Delft University of Technology
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft Institute of Applied Mathematics

Bayesian inference on the generalized Heffernan and Tawn model

A reflection on a pragmatic multivariate extreme value model

A thesis submitted to the
Delft Institute of Applied Mathematics
in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE
in
APPLIED MATHEMATICS

by

THIJS WILLEMS

Delft, the Netherlands,
November 2016

DELFT UNIVERSITY OF TECHNOLOGY

MASTER THESIS

Bayesian inference on the generalized Heffernan and Tawn model

Author:
Thijs WILLEMS

Supervisor:
Dr. Juan-Juan CAI

Supervising professor:
Prof. dr. ir. Geurt JONGBLOED

Other Committee member(s):
Dr. Pasquale CIRILLO

In collaboration with Shell:
PhD P. Jonathan
PhD D. Randell
PhD S. Bierman

A thesis submitted in fulfillment of the requirements for the degree of Master of Science

in the

Department of Applied Statistics
Delft Institute of Applied Mathematics

Thursday 17th November, 2016

"The tall waves were resounding, no one could tell why. Whichever wave you looked at, each one was trying to rise higher than all the rest and to chase and crush the next one; after it a third as fierce and hideous flew noisily, with a glint of light on its white crest. The sea has no sense and no pity."

Anton Chekhov (1890, *Gusev*)

Preface

Before you lies the culmination of this thesis project. Nine months of mathematical research and intensive coding have led to this report and the results presented in it. If Prometheus handed me his fire — a homage to the professors who's lectures I have enjoyed over the last couple of years — this project felt like poring a can of Shell V-power diesel onto that firing, resulting into a burst of flames.

The story of this thesis project starts a little over one year ago. It was a sunny autumn day when I was visiting Delft for the weekend, during my stay at EPFL in Lausanne, Switzerland. I had scheduled an appointment with professor Jongbloed to discuss the possibility of writing a thesis on a statistical subject. Expecting a large pile of project proposals, I was caught of balance when he said: "Rather than telling you what we have in store, why don't you tell me what you are looking for?". Giving the question some thought, I came up with two preferences to narrow the scope: the subject should be related to *extreme value theory* and the project should be commissioned by a *large engineering company*. As a stroke of fate, Stijn Bierman from Shell's Statistics and Chemometrics department had recently approached the TU Delft and was willing to supervise a master project on extreme value theory. I seized the opportunity as it was a perfect fit to my preferences.

At the offset of the project, I suspected it was going to be challenging. My background in statistics was limited to the mere basics, and as a civil engineering graduate this was going to be my first mathematical research project. On the upside, it created an exciting opportunity to learn a lot of new mathematics. The lessons I learned and the knowledge I gained during this project have enriched me personally and made me a more all-round mathematician.

First of all, I would like to express my sincere gratitude to Stijn Bierman, David Randell and Philip Jonathan from Shell for the amazing opportunity to work alongside them on cutting edge applied statistics. Their relentless support was indispensable to address the numerous (coding) challenges throughout this thesis project. Secondly, I am very grateful to Juan-Juan Cai — my daily supervisor from the TU Delft — for her support during our (bi-)weekly meetings. Decoding the incomprehensible paper by Heffernan and Tawn (2004) and her constructive and critical attitude towards assertions in my report have been a great help. Thirdly, a word of thanks to Pasquale Cirillo and Geurt Jongbloed for taking on the (non-)enviable task of reviewing my work as members of my thesis committee. Finally, a brief acknowledgement to my parents, girlfriend and roommates for their informal support and sticking up with me at times the project got the better of me.

Abstract

Multivariate extreme value modeling has gained traction in a wide range of applications to account for extremal dependence. To accomodate covariates, Jonathan et al. (2014) presented a generalization of the multivariate extreme value model proposed by Heffernan and Tawn (2004). Statistical inference and communicating uncertainty for the generalized Heffernan and Tawn model is challenging because of the large number of parameters and the non-linear relationship between certain model parameters.

In this thesis, a novel Bayesian approach is presented to address these issues. The manifold Metropolis adjusted Langevin algorithm (mMALA) proposed by Girolami and Calderhead (2011) is adopted as the standard Metropolis-Hastings algorithm is shown to yield disproportionate (auto)correlation in the posterior samples. The expected Fisher information matrix for the Heffernan and Tawn is derived as it defines a suitable metric on the statistical manifold associated to the parameter space. This metric is exploited by the mMALA to ensure faster convergence and superior mixing of the Markov chains.

Properties of the negative log-likelihood function and the maximum likelihood estimator for the Heffernan and Tawn model parameters are studied. The leading sources of bias in the maximum likelihood estimator are identified. The full observed– and expected Fisher information matrix are shown to be positive semi-definite for only a subspace of the parameter space.

The fact that the Heffernan and Tawn model can accommodate both asymptotic dependent– and asymptotic independent data is a distinctive feature compared to other multivariate extreme value models. However, the model as proposed by Heffernan and Tawn (2004) requires a different parameterization for either class of extremal dependence. A marginal transformation to the Laplace scale was suggested by Keef et al. (2013) to obtain a unified parameterization. This transformation is adopted and the proposed Bayesian inference methodology is shown to accommodate both classes of extremal dependence.

The contributions set forth in this thesis contribute to an enhanced understanding of the Heffernan and Tawn model. New and complex applications of the (generalized) Heffernan and Tawn model open up as the proposed Bayesian approach provides a natural framework to quantify and communicate uncertainty.

List of Figures

List of Tables

List of Algorithms

List of Abbreviations

EVT	**E**xtreme **V**alue **T**heory
GEV	**G**eneralized **E**xtreme **V**alue (distribution)
GP	**G**eneralized **P**areto (distribution)
MALA	**M**etropolis **a**djusted **L**angevin **a**lgorithm
MDA	**M**aximum **D**omain (of) **A**ttraction
MLE	**M**aximum **L**ikelihood **E**stimate
MSE	**M**ean **S**quared **E**rror
POT	**P**eaks **O**ver **T**hreshold (approach)
smMALA	**s**implified **m**anifold **M**etropolis **a**djusted **L**angevin **a**lgorithm

List of Symbols

Symbol	**Description**
Y	Random variable
y	Realization from Y
$\mathbf{Y}$	Random vector
$\boldsymbol{y}$	Realization from $\mathbf{Y}$
X	Covariate
x	Realization from X

Function	**Description**
$a(\cdot)$	Location normalizing function
$b(\cdot)$	Scale normalizing function
$f_Y(\cdot)$	Probability density function
$f_\Theta(\cdot)$	Prior density function
$f_{Y\|\Theta}(\cdot)$	Likelihood function
$f_{\Theta\|Y}(\cdot)$	Posterior density function
$F_Y(\cdot)$	Cumulative distribution function
$\tilde{F}_Y(\cdot)$	Empirical CDF
$\breve{F}_Y(\cdot)$	Empirical CDF with Pareto tail
$\bar{\ell}_{\text{HT}}(\cdot)$	Negative log-likelihood function Heffernan and Tawn model
$V(\cdot)$	Exponent measure
$T(\cdot)$	Probability integral transform

Distribution	**Name**
$\mathcal{U}_{[a,b]}$	Uniform distribution
$\mathcal{N}\left(\mu, \sigma^2\right)$	Gaussian distribution
$\mathcal{G}\left(a, b\right)$	Gamma distribution
$G_\xi\left(\xi, \sigma, \tau\right)$	Generalized Extreme Value distribution
$G_u\left(\xi, \sigma_u\right)$	Generalized Pareto distribution
$G_{\|i}\left(\cdot\right)$	Limit distribution in Heffernan and Tawn model
$\mathcal{C}\left(p_1, p_2\right)$	Copula function (bivariate)
$G(\cdot)$	Arbitrary distribution function
$H(\cdot)$	Spectral distribution function

Parameter	**Name**
ξ	Shape parameter
σ	Scale parameter GEV
τ	Location parameter GEV
σ^*	Modified scale parameter GP
α	Location parameter HT model
β	Scale parameter HT model
μ	Mean residual distribution HT model
ψ^2	Variance residual distribution HT model
ρ	Dependence parameter Case 1 and Case 2 data
ζ	Weight coefficient spline function
θ	Arbitrary parameter
$\boldsymbol{\theta}$	Set of arbitrary parameters
$\eta_\theta^{(\cdot)}$	Hyper-parameter Bayesian framework
Miscellaneous	**Description**
p	Non-exceedance proability
t_Y	Return period
u	Quantile which serves as threshold
y^F	Right end point of Y
$\mathcal{J}(\cdot)$	Observed Fisher information matrix
$\mathcal{I}(\cdot)$	Expected Fisher information matrix
$\hat{\chi}(\cdot)$	Measure for Extremal Dependence
$\hat{\bar{\chi}}(\cdot)$	Measure for Extreme Independence
$\hat{\theta}_{\text{MLE}}$	Maximum likelihood estimate
$\hat{\theta}_{\text{MAP}}$	Maximum a-posteriori estimate
Set	**Description**
A	Extreme set
$\Omega_{\boldsymbol{\theta}}$	Parameter space
$\mathbb{N}$	Set of integers
$\mathbb{R}$	Set of real numbers

Contents

Chapter 1

Introduction

Extreme events have a severe impact but a small probability of occurring. The unfathomable impact of catastrophes in the oil and gas industry, such as the Piper Alpha (1988) and Deep Water Horizon (2010) accidents, led to a high standard for risk management among offshore engineers. Design criteria for offshore structures should take extreme events into account as these events tend to govern structural failure mechanisms. In offshore applications, metocean[1] engineers are responsible for the design of offshore structures. Design criteria for oil rigs or floating structures are based on the *return levels* of different metocean quantities, such as *significant wave height*, *current speed* and *wind speed*. This thesis project is commissioned by Shell. A team of statisticians within Shell, led by Philip Jonathan, is continuously improving its multivariate extreme value models to come up with better estimates for design criteria.

