametric regressors.

8.2 Simulation study

In this Section we check the accuracy of the MLE and the robust estimates of the parameter β and σ, applying some simulation study. Here we also aim to confirm or deny the suitability of the generalized linear model with generalized Pareto error distribution for the real data from above.

First, we compute estimates from these real data, taking 1000 observations and 5-dimensional regressors. As they are the reference parameters for the simulation discussed here, we repeat them:

```
> betaRob
[1] -7.682626  19.562721  52.655627  -7.812973 -18.443984
> scaleRob
[1] 11.13159
```

For the comparison of the mean, minimum and maximum of the shape estimators we refer to the Table 8.1.

In order to get some appropriate performance time for the simulation study, we take the random sample of 100 numbers from 1 to 1000. Using it, we randomly choose 100 regressors from the real data and 100 shape parameters from 1000-dimensional robust estimate `shapeRob`. For simplicity we denote the regressors matrix as X_{100}. Then we plug these robust estimates and the `scaleRob` in the GPD and simulate random sample of 100 variables. Note, that each random variable is simulated from the different value of the shape parameter, according to the respective value of the link function in the GLM regression. We treat this random sample as new ideal observation vector y_{id} and then, we contaminate it, replacing around 5% of its values by the data simulated from GPD with totally different parameters, i.e. we get y_{cont}. Further, we proceed with both types of the observations, ideal and contaminated.

We use pairs of the regression matrix and observation vector, i.e. (X_{100}, y_{id}) and $(X_{100}, y_{\text{cont}})$ to estimate parameter β and scale σ. We take values `betaRob` and `sigmaRob` as the starting estimators, and first compute the MLE. Then again, we through 5% of the observations with the largest impact out of the data and compute robust estimates.

For this simulation we do 100 runs to get 100 values of the estimates computed from 100 different synthetic data pairs (X_{100}, y_{id}) and $(X_{100}, y_{\text{cont}})$. Unfortunetly we cannot apply k-step optimally robust estimation here, since then we would have to compute 100 optimal influence functions for each run, and it would cost too much time.

Finally, we calculate MSE of the obtained estimators comparing to `betaRob`, computed from the real data and used as starting estimator for the synthetic data simulation. The resulting MSEs are illustrated in the following table:

	ideal	contaminated
MLE	1.601	2.35e+18
Robust	2.497	83051.25

TABLE 8.3: MSE of the estimates obtained from the real and synthetic data

From this table we conclude, that MLE and robust estimation perform very well and similar to each other, as we expected. Since ML estimator is not robust, its MSE is quite large in the case of the contaminated data, whereas applying robust estimation to it evidently reduces MSE.

We can confirm this conclusion looking at absolute difference between means of the shape parameter computed after simulations, and the `mean(shapeRob)` from the original data:

	ideal	contaminated
MLE	3.656e-05	3.759e-05
Robust	3.642e-05	3.641e-05

TABLE 8.4: Absolute mean difference between estimates of the shape for real and synthetic data

One should note, that performance of the robust estimation is the same on the shape scale in both, ideal and contaminated situation, when on the β-scale the difference was observed and almost disappeared after use of the link function.

For the better illustration of these results we also draw the box plots for the scale and 5 coordinates of the parameter β for each of four cases, see Figure 8.10 - Figure 8.13.

From Figure 8.10 we observe that for the MLE in the ideal situation there are only few, visible, but not huge, outliers for each coordinate of the estimated parameters. Next, since robust procedure we use takes out of the data some influential observations, the slightly bigger number of the outliers on Figure 8.11 was predictable. What is more dramatic but also expected, are the amount and sizes of the outliers occurred from the MLE in the case of the contaminated data, Figure 8.12. The good point is that this amount and the sizes of the outliers are greatly reduced by applying the robust estimator, what one can see on Figure 8.13

For another analysis of the each coordinate of the estimated parameter β, we draw the histograms for the difference between the estimated regression parameter in each situation and `betaRob`, taken as the starting estimator.

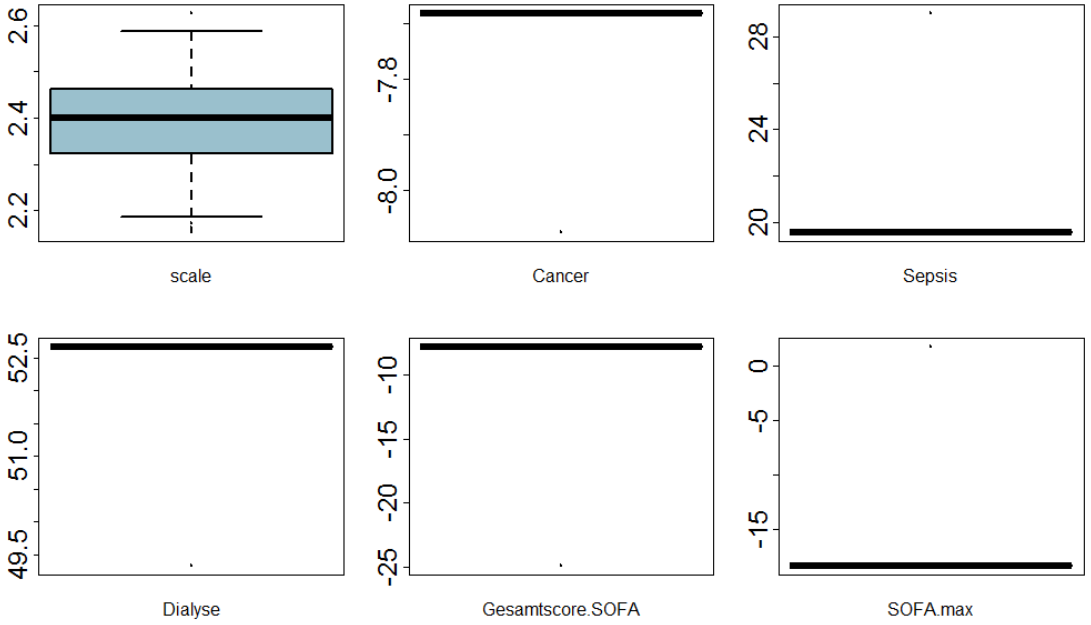

FIGURE 8.10: Scale and regressor parameters MLE for ideal observations

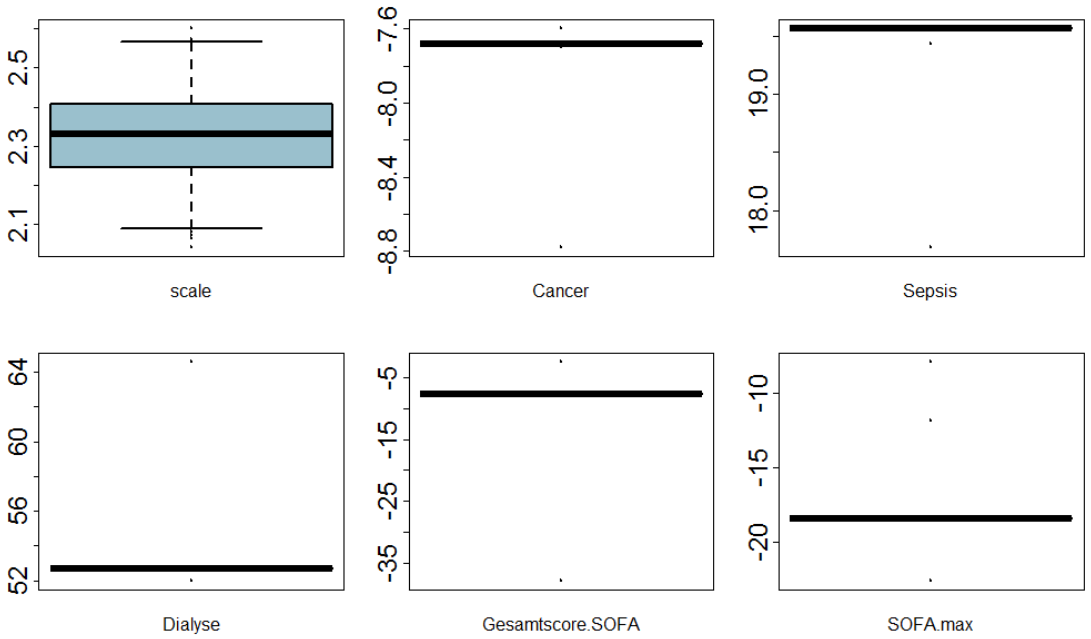

FIGURE 8.11: Scale and regressor parameters robust estimates for ideal observations

From the histograms on Figures 8.14-8.17 we only confirm previous conclusions, that MLE performs almost perfectly in the ideal situation, but does not handle outliers in the contaminated case, when performance of the robust estimator is good enough in the absence and presence of the outliers in the data.

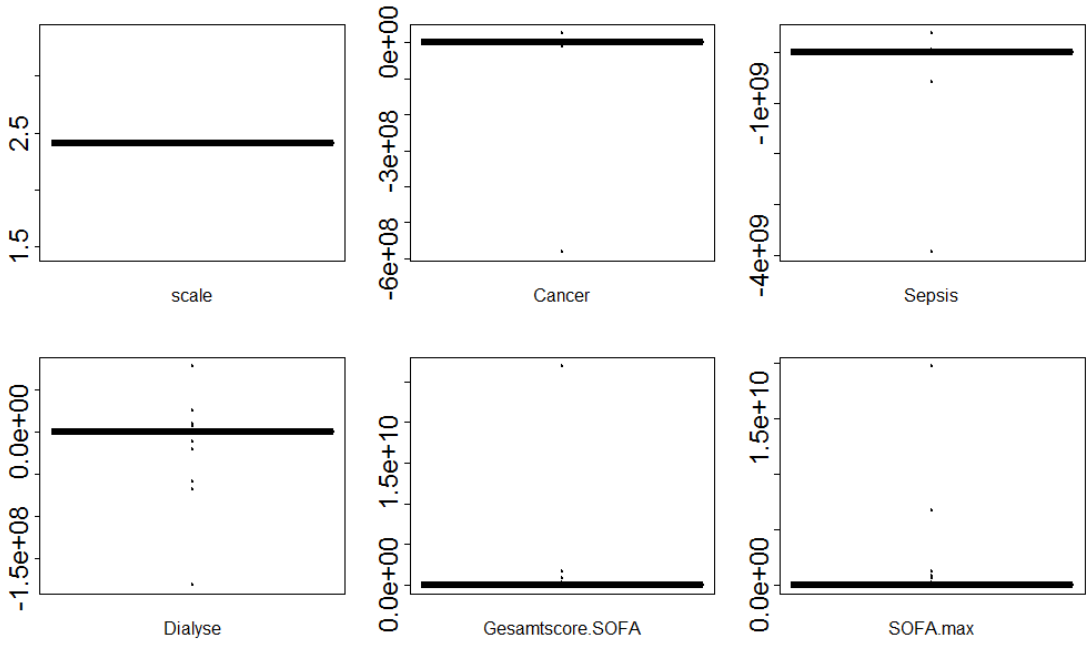

FIGURE 8.12: Scale and regressor parameters MLE for contaminated observations

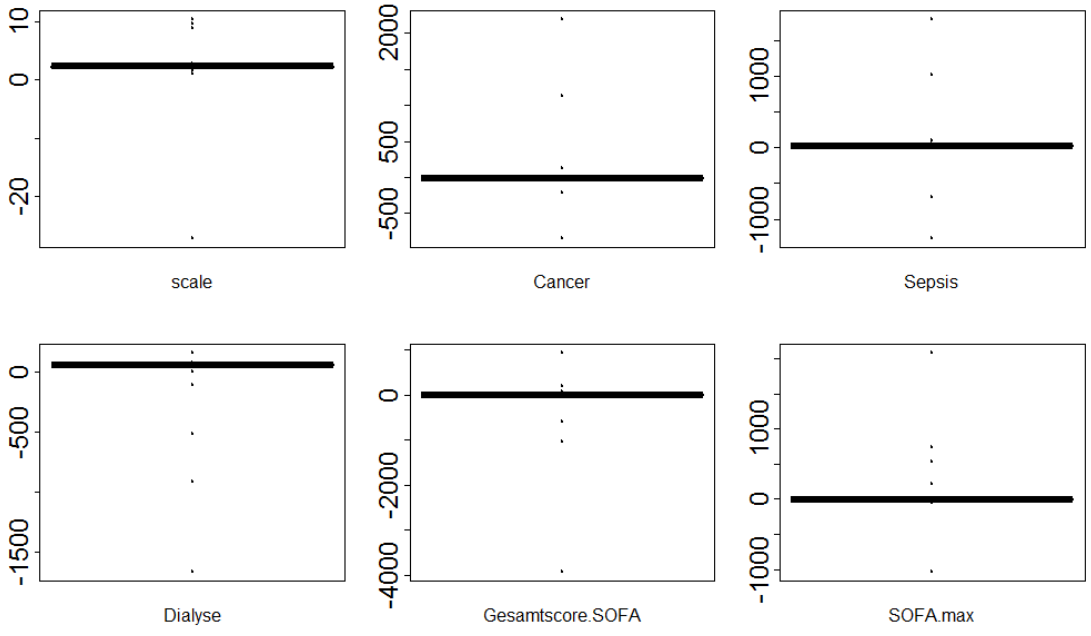

FIGURE 8.13: Scale and regressor parameters robust estimates for contaminated observations