When dependence between different metocean quantities is apparent, characterizing the extremes of an entire offshore environment requires a multivariate extreme value model. The aim is to estimate the probability that a multivariate random variable $\mathbf{Y}$ with cumulative distribution function $F_{\mathbf{Y}}$ is extreme, i.e. $\Pr(\mathbf{Y} \in A)$ for an *extreme set* A. First of all, this requires models for each of the marginal distributions. Additional complexity can be introduced by accounting for *temporal dependence* (Chavez-Demoulin and Davison 2012) and *spatial dependence* (Davison et al. 2012). A generic methodology to incorporate covariate effects in an univariate extreme value model was proposed by Randell et al. (2016). A second aspect of a multivariate extreme value model accounts for *extremal dependence*. This determines whether extreme events of different quantities have a tendency to occur together. Financial markets (Tawn et al. 2003), river networks (Davison et al. 2015a) and offshore environments (Johannessen et al. 2002) are practical examples where extremal dependence among

[1] Metocean: A contraction of meteorological and oceanographic.

extreme events is acknowledged. A visual representation of a multivariate extreme value model framework is shown in Figure 1.1.

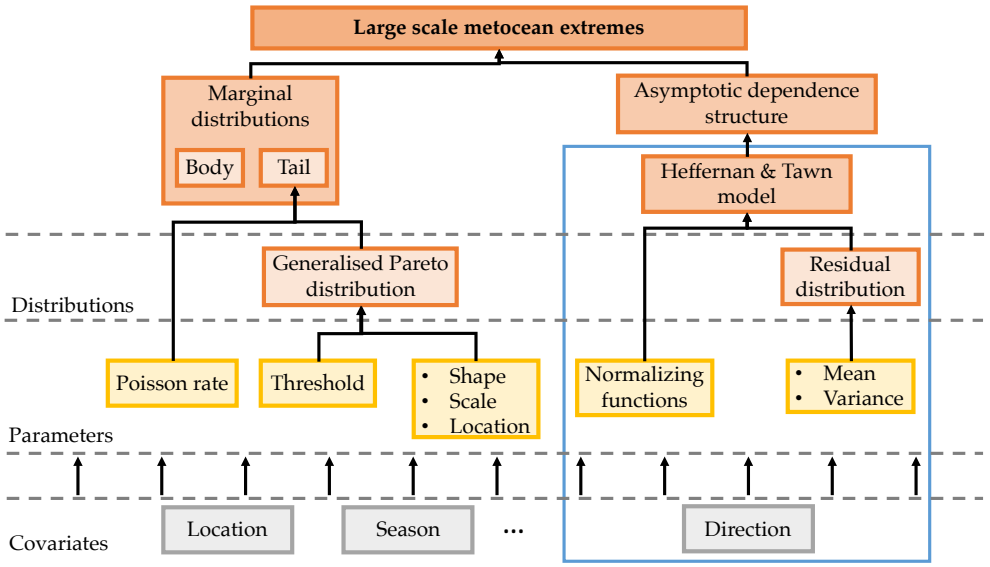


FIGURE 1.1: Structure of the *large scale metocean extremes* model. The scope of this thesis project is marked by the blue box.

The multivariate extreme value model proposed by Heffernan and Tawn (2004) is widely used among practitioners to characterize the joint tail of a distribution. Many results in extreme value theory rely on the assumption that data is *independent* and *identically* distributed. Jonathan et al. (2008) and Raghupathi et al. (2016) show that the identicality assumption in the presence of *covariate effects* leads to *biased* parameter estimates. The inclusion of additional information in a model, such that the response variable $\mathbf{Y}$ is explained by an explanatory variable or *covariate* $\mathbf{X}$, can resolve this issue. Seasonality, directionality or location are examples of covariates which are commonly used in extreme value analysis with metocean applications. The response of a floating structure to the forces exerted by the environment is referred to as *weathervaning*:

> *"Weathervaning is the process by which a floating structure passively varies its heading in response to time-varying environmental actions."*
>
> - ISO 19901-7:2013, *Petroleum and natural gas industries*

Accounting for weathervaning is particularly important as the angle of incidence of a particular metocean quantity relative to the vessel heading significantly affects the structural response. The scope of this project is restricted to the *direction* covariate $\mathbf{X} \in [-\pi, \pi]^d$ because it is regarded to be the dominant covariate related to *weathervaning*.

1.1 Problem statement

Past efforts to account for covariate effects in the extremal dependence structure rely on binning the data in directional sectors. For each sector, a constant Heffernan and Tawn model is fitted. Although this approach works in practice, it has several downsides. First of all, defining appropriate sectors is not obvious. Secondly, the assumption that a model is constant within a particular sector can lead to biased parameter estimates. Thirdly, by construction, the parameters will be step functions with respect to the covariates, while it would be more natural to consider smooth functions.

To address these issues, Jonathan et al. (2014) propose a spline parameterization to generalize the Heffernan and Tawn model in order to accomodate covariates. Statistical inference for the model is a tedious job that requires cross validating the entire model to obtain an optimal smooth model. In addition, it is very time consuming to quantify uncertainty regarding the parameter estimates, as the proposed methodology relies on re-sampling from the data and refitting the model for each new sample. These issues motivate the development of a new methodology for statistical inference for the generalized Heffernan and Tawn model proposed by Jonathan et al. (2014).

The research question for this project is:

How to estimate the parameters of the generalized Heffernan and Tawn model, such that the methodology is both robust and easy to fit?

To narrow the scope of the project the multivariate extreme value model proposed by Heffernan and Tawn (2004) — which will be referred to as the Heffernan and Tawn model — is adopted. In addition, a directional covariate is considered. Without going into too much detail, a glimpse of the statistical challenge is revealed. The model proposed by Heffernan and Tawn (2004) is a semi-parametric regression model that regresses on a *conditioning variable* $T_{\mathrm{L}}(Y_i)$ being extreme, where T_{L} denotes a marginal transformation to the Laplace scale. Estimating $\Pr\{T_{\mathrm{L}}(\mathbf{Y}) \in A \mid \mathbf{X} = \mathbf{x}\}$ under the Heffernan and Tawn model when the parameters of the model are assumed to be a smooth function of a covariate $\mathbf{X}$, requires fitting, for each $i = 1, \ldots, d$:

$$T_{\mathrm{L}}(\mathbf{Y}_{-i}) \mid T_{\mathrm{L}}(Y_i) = y, \mathbf{X} = \mathbf{x} \sim \mathcal{N}\left\{\boldsymbol{\alpha}_{|i}(\mathbf{x})\, y + y^{\boldsymbol{\beta}_{|i}(\mathbf{x})} \boldsymbol{\mu}_{|i}(\mathbf{x}), y^{2\boldsymbol{\beta}_{|i}(\mathbf{x})} \boldsymbol{\psi}^2_{|i}(\mathbf{x})\right\}.$$

Constraints that ensure a stochastic ordering of the conditional quantiles under the Heffernan and Tawn model have been proposed by Keef et al. (2013). These constraints in conjunction with both the constant– and generalized Heffernan and Tawn model, define the four different models shown in Figure 1.2.

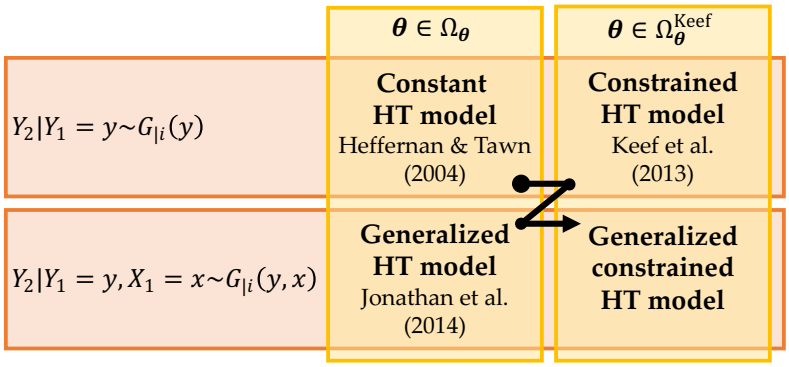


FIGURE 1.2: Roadmap to a generic and mathematically sound multivariate extreme value model.

1.2 Research contributions

The knowledge gap that is identified is shown in Figure 1.3. The methodology presented in this thesis addresses the knowledge gap. The developed Matlab routines can be implemented by Shell into the all encompassing large scale metocean extremes model, shown in Figure 1.1.

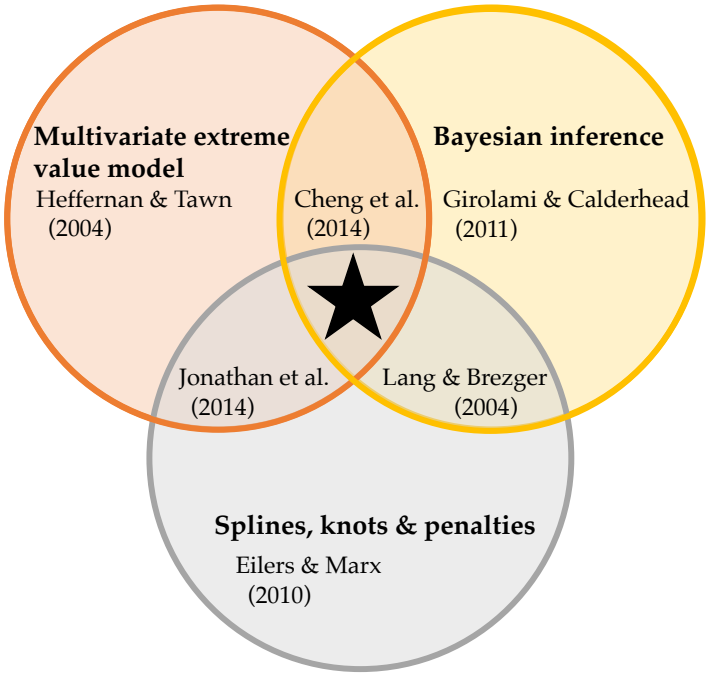


FIGURE 1.3: Indication of the knowledge gap that this thesis project addresses (★), in relationship to the relevant mathematical concepts and key references.

The main contributions of this project are:

1. **Study the finite sample properties of the maximum likelihood estimator the Heffernan and Tawn model.**

The influence of threshold selection and sample size are two key sources of uncertainty in extreme value theory, both of which remain unaddressed in the literature. The results of a simulation study to asses the sensitivity of parameter estimates with respect to sample size, threshold selection and strength of dependence in the simulated data, are presented to provide a better understanding of the likelihood function of the Heffernan and Tawn model.

2. **Present a Bayesian inference framework to estimate the parameters of the Heffernan and Tawn model.**
   Quantifying uncertainty — in particular for the generalized Heffernan and Tawn model proposed by Jonathan et al. (2014) — is not trivial. This calls for a Bayesian inference framework. The Bayesian model proposed by Cheng et al. (2014) is unsatisfactory as it requires strong prior information and requires an undesirable adjustment to the likelihood function. A full Bayesian inference framework that addresses both issues is presented to jointly estimate the parameters of the Heffernan and Tawn model. A logistic transformation of the parameters of the Heffernan and Tawn model is proposed to accommodate asymptotically dependent data.

3. **Extend the proposed Bayesian inference framework such that covariate effects can be accounted for.**
   Under the assumption that the parameters of the Heffernan and Tawn model are smooth functions of covariates, penalized basis splines can accommodate covariate effects in the parameterization of the Heffernan and Tawn model. Bayesian inference on the semi-parametric P-splines model follows Lang and Brezger (2004).

1.3 Thesis outline

A brief introduction to extreme value theory is presented in Chapter 2. Univariate extreme value models are introduced in Section 2.1, and multivariate models are introduced in Section 2.2. The aim of this chapter is to introduce the mathematical context of this thesis. Chapter 3 is dedicated to the model proposed by Heffernan and Tawn (2004). A description of the Heffernan and Tawn model is presented in Section 3.1. Issues regarding statistical inference for the model when the sample size is finite, are addressed in Section 3.2.

The answer to the research question is presented in Chapter 4. The most significant academic contribution of this project is the implementation of a Bayesian inference framework for the generalized Heffernan and Tawn model proposed by Jonathan et al. (2014). Bayesian statistics is briefly introduced in

Section 4.1. The proposed framework is demonstrated for the cases shown in Figure 1.2:

1. Constant Heffernan and Tawn model, see Section 4.2,
2. Constrained Heffernan and Tawn model, see Section 4.3,
3. Generalized Heffernan and Tawn model, see Section 4.4.

Inference on the generalized constrained Heffernan and Tawn model is omitted because of unresolved challenges related to inference on the aforementioned models. The simulation study considers both asymptotic independent– and asymptotic dependent data.

Finally, concluding remarks, a brief discussion on the methodology and several recommendations are presented in Chapter 5. The outline of this thesis is summarized in Figure 1.4. The literature review in Chapter 2 and Section 3.1 solely relies on the work of others. Although the constant Heffernan and Tawn model (Heffernan and Tawn 2004), constrained Heffernan and Tawn model (Keef et al. 2013) and generalized Heffernan and Tawn model (Jonathan et al. 2014) have been introduced by others, the results presented in Section 3.2 and Chapter 4 are new and reflect my own work.

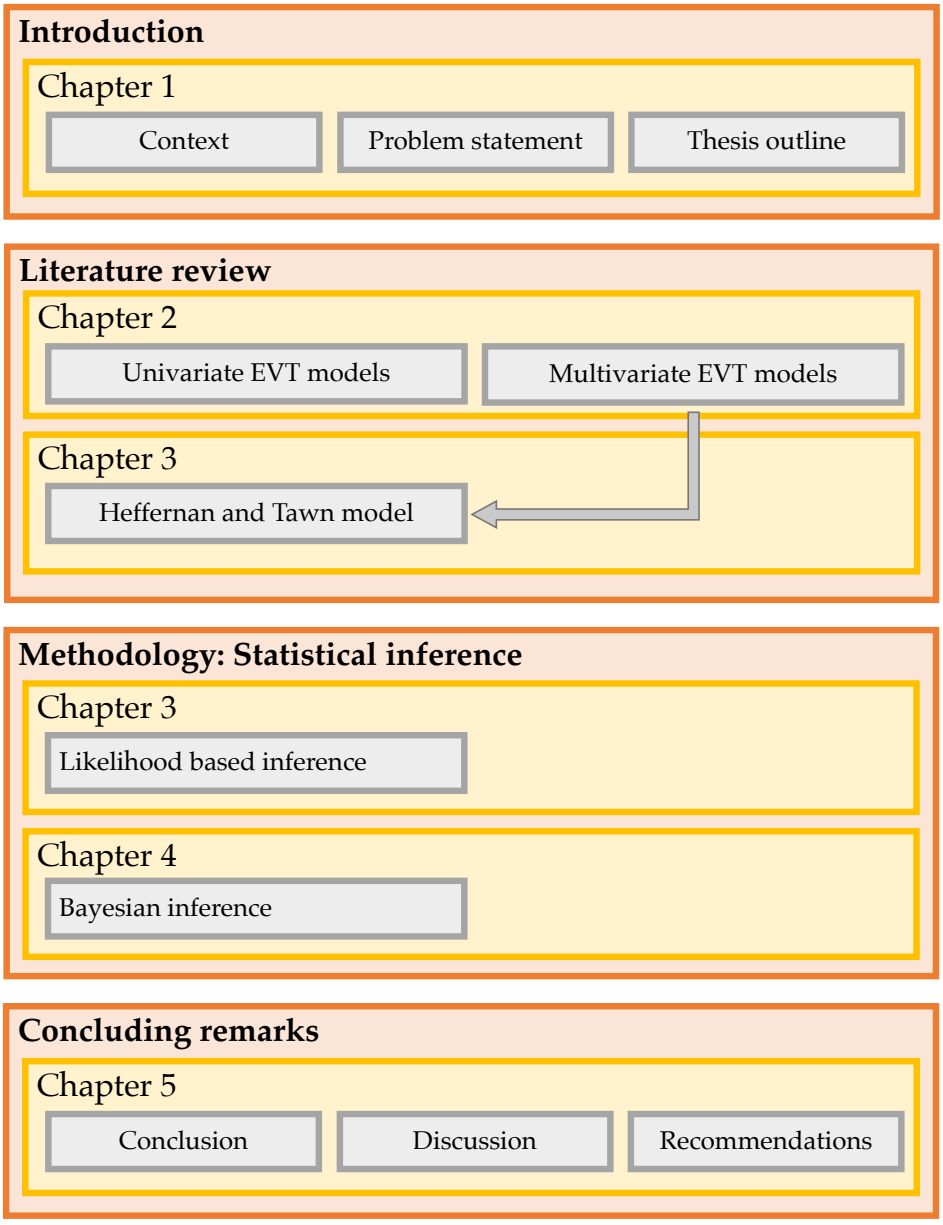


FIGURE 1.4: Outline of this thesis report.

Chapter 2

Extreme Value Theory: an Introduction

> As far as the laws of mathematics refer to reality, they are not certain, and as far as they are certain, they do not refer to reality.
>
> — Albert Einstein

Extreme events have a small probability of occurring, but their impact is an order of magnitude larger than what is typically observed. The combination of spring tide and extreme wind gusts that led to the 1953 flooding of the Netherlands is an example of the severe consequences of extreme *metocean* events and motivates why thorough understanding of the distribution of extreme events is a necessity.

Over the last couple of decades, *extreme value theory* (EVT) has received a lot of attention from both academia and practitioners. Applications of extreme value statistics range from environmental catastrophe modeling to financial stress testing. Regardless the application, the central challenge usually concerns the estimation of probabilities associated to events far worse than anything that has ever been recorded, with an associated *return period*[1] that is much larger than the time span for which data is available, i.e. estimate the 100-year return level when only 10 years worth of data is available.

Extrapolation beyond the observed data is a non-trivial task. There are four key-issues related to the statistical analysis of extremes, which will we addressed in this thesis. These issues being:

1. Parameter estimation for extreme value models,

[1] See Section 2.1.1 for a formal definition.

2. Quantifying and communicating uncertainty regarding parameter estimates or return levels,
3. Model diagnostics and goodness-of-fit, and,
4. How to use the available data optimally, such that model uncertainty can be reduced.

This chapter provides an introduction to extreme value theory. The goal of this chapter is to provide the reader with the relevant concepts from extreme value theory, on which the subsequent chapters are based. Univariate extreme value models are introduced in Section 2.1. Section 2.2 provides an overview of the higher dimensional extensions of the univariate paradigms introduced in Section 2.1.

Relevant definitions and theorems are provided throughout this chapter, but full proofs are omitted. The works by Galambos (1978), Resnick (1987), Beirlant et al. (2004), De Haan and Ferreira (2006), and Reiss and Thomas (2007) provide full coverage of the fundamentals of extreme value theory and the proofs of the theorems presented in this chapter. The book by Coles (2001) is particularly accessible for inexperienced readers. Unless specifically stated otherwise, the introduction to the main concepts of extreme value theory presented in this chapter follow Coles (2001) and De Haan and Ferreira (2006).

The outline of this chapter is summarized in Figure 2.1.

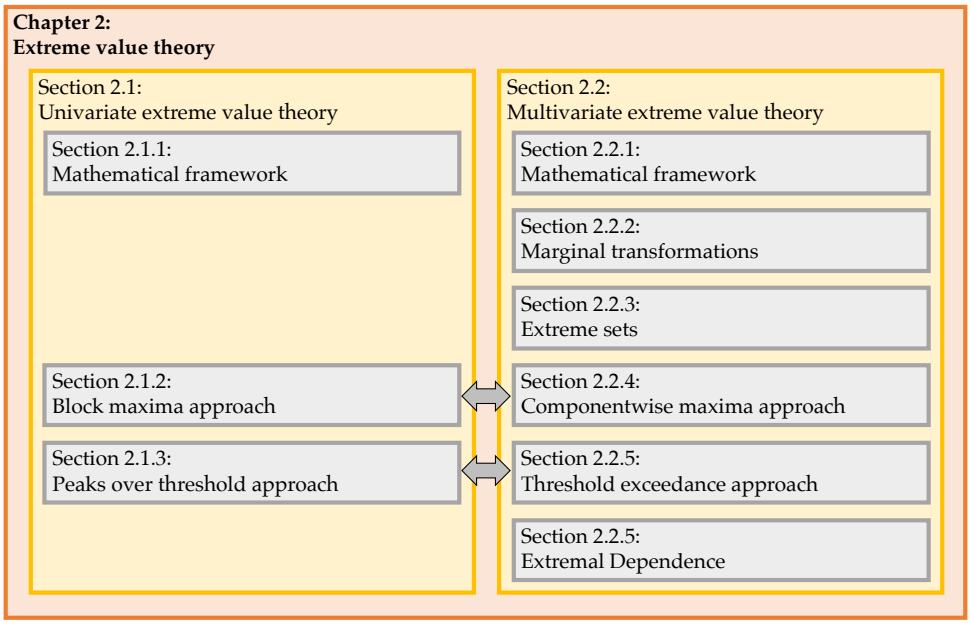


FIGURE 2.1: Outline of Chapter 2.