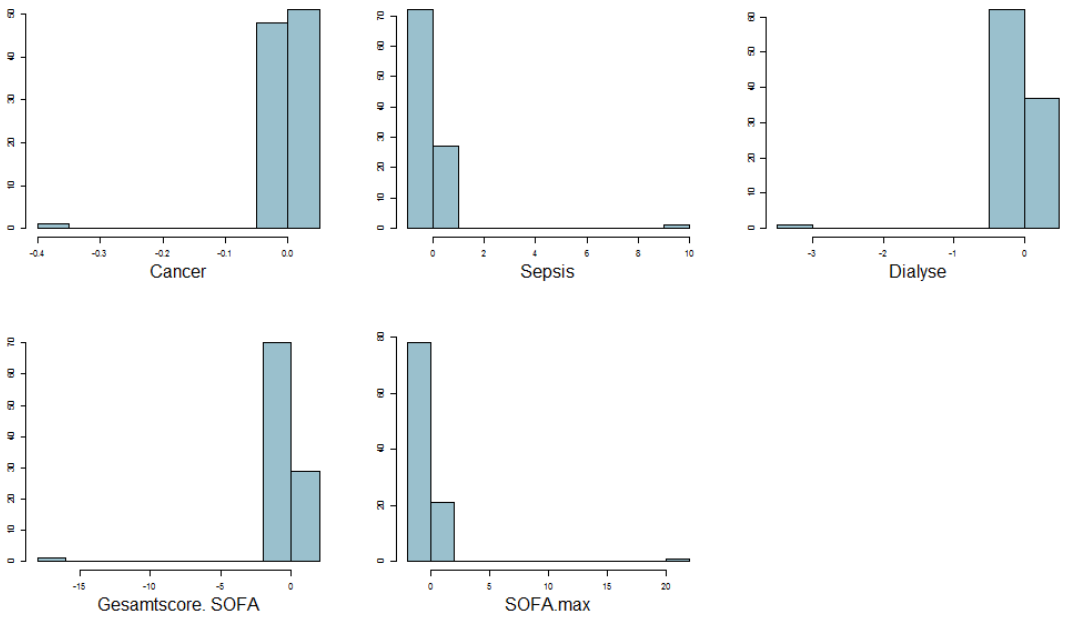

FIGURE 8.14: MLE for ideal observations

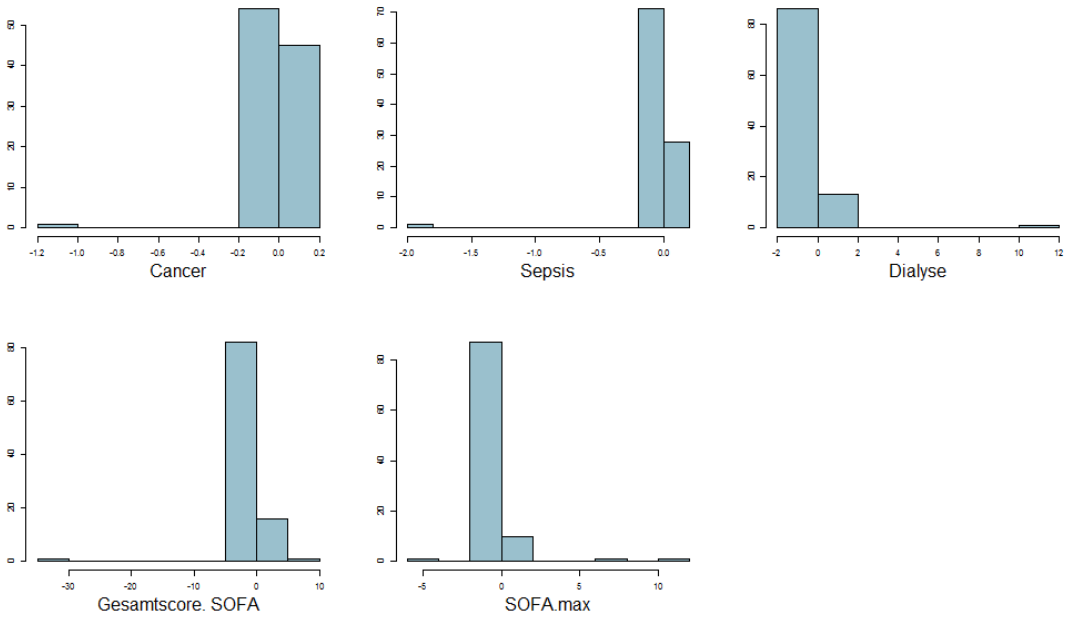

FIGURE 8.15: Robust estimates for ideal observations

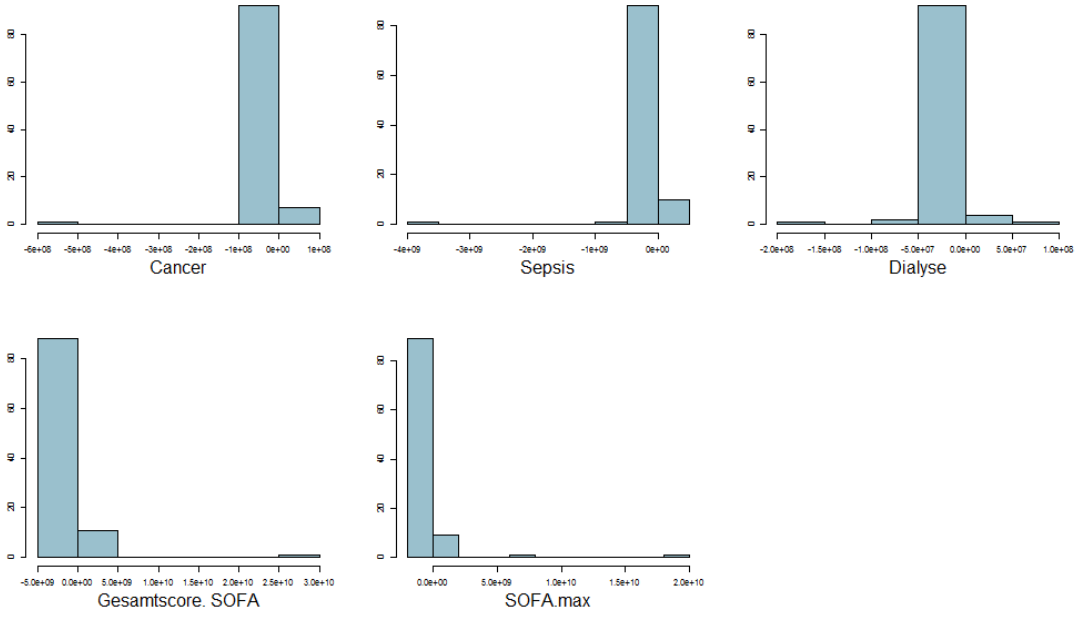

FIGURE 8.16: MLE for contaminated observations

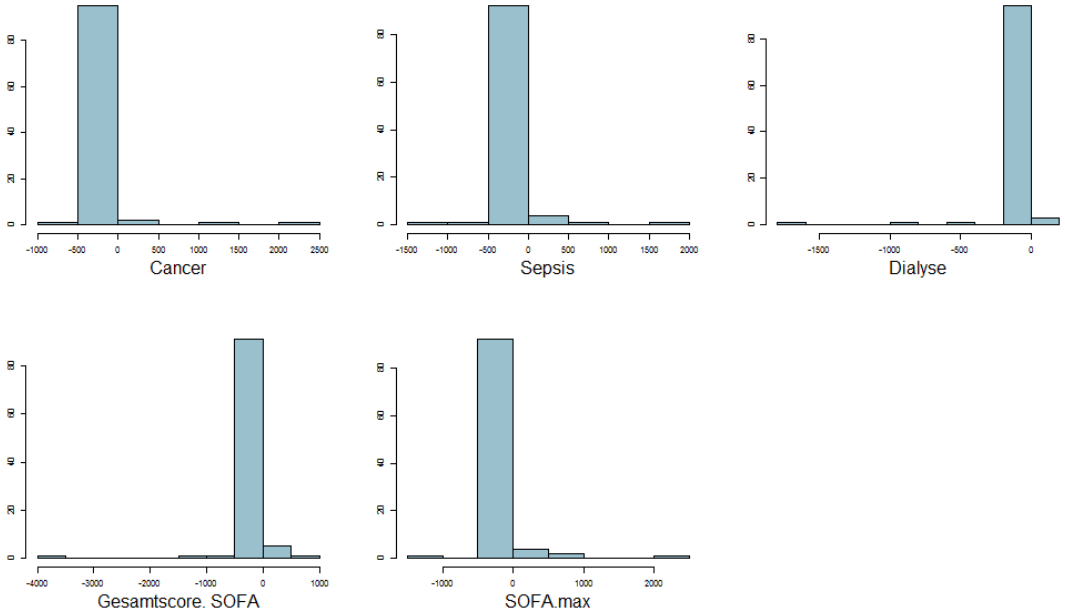

FIGURE 8.17: Robust estimates for contaminated observations

Chapter 9

Conclusions

The main focus of this thesis belongs to the regression models with extreme value error distributions. Working with real data we usually suspect that it can be contaminated by some proportion of outliers. Therefore, for the variety of application domains, classical estimation and inference is not always reliable. Hence, we have set a goal to develop robust procedures for the systems which contain extreme events, i.e. apply robust statistics to extreme value theory. The main challenge we faced, came from the choice of the asymmetric error distributions and caused significant complication of the aimed robustification.

Nevertheless, we achieved the desired goal for two types of the regression models. First, for dynamical regression models, more specifically, state-space models, we used robust versions of the Kalman filter and reworked classical Kalman smoother and extended Kalman filter in a robust way for different types of outliers, i.e. spiky outliers, AOs and IOs. Here, we were first to compute general IO-robust filter and introduce new idea for the robust smoother.

To assess the performance of new procedures, we applied it at real data and stylized outlier situation. Thence, we concluded that our procedures perform very well in the situations they were created for. Moreover, they cover wider variety of outlier situations comparing to other existing approaches and win in efficiency in all contamination situations.

Our invented procedures are recursive, therefore they are quite fast and convenient for online using. We implemented them in `R` in the framework of the package `robKalman` (last developer version available in `R-Forge` is 0.3).

Still, there are some open issues in our procedures, which are topics for further research, e.g. IO-robust smoother have to be essentially improved. Besides, after checking filters

in the case of some non observed aspects, i.e. when the observation matrix of the model is non invertible, we conclude that all filters cannot cope with this situation. Another open issue we left for the further considerations is the hybrid filter and smoother, which can be used for the mixed situations, i.e. presence of both types of outliers in the data. Although, the first attempts were made by Ruckdeschel (2010c, Ch. 5).

Next, we devoted one Section to the model diagnostics, which draws diagnostic plots to see different aspects of the taken model. We introduced our new `R`-package `RobExtremes`, which provides infrastructure for optimally robust estimation in scale-shape models, covering GEVD and GPD. Moreover, it implements general LD estimators, including the high-breakdown point estimators, and applies interpolation technique.

Further, we reviewed generalized published results on robustness properties of some estimators for the GPD parametric model, to cover GEVD case, more precisely, for the classical moment based and Cramér-von-Mises minimum distance estimators.

The second type of regression models, covered by this thesis, were generalized linear models, studied in some more general form, i.e. with extreme value error distributions, i.e. GEVD and GPD, instead of distributions from the exponential family. In order to obtain some robustness for these models, we created sufficient conditions of L_2 differentiability for them, deriving smoothness in terms of it. We generalized theory of Rieder (1994) on L_2 differentiability for linear regression models, considering cases of stochastic and deterministic regressors separately. Moreover, we computed corresponding L_2 derivatives and the Fisher information matrices for each case. Our important achievement here was making our approach cover higher dimensional error distributions and case of regressors of possibly different length for each parameter, what was new.