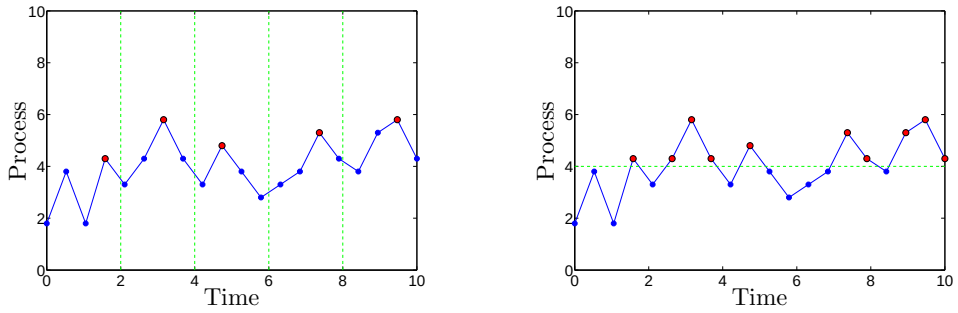
2.1 Univariate extreme value theory

Extreme events have a small probability of occurring. These events stand out from the bulk of the observations, because they are either much smaller or much larger in magnitude than what is typically observed. Attention is restricted to maxima as minima of metocean extremes do not affect design criteria. The theorems presented in in this chapter apply equally well to minima.

There is no universal rule to characterize extreme events, but two different paradigms are commonly used in practice to define extreme events:

1. Block Maxima approach, and,
2. Peaks Over Threshold approach.

See Figure 2.2 for an intuitive definition of both approaches. The red dots in Figure 2.2(A) form the set of block maxima, where the different blocks are separated by the green lines. On the other hand, the red dots in Figure 2.2(B) are the set of threshold exceedances, where the green line defines the threshold. Both paradigms provide a rich framework for characterizing extremes. It turns out that both approaches have a strong connection, although they seem very different at first sight. Section 2.1.2 and 2.1.3 provide a formal mathematical introduction to the Block Maxima approach and Peaks Over Threshold approach respectively, and serve as a proper starting point for the more advanced concepts introduced in subsequent chapters.



(A) Block maxima approach (B) Peaks over threshold approach

FIGURE 2.2: Two different approaches to define extreme events for an arbitrary sequence of observations.

2.1.1 Mathematical framework

Let Y be a random variable with cumulative distribution function F_Y. Unless specifically stated otherwise, throughout this thesis, the random variable Y is assumed to be continuous. Continuity of Y guarantees that the probability density function f_Y exists and is defined for all *continuity points* of F_Y.

A formal definition of the random variable Y states that $Y\colon \Omega \to E$ is a $(\mathcal{F}, \mathcal{E})$-measurable function for a generic probability space $(\Omega, \mathcal{F}, \Pr)$ and measurable space $(E, \mathcal{E})$. Within the context of this thesis, $E = \mathbb{R}$ the set of all real numbers and $\mathcal{E}$ is the Borel σ-algebra of E, which is a trivial framework for many probabilistic and statistical applications.

A sequence $Y_1, \ldots, Y_n$ is denoted by $\{Y_l\}_{1 \leq l \leq n}$, where $n \in \mathbb{N}$ denotes the sample size. Throughout this report, the index l will always refer to the index of a particular element of a sample. For the time being, assume that the sequence of random variables $Y_1, \ldots, Y_n$ are *identically distributed*, such that Y_l has the same cumulative distribution function F_Y for each $l = 1, \ldots, n$. Furthermore, assume that the random variables $Y_1, \ldots, Y_n$ are *independent*.

Define the *right endpoint* of F_Y by $y^{\mathrm{F}} \coloneqq \sup\{y\colon F_Y(y) < 1\}$. Then

$$\max(Y_1, \ldots, Y_n) \xrightarrow{p} y^{\mathrm{F}}, \quad \text{as} \quad n \to \infty.$$

The *left-continuous inverse* of the cumulative distribution function F_Y is defined by $F_Y^{\leftarrow}(y) \coloneqq \inf\{t\colon F_Y(t) \geq y\}$. The *quantile* function U associated to the *return period* t_Y is defined by

$$U(t_Y) \coloneqq F_Y^{\leftarrow}\left(1 - \frac{1}{t_Y}\right), \quad \text{for} \quad t_Y > 1.$$

Define the *non-exceedance probability* for the return period t_Y by $p \coloneqq \Pr(Y \leq y) = 1 - 1/t_Y$.

Assuming that the data are identically distributed is a common starting point in order to constrain a problem and obtain practical results. The concept of *weakly identically distributed* data is introduced to provide a formal framework to characterize random variables when covariate effects are apparent. A random variable is said to exhibit *covariate effects* if the associated probability distribution is a function of explanatory variables or covariates. Observations from random variables that are weakly identically distributed can still be assumed to admit the same parametric probability distribution, but the parameterization of the probability density function is no longer assumed to be constant. Under the assumption that the model parameters are smooth functions of the covariate X, additional information can be incorporated in the model. This resolves the issue of biased parameter estimates when covariate effects are wrongly neglected, as pointed out by Jonathan et al. (2008) and Raghupathi et al. (2016).

Definition 2.1.1 (Weakly identically distributed)**.**
Let X denote a covariate, and Y a random response variable. Let a sequence $Y_1, \ldots, Y_n$ of random variables with distribution functions $F_1, \ldots, F_n$ be parameterized by $\boldsymbol{\theta}(X_1), \ldots, \boldsymbol{\theta}(X_n)$,

where $\boldsymbol{\theta}$ is a smooth function of a covariate X. A random variable is said to be weakly identically distributed *if for any $x \in \Omega_X$, and $i, j \in \{1, \ldots, n\}$, $F_i \{y_i \mid \boldsymbol{\theta}(x)\} = F_j \{y_j \mid \boldsymbol{\theta}(x)\}$ holds.*

2.1.2 Block maxima approach

A natural starting point for this introduction to extreme value theory, is to consider the paradigm proposed by Fisher and Tippett (1928). The mathematical framework introduced in the previous section serves as a starting point. Let M_n denote the *partial maximum* of a sequence of random variables, i.e. $M_n = \max\{Y_1, \ldots, Y_n\}$, where M_n is referred to as the block maximum. Under the assumption that $Y_1, \ldots, Y_n$ are independent, the limiting distribution of M_n is degenerate, since

$$\Pr(M_n \leq y) = \Pr(Y_1 \leq y, \ldots, Y_n \leq y) = F_Y^n(y), \tag{2.1}$$

which converges to either 0 or 1 as n tends to infinity, for $y < y^{\mathrm{F}}$ and $y \geq y^{\mathrm{F}}$ respectively. Luckily, studying maxima does not stop here, because similar to the the central limit theorem which states that normalizing partial sums yields the non-degenerate Gaussian distribution in the limit, normalizing maxima turns out to resolve the degeneracy issue as well.

Assume that there exist proper normalizing constants $a_n \in \mathbb{R}$ and $b_n > 0$ for $n \in \mathbb{N}$, referred to as the *location normalizing constant* and the *scale normalizing constant* respectively. Suppose that $(M_n - a_n)/b_n$ has a limit distribution G. That means $\Pr\{(M_n - a_n)/b_n \leq y\} \approx G(y)$ for large n and similarly $\Pr(M_n \leq y) \approx G\{(y - a_n)/b_n\}$. Now, for any $k \in \mathbb{N}$, the independence assumption yields

$$\Pr(M_{n \cdot k} \leq y) = \Pr(M_n \leq y)^k \approx G^k\left(\frac{y - a_n}{b_n}\right) \approx G\left(\frac{y - a_{n \cdot k}}{b_{n \cdot k}}\right),$$

which shows that $G^k(\cdot) \approx G(\cdot)$ up to re-scaling y by proper normalizing constants.

The limit distribution function G is said to be *max-stable*, if $G^k(x) = G(b_k x + a_k)$ for $k \in \mathbb{N}$ and normalizing constants $a_k \in \mathbb{R}$ and $b_k > 0$. It can be shown that if the limit distribution G of the normalized partial maxima $(M_n - a_n)/b_n$ exists, G must be max-stable. In addition, distribution functions G and G^* are of the *same type* if there exist constants $a \in \mathbb{R}$ and $b > 0$ such that $G^*(by + a) = G(y)$ for all $y \in \mathbb{R}$.

For proper normalizing constants, the normalized partial maxima $(M_n - a_n)/b_n$ have a non-degenerate limit distribution G, i.e.

$$\lim_{n\to\infty} F_Y^n(b_n y + a_n) = G(y). \tag{2.2}$$

If (2.2) holds, the cumulative distribution function F_Y is said to be in the *domain of attraction* of G, which is denoted by $F_Y \in \mathcal{D}(G)$. The limit distribution G encompasses a broad family of distributions which is referred to as the family of *extreme value distributions*, a proposed by Fréchet (1927). Fisher and Tippett (1928) and Gnedenko (1943) formalized these concepts and are accredited for the Extremal Type Theorem, which provides a parametric expression for the limit distribution G, see Theorem 2.1.2. Von Mises (1936) and Jenkinson (1955) proposed improvements regarding the parameterization of the extreme value distribution.

Theorem 2.1.2 (Extremal types theorem (Fisher and Tippett 1928; Gnedenko 1943))**.**
If there exist sequences of constants $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$ such that $a_n \in \mathbb{R}$ $b_n > 0$ for all $n \in \mathbb{N}$, and as $n \to \infty$,

$$\Pr\left(\frac{M_n - a_n}{b_n} \leq y\right) \xrightarrow{d} G(y) \tag{2.3}$$

for some non-degenerate limit distribution G, then G is said to be of the same type as the generalized extreme value distribution G_ξ, defined by

$$G_\xi(y) = \exp\left\{-\left(1 + \xi\frac{y-\tau}{\sigma}\right)_+^{-1/\xi}\right\}, \tag{2.4}$$

for some $\tau, \xi \in \mathbb{R}$ and $\sigma > 0$, defined for all x in the set $\{x\colon 1 + \xi(x-\tau)/\sigma > 0\}$. The parameters ξ, τ, σ are referred to as the shape–, location– and scale parameter respectively. The convention $(\cdot)_+ = \max\{\cdot, 0\}$ is used. The special case where $\xi = 0$, is interpreted as the limit of $\xi \to 0$.

Conversely, each of these distributions G may appear as the limit for the distribution of $(M_n - a_n)/b_n$ and does so when G itself is the distribution of Y.

A strong implication of Theorem 2.1.2 is that a distribution belongs to the family of extreme value distributions if and only if the distribution is max-stable (Coles 2001).

Although Theorem 2.1.2 yields a very rich framework, it has some limitations. First of all, not every sequence of properly normalized partial maxima yields a limit distribution G that exists. Consider for example the partial maxima of a sequence of Poisson distributed random variables. Secondly, as shown in (2.2), the speed of convergence of M_n to the limiting distribution depends on

the cumulative distribution function F_Y. For certain distributions, such as the Gaussian distribution, convergence is slow. This raises some concerns about the validity of Theorem 2.1.2 for small samples.

Three sub-classes of the family of extreme value distributions arise naturally by considering $\xi > 0, \xi = 0$ and $\xi < 0$ in (2.4), and are given by (2.5), (2.6) and (2.7). Properties and examples for each of the three different classes are provided below.

- $\xi > 0$: **Fréchet class of distributions**.
  For $\xi > 0$, it follows that $G_\xi(y) < 1$ for all $y \in \mathbb{R}$, and hence the right endpoint y^{F} of the distribution is infinity. In addition, the distribution has a heavy right tail that admits a power law, since, for $y \to \infty$, it follows that $1 - G_\xi(y) \sim (\xi y)^{-1/\xi}$. Consequently, the moments of the distribution of order greater than $1/\xi$ do not exist. For example the Cauchy and Pareto distribution are both in the Fréchet class.

$$G_\xi(y) \coloneqq \begin{cases} 0 & \text{if} \quad y \leq \tau - \sigma/\xi, \\ \exp\left\{-\left(\frac{y-\tau}{\sigma}\right)^{-1/\xi}\right\} & \text{if} \quad y > \tau - \sigma/\xi. \end{cases} \tag{2.5}$$

- $\xi = 0$: **Gumbel class of distributions**.
  Distributions in the Gumbel class are regarded to be light-tailed, since $1 - G_\xi(y) \sim \exp(-y)$ as $y \to \infty$ and all moments exist. The right endpoint y^{F} of the distribution is either finite or infinite. For example the Gaussian and Gamma distribution are both in the Gumbel class.

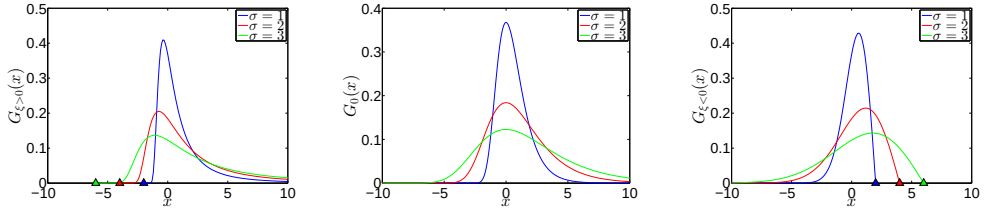
$$G_\xi(y) \coloneqq \exp\left\{-\exp\left(-\frac{y-\tau}{\sigma}\right)\right\}, \quad \forall y \in \mathbb{R} \tag{2.6}$$

- $\xi < 0$: **reverse-Weibull class of distributions**.
  The right endpoint is given by $y^{\mathrm{F}} = -1/\xi$. Since y^{F} is finite, the distribution is said to be short-tailed. For example the Uniform and Beta distribution are both in the reverse-Weibull class.

$$G_\xi(y) \coloneqq \begin{cases} \exp\left\{-\left(-\frac{y-\tau}{\sigma}\right)^{-1/\xi}\right\} & \text{if} \quad y < \tau - \sigma/\xi, \\ 1 & \text{if} \quad y \geq \tau - \sigma/\xi. \end{cases} \tag{2.7}$$

These three classes are also referred to as the Fréchet–, Gumbel– and reverse-Weibull– domain of attraction. For example, $F_Y \in \mathcal{D}(G_{\xi>0})$ means that for the normalized maxima $(M_n - a_n)/b_n$, the limit distribution G in (2.2) is the Fréchet distribution. Figure 2.3 shows the probability densities for each of the three different classes of the generalized extreme value distributions, for $\tau = 0$. The edge of the support of the Fréchet and reverse-Weibull distribution

are marked by small triangles, and decrease and increase respectively, as σ increases. A sufficient, but not necessary, condition for F_Y to belong to either one



(A) Fréchet (B) Gumbel (C) reverse-Weibull

FIGURE 2.3: Probability densities for the three different classes within the family of generalized extreme value distribution.

of the three maximum domains of attraction was proved by Von Mises (1936). The condition is commonly referred to as the *von Mises condition*, see Theorem 2.1.3.

Theorem 2.1.3 (von Mises condition (Von Mises 1936))**.**
Let F_Y be a distribution function and y^{F} its right endpoint. Suppose the second derivative $F_Y''(y)$ exists and the first derivative $F_Y'(y)$ is positive for all y in some left neighborhood of y^{F}. If

$$\lim_{y \uparrow y^{\mathrm{F}}} \left(\frac{1 - F_Y}{F_Y'} \right)' (y) = \lim_{y \uparrow y^{\mathrm{F}}} r'(y) = \xi, \tag{2.8}$$

or equivalently,

$$\lim_{y \uparrow y^{\mathrm{F}}} \frac{\{1 - F_Y(y)\} F_Y''(y)}{\{F_Y'(y)\}^2} = -\xi - 1, \tag{2.9}$$

then F_Y is in the maximum domain of attraction of the generalized extreme value distribution.

where $r(y) := \{1 - F_Y(y)\} / f_Y(y)$ in (2.8) is referred to as the *reciprocal hazard function*. Applying (2.2) to Theorem 2.1.3, for a sufficiently smooth cumulative distribution function F_Y, with right endpoint y^{F}, for

$$a_n = F^{-1}(1 - 1/n), \quad b_n = r(b_n) \quad \text{and} \quad \xi = \lim_{y \to y^{\mathrm{F}}} r'(x),$$

results in the limit distribution of $(M_n - a_n)/b_n$ being the generalized extreme value (GEV) distribution with shape parameter ξ.

The Von Mises condition stated in Theorem 2.1.3 provides explicit expressions for a_n, b_n and ξ if the cumulative distribution F_Y is known, see Table 2.1 for some examples. Since the cumulative distribution F_Y is generally unknown in practice, the ramifications of Theorem 2.1.3 are limited to theoretical applications. A stronger result than the Von Mises condition is given by Theorem 2.1.4, which provides a sufficient and necessary condition for the cumulative

TABLE 2.1: Explicit expressions for the normalizing constants a_n, b_n and shape parameter ξ, for different probability distributions.

Distribution	a_n	b_n	ξ	$\mathcal{D}(\cdot)$
Fréchet	n	n	1	Fréchet
Exponential	$\log(n)$	1	0	Gumbel
Gaussian	§	$\frac{1}{\sqrt{2\log(n)}}$	0	Gumbel
Uniform	$1-1/n$	$1/n$	-1	rev.-Weibull

§: $\sqrt{2\log(n)} - \frac{1}{2}\frac{1}{\sqrt{2\log(n)}}\left[\log\{\log(n)\} + \log(4\pi)\right]$

distribution function F_Y to belong to the domain of attraction of the generalized extreme value distribution.

Theorem 2.1.4 (De Haan and Ferreira (2006))**.**
The distribution function F_Y is in the domain of attraction of the extreme value distribution G_ξ if and only if

1. *for $\xi > 0$:*

$$\lim_{t\to\infty} \frac{1 - F_Y(ty)}{1 - F_Y(t)} = y^{-1/\xi}, \quad \forall y > 0 \quad \textit{and} \quad y^{\mathrm{F}} = \infty.$$

2. *for $\xi = 0$:*

$$\lim_{t\uparrow y^{\mathrm{F}}} \frac{1 - F_Y\{t + y f_Y(t)\}}{1 - F_Y(t)} = \mathrm{e}^{-y}, \quad \forall y \in \mathbb{R}, \tag{2.10}$$

for a suitable positive function f_Y. If (2.10) *holds for some function f_Y, then*

$$\int_t^{y^{\mathrm{F}}} 1 - F_Y(s)\mathrm{d}s < \infty \quad \textit{for} \quad t < y^{\mathrm{F}} \quad \textit{and} \quad f_Y(t) \coloneqq \frac{\int_t^{y^{\mathrm{F}}} 1 - F_Y(s)\,\mathrm{d}s}{1 - F_Y(t)}.$$

3. *for $\xi < 0$:*

$$\lim_{t\to\infty} \frac{1 - F_Y(y^{\mathrm{F}} - ty)}{1 - F_Y(y^{\mathrm{F}} - t)} = y^{-1/\xi}, \quad \forall y > 0, \quad \textit{and} \quad y^{\mathrm{F}} < \infty.$$

Relaxing the identicality assumption affects the limiting result (2.3) in Theorem 2.1.2. Very little has been published on the subject of non-identically distributed extremes. A result by Mejzler (1956) shows how convexity of the limit distribution G in (2.3) holds even when the identicality assumption in Theorem 2.1.2 is relaxed.