We checked suitability of the introduced L_2 differentiability conditions on the various models, including GEVD and GPD joint shape-scale models. We discussed in details the way we obtained the appropriate componentwise link function and proved, that our choice satisfies all conditions we required for the L_2 differentiability of the model.

Important part of this thesis was focused on the fixed point iteration algorithm for the computation of the optimally robust influence curve. Our version of this algorithm differ from the similar algorithms by using another techniques to get some intermediate values. Here, we not only discussed it step by step and pointed out its week stages, but also implemented it in `R` under the name `FixPglm`. Here we did not aim to provide the most flexible implementation, but rather sketch how it should be done and retain points of particular importance. We tested function `FixPglm` for the Binomial case, with the `R`-data "carrots", and analyzed time of its performance. We concluded that for the exemplary implementation, `FixPglm` showed quite good results.

In the third part of the thesis we discussed three applications, which were studied in the framework of the project "Robust Risk Estimation", i.e. operational risk, hospitalization times and hydrological river discharge data. For each of them, we discussed main issues, presented the available data and pointed the specific problems, which can be solved by using our results. Then, we applied function `FixPglm` to the real data set taken from Jena university hospital ICU.

We have fitted these data with the GLM with generalized Pareto error distribution and estimated its shape parameter by means of the link function and regression parameter estimation. The most important and difficult stage here was to compute the appropriate link function, but after several attempts, we got the desired link. Another significant for the procedure performance choice we made related to the clipping hight, but we have found good approximation for it.

To the regression parameter we applied three types of estimation, MLE, robust estimator using skipping technique and k-step procedure with $k = 1, 2$. With various illustrations we compared performance of all estimators and analyzed the data pointing out the most influential observations and parameters, e.g. coordinate responsible for special treatment `Dialyse` was the most influential from all five chosen parameters. Duration of the function `FixPglm` performance showed quite satisfactory results.

Finally, we made some simulation study which confirmed effectiveness of applied approach and demonstrated behavior of MLE and robust estimators on the ideal and contaminated synthetic data. Here all our expectations were fulfilled, i.e. MLE performed almost perfectly in the ideal situation, but could not handle outliers in the contaminated case, when performance of the robust estimator was good enough in both cases.

Appendix A

Robustness properties of the GEVD estimators

In this Section $Q_{\sigma,\xi}$ denotes GEVD c.d.f. (3.1) with known location parameter μ.

A.1 Proof of Theorem 3.18

To get influence function of the method of moments estimator we need to compute Jacobian matrix

$$D = \begin{pmatrix} \frac{\partial \hat{\boldsymbol{\xi}}}{\partial m_1} & \frac{\partial \hat{\boldsymbol{\xi}}}{\partial m_2} \\ \frac{\partial \hat{\boldsymbol{\sigma}}}{\partial m_1} & \frac{\partial \hat{\boldsymbol{\sigma}}}{\partial m_2} \end{pmatrix} = \begin{pmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{pmatrix}.$$

Since we do not have explicit form of the estimators $\hat{\boldsymbol{\sigma}}$ and $\hat{\boldsymbol{\xi}}$, we compute Jacobian matrix as the inverse matrix, where all terms can be explicitly calculated, i.e.

$$D = \begin{pmatrix} \frac{\partial m_1}{\partial \hat{\boldsymbol{\xi}}} & \frac{\partial m_1}{\partial \hat{\boldsymbol{\sigma}}} \\ \frac{\partial m_2}{\partial \hat{\boldsymbol{\xi}}} & \frac{\partial m_2}{\partial \hat{\boldsymbol{\sigma}}} \end{pmatrix}^{-1}.$$

Expressions in (3.16) display first two theoretical moments of GEVD, which we repeat here

$$m_1 = \frac{\sigma(g_1 - 1)}{\xi}, \quad m_2 = \sigma^2 \frac{g_2 - 2g_1 + 1}{\xi^2}$$

where $g_k := \Gamma(1 - k\xi)$, $k = 1, 2$ and $\xi < 0.5$.

For simplicity, we introduce some additional notations:

$$a = \frac{\partial m_1}{\partial \hat{\boldsymbol{\xi}}} = \sigma \frac{g_1' \xi - g_1}{\xi^2}; \quad b = \frac{\partial m_1}{\partial \hat{\boldsymbol{\sigma}}} = \frac{g_1 - 1}{\xi};$$

$$c = \frac{\partial m_2}{\partial \hat{\boldsymbol{\xi}}} = \sigma \frac{(g_2' - 2g_1')\xi - 2(g_2 - 2g_1 + 1)}{\xi^3}; \quad d = \frac{\partial m_1}{\partial \hat{\boldsymbol{\sigma}}} = 2\sigma \frac{g_2 - 2g_1 + 1}{\xi^2}.$$

Then, Jacobian matrix D can be computed in the following way

$$D = \begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -c \\ -b & a \end{pmatrix} = \begin{pmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{pmatrix}$$

We start with the calculation of the denominator, i.e.

$$ad - bc = \frac{2\sigma^2}{\xi^4}(g_1'\xi - g_1)(g_2 - 2g_1 + 1) - \frac{\sigma^2}{\xi^4}(g_1 - 1)((g_2' - 2g_1')\xi - 2(g_2 - 2g_1 + 1)) =$$

$$= \frac{\sigma^2}{\xi^4}(2(g_1'\xi - 1)(g_2 - 2g_1 + 1) + \xi(1 - g_1)(g_2' - 2g_1')).$$

Next, we compute each component of the Jacobian matrix D and get the following terms

$$d_{11} = \frac{d}{ad - bc} = \frac{2\xi^2(g_2 - 2g_1 + 1)}{\sigma(2(g_1'\xi - 1)(g_2 - 2g_1 + 1) + \xi(1 - g_1)(g_2' - 2g_1'))};$$

$$d_{12} = \frac{-c}{ad - bc} = \frac{-\xi((g_2' - 2g_1')\xi - 2(g_2 - 2g_1 + 1)}{\sigma(2(g_1'\xi - 1)(g_2 - 2g_1 + 1) + \xi(1 - g_1)(g_2' - 2g_1'))};$$

$$d_{21} = \frac{-b}{ad - bc} = \frac{-\xi^3(g_1 - 1)}{\sigma(2(g_1'\xi - 1)(g_2 - 2g_1 + 1) + \xi(1 - g_1)(g_2' - 2g_1'))};$$

$$d_{22} = \frac{a}{ad - bc} = \frac{\xi^2(g_1'\xi - g_1)}{\sigma(2(g_1'\xi - 1)(g_2 - 2g_1 + 1) + \xi(1 - g_1)(g_2 - 2g_1'))}.$$

Influence functions of the first two moments for GEVD are correspondingly

$$IF(x, m_1, Q_{\sigma,\xi}) = x - m_1, \quad \text{and} \quad IF(x, m_2, Q_{\sigma,\xi}) = x^2 - m_2.$$

Therefore, by the delta method, influence function of the method of moments estimator can be calculated from the following expression, plugging obtained Jacobian matrix D in it:

$$IF(x, \text{MOM}, Q_{\sigma,\xi}) = D(IF(x, m_1, Q_{\sigma,\xi}), IF(x, m_2, Q_{\sigma,\xi}))^{\text{T}} =$$

$$= D(x - m_1, x^2 - m_2)^{\text{T}}.$$

□

A.2 Proof of Theorem 3.19

To calculate influence function of the Cramér-von-Mises MDE we follow method of Horbenko (2011, Sec. 6.2), which is originally based on the results presented in Rieder (1994, Ex. 4.2.15, Thm. 6.3.8). To do so, we link to the definition of the *Cramér-von-Mises differentiability* of the parametric model, taken from Rieder (1994, Def. 2.3.11), with the corresponding *Cramér-von-Mises derivative* Δ_θ and *Cramér-von-Mises information matrix* $\mathcal{J}_\theta = \int \Delta_\theta \Delta_\theta^{\mathrm{T}} dQ_\theta$.

In Section 3.2.1, we have checked, that GEVD(μ, σ, ξ) is L_2 differentiable. Moreover, by Rieder (1994) we know that L2-differentiability implies Cramér-von-Mises-differentiability, therefore GEVD is also Cramér-von-Mises-differentiable. Cramér-von-Mises derivative for GEVD can be obtained as derivative of the cumulative distribution function $Q_{\sigma,\xi}(x)$ with respect to the unknown parameters σ and ξ, i.e. $\Delta_\theta = (\Delta_\xi, \Delta_\sigma)^{\mathrm{T}}$, with the following terms

$$\Delta_\xi(x) = \frac{\partial}{\partial \xi} Q_{\sigma,\xi}(x) = \exp\Big(-(1+\xi\frac{x-\mu}{\sigma})^{-\frac{1}{\xi}}\Big)\Big(\frac{1}{\xi}(1+\xi\frac{x-\mu}{\sigma})^{-\frac{1}{\xi}}\ln(1+\xi\frac{x-\mu}{\sigma})^{-\frac{1}{\xi}} + \\ + \frac{x}{\xi}(1+\xi\frac{x-\mu}{\sigma})^{-\frac{1}{\xi}-1}\Big),$$

$$\Delta_\sigma(x) = \frac{\partial}{\partial \sigma} Q_{\sigma,\xi}(x) = -\exp\Big(-(1+\xi\frac{x-\mu}{\sigma})^{-\frac{1}{\xi}}\Big)\frac{x-\mu}{\sigma^2}(1+\xi\frac{x-\mu}{\sigma})^{-\frac{1}{\xi}-1}.$$

Following Rieder (1994, Ex. 4.2.15, Thm. 6.3.8), we obtain influence function of the Cramér-von-Mises MDE in terms of the Cramér-von-Mises derivative and information matrix

$$IF(x, \mathrm{MDE}, Q_{\sigma,\xi}) = \mathcal{J}_\theta^{-1}\Big(\int_x^\infty (1-Q_{\sigma,\xi}(y))\Delta_\theta(y)dQ_{\sigma,\xi}(y) - \int_0^x Q_{\sigma,\xi}(y)\Delta_\theta(y)dQ_{\sigma,\xi}(y)\Big) =$$

$$= \mathcal{J}_\theta^{-1}\Big(\int_0^\infty \Delta_\theta(y)dQ_{\sigma,\xi}(y) - \int_0^x \Delta_\theta(y)dQ_{\sigma,\xi}(y) - \int_0^\infty Q_{\sigma,\xi}(y)\Delta_\theta(y)dQ_{\sigma,\xi}(y)\Big) =$$

$$= \mathcal{J}_\theta^{-1}\Big(-\int_0^x \Delta_\theta(y)dQ_{\sigma,\xi}(y) + \int_0^\infty \Delta_\theta(y)(1-Q_{\sigma,\xi}(y))dQ_{\sigma,\xi}(y)\Big),$$

For simplicity, we denote $z := \frac{x-\mu}{\sigma}$. Since for GEVD we require that $1+\xi z > 0$, then for positive shapes influence function of the Cramér-von-Mises MDE can be computed by the formula

$$IF(z, \mathrm{MDE}, Q_{\sigma,\xi}) = \mathcal{J}_\theta^{-1}\Big(-\int_{-1/\xi}^z \Delta_\theta(y)dQ_{\sigma,\xi}(y) + \int_{-1/\xi}^\infty \Delta_\theta(y)(1-Q_{\sigma,\xi}(y))dQ_{\sigma,\xi}(y)\Big), \tag{A.1}$$

First, we compute Cramér-von-Mises information matrix $\mathcal{J}_\theta$, i.e.

$$\mathcal{J}_{(\sigma,\xi)} = \begin{pmatrix} \int \Delta_\xi^2 dQ_{\sigma,\xi} & \int \Delta_\xi \Delta_\sigma dQ_{\sigma,\xi} \\ \int \Delta_\xi \Delta_\sigma dQ_{\sigma,\xi} & \int \Delta_\sigma^2 dQ_{\sigma,\xi} \end{pmatrix} = \begin{pmatrix} J_{\xi\xi} & J_{\xi\sigma} \\ J_{\xi\sigma} & J_{\sigma\sigma} \end{pmatrix}.$$

To simplify further calculations, we introduce another notation, $u := (1+\xi z)^{-\frac{1}{\xi}}$, and get the following relations with it

$$z = \frac{1}{\xi}(u^{-\xi} - 1), \quad dz = -u^{-\xi-1}du, \quad Q_{\sigma,\xi}(z) = \exp(-u),$$

$$dQ_{\sigma,\xi}(z) = \frac{1}{\sigma}u^{\xi-1}\exp(-u)(-u^{-\xi-1})du = -\frac{1}{\sigma}\exp(-u)du.$$

Then, corresponding Cramér-von-Mises derivative functions can be rewritten w.r.t. the variable u, i.e.

$$\Delta_\xi(u) = \exp(-u)\Big(\frac{1}{\xi}u\ln u + \frac{1}{\xi^2}(u^{-\xi}-1)u^{\xi+1}\Big) = \frac{1}{\xi}\exp(-u)\Big(u\ln u + \frac{1}{\xi}(u-u^{\xi+1})\Big),$$

$$\Delta_\sigma(u) = -\exp(-u)\frac{1}{\xi\sigma}(u^{-\xi}-1)u^{\xi+1} = \exp(-u)\frac{1}{\xi\sigma}(u^{\xi+1}-u).$$

Since suitable values of z are restricted by $1+\xi z > 0$, we get that variable u has to be positive, i.e. $u > 0$. Here we do some additional calculations for future reference

$$\int_0^\infty \exp(-3u)u^2 du = \frac{\Gamma(3)}{3^3} = \frac{2}{27}, \tag{A.2}$$

$$\int_0^\infty \exp(-3u)u^{\xi+2} du = \frac{\Gamma(\xi+3)}{3^{\xi+3}}, \tag{A.3}$$

$$\int_0^\infty \exp(-3u)u^{2\xi+2} du = \frac{1}{3^{2\xi+3}}\int_0^\infty \exp(-3u)(3u)^{2\xi+2} d(3u) = \frac{\Gamma(2\xi+3)}{3^{2\xi+3}}. \tag{A.4}$$

We start computation of the Cramér-von-Mises information matrix with the easiest term $J_{\sigma\sigma}$, i.e.

$$J_{\sigma\sigma} = \int_{-1/\xi}^\infty \Delta_\sigma^2(z) dQ_{\sigma,\xi}(z) = \frac{1}{\xi^2\sigma^2}\int_0^\infty \exp(-2u)(u^{\xi+1}-u)^2(-\frac{1}{\sigma}\exp(-u))du =$$

$$= -\frac{1}{\xi^2\sigma^3}\int_0^\infty \exp(-3u)(u^{2\xi+2} - 2u^{\xi+2} + u^2)du.$$

Applying calculations (A.2)-(A.4) we get the first term of the matrix

$$J_{\sigma\sigma} = -\frac{1}{\xi^2\sigma^3}\left(\frac{\Gamma(2\xi+3)}{3^{2\xi+3}} - 2\frac{\Gamma(\xi+3)}{3^{\xi+3}} + \frac{2}{27}\right) =$$

$$= \frac{1}{27\xi^2\sigma^3}\left(-\frac{1}{3^{2\xi}}\Gamma(2\xi+3) + \frac{2}{3^{\xi}}\Gamma(\xi+3) - 2\right). \tag{A.5}$$

Next, we calculate term $J_{\xi\xi}$ of the Cramér-von-Mises information matrix as follows

$$J_{\xi\xi} = \int_{-1/\xi}^{\infty} \Delta_{\xi}^2(z) dQ_{\sigma,\xi}(z) = \frac{1}{\xi^2}\int_0^{\infty} \exp(-2u)(u\ln u +$$

$$+\frac{1}{\xi}(u - u^{\xi+1}))^2(-\frac{1}{\sigma}\exp(-u))du = -\frac{1}{\xi^2\sigma}\int_0^{\infty}\exp(-3u)u^2(\ln u)^2 du +$$

$$+\frac{2}{\xi^3\sigma}\int_0^{\infty}\exp(-3u)(u^{\xi+2}-u^2)\ln u du - -\frac{1}{\xi^4\sigma}\int_0^{\infty}\exp(-3u)(u-u^{\xi+1})^2 du \tag{A.6}$$

Each summand in the last expression we compute separately. For the second summand we get the following result

$$\int_0^{\infty}\exp(-3u)u^{\xi+2}\ln u du = \frac{1}{3^{\xi+3}}\int_0^{\infty}\exp(-3u)(3u)^{\xi+2}\ln(3u)d(3u) -$$

$$-\frac{\ln 3}{3^{\xi+3}}\int_0^{\infty}\exp(-3u)(3u)^{\xi+2}d(3u) = \frac{1}{3^{\xi+3}}\Gamma'(\xi+3) - \frac{\ln 3}{3^{\xi+3}}\Gamma(\xi+3) \tag{A.7}$$

Then, plugging $\xi = 0$ in the equality (A.7) we get

$$\int_0^{\infty}\exp(-3u)u^2\ln u du = \frac{1}{27}\Gamma'(3) - \frac{\ln 3}{27}\Gamma(3) = \frac{\Gamma'(3) - 2\ln 3}{27}.$$

Combining last results together we get the second summand of the expression (A.6) calculated as

$$\frac{2}{\xi^3\sigma}\int_0^{\infty}\exp(-3u)(u^{\xi+2}-u^2)\ln u du =$$

$$= \frac{2}{\xi^3\sigma}\Big(\frac{\Gamma'(\xi+3) - \ln 3\Gamma(\xi+3)}{3^{\xi+3}} - \frac{\Gamma'(3)-2\ln 3}{27}\Big).$$

The first term of the expression (A.6) is calculated in the following way

$$-\frac{1}{\xi^4\sigma}\int_0^{\infty}\exp(-3u)u^2(\ln u)^2 du = -\frac{1}{\xi^4\sigma}\Big(\int_0^{\infty}\exp(-3u)u^2(\ln 3u)^2 du -$$

$$-2\ln 3\int_0^{\infty}\exp(-3u)u^2\ln u du - (\ln 3)^2\int_0^{\infty}\exp(-3u)u^2 du\Big) =$$

$$= -\frac{1}{\xi^4\sigma}\Big(\frac{\Gamma''(3)}{27} - 2\ln 3\frac{\Gamma'(3)-2\ln 3}{27} - (\ln 3)^2\frac{2}{27}\Big) =$$

$$= -\frac{1}{27\xi^2\sigma}(\Gamma''(3) - 2\ln 3\Gamma'(3) + 2(\ln(3))^2).$$

The last summand of (A.6) is easy to get from the calculations (A.2)-(A.4), i.e.

$$-\frac{1}{\xi^4\sigma}\int_0^\infty \exp(-3u)(u-u^{\xi+1})^2du = \frac{1}{\xi^4\sigma}\Big(-\frac{\Gamma(2\xi+3)}{3^{2\xi+3}}+2\frac{\Gamma(\xi+3)}{3^{\xi+3}}-\frac{2}{27}\Big).$$

Plugging all these calculations in expression (A.6) we get next term of the Cramér-von-Mises information matrix

$$\begin{aligned}J_{\xi\xi} &= \frac{1}{27\xi^2\sigma}(-\Gamma''(3)+2\log 3\Gamma'(3)-2(\log(3))^2)+\frac{2}{\xi^3\sigma}\Big(\frac{\Gamma'(\xi+3)-\log 3\Gamma(\xi+3)}{3^{\xi+3}}-\\ &\quad -\frac{\Gamma'(3)-2\log 3}{27}\Big)+\frac{1}{\xi^4\sigma}\Big(-\frac{\Gamma(2\xi+3)}{3^{2\xi+3}}+2\frac{\Gamma(\xi+3)}{3^{\xi+3}}-\frac{2}{27}\Big)=\\ &= \frac{1}{27\xi^2\sigma}\Big(-\Gamma''(3)+2\log 3\Gamma'(3)-2(\log(3))^2+\frac{2}{\xi 3^\xi}(\Gamma'(\xi+3)-\log 3\Gamma(\xi+3))-\\ &\quad -\frac{2}{\xi}(\Gamma'(3)-2\log 3)-\frac{1}{\xi^2 3^{2\xi}}\Gamma(2\xi+3)+\frac{2}{\xi^2 3^\xi}\Gamma(\xi+3)-\frac{2}{\xi^2}\Big). \qquad (A.8)\end{aligned}$$

To complete Cramér-von-Mises information matrix $\mathcal{J}_{(\sigma,\xi)}$ term $J_{\xi\sigma}$ is missing, therefore

$$\begin{aligned}J_{\xi\sigma} &= \int_{-1/\xi}^\infty \Delta_\xi(z)\Delta_\sigma(z)dQ_{\sigma,\xi}(z)=\\ &= -\frac{1}{\xi^2\sigma^2}\int_0^\infty \exp(-3u)(u\log u+\frac{1}{\xi}(u-u^{\xi+1}))(u^{\xi+1}-u)du=\\ &= -\frac{1}{\xi^2\sigma^2}\int_0^\infty \exp(-3u)(u^{\xi+2}-u^2)\log u du+\frac{1}{\xi^3\sigma^2}\int_0^\infty \exp(-3u)(u-u^{\xi+1})^2du.\end{aligned}$$

Using previous calculations (A.2)-(A.4) and equation (A.7) we get

$$\begin{aligned}J_{\xi\sigma} &= -\frac{1}{\xi^2\sigma^2}\Big(\frac{\Gamma'(\xi+3)-\log 3\Gamma(\xi+3)}{3^{\xi+3}}-\frac{\Gamma'(3)-2\log 3}{27}\Big)+\\ &\quad +\frac{1}{\xi^3\sigma^2}\Big(\frac{\Gamma(2\xi+3)}{3^{2\xi+3}}-2\frac{\Gamma(\xi+3)}{3^{\xi+3}}+\frac{2}{27}\Big)=\\ &= \frac{1}{27\xi^2\sigma^2}\Big(\frac{1}{3^\xi}(\Gamma'(\xi+3)-\log 3\Gamma(\xi+3))-\Gamma'(3)+2\log 3+\\ &\quad +\frac{1}{\xi 3^{2\xi}}\Gamma(2\xi+3)-\frac{2}{\xi 3^\xi}\Gamma(\xi+3)+\frac{2}{\xi}\Big). \qquad (A.9)\end{aligned}$$

Inverse of the Cramér-von-Mises information matrix $I_{(\sigma,\xi)}$ then can be calculated as follows

$$\mathcal{J}^{-1}_{(\sigma,\xi)} = \begin{pmatrix} J_{\xi\xi} & J_{\xi\sigma}\\ J_{\xi\sigma} & J_{\sigma\sigma}\end{pmatrix}^{-1} = \frac{1}{J_{\xi\xi}J_{\sigma\sigma}-J^2_{\xi\sigma}}\begin{pmatrix} J_{\sigma\sigma} & J_{\xi\sigma}\\ -J_{\xi\sigma} & J_{\xi\xi}\end{pmatrix}^{-1},$$

plugging expressions (A.5), (A.8) and (A.9) in. We omit writing inverse matrix explicitly due to its cumbersome form.