Theorem 2.1.5 (Mejzler (1956))**.**
Suppose $Y_1, \ldots, Y_n$ are independent random variables with distribution functions $F_1, \ldots, F_n$ respectively. Suppose there exist sequences $a_n \in \mathbb{R}$ and $b_n > 0$ for $n \in \mathbb{N}$ such that the

normalized partial maxima $(M_n - a_n)/b_n$ *have a non-degenerate limit distribution, which we call* G*. Suppose that as* $n \to \infty$*,*

$$|\log a_n| + |b_n| \to \infty$$

and both

$$\frac{b_{n+1}}{b_n} \to 1, \quad \textit{and} \quad \frac{a_{n+1} - a_n}{b_n} \to 0.$$

Then

$$-\log G(y) \text{ is convex if } y^{\mathrm{F}} = \infty, \tag{2.11}$$

and

$$\log G\left\{y^{\mathrm{F}} - \exp(-y)\right\} \quad \text{is convex if} \quad y^{\mathrm{F}} < \infty, \tag{2.12}$$

Conversely, any distribution function G *satisfying* (2.11) *and* (2.12) *occurs as a limit in the given set-up.*

As Kourbatov (2014) points out, "Mejzler's theorem states that the limiting distribution of properly normalized non-identically distributed independent random variables, if the limiting distribution exists at all, can be any distribution with a log-concave cumulative distribution function". See Bagnoli and Bergstrom (1989) for an overview of log-concave cumulative distributions functions and Dümbgen and Rufibach (2009) for an elaborate study on the properties of log-concave densities.

2.1.3 Peaks over threshold approach

The Peaks Over Threshold approach offers an alternative methodology to modeling extremes. As the name of this approach suggests, rather than looking at the block maxima, this approach focuses on the stochastic behavior of exceedances of a high threshold u. Theorem 2.1.6 provides the connection between the concepts introduced in the previous section and the Peaks Over Threshold approach.

Theorem 2.1.6 (Balkema and De Haan (1974) and Pickands (1975))**.**
For $\xi \in \mathbb{R}$ *the following statements are equivalent:*

1. *There exist normalizing constants* $a_n \in \mathbb{R}$ *and* $b_n > 0$ *such that*

$$\lim_{n\to\infty} F_Y^n(b_n y + a_n) = G_\xi(y),$$

with $G_\xi(y)$ *defined as in Theorem 2.1.2.*

2. *There is a positive function f such that*

$$\lim_{u \uparrow y^{\mathrm{F}}} \frac{1 - F_Y\{u + yf(u)\}}{1 - F_Y(u)} = \left(1 + \xi \frac{y}{\sigma_u}\right)_+^{-1/\xi} \tag{2.13}$$

for all $y > 0$ in the set $\{y \colon 1 + \xi(y - \tau)/\sigma > 0\}$ and $\sigma_u \coloneqq \sigma + \xi(u - \tau) > 0$.

For a random variable Y with cumulative distribution function $F_Y \in \mathcal{D}(G_\xi)$, the distribution of the independent threshold exceedances $Y_1 - u, \ldots, Y_n - u$ of Y for a high threshold u, conditional on Y exceeding u, is given by (2.14). This distribution is referred to as the generalized Pareto (GP) distribution, and is given by

$$G_u(y) \coloneqq \lim_{u \uparrow y^{\mathrm{F}}} \Pr\left(\frac{Y - u}{f(u)} \leq y \,\middle|\, Y > u\right) = 1 - \left(1 + \xi \frac{y}{\sigma_u}\right)_+^{-1/\xi}, \quad \text{for} \quad y > 0, \tag{2.14}$$

Beware that G_u refers to the generalized Pareto distribution, while G_ξ refers to the generalized extreme value distribution.

The limit distribution $G_u(y)$ does not fully characterize the stochastic behavior of the threshold exceedances. The arrival times of threshold exceedances is itself a random process. Modeling the exceedances as observations from a point process perspective, such that the distribution of each observations is G_u, has advantages compared to the block maxima approach introduced in the previous section. The most apparent advantages is that this model utilizes the information regarding all threshold exceedances, rather than just the block maxima.

Using a representation proposed by Rényi (1953), the homogeneous Poisson process arises as the limit of the distribution of the *extreme-order statistics*, i.e. the sequence of observations ordered in descending order. Let the *exceedance probability*

$$\bar{p} = \left(1 + \xi \frac{u - \tau}{\sigma}\right)^{-1/\xi} = \left(\frac{\sigma_u}{\sigma}\right)^{-1/\xi},$$

define the rate of the Poisson process. Homogeneity of the Poisson process is guaranteed because $\bar{p}$ is constant for given parameters ξ, τ and σ. There is a strong relationship between the generalized extreme value distribution and the point process representation introduced in this section, see Derivation 2.1.7.

Derivation 2.1.7 (The link between the generalized extreme value distribution and the generalized Pareto distribution.)**.**

Let N be a Poisson random variable with rate parameter $\bar{p}$, and let $Y_1, \ldots, Y_N$ be generalized Pareto distributed. The generalized extreme value distribution arises as

the distribution of normalized partial maxima $M_N \coloneqq \max\{Y_1, \ldots, Y_N\}$*, since*

$$\begin{aligned}\Pr(M_N \leq y) &= \sum_{n=1}^{\infty} \Pr(N = n) \cdot \Pr(Y_1 \leq y, \ldots, Y_n \leq y), \\ &= \exp\left\{ -\left(1 + \xi \frac{y + u - \tau}{\sigma}\right)_+^{-1/\xi} \right\}, \\ &= G_\xi(y + u)\end{aligned}$$

See Appendix A.1 for the intermediate steps of this derivation.

Intuitively, the high threshold u marks the transition from the body to the tail of the distribution. The question remains how to properly choose the threshold u. Determining what an appropriate choice for u is, is not straightforward, but a combination of heuristics provide a satisfactory framework for threshold selection. Two common heuristics for threshold selection focus on finding a quantile u associated to the non-exceedance probability p, such that:

1. The parameters of the generalized Pareto distribution show stability, and,
2. The *mean residual life* is linear in u with slope $\xi/(1-\xi)$,

where the mean residual life of Y is defined as

$$\mathrm{E}[Y - u \mid Y > u] = \frac{\sigma + \xi u}{1 - \xi}, \quad \text{for} \quad \xi < 1. \tag{2.15}$$

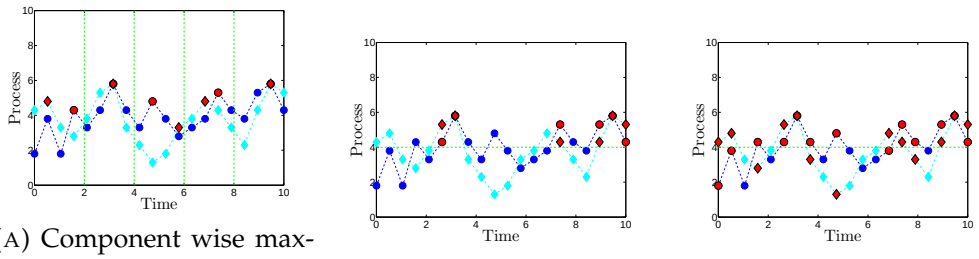
2.2 Multivariate extreme value theory

Studying the interaction between different random variables might reveal features of the data that would remain concealed if the variables are modeled independently. Characterizing dependence is particularly important in systems with an exposure to high dimensional random phenomena, such as a portfolio of financial assets (Tawn et al. 2003), or a complex river network (Davison et al. 2015a). Univariate extreme value models are unsatisfactory when the data exhibits *extremal dependence*, which means that extreme observations in different variables tend to occur simultaneously. See Section 2.2.6 for a formal introduction to the concept of extremal dependence.

The concepts presented in Section 2.2.1 aim at providing a proper mathematical framework for the univariate paradigms introduced in Section 2.1 to be extend to higher dimensions. The extensions of the block maxima approach and the threshold exceedance approach to higher dimensions are introduced in Section 2.2.4 and 2.2.5 respectively. An intuitive definition of both approaches is presented in Figure 2.4.

There are two additional characterizations of multivariate extremes. A third approach relies on transforming the data to pseudo-polar coordinates and modeling the resulting data as a point process, which is briefly introduced in 2.2.5 and viewed here as a special case of the threshold exceedance approach.

Heffernan and Tawn (2004) proposed a fourth approach which plays a central role in this thesis project. The idea evolves around assuming a semi-parametric model for different conditional probability distributions, that together define the joint distribution of a multivariate extreme value model. Rather than only considering observations that are extreme in all components of a random vector, observations with at least one component being extreme can be considered as well. As shown in Figure 2.4(C) by the red dots, this can yield a considerable gain in the number of observations that are considered to be extreme. See Chapter 3 for a formal introduction to the Heffernan and Tawn model. A sim-



(A) Component wise maxima (B) Threshold exceedances (C) Heffernan and Tawn

FIGURE 2.4: Two dimensional equivalent of the univariate extremes paradigms. The red dots indicate extreme events under the different models.

ulation study to compare the performance of the different methods in return level estimation was presented by Zheng et al. (2014). Summarizing, the three different approaches to modeling multivariate extremes are:

1. Componentwise Maxima approach, see Section 2.2.4,
2. Threshold Exceedance approach, see Section 2.2.5,
3. Heffernan and Tawn approach, see Chapter 3.

2.2.1 Mathematical framework

Before introducing the different approaches to modeling multivariate extremes, some additional concepts need to be introduced to serve as a common framework. Let $\mathbf{Y} = (Y_1, \ldots, Y_d) \in \mathbb{R}^d$ for $d \geq 2$ denote a finite dimensional random variable with cumulative distribution function $F_{\mathbf{Y}}$. Define the index set $I \coloneqq \{i \colon 1 \leq i \leq d\}$. Throughout this report, the index i is the primary index to refer to a particular component of a vector.

A sequence of length n of random vectors $\mathbf{Y}$ is denoted by

$$\mathbf{Y}_1, \ldots, \mathbf{Y}_n = (Y_{11}, \ldots, Y_{d1}), \ldots, (Y_{1n}, \ldots, Y_{dn}),$$

in accordance with the univariate notation. From here onward, the first index always refers to a particular component of $\mathbf{Y}$, and the second index refers to sample index.

The marginal distribution for each element of $\mathbf{Y}$ is denoted by F_{Y_i}, which will be abbreviated to F_i if it is clear from the context that F_i is the cumulative distribution function of Y_i.