To complete expression (A.1) we calculate integral two integrals for the scale and shape parameters, i.e.

$$\int_{-1/\xi}^{\infty}(1-Q_{\sigma,\xi}(z))\Delta_{\xi}(z)dQ_{\sigma,\xi}(z)=$$

$$=-\int_{0}^{\infty}(1-e^{-u})\frac{1}{\xi}e^{-u}\Big(u\log u+\frac{1}{\xi}(u-u^{\xi+1})\Big)\Big(-\frac{1}{\sigma}e^{-u}\Big)du=$$

$$=\frac{1}{\xi\sigma}\Big(\int_{0}^{\infty}e^{-2u}(u\log u+1/\xi(u-u^{\xi+1}))du-\int_{0}^{\infty}e^{-3u}(u\log u+1/\xi(u-u^{\xi+1}))du\Big)=$$

$$=\frac{1}{\xi\sigma}\Big(\frac{1}{4}(\Gamma'(2)-\log 2\Gamma(2))+\frac{1}{4\xi}\Gamma(2)-\frac{1}{2^{\xi+2}\xi}\Gamma(\xi+2)-$$

$$-\frac{1}{9}(\Gamma'(2)-\log 3\Gamma(2))+\frac{1}{9\xi}\Gamma(2)-\frac{1}{3^{\xi+2}\xi}\Gamma(\xi+2)\Big)=$$

$$=\frac{1}{\xi\sigma}\Big(\Big(\frac{1}{3^{\xi+2}}-\frac{1}{2^{\xi+2}}\Big)\frac{1}{\xi}\Gamma(\xi+2)+\frac{5}{36\xi}+\frac{5}{36}\Gamma'(2)+\Big(\frac{\log 3}{9}-\frac{\log 2}{4}\Big)\Big).$$

Analogically we get it for the scale

$$\int_{-1/\xi}^{\infty}(1-Q_{\sigma,\xi}(z))\Delta_{\sigma}(z)dQ_{\sigma,\xi}(z)=-\int_{0}^{\infty}(1-e^{-u})e^{-u}\frac{1}{\xi\sigma}(u^{\xi+1}-u)\Big(-\frac{1}{\sigma}e^{-u}\Big)du=$$

$$=\frac{1}{\xi\sigma^{2}}\Big(\int_{0}^{\infty}e^{-2u}(u^{\xi+1}-u)du-\int_{0}^{\infty}e^{-3u}(u^{\xi+1}-u)du\Big)=$$

$$=\frac{1}{\xi\sigma^{2}}\Big(\frac{1}{2^{\xi+2}}\Gamma(\xi+2)-\frac{1}{4}\Gamma(2)-\frac{1}{3^{\xi+2}}\Gamma(\xi+2)+\frac{1}{9}\Gamma(2)\Big)=$$

$$=\frac{1}{\xi\sigma^{2}}\Big(\Big(\frac{1}{2^{\xi+2}}-\frac{1}{3^{\xi+2}}\Big)\Gamma(\xi+2)-\frac{5}{36}\Big).$$

The last calculation to be doen concerns the second integral in the expression (A.1), which we calculate separately for shape and scale

$$\int_{-\frac{1}{\xi}}^{z}\Delta_{\xi}(y)dQ_{\sigma,\xi}(y)=-\int_{u}^{\infty}\frac{1}{\xi}e^{-s}(s\log s+\frac{1}{\xi}(s-s^{\xi+1})(-\frac{1}{\sigma}e^{-s})ds=$$

$$=\frac{1}{4\xi\sigma}\Big(\int_{u}^{\infty}e^{-2s}s\log sds+\int_{u}^{\infty}\frac{1}{\xi}e^{-2s}(s-s^{\xi+1})ds\Big)=$$

$$=\frac{1}{4\xi\sigma}(\Gamma'(2,u)-\log 2\Gamma(2,u))+\frac{1}{\xi^{2}\sigma}\Big(\frac{1}{4}\Gamma(2,u)-\frac{1}{2^{\xi+2}}\Gamma(\xi+2,u)\Big)=$$

$$=-\frac{1}{2^{\xi+2}\xi^{2}\sigma}\Gamma(\xi+2,u)+\frac{1}{4\xi\sigma}((1/\xi-\log 2)\Gamma(2,u)+\Gamma'(2,u)).$$

where $\Gamma(2,u)$ denotes the *incomplete Gamma function* and

$$\int_{-\frac{1}{\xi}}^{z} \Delta_\sigma(y) dQ_{\sigma,\xi}(y) = -\int_u^\infty e^{-s} \frac{1}{\xi\sigma}(s^{\xi+1} - s)(-\frac{1}{\sigma}e^{-s})ds =$$

$$= \frac{1}{\xi\sigma^2}\Big(\frac{1}{2^{\xi+2}}\Gamma(\xi+2,u) - \frac{1}{4}\Gamma(2,u)\Big).$$

We get IF of Cramér-vom-Mises minimum distance estimator of the form

$$IF(x, MDE, Q_{\sigma,\xi}) = I^{-1}_{(\sigma,\xi)}\Big(\varphi_\xi(x), \varphi_\sigma(x)\Big)^{\mathrm{T}},$$

with Cramér-von-Mises information matrix obtained by (A.5),(A.8) and (A.9) and functions

$$\varphi_\xi(x) = \frac{1}{\xi\sigma}\Big(\frac{1}{2^{\xi+2}\xi}\Gamma\Big(\xi+2, (1+\xi\frac{x-\mu}{\sigma})^{-\frac{1}{\xi}}\Big) - \frac{1}{4}\Big((1/\xi - \log 2)\Gamma(2,u) +$$

$$+\Gamma'\Big(2, (1+\xi\frac{x-\mu}{\sigma})^{-\frac{1}{\xi}}\Big)\Big) + \Big(\frac{1}{3^{\xi+2}} - \frac{1}{2^{\xi+2}}\Big)\frac{1}{\xi}\Gamma(\xi+2) + \frac{5}{36\xi} + \frac{5}{36}\Gamma'(2) + \frac{\log 3}{9} - \frac{\log 2}{4}\Big)$$

and

$$\varphi_\sigma(x) = \frac{1}{\xi\sigma^2}\Big(\frac{1}{4}\Gamma\Big(2, (1+\xi\frac{x-\mu}{\sigma})^{-\frac{1}{\xi}}\Big) - \frac{1}{2^{\xi+2}}\Gamma\Big(\xi+2, (1+\xi\frac{x-\mu}{\sigma})^{-\frac{1}{\xi}}\Big) +$$

$$+\Big(\frac{1}{2^{\xi+2}} - \frac{1}{3^{\xi+2}}\Big)\Gamma(\xi+2) - \frac{5}{36}\Big).$$

□

Appendix B

L_2 differentiability for GLM

B.1 Proof of Lemma 6.3

First, we take some non zero k'-dimensional sequence h_n, which converges to zero, i.e. $h_n \neq 0$ and $h_n \to 0, n \to \infty$ in $\mathbb{R}^{k'}$. We denote $\vartheta_n := l(\theta_0 + h_n)$ and $\theta_0 := l(\theta_0)$.

By using smoothness of the link function l, we get corresponding expression for the parameter

$$\vartheta_n = l(\theta_0 + h_n) = \vartheta_0 + \dot{l}(\theta_0)h_n + r(\theta_0, h_n), \tag{B.1}$$

with the remainder function r, which satisfies the following convergence

$$\frac{r(\theta_0, h_n)}{|h_n|} \to 0, n \to \infty. \tag{B.2}$$

We suppose, that probabilities Q_{ϑ_n} are dominated by some measure ν. Therefore, we denote corresponding absolutely continuous densities as q_{ϑ_n}, s.t. $dQ_{\theta_n} = q_{\theta_n} d\nu$.

In Section 6.1 we assumed that parametric model $\mathcal{Q}$ is L_2 differentiable, therefore, by the Definition 2.6 for the expression

$$R_n := \int (\sqrt{q_{\vartheta_n}} - \sqrt{q_{\vartheta_0}}(1 + \frac{1}{2}(\Lambda^{\mathcal{Q}}_{\vartheta_0})^{\mathrm{T}}(\vartheta_n - \vartheta_0)))^2 d\nu, \quad \text{holds} \quad \frac{R_n}{|\theta_n - \theta_0|^2} \to 0, n \to \infty \tag{B.3}$$

From the other side, plugging expression (B.1) in the same term R_n, we can rewrite it in the following way:

$$R_n = \int (A_n - B_n)^2 d\nu, \quad \text{where}$$

$$A_n := \sqrt{q_{\theta_n}} - \sqrt{q_{\theta_0}}(1 + \frac{1}{2}(\Lambda^{Q}_{\theta_0})^{\mathrm{T}}\dot{l}(\vartheta_0)h_n) \quad \text{and} \quad B_n := \frac{1}{2}\sqrt{q_{\theta_0}}(\Lambda^{Q}_{\theta_0})^{\mathrm{T}} r(\vartheta_0, h_n).$$

By using well-known Cauchy-Schwarz inequality, modified in the next way

$$A_n^2 = (A_n - B_n + B_n)^2 \leq 2(A_n - B_n)^2 + 2B_n^2$$

and applying integration w.r.t. the dominating measure ν, we get the following inequality

$$\int A_n^2 d\nu \leq 2\int (A_n - B_n)^2 d\nu + 2\int B_n^2 d\nu = 2R_n + 2\int B_n^2 d\nu \leq$$

$$\leq 2R_n + \frac{1}{2}|r(\vartheta_0, h_n)|^2 \int q_{\vartheta_0} |\Lambda_{\vartheta_0}^{\mathcal{Q}}|^2 d\nu \leq 2R_n + \frac{1}{2}|I_{\vartheta_0}^{\mathcal{Q}}||r(\vartheta_0, h_n)|^2.$$

Therefore, applying (B.1), (B.2), and (B.3), we get that

$$\frac{1}{|h_n|^2}\int A_n^2 d\nu = \frac{2R_n}{|h_n|^2} + \frac{1}{2}|I_{\vartheta_0}^{\mathcal{Q}}|\frac{|r(\vartheta_0, h_n)|^2}{|h_n|^2} =$$

$$= \frac{2R_n}{|\vartheta_n - \vartheta_0|^2}\frac{(\dot{l}(\vartheta_0)h_n + r(\vartheta_0, h_n))^2}{|h_n|^2} + \frac{1}{2}|I_{\vartheta_0}^{\mathcal{Q}}|\frac{|r(\vartheta_0, h_n)|^2}{|h_n|^2} = o(1).$$

If we return expression under the notation A_n to the last equality, we get exactly needed condition (2.9) from the Definition 2.6, therefore, parametric model $\tilde{\mathcal{Q}}$ is L_2 differentiable in $\theta_0 \in \Theta'$. □

B.2 Proof of Theorem 6.1

We take some sequence in $\mathbb{R}^p$, which converges to zero, i.e. $s_n \to 0, n \to \infty$ and s.t. $\tilde{s}_n = s_n/|s_n| \to \tilde{s}_0$ for some $\tilde{s}_0$ with $|\tilde{s}_0| = 1$.

Here we introduce some additional notations $\vartheta_s := l(\theta_s)$, $\theta_s := x^{\mathrm{T}}(\beta_o + s)$ and $\dot{l}_s := \dot{l}(\theta_s)$.

As it was in the previous proof, we suppose that probabilities Q_{ϑ_n} are dominated by some measure ν and denote corresponding densities as q_{ϑ_n}.

Using similar notations as in the proof of the chain rule B.1, we state that by Definition 2.6, generalized linear model $\mathcal{P}$ is L_2 differentiable at every $\beta \in \mathbb{R}^p$ if holds the following convergence

$$\frac{1}{|s_n|^2}\int\int \tilde{A}_n^2 \nu(dy)K(dx) \to 0, \quad n \to \infty,$$

for similar expression to A_n from above, only taking up the dependence on x, i.e.

$$\tilde{A}_n = \tilde{A}_n(x, y) := \sqrt{q_{\vartheta_n}} - \sqrt{q_{\vartheta_0}}(1 + \frac{1}{2}(\Lambda_{l(x^{\mathrm{T}}\beta_0)}^{\mathcal{Q}})^{\mathrm{T}}\dot{l}(x^{\mathrm{T}}\beta_0) \cdot_\pi x^{\mathrm{T}}s_n). \tag{B.4}$$

Applying the chain rule pointwise in (x, y) and the Hájek condition (H.1), leads to the pointwise existence (for P_β-a.e. (x, y)) of L_2 derivative of the form (6.7).

From the last steps of the proof of Lemma 6.3 we obtain

$$\int \tilde{A}_n^2 \nu(dy) = |x^{\mathrm{T}} s_n|^2 (z(x^{\mathrm{T}} s_n))^2.$$

for some function $z(s) \to 0$, K-a.e. x and s small enough.

Therefore, for K-a.e. fixed x we get that

$$\tilde{A}_n'(x) := \frac{1}{|s_n|^2} \int \tilde{A}_n^2 \nu(dy).$$

From the other side, we can use the Cauchy-Schwarz inequality $(a-b)^2 \le 2(a^2+b^2)$ for the $\tilde{A}_n'(x)$ as follows

$$\tilde{A}_n'(x) \le \frac{2}{|s_n|^2} \int (\sqrt{q_{\vartheta_{s_n}}} - \sqrt{q_{\vartheta_0}})^2 \nu(dy) + \frac{1}{2|s_n|^2} \int q_{\vartheta_0} ((\Lambda^{\mathcal{Q}}_{l(x^{\mathrm{T}}\beta_0)})^{\mathrm{T}} \dot{l}(x^{\mathrm{T}}\beta_0) \cdot_\pi x^{\mathrm{T}} s_n)^2 \nu(dy). \tag{B.5}$$

Then, we apply well-known fundamental theorem of calculus for absolutely continuous functions to the first summand, using Lebesgue measure λ, fixed $x \in \mathbb{R}^p$ and $u \in [0;1]$. For K-a.e. fixed x we obtain

$$\frac{1}{|s_n|^2} \int (\sqrt{q_{\vartheta_{s_n}}} - \sqrt{q_{\vartheta_0}})^2 d\nu = \frac{1}{|s_n|^2} \int \Big(\int_0^1 \frac{1}{2} \sqrt{q_{\vartheta_{us_n}}} ((\dot{l}_{us_n})^{\mathrm{T}} \Lambda^{\mathcal{Q}}_{\vartheta_{us_n}} \cdot_\pi x^{\mathrm{T}} s_n) \lambda(du) \Big)^2 d\nu \le$$

$$\le \frac{1}{4|s_n|^2} \int \int_0^1 q_{\vartheta_{us_n}} ((\dot{l}_{us_n})^{\mathrm{T}} \Lambda^{\mathcal{Q}}_{\vartheta_{us_n}} \cdot_\pi x^{\mathrm{T}} s_n)^2 \lambda(du) d\nu = \frac{1}{4} \tilde{s}_n^{\mathrm{T}} \int_0^1 I^{\mathcal{Q}}_{\vartheta_{us_n}}(x) \lambda(du) \tilde{s}_n =$$

$$= \frac{1}{4|s_n|} \tilde{s}_n^{\mathrm{T}} \int_0^{|s_n|} I^{\mathcal{Q}}_{\vartheta_{us_n}}(x) \lambda(du) \tilde{s}_n =: B_n(x)$$

Additionally, we introduce the term $B_0 = \frac{\tilde{s}_n^{\mathrm{T}}}{4} I^{\mathcal{Q}}_{\vartheta_0}(x) \tilde{s}_n$.

By using conditions (ii) and (iii), we obtain that in integral $\int B_n(x) K(dx)$ is finite eventually in n. Therefore, applying condition (iii) and Fubini theorem, we get the following expression

$$\int B_n(x) K(dx) = \frac{1}{4} \int_0^{|s_n|} \int |I^{\mathcal{Q}}_{\vartheta_{us_n}}(x)| K(dx) \lambda(du) = \int B_0(x) K(dx) + o(1).$$

Next, using Vitali's theorem (see Rieder (1994, Prop. A.2.2)), we conclude that B_n is uniformly integrable (w.r.t. K). Moreover, from the inequality (B.5) we obtain, that $\tilde{A}_n'(x) \le 2B_n(x) + 2B_0(x)$. Therefore, $\tilde{A}_n'(x)$ is also uniformly integrable (w.r.t. K).

Hence, again by Vitali's theorem, $\int \tilde{A}'_n(x)K(dx) \to 0$, what is exactly condition (2.9) from the Definition 2.6. Hence, by Definition 2.6, generalized linear model $\mathcal{P}$ is L_2 differentiable at every $\beta \in \mathbb{R}^p$.

Continuity (2.11) also follows from Vitali's theorem, as it is just continuity of the Fisher information just proven.

If we replace in the proof B_n and B_0 by the terms $|I^{\mathcal{Q}}_{\vartheta_{st}}||\dot{l}_{st}|^2|x|^2$ and $|I^{\mathcal{Q}}_{\vartheta_0}||\dot{l}_0|^2|x|^2$ respectivly, we get the proof of Remark 6.2.

□

B.3 Proof of Theorem 6.5

Argument for the condition (4.17) from the Definition 4.3 we reproduce from Rieder (1994, Thm. 2.3.7). Since parametric model $\mathcal{Q}$ is assumed to be L_2 differentiable, by definition (2.9), with the densities q_ϑ, we get that

$$|\int(\sqrt{q_{\vartheta+h}}-\sqrt{q_\vartheta}(1+\frac{1}{2}(\Lambda^{\mathcal{Q}}_\vartheta)^{\mathrm{T}}h)\sqrt{q_\vartheta}d\nu|^2 \le \int|\sqrt{q_{\vartheta+h}}-\sqrt{q_\vartheta}(1+\frac{1}{2}(\Lambda^{\mathcal{Q}}_\vartheta)^{\mathrm{T}}h)|^2d\nu = o(|h|^2),$$

which leads to the following property

$$\mathbf{E}_\vartheta(\Lambda^{\mathcal{Q}}_\vartheta)^{\mathrm{T}}h \ge \int(\sqrt{q_{\vartheta+h}}-\sqrt{q_\vartheta})\sqrt{q_\vartheta}d\nu + o(|h|) =$$

$$= \int\sqrt{q_{\vartheta+h}}\sqrt{q_\vartheta}d\nu - 1 + o(|h|) = -\frac{1}{2}\int(\sqrt{q_{\vartheta+h}}-\sqrt{q_\vartheta})^2d\nu + o(|h|) =$$

$$= -\frac{1}{2}h^{\mathrm{T}}I^{\mathcal{Q}}_\vartheta h + o(|h|^2) + o(|h|) = o(|h|).$$

Therefore, we conclude that $\mathbf{E}_\vartheta\Lambda^{\mathcal{Q}}_\vartheta = 0$, and then the first condition from Definition 4.3 is satisfied, i.e. $\mathbf{E}_{n,i,\beta_0}\Lambda^{\mathcal{P}}_{n,i,\beta_0} = 0$.

Lindeberg condition (4.18) is fulfilled automatically, so it is left to show that condition (4.19) is also satisfied by the GLM $\mathcal{P}$.

We denote the $Q_{\vartheta_{n,i,t_n}}$-null set as $N_{n,i}$ and suppose that two Hájek conditions (H.1) and (H.2) hold for all $y \in N^c_{n,i}$. Then we let $N = \bigcup_n \bigcup_{i=1}^{i_n} N_{n,i}$. Then from (H.1) and the chain rule Lemma 6.3 applied pointwise (in $y \in N^c$), analogically to the case of stochastic regressors, we obtain (pointwise) existence and form and of the L_2 derivative.

We inherit notation $\tilde{A}_n$ from the previous proof, only replacing sequence s_n by t_n. Moreover, here we assume that $\tilde{A}_n$ from (B.4) takes up the dependence on the regressors $x_{n,i}$, i.e. $\tilde{A}_{n,i} = \tilde{A}_n(x_{n,i})$.

Then for every fixed i we get that

$$\tilde{A}'_{n,i} := \int \tilde{A}^2_{n,i} \nu(dy) \to 0 \quad \text{as} \quad t_n \to 0.$$

If we show that convergence $\lim_{n\to\infty} \sup_{|t|\le b} \sum_{i=1}^{i_n} \int \tilde{A}^2_{n,i} \nu(dy) = 0$ holds, then clearly condition (4.19) is also fulfilled.

We use similar trick as in the previous proof. For some fixed value i we apply fundamental theorem of calculus for absolutely continuous functions to get the following

$$\tilde{A}'_{n,i} = \int (\sqrt{q_{\vartheta_{n,i,t_n}}} - \sqrt{q_{\vartheta_{n,i,0}}})^2 d\nu \le \frac{1}{2|t_n|} \int_0^{|t_n|} t_n^{\mathrm{T}} I^{\mathcal{P}}_{n,i,ut} t_n \lambda(du) =: B_{n,i},$$

with Lebesgue measure λ and $u \in [0;1]$. As before, we also denote $B_{0,i} = \frac{1}{4} t_n^{\mathrm{T}} I^{\mathcal{P}}_{n,i,0} t_n$ and note that $\sum_{i=1}^{i_n} I^{\mathcal{P}}_{n,i,0} = I^{\mathcal{P}}_{n,\beta_0}$, therefore $t_n^{\mathrm{T}} I^{\mathcal{P}}_{n,i,0} t_n = |t|^2 \le b$.

Then by the condition (iii) from the theorem we get that

$$\sum_{i=1}^{i_n} B_{n,i} = \sum_{i=1}^{i_n} B_{0,i} + 0(1) = \frac{|t|}{4} + o(1).$$

We again apply the Vitali's theorem, which leads to the uniform integrability of $B_{n,i}$ w.r.t. the counting measure. Since we got that $\tilde{A}'_{n,i} \le 2B_{n,i} + 2B_{0,i}$ holds, we obtain that $\tilde{A}'_{n,i}$ is also uniformly integrable. Hence, by using Vitali's theorem again, we get that $\sum_{i=1}^{i_n} \tilde{A}'_{n,i} \to 0$, what is exactly condition (4.19) from the Definition 4.3. Therefore, by Definition 4.3 generalized linear model $\mathcal{P}$ is L_2 differentiable.

Continuity (4.20) also follows from Vitali's theorem, as it is just continuity of the Fisher information just shown.

Again, if we replace in the proof $B_{n,i}$ and $B_{0,i}$ by the terms $|I^{\mathcal{Q}}_{n,i,t}||\dot{l}_{n,i,t}|^2|x_{n,i}|^2$ and $|I^{\mathcal{Q}}_{n,i,0}||\dot{l}_{n,i,0}|^2|x_{n,i}|^2$ respectively, we get the proof of Remark 6.6.

□

B.4 Link function for GEVD joint shape-scale model

In Section 6.2, giving examples of the GEVD and GPD joint shape-scale models, we designed link function for shape of GEVD of the form $l_\xi(\theta_\xi) = \log(f(\log(x_\xi)^{\mathrm{T}}\beta_\xi))$, where function f is the following

$$f(x) = (x^2/2 + x + 1)\mathbb{1}(x > 0) + (a_1(\log(a_2 - x))^{-2} + a_3)\mathbb{1}(x \leq 0)$$

for some $a_1, a_2, a_3 > 0$. We need function f to be continuously differentiable in 0 and $f(x) > e^{-1/2}$, therefore

$$\frac{a_1}{(\log(a_2))^2} + a_3 = \frac{2a_1}{a_2(\log(a_2))^3} = 1, \quad \frac{a_1}{(\log(a_2 - x))^2} + a_3 > e^{-1/2}, \forall x < 0.$$

In the last inequality we have that $a_1(\log(a_2 - x))^{-2} > 0$, so the choice of the constant a_3 will be $a_3 = e^{-1/2} \approx 0.6063$ to ensure this inequality. Then, we solve system of two first equations with two unknowns, and we get $a_2^{a_2} = e^{2(1-e^{-0.5})}$, so $a_2 \approx 1.624$ and $a_1 = 0.5a_2(\log(a_2))^3 \approx 0.00926$. Hence, function f approximately turns to

$$f(x) = (x^2/2 + x + 1)\mathbb{1}(x > 0) + (0.00926(\log(1.624 - x))^{-2} + 0.6063)\mathbb{1}(x \leq 0).$$

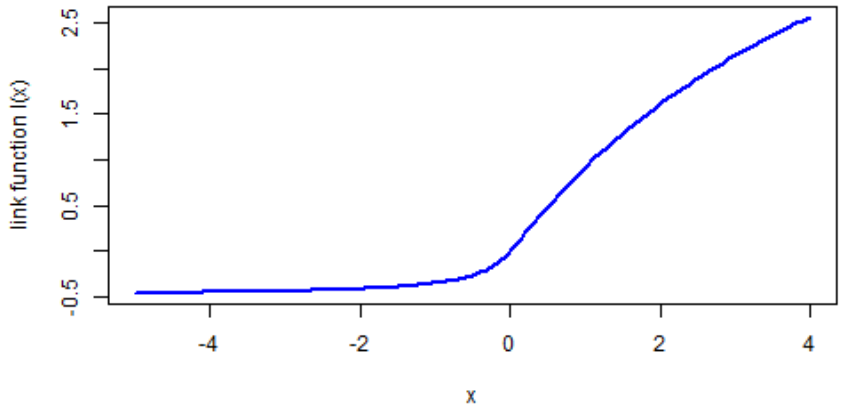

FIGURE B.1: Link function for the shape of GEVD

We have mentioned that usually shape is varying in the interval $(0, 2)$. It is visible from the Figure B.1, that the argument of the link function $\beta\log(x_{t-1})$, then falls to the corresponding interval $(-\infty, \sqrt{1 - 2(1 - e^2)} - 1 \approx 2.712)$. Therefore, we conclude that taking $\beta = 1$ our link $l = \log(f(\beta\log(x_{t-1}))) < 2$ as long as $x_{t-1} < 15$, and $l < 3$ for $x_{t-1} < 193$.

Next, we show that our choice of link function for GEVD, fulfills conditions (ii) and (iii) of the Theorem 6.1. First, we calculate derivative $\dot{l} = \dot{f}/f$ and obtain

$$\dot{l} = (x+1)/(x^2/2 + x + 1)\mathbb{1}(x > 0) + 2a_1(a_2 - x)^{-1}(\log(a_2 - x))^{-3}\mathbb{1}(x \leq 0).$$

Hence, for large x, $\dot{l}$ behaves like $2/x$, while for $x < 0$, it essentially behaves like $-x^{-1}(\log(-x))^{-3}$.