In general, bold capital letters such as $\mathbf{Y}$ are used to denote random vectors, while lowercase bold letters such as $\boldsymbol{y}$ are realizations of the associated random vector. Following the vector notation proposed by Heffernan and Resnick (2007, Appendix 1), operators are assumed to apply component-wise.

2.2.2 Marginal transformations

It is beneficial to ensure that the marginal distributions of a multivariate random variable are similar when studying multivariate extremes. The *probability integral transform* is a two-step transformation that allows data from any arbitrary distribution to be transformed to any desirable scale. First, the random variable Y with cumulative distribution function F_Y is transformed to the uniform scale by $F_Y(Y)$. Secondly, the uniform data can be transformed to an arbitrary scale by any monotone increasing transformation $T(Y)$. An important feature of the probability integral transform is that the dependence structure is unaffected by the marginal transformations.

If the marginal distribution of Y is explicitly known and the transformation is monotone increasing, the probability integral transform is a bijection between the data on the original scale and the data on the new scale. For the simulated data used in this thesis, the cumulative distributions functions are specified. Hence data will be transformed based on the true cumulative probability distribution.

If on the other hand, the marginal distribution is unknown or the parameters that define the distribution have to be estimated, the empirical distribution function with Pareto tail $\breve{F}$ proposed by Coles and Tawn (1994) can be used. Under the assumption that the tail of a distribution is well approximated by the generalized Pareto distribution, the cumulative distribution function F_Y is well approximated by

$$\breve{F}_Y(y) \coloneqq \begin{cases} \tilde{F}_Y(y) & \text{for} \quad y \leq u \\ 1 - \left\{1 - \tilde{F}_Y(u)\right\} \{1 - G_u(y-u)\} & \text{for} \quad y > u \end{cases}, \tag{2.16}$$

where the function $\tilde{F}$ denotes the empirical cumulative distribution function. The generalized Pareto distribution G_u defined by (2.14) can be evaluated based on the maximum likelihood estimates $\hat{\xi}$ and $\hat{\sigma}_u$.

Substituting the tail of the empirical cumulative distribution function $\tilde{F}$ with a continuous Pareto tail, allows $\breve{F}_Y$ to be inverted such that the aforementioned bijectional property is preserved. The empirical cumulative distribution function $\tilde{F}_Y$ is a step function by construction, and since the number of observations in the tail of the distribution is small, $\tilde{F}^{\leftarrow}\colon T(Y) \to Y$ becomes a surjective mapping.

The Gumbel–, Laplace–, Fréchet– and Pareto scale are commonly used in the context of extreme value analysis to ensure that the marginal distributions of a multivariate random variable share a common scale. The transformations are to the unit scale of a particular distribution.

- **Gumbel scale**
  Heffernan and Tawn (2004) propose to work on the Gumbel scale because they exploit the property that $T_{\mathrm{G}}(Y)$ has an exponential upper tail. The transformation T_{G} is defined by the inverse of (2.6), i.e.:

$$T_{\mathrm{G}}(Y) \coloneqq -\log\left[-\log\left\{\breve{F}_Y(Y)\right\}\right]. \tag{2.17}$$

- **Laplace scale**
  Transforming marginals to the Laplace scale — rather than the Gumbel scale — was proposed by Keef et al. (2013). An important advantage of

the Laplace distribution is the fact that it is symmetric and has exponentially decaying tails. In addition, symmetry allows the parameterization of the Heffernan and Tawn model for both positive and negative *extremal dependence*[2] to be unified. The transformation to the Laplace scale is given by:

$$T_{\text{L}}(Y) \coloneqq \begin{cases} \log\left\{2\breve{F}_Y(Y)\right\}, & Y \leq \breve{F}_Y^{-1}(1/2), \\ -\log\left[2\left\{1-\breve{F}_Y(Y)\right\}\right], & Y > \breve{F}_Y^{-1}(1/2); \end{cases}. \tag{2.18}$$

- **Fréchet scale**
  The Fréchet transformation is commonly used in theoretical applications, see for example Coles (2001). The limiting joint distribution of the extremes of a multivariate random variable with Fréchet marginal distributions, gives rise to the family of multivariate extreme value distributions. The data is transformed to the Fréchet scale by the inverse of (2.5), i.e.:

$$T_{\text{F}}(Y) \coloneqq \frac{1}{-\log\left\{\breve{F}_Y(Y)\right\}}. \tag{2.19}$$

- **Pareto scale**
  Das and Resnick (2011) and Mitra and Resnick (2013) work on the Pareto scale because "it facilitates the use of tools from standard regular variation theory", (Das and Resnick 2011). The transformation to the Pareto scale is given by:

$$T_{\text{P}}(Y) \coloneqq \frac{1}{1-\breve{F}_Y(Y)}. \tag{2.20}$$

2.2.3 Extreme sets

The threshold u marks the beginning of the tail of a distribution, such that observations that fall in the set $\{y \in \mathbb{R} \colon y > u\}$ are regarded to be extreme. This provides a useful framework for univariate extreme value models, but for high dimensional problems this definition no longer suffices. Defining an *extreme set* A allows observations $\boldsymbol{y}$ of a multivariate random variable $\mathbf{Y}$ to be regarded as extreme if $\boldsymbol{y} \in A$.

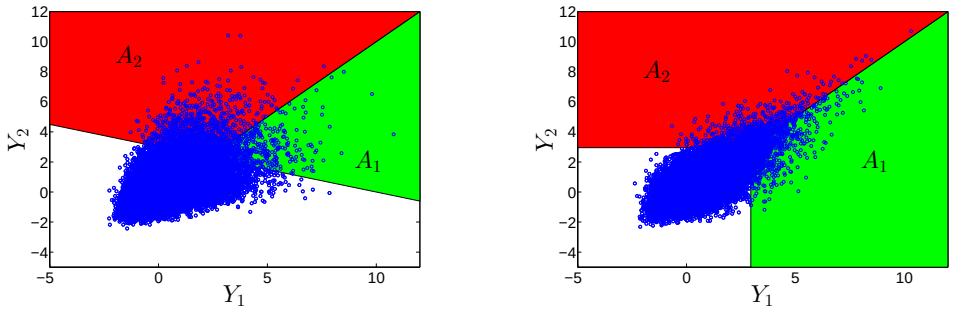
Definition 2.2.1 (Extreme set (Heffernan and Tawn 2004))**.**
Let the set $A \subset \mathbb{R}^d$ be an extreme set *such that at least one component of $\boldsymbol{y}$: $\{y_1, \ldots, y_d\}$ is extreme. The set A can be partitioned into subsets A_i, such that $A \coloneqq \bigcup_{i=1}^{d} A_i$ and for any $i \in \{1, \ldots, d\}$ the A_i is the subset of A for which component y_i is largest on*

[2]Extremal dependence is introduced in Section 2.2.6.

quantile scale, i.e.

$$A_i \coloneqq A \cap \left\{ \boldsymbol{y} \in \mathbb{R}^d \colon F_i\left(y_i\right) > F_j\left(y_j\right) \quad \textit{for all} \quad j \in I, j \neq i \right\}, \quad \forall i \in I.$$

See Figure 2.5 for two practical examples of an extreme set A in a bivariate setting, and how A_1 and A_2 make up for A. The nature of the data considered in Figure 2.5 is formally introduced in Section 3.2.1.



(A) $A = \{T_{\text{L}}\left(\boldsymbol{y}\right) : 0.3T_{\text{L}}\left(y_1\right) + T_{\text{L}}\left(y_2\right) > 6\}.$ (B) $A = \{T_{\text{L}}\left(\boldsymbol{y}\right) : T_{\text{L}}\left(y_1\right) > u, T_{\text{L}}\left(y_2\right) > u\}$

FIGURE 2.5: Examples of *extreme sets* A in a bivariate setting. The data is bivariate Gaussian (A) and generalized extreme value distributed with symmetric logistic dependence function (B), both transformed to the Gumbel scale. The threshold u is equal to the 95% marginal quantile.

2.2.4 Componentwise maxima approach

Similar to the univariate case, stochastic properties of extremes in a multivariate setting rely on the limiting distribution of component wise maxima. However, defining block maxima for random vectors is not trivial. One variable being extreme does not require the other variable(s) to be extreme as well, as shown in Figure 2.4(A).

Without loss of generality, the concepts introduced in the section focus on a bivariate setting because in higher dimensions the expressions become unwieldy. Let $\mathbf{Y}$ be a bivariate random vector in accordance with the properties and notation presented in Section 2.2.1, such that $I = \{1, 2\}$.

The definition for *partial maxima* is understood component wise, i.e.

$$M_{in} \coloneqq \max\left(Y_{i1}, \ldots, Y_{in}\right) \quad \text{for} \quad i \in I,$$

The *vector of component wise maxima* is defined as $\mathbf{M}_n = (M_{1n}, M_{2n})$. The central challenge at this point is to characterize non-degenerate limiting distributions of re-scaled pairs $\left(Y_{11}, Y_{21}\right), \ldots, \left(Y_{1n}, Y_{2n}\right)$, if they exist at all. The following steps should provide a rigorous answer to this challenge:

1. Re-scale the marginal distributions to the Fréchet scale.
2. Show that if a non-degenerate joint limiting distribution exists, then it must be max-stable.
3. Identify a functional for the family of max-stable distributions.

Start with transforming the data to the Fréchet scale by $T_{\mathrm{F}}(\mathbf{Y})$ such that $F_i \in \mathcal{D}(G_\xi)$ for all $i \in I$. The generalized extreme value distribution stated in (2.4), reduces to the unit Fréchet distribution if $\xi = 1, \tau = 1$ and $\sigma = 1$. This guarantees max-stability of the marginal distributions, as mentioned in Section 2.1.2. Max stability of the marginal distributions is desirable because it guarantees that the bivariate limit distribution of the partial maxima M_{in} scaled by n is also max stable. Tawn (1988) shows that the limiting joint distribution of $(M_{1n}/n, M_{2n}/n)$ as $n \to \infty$ gives rise to the *bivariate extreme value distribution*, see Definition 2.2.2.