As we mentioned in the example of Section 6.2, all terms of the Fisher information matrix for GEVD are dominated by the term $\Gamma(2\xi + 1)$. We use well-known Stirling approximation, i.e., $\Gamma(x) \approx \sqrt{2\pi}\,\exp(x(\log(x) - 1/2))$ and, due to the iterated logarithm in the link function, we get that $\Gamma(2l_\xi(\theta_\xi)) \approx \beta_\xi \log(x_\xi)$. Therefore, by equivariance in location μ and scale σ, condition (ii) turns to the finiteness of the following terms

$$B_1(\xi) := \frac{4}{\beta_\xi} \int \log(x_\xi) K(dx) \quad \text{for } \beta_\xi > 0 \tag{B.6}$$

and

$$B_2(\xi) := \frac{1}{\beta_\xi} \int \frac{\log(x_\xi)}{(\log(-\beta_\xi) + \log(\log(x_\xi))^6}\, K(dx) \qquad \text{for } \beta_\xi < 0 \tag{B.7}$$

Finiteness of (B.6) and (B.7) follows from finiteness of $\mathbf{E}(\min\{1, (\log x)^k\})$ for $x \sim \mathrm{GEVD}(0, 1, \xi)$, $k \in \mathbf{N}$, which is based on the fact that expectation of the random variable is the integral of its quantile, i.e.

$$\mathbf{E}_\xi(\min\{1, (\log x)^k\}) = \int_{u_0}^1 \Big(\log(((-\log y)^{-\xi} - 1)/\xi)\Big)^k dy$$

for $u_0 > \exp(-(1+\xi)^{-\frac{1}{\xi}})$ so that $((-\log y)^{-\xi} - 1)/\xi > 1$ for $y > u_0$. We use well-known inequality $-\log(x) < (1-x)/x$ for $x \in (0, 1)$ to bound quantile in the following way

$$\frac{((-\log y)^{-\xi} - 1)}{\xi} < \frac{(((1-y)/y)^{-\xi} - 1)}{\xi} < \frac{(1/y - 1)^{-\xi}}{\xi},$$

hence, finiteness of the expectation is equivalent to the followowing $\int_0^1 (-\log(y))^k dy < \infty$, what is true, since after some transformation one can see that $\int_0^1 (-\log(y))^k dy = \Gamma(k-1)$. Therefore, condition (ii) of the Theorem 6.1 is fulfilled by the chosen link.

Reconsidering (B.6) and (B.7) at $\xi + s$, for $|s| < h$, $h < 1$, we see that $\sup_{|s|<h} B_i(\xi + s) < \infty$ for $i = 1, 2$, hence condition (iii) of the Theorem 6.1 follows from dominated convergence and continuity of Fisher information $I_{\xi\xi}$ in ξ.

□

Bibliography

B. Anderson and J. Moore. *Optimal control. Linear quadratic methods.* Prentice Hall, 1990.

F. Anscombe. Rejection of outliers. *Technometrics*, 2:123–147, 1960.

A. Balkema and L. de Haan. Residual life time at great age. *Annals of Probability*, 2: 792–804, 1974.

J. Beirlant, P. Vynckier, and J. Teugels. Excess functions and estimation of the extreme-value index. *Bernoulli*, 2(4):293–318, 1996.

L. Bortkiewicz. Variationsbreite und mittlerer fehler. *Sitzungsber. Berli. Math. Ges.*, 21:3–11, 1922.

P. Brockwell and R. Davis. *Introduction to Time Series and Forecasting. Second Edition.* Springer-Verlag, New York, 2002.

E. Castillo. *Extreme Value Theory in Engineering.* Academic Press, 1988.

C. Chatfield. *The Analysis of Time Series: An Introduction. Fifth Edition.* Chapman & Hall/CRC, 1996.

T. Cipra and R. Romera. Robust kalman filter and its application in time series analysis. *Kybernetika*, 27:481–494, 1991.

S. Coles. *An Introduction to Statistical Modeling of Extreme Values.* Springer, 2001.

D. Cox. Statistical analysis of time series: some recent developments. *Scandinavian Journal of Statistics*, 8:93–115, 1981.

L. de Haan and A. Ferreira. *Extreme Value Theory: An Introduction.* Springer, 2006.

C. Dethlefsen, S. Lundbye-Christensen, and A. Christensen. *sspir: State Space Models in R.*, 2009. URL `http://CRAN.R-project.org/package=sspir`. R package version 0.2.8.

D. Donoho and I. Johnstone. Ideal patial adaptation via wavelet shrinkage. *Biometrika*, 81:425–455, 1994.

D. Dunkan and S. Horn. Linear dynamic recursive estimation from the viewpoint of regression analysis. *J. Am. Stat. Assoc.*, 67:815–821, 1972.

D. Dupuis and C. Field. Robust estimation of extremes. *Canadian Journal of Statistics.*, 26(2):199–216, 1998.

D. Dupuis and S. Morgenthaler. Robust weighted likelihood estimators with an application to bivariate extreme value problems. *The Candidat Journal of Statistics*, 30(1): 17–36, 2002.

P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling Extremal Events: For Insurance and Finance.* Springer, 1997.

A. Ershov and R. Lipster. Robust kalman filter in discrete time. *Autom. Remote Control*, 39:359–367, 1978.

L. Fahrmeir. Maximum likelihood estimation in misspecified generalized linear models. *Statistics*, 21(4):487–502, 1990.

L. Fahrmeir and H. Kaufmann. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, 13(1): 342–368, 1985.

L. Fahrmeir and G. Tutz. *Multivariate Statistical Modelling Based on Generalized Linear Models. 2nd Edn.* Springer., 2001.

M. Falk, J. H´usler, and R. Reiss. *Laws of Small Numbers: Extremes and Rare Events.* Springer, 2011.

T. Fernando. Kalman filtering in r. *Journal of Statistical Software*, 39(2), 2011.

A. Ferreira and L. de Haan. *Extreme Value Theory: An Introduction.* Springer, 2006.

A. Ferreira and L. de Haan. On the block maxima method in extreme value theory. arXiv:1310.3222v1, 2013. working paper.

C. Field and E. Ronchetti. *Small sample asymptotics, Vol. 13 of IMS Lecture Notes - Monograph Series.* Institute of Mathematical Statistics., Hayward, CA., 1990.

R. Firsher and L. Tippett. Limiting forms of the frequency distribution of the largest and smallest member of a sample. *Proc. Camb. Phil. Soc.*, 24:180–190, 1928.

A. Fox. Outliers in time series. *J. R. Stat. Soc., Ser. B*, 34:350–363, 1972.

M. Fréchet. Sur la loi de probabilitc de l'écart maximum. *Ann. SOC. Polon. Math. Cracovie*, 6:93–116, 1927.

R. Fried, J. Einbeck, and U. Gather. Weighted repeated median smoothing and filtering. *J. Amer. Statist. Assoc.*, (480):1300–1308, 2007.

E. Gilleland and R. Katz. New software to analyze how extremes change over time. *Eos*, 92(2):13–14, 2011.

E. Gilleland, M. Ribatet, and A. Steohenson. A software review for extreme value analysis. *Extremes.*, (16):103–119, 2013.

B. Gnedenko. Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of Mathematics*, 44:423–453, 1943.

E. Gumbel. *Statistics of Extremes.* Columbia University Press, 1958.

J. Hájek. Local asymptotic minimax and admissibility in estimation. *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability.*, 1:175–194, 1972.

F. Hampel. *Contributions to the theory of robust estimation.* Dissertation, University of California, Berkely, CA., 1968.

F. Hampel. A general definition of qualitative robustness. *The Annals of Mathematical Statistics*, 42:1887–1896, 1971.

F. Hampel. The influence curve and its role in robust estimation. *The Annals of Mathematical Statistics*, 69:383–393, 1974.

F. Hampel, E. Ronchetti, P. Rousseeuw, and W. Stahel. *Robust statistics. The approach based on influence functions.* Wiley Series in Probability and Mathematical Statistics. Wiley., 1986.

L. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica.*, (50):1029–1054, 1982.

H. Harter. A bibliography of extreme-value theory. *Internat. Statist. Rev.*, 46:279–306, 1978.

J. Heffernan and A. Stephenson. *ismev: An Introduction to Statistical Modeling of Extreme Values.*, 2012. URL `http://CRAN.R-project.org/package=ismev`. R package version 1.39.

J. Helske. *KFAS: Kalman Filter and Smoothers for Exponential Family State Space Models.*, 2010. URL `http://CRAN.R-project.org/package=KFAS`. R package version 0.6.0.

B. Hill. A simple general approach to inference about the tail of a distribution. *Annals of Statistics*, 3(5):1163–1174, 1975.

N. Horbenko. *Robuste Ansätze für Operationelle Risiken von Banken.* PhD thesis, Technical University of Kaiserslautern, Department of Mathematics, Kaiserslautern, November 2011.

R. Hosking and T. Wallis. Regional frequency analysis: An approach based on l-moments. *Cambridge University Press.*, 1997.

P. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35:73–101, 1964.

P. Huber. A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, 36:1753–1758, 1965.

P. Huber. *Robust Statistics.* New York: John Wiley & Sons, Inc., 1981.

M. Hubert, P. Rousseeuw, and S. Van Aelst. Multivariate outlier detection and robustness. *Handbook of Statistics.*, Volume 23: Data Mining and Computation in Statistics (C.R. Rao, E.J. Wegman, J.L. Solka, Eds.):263–302, 2005.

I. Ibragimov and Y. Linnik. *Independent and stationary sequences of random variables. Wolters-Noordhoff Series of Monographs and Textbooks on Pure and Applied Mathematics.* Wolters-Noordhoff Publishing Company., Groningen., 1971.

R. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering-Transactions of the ASME*, 82, 1960.

R. Kalman and R. Bucy. New result in linear filtering and prediction theory. *Journal of Basic Engineering*, 95, 1961.

S. Kassam and H. Poor. Robust techniques for signal processing: A survey. *Proc. IEEEl*, 73(3):433–481, 1985.

M. Kohl. *Numerical Contributions to the Asymptotic Theory of Robustness.* PhD thesis, University of Bayreuth, Bayreuth, September 2005.

M. Kohl. *RobRex: Optimally robust influence curves for regression and scale*, 2013. URL `http://robast.r-forge.r-project.org/`. R package version 1.0.

M. Kohl and P. Ruckdeschel. R-package distrex: Extensions of package distr. *CRAN.*, 2005.

M. Kohl and P. Ruckdeschel. R package distrMod: S4 classes and methods for probability models. *Journal of Statistical Software*, 35(10):1–27, 2010. URL `http://www.jstatsoft.org/v35/i10/`.

M. Kohl and P. Ruckdeschel. *RobLox: Optimally robust influence curves and estimators for location and scale*, 2013a. URL `http://robast.r-forge.r-project.org/`. R package version 1.0.

M. Kohl and P. Ruckdeschel. *RandVar: Implementation of random variables*, 2013b. URL `http://robast.r-forge.r-project.org/`. R package version 1.0.

M. Kohl and P. Ruckdeschel. *ROptEst: Optimally robust estimation*, 2013c. URL `http://robast.r-forge.r-project.org/`. R package version 1.0.

M. Kohl and P. Ruckdeschel. *ROptRegTS: Optimally robust estimation for regression-type models*, 2013d. URL `http://robast.r-forge.r-project.org/`. R package version 1.0.

M. Kohl and P. Ruckdeschel. *RobAStBase: Robust Asymptotic Statistics*, 2013e. URL `http://robast.r-forge.r-project.org/`. R package version 1.0.

S. Kotz and S. Nadarajah. *Extreme Value Distributions: Theory and Applications.* London: Imperial College Press, 2000.

H. Künsch. State space models and hidden markov models. *In: Barndorff-Nielsen OE, Cox DR and Klüppelberg C (Eds.) Complex Stochastic Systems*, pages 109–173, 2001.

M. Leadbetter, G. Lindgren, and H. Rootzen. *Extremes and Related Properties of Random Sequences and Series.* Springer Verlag, New York, 1983.

L. LeCam. On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *The Annals of Mathematical Statistics*, 41(3):802–828, 1970.

D. Luethi, P. Erb, and S. Otziger. *FKF: Fast Kalman Filter.*, 2010. URL `http://CRAN.R-project.org/package=FKF`. R package version 0.1.1.

R. Maronna, D. Martin, and V. Yohai. *Robust Statistics: Theory and Methods.* John Wiley & Sons, Inc., 2006.

P. McCullagh and J. Nedler. *Generalized linear models.* Chapman Hall, 1989.

J. Nedler and R. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society.*, Series A:370–384, 1972.

M. Nikulin. *Hellinger distance.* Encyclopedia of Mathematics, 2001.

G. Petris. An r package for dynamic linear models. *Journal of Statistical Software.*, 36 (12):1–16, 2010. URL `http://www.jstatsoft.org/v36/i12/`.

B. Pfaff and A. McNeil. *evir: Extreme Values in R.*, 2012. URL `http://CRAN.R-project.org/package=evir`. R package version 1.7-3.

J. Pfanzagl. *Asymptotic expansions for general statistical models.* With the assist. of W. Wefelmeyer., Vol. 31 of Lecture Notes in Statistics. Springer-Verlag., 1985.

P.Gilbert. *Brief User's Guide: Dynamic Systems Estimation.*, 2011. URL `http://CRAN.R-project.org/package=dse`. R package vignette, version 2009.12-1.

J. Pickands. Statistical inference using extreme order statistics. *Annals of Statistics.*, 3 (1):119–131, 1975.

D. Pollard, E. Torgersen, and G. L. Yang. *Festschrift for Lucien Le Cam.* Research Papers in Probability and Statistics., 1997.

D. Pupashenko. Robust kalman smoothing for dynamic vehicle data. Master thesis, University of Kaiserslautern, Kaiserslautern, August 2011.

D. Pupashenko, P. Ruckdeschel, and M. Kohl. L2 differentiability of generalized linear models. *Statistics and Probability Letters.*, 2014. URL `arXiv:1407.5798`.

R. Reiss and M. Thomas. *Statistical Analysis of Extreme Values: with Applications to Insurance, Finance, Hydrology and Other Fields.* Springer, 2007.

M. Ribatet. *Generalized Pareto Distribution and Peaks Over Threshold.*, 2007. URL `http://r-forge.r-project.org/projects/pot/`. R package version 1.0.5.

M. Ribatet and R. Singleton. *SpatialExtremes: Modelling Spatial Extremes.*, 2013. URL `http://CRAN.R-project.org/package=SpatialExtremes`. R package version 2.0-0.

H. Rieder. *Robust asymptotic statistics.* Springer Series in Statistics. Springer., 1994.

H. Rieder. A motivation for $1/\sqrt{n}$-shrinking-neighborhoods. *Metrika*, 63(3):295–307, 2006.

H. Rieder, M. Kohl, and P. Ruckdeschel. The costs of not knowing the radius. *Statistical Methods and Applications*, 17(1):13–40, 2008.

P. Rousseeuw, C. Croux, V. Todorov, A. Ruckstuhl, M. Salibian-Barrera, T. Verbeke, M. Koller, and M. Maechler. *robustbase: Basic Robust Statistics*, 2012. URL `http://CRAN.R-project.org/package=robustbase`. R package version 0.8-1-1.

P. Ruckdeschel. Robust kalman filtering. pages 483–516, 2000.

P. Ruckdeschel. *Ansätze zur Robustifizierung des Kalman-Filters.* Dissertation, Universität Bayreuth, Bayreuth, 2001.

P. Ruckdeschel. Higher order asymptotics for the mse of the sample median on shrinking neighborhoods. *ArXiv 1006.0045*, 2010a.

P. Ruckdeschel. Higher order expansion for the mse of m-estimators on shrinking neighborhoods. *ArXiv 1006.0037.*, 2010b.

P. Ruckdeschel. Optimally robust kalman filtering. *Techn. Report 185, Fraunhofer ITWM Kaiserslautern*, 2010c. URL `http://www.itwm.fraunhofer.de/fileadmin/ITWM-Media/Zentral/Pdf/Berichte_ITWM/2010/bericht_185.pdf`.

P. Ruckdeschel. Optimally robust covariances. Technical report, Fraunhofer ITWM, Kaiserslautern, Germany, May 2014.

P. Ruckdeschel and N. Horbenko. Robust estimators in generalized pareto models. *Forschungsbericht, Fraunhofer ITWM, Kaiserslautern.*, 2010.

P. Ruckdeschel and N. Horbenko. Yet another breakdown point notion: Efsbp - illustrated at scale-shape models. *Metrika.*, 75(8):1025–1047, 2012.

P. Ruckdeschel and N. Horbenko. Optimally-robust estimators in generalized pareto models statistics. *Statistics.*, 47(4):762–791, 2013.

P. Ruckdeschel and M. Kohl. R-package distrmod: Object orientated implementation of probability models. *CRAN.*, 2008.

P. Ruckdeschel, M. Kohl, T. Stabla, and F. Camphausen. R-package distr: Object oriented implementation of distributions. *CRAN.*, 2005.

P. Ruckdeschel, M. Kohl, T. Stabla, and F. Camphausen. S4 classes for distributions. *R News*, 6(2):2–6, May 2006. URL `http://www.uni-bayreuth.de/departments/math/org/mathe7/DISTR/distr.pdf`.

P. Ruckdeschel, M. Kohl, A. Hueller, and E. Feistl. R-package distrteach: Extensions of package distr for teaching stochastics/statistics in secondary school. *CRAN.*, 2008a.

P. Ruckdeschel, M. Kohl, T. Stabla, and F. Camphausen. S4 classes for distributions—a manual for packages distr, distrsim, distrtest, distrex, distrmod, and distrteach. Technical report, Fraunhofer ITWM, Kaiserslautern, Germany, July 2008b. URL `http://r-forge.r-project.org/plugins/scmsvn/viewcvs.php/*checkout*/pkg/distrDoc/inst/doc/distr.pdf?root=distr`.

P. Ruckdeschel, M. Kohl, and N. Horbenko. *RobExtremes: Optimally robust estimation for extreme value distributions*, 2013. URL `http://robast.r-forge.r-project.org/`. R package version 1.0.

P. Ruckdeschel, M. Kohl, B. Spangl, Sascha Desmettre, D. Pupashenko, M. Pupashenko, and N. Horbenko. Five diagnostic plots based on robust statistics. *In preparation.*, 2014a.

P. Ruckdeschel, B. Spangl, and D. Pupashenko. Robust kalman tracking and smoothing with propagating and non-propagating outliers. *Statistical Papers*, 55:93–123, 2014b.

K. Schettlinger. *Signal and variability extraction for online monitoring in intensive care.* Dissertation, TU Dortmund, Dortmund, 2009.

I. Schick and S. Mitter. Robust recursive estimation in the presence of heavy-tailed observation noise. *Kybernetika*, 22(2):1045–1080, 1994.

M. Schlather. *Glattheit von Generalisierten Linearen Modellen und statistische Folgerungen.* Dissertation, Universität Bayreuth, Bayreuth, 1994.

R. Shumway and D. Stoffer. An approach to time series smoothing and forecasting using the em algorithm. *Journal of Time Series Analysis*, 3(4), 1982.

B. Spangl. *On Robust Spectral Density Estimation.* Dissertation, Technischen Universität Wien, Wien, 2008.

A. Stephenson. evd: Extreme value distributions. *R News*, 2(2), 2002. URL `http://CRAN.R-project.org/doc/Rnews/`.

N. Stockinger and R. Dutter. Robust time series analysis: A survey. *Kybernetika*, 23: Supplement, 1987.

M. Taniguchi and Y. Kakizawa. *Asymptotic theory of statistical inference for time series.* Springer Series in Statistics. Springer., 2002.

L. Tippett. On the extreme individuals and the range of samples taken from a normal population. *Biometrika*, 17:364–387, 1925.

T. Todorov. rrcov: Scalable robust estimators with high breakdown point. *R package.*, 2009.

V. Todorov and P. Filzmoser. An object-oriented framework for robust multivariate analysis. *Journal of Statistical Software*, 32(3):1–47, 2009.

J. Tukey. A survey of sampling from contaminated distributions. *Contributions to Probability and Statistics*, 1, 1960.

J. Tukey. The future of data analysis. *The Annals of Mathematical Statistics*, 35:1–67, 1962.

A. van der Vaart. *Asymptotic Statistics.* Cambridge University Press., 1998.

E. Wan and R. van der Merwe. *The Unscented Kalman Filter.* John Wiley & Sons, Inc., 2002.

J. Wang, R. Zamar, A. Marazzi, V. Yohai, M. Salibian-Barrera, R. Maronna, E. Zivot, D. Rocke, D. Martin, M. Maechler, and K. Konis. *robust: Robust Library*, 2014. URL `http://CRAN.R-project.org/package=robust`. R package version 0.4-16.

W. Weibull. A statistical theory of the strength of materials. ingeniors vetenskaps akademien. *Handlingar*, 151(3):45–55, 1939.

W. Weibull. A statistical distribution function of wide applicability. *J. Appl. Mech.*, 18: 293–297, 1951.

J. Wolfowitz. Estimation by the minimum distance method . *The Annals of Mathematical Statistics*, 5:9–23, 1953-1954.

D. Wuertz. *fExtremes: Rmetrics - Extreme Financial Market Data.*, 2013. URL `http://CRAN.R-project.org/package=fExtremes`. R package version 3010.81.

Wissenschaftlicher Werdegang

09.2000 - 06.2004	Allgemeine Hochschulreife an Kiew Naturwissenschaftliches Gymnasium Nr. 145, Ukraine
09.2004 - 06.2008	Bachelor-Studium an der Nationalen Taras-Schewtschenko-Universität Kiew, Ukraine
09.2008 - 06.2010	Master-Studium an der Nationalen Taras-Schewtschenko-Universität Kiew, Ukraine
10.2009 - 08.2011	Master-Studium an der Technischen Universität Kaiserslautern
10.2011 - 07.2014	Doktorand bei Dr. habil. Peter Ruckdeschel an der Technischen Universität Kaiserslautern

Fachliche Veröffentlichungen

2014 **D.Pupashenko**, P.Ruckdeschel und M.Kohl - *L2 Differentiability of Generalized Linaer Models*, Statistics and Probability Letters (eingereicht).

2014 P.Ruckdeschel, B.Spangl und **D. Pupashenko** - *Robust Kalman tracking and smoothing with propagating and non-propagating outliers*, Statistical Papers, Vol. 55, Nr. 1, Seiten 93-123.

2013 P. Ruckdeschel, M. Kohl, N. Horbenko, G. Kroisandt, **D. Pupashenko**, und M. Pupashenko - *RobExtremes: Optimally robust estimation for extreme value distributions (v. 0.9)*, Open Source Software - R-package published on r-forge.

2013 **D.Pupashenko**, S.Shklyar and A.Kukush - *Asymptotic properties of Corrected Score estimator in autoregressive model with meas-urement errors*, Teor. Imovir. ta Matem. Statyst., Nr.89, Seiten 156-166.

Scientific Background

09.2000 - 06.2004	High school at the Scientific High School Nr. 145 in Kiev, Ukraine
09.2004 - 06.2008	Bachelor-Studies at the National Taras-Schewtschenko-University Kiev, Ukraine
09.2008 - 06.2010	Master-Studies at the National Taras-Schewtschenko-University Kiev, Ukraine
10.2009 - 08.2011	Master-Studies at the University of Kaiserslautern
10.2011 - 07.2014	PhD student with supervisor Dr. habil. Peter Ruckdeschel at the University of Kaiserslautern

Publications

2014 **D.Pupashenko**, P.Ruckdeschel and M.Kohl - *L2 Differentiability of Generalized Linaer Models*, Statistics and Probability Letters (submitted).

2014 P.Ruckdeschel, B.Spangl and **D. Pupashenko** - *Robust Kalman tracking and smoothing with propagating and non-propagating outliers*, Statistical Papers, Vol. 55, Nr. 1, Pages 93-123.

2013 P. Ruckdeschel, M. Kohl, N. Horbenko, G. Kroisandt, **D. Pupashenko**, and M. Pupashenko - *RobExtremes: Optimally robust estimation for extreme value distributions (v. 0.9)*, Open Source Software - R-package published on r-forge.

2013 **D.Pupashenko**, S.Shklyar and A.Kukush - *Asymptotic properties of Corrected Score estimator in autoregressive model with measurement errors*, Teor. Imovir. ta Matem. Statyst., Nr.89, Pages.156-166.