Definition 2.2.2 (Bivariate extreme value distribution (Resnick 1987; Tawn 1988))**.**

To define the bivariate extreme value distribution, consider random vectors (Y_{1l}, Y_{2l}) *for* $l \in \{1, \ldots, n\}$*, with standard Fréchet marginal distributions. Then*

$$\Pr\left(M_{1n}/n \leq y_1, M_{2n}/n \leq y_2\right) \xrightarrow{d} G\left(y_1, y_2\right), \tag{2.21}$$

where G is a non-degenerate distribution functions, and G has the form

$$G_{\boldsymbol{\xi}}\left(y_1, y_2\right) = \exp\left\{-V\left(y_1, y_2\right)\right\}, \quad \textit{for} \quad y_1, y_2 > 0, \tag{2.22}$$

where $V(y_1, y_2)$ *is called the* exponent measure *and is defined by*

$$V\left(y_1, y_2\right) = 2\int_0^1 \max\left(\frac{w}{y_1}, \frac{1-w}{y_2}\right) \mathrm{d}H\left(w\right) \tag{2.23}$$

and H *is a spectral distribution function on* $[0, 1]$ *satisfying the mean constraint*

$$\int_0^1 w \mathrm{d}H(w) = 1/2. \tag{2.24}$$

Mind the subtle difference in notation, as G_ξ refers to the univariate generalized extreme value distribution defined by (2.4), and $G_{\boldsymbol{\xi}}$ defined by (2.22) denotes its bivariate counterpart. Definition 2.2.2 can be generalized for $d \geq 2$ dimensions, see De Haan and Ferreira (2006).

The exponent measure V is said to be *homogeneous of order -1*, which means that for any constant $c > 0$ the equality $V(c^{-1}y_1, c^{-1}y_2) = cV(y_1, y_2)$ holds. This is particularly useful because max-stability of the bivariate extreme value distribution is guaranteed since $G^n(y_1, y_2) = G\left(n^{-1}y_1, n^{-1}y_2\right)$.

Asymptotic independence arises for the discrete spectral measure $H = 0.5$ for $w = 0$ and $w = 1$. This yields the exponent measure $V(y_1, y_2) = y_1^{-1} + y_2^{-1}$, which leads to the bivariate extreme value distribution being defined by

$$G_{\boldsymbol{\xi}}(y_1, y_2) = \exp\left\{-\left(y_1^{-1} + y_2^{-1}\right)\right\}, \quad y_1, y_2 > 0.$$

Exact dependence on the other hand, arises when $H = 1$ for $w = 0.5$. In that case, the exponent measure $V(y_1, y_2) = \max\left(y_1^{-1}, y_2^{-1}\right)$, such that the bivariate extreme value distribution is given by

$$G_{\boldsymbol{\xi}}(y_1, y_2) = \exp\{-\max(y_1^{-1}, y_2^{-1})\}, \quad \text{for} \quad y_1, y_2 > 0.$$

Although $G_{\boldsymbol{\xi}}$ in (2.22) characterizes the non-degenerate limit distributions, this class is still very broad. The only constraint on the exponent measure $V(y_1, y_2)$ is given by (2.23) and (2.24). Characterizing the entire family of limiting distributions G with a single parametric model, just like (2.4) for the univariate case, is impossible.

An implication of Theorem 2.2.2 is that $G_{\boldsymbol{\xi}}$ in (2.4) has a bijectional relationship with the class of spectral distribution functions H that satisfy (2.24). This motivated the development of parametric models that cover sub-families of $G_{\boldsymbol{\xi}}$. The bijectional relationship between $G_{\boldsymbol{\xi}}$ and H ensures that it is sufficient to characterize H. Several parametric families for H have been proposed, such as the model by Hüsler and Reiss (1989), the bilogistic model by Joe et al. (1992) or the Dirichlet model by Coles et al. (1991). See Appendix A.2 for additional information regarding these models.

2.2.5 Threshold exceedance approach

The *threshold exceedance approach* is a high dimensional extension of the Peaks Over Threshold paradigm. It allows the utilization of available data to be improved. In addition, the issue that componentwise maxima might not correspond to actual observations was neglected so far. This is resolved by considering threshold exceedances rather than block maxima, as shown in Figure 2.4(B) all threshold exceedances are proper extreme events.

Start with the familiar set-up. For a bivariate random variable $\mathbf{Y}$ with cumulative distribution function $F_{\mathbf{Y}}$ and marginal distributions F_1 and F_2. Assume that the tail of each of the marginal distributions above the high thresholds u_1 and u_2 can be approximated by the generalized Pareto distribution. Transform each of the marginal distributions to the Fréchet scale. For $y_1 > u_1$ and $y_2 > u_2$ the joint tail of $T_{\text{F}}(\mathbf{Y})$ is well approximated by the bivariate extreme value distribution defined by (2.22).

The thresholds u_1 and u_2 partition $\mathbb{R}^2$ into four quadrants. A disadvantage of the threshold exceedance approach is that the model is only valid in the region $[u_1, \infty) \times [u_2, \infty)$ where both Y_1 and Y_2 are extreme. Fitting a model to the joint tail of a distribution can be difficult due to sparse data . Censored likelihood functions have been proposed to address this issue.

The *point process approach* is an alternative characterization of the threshold exceedance approach. For a bivariate random variable, the transformed variables $T_{\mathrm{F}}(Y_1)$ and $T_{\mathrm{F}}(Y_2)$ have Fréchet marginal distributions. Define a sequence of point processes by

$$\mathcal{P}_n \coloneqq \left\{ \left(\frac{T_{\mathrm{L}}(Y_1)}{n}, \frac{T_{\mathrm{L}}(Y_2)}{n} \right) : i \in I \right\}, \quad n \in \mathbb{N}.$$

Under the conditions for convergence of the componentwise maxima, stated in Section 2.2.4, it can be shown that $\mathcal{P}_n$ converges to a Poisson process $\mathcal{P}$ as $n \to \infty$. Define the pseudo-polar coordinates (r, w) by

$$r = T_{\mathrm{F}}(Y_1) + T_{\mathrm{F}}(Y_2) \quad \text{and} \quad w = \frac{T_{\mathrm{F}}(Y_1)}{T_{\mathrm{F}}(Y_1) + T_{\mathrm{F}}(Y_2)}.$$

The *intensity function* Λ that defines the Poisson process $\mathcal{P}$ is given by

$$\Lambda(\mathrm{d}r, \mathrm{d}w) = \frac{\mathrm{d}r}{r^2} \times 2Q(\mathrm{d}w).$$

A major advantage of this approach is that it allows the probability of an extreme set to be estimated when the intersection between the extreme set A and the set of threshold exceedances is void. If the extreme set $A \subset [u_1, \infty) \times [u_2, \infty)$, for u_1 and u_2 sufficiently large, then for some constant $c > 0$,

$$\Pr\{T_{\mathrm{F}}(\mathbf{Y}) \in cA) \approx \frac{1}{c} \Pr(\mathbf{Y} \in A)$$

and

$$\Pr\{T_{\mathrm{G}}(\mathbf{Y}) \in c + A) \approx \exp(-c) \Pr(\mathbf{Y} \in A).$$

2.2.6 Extremal dependence

Dependence governs the joint behavior of a multivariate random variable. A similar concept is introduced to characterize dependence among extreme events. There are two different classes of *extremal dependence*:

1. Asymptotic dependence, and,
2. Asymptotic independence.

As Davison et al. (2015b) state, "it is important to detect the appropriate dependence class because most models for bivariate extremes encompass one type of dependence, or the other, but not both". A measure for the extremal dependence among the different components of a random vector $\mathbf{Y}$ is given by the limit of $\chi(p)$ for non-exceedance probability $p \to 1$, where for any $i, j \in I$ such that $i \neq j$,

$$\chi(p) \coloneqq \Pr\{F_i(Y_i) > p \mid F_j(Y_j) > p\}. \tag{2.25}$$

Provided that the limit $\chi \coloneqq \lim_{p\to 1} \chi(p)$ exists, $\chi \in [0,1]$ defines a measure for the extremal dependence. If $\chi = 0$, the random variables Y_i and Y_j are said to be *asymptotically independent* and if $\chi > 0$ they are *asymptotically dependent*.

For asymptotic independence $\chi = 0$. The rate at which the limit $\lim_{p\to 1} \chi(p)$ converges to 0 is not an appropriate measure for the strength of asymptotic independence. Hence a measure for extremal independence is defined by $\bar{\chi} \coloneqq \lim_{p\to 1} \bar{\chi}(p) \in (-1, 1]$, where

$$\bar{\chi}(p) \coloneqq \frac{2\log(1-p)}{\log \Pr\{F_i(Y_i) > p, F_j(Y_j) > p\}} - 1, \tag{2.26}$$

provided that the limit exists. A summary of the properties of χ and $\bar{\chi}$ is presented in Table 2.2. To extend the applicability of these concepts to practical

Asymptotic Dependence	Asymptotic Independence		
• $\chi \neq 0$	• $\chi = 0$		
• $\bar{\chi} = 1$	• $\bar{\chi} \neq 0$		
	Negative association	Exact Independence	Positive association
	• $\bar{\chi} < 0$	• $\bar{\chi} = 0$	• $\bar{\chi} > 0$

TABLE 2.2: Relationship between different classes of extremal dependence.

applications, estimators for χ and $\bar{\chi}$ have been developed. Define a bivariate *copula* function by

$$\mathcal{C}(p,p) \coloneqq F_{\mathbf{Y}}\left\{F_i^{-1}(p), F_j^{-1}(p)\right\}, \qquad \text{for} \quad 0 < p < 1. \tag{2.27}$$

The joint survival distribution in terms of the associated bivariate copula is given by

$$\bar{\mathcal{C}}(p,p) \coloneqq \Pr\{F_i(Y_i) > p, F_j(Y_j) > p\} = 1 - 2p + \mathcal{C}(p,p). \tag{2.28}$$

The estimators $\hat{\chi}$ and $\hat{\bar{\chi}}$ rely on the empirical copula function $\hat{\mathcal{C}}$ defined by (2.29). The copula function $\mathcal{C}(p,p)$ defined by (2.27) can be approximated by

the empirical copula function, see AghaKouchak et al. (2013), defined by

$$\hat{\mathcal{C}}\left(p,p\right) \coloneqq \frac{1}{n+1}\sum_{l=1}^{n} \mathbb{1}_{\max\left\{\mathrm{rnk}(y_{il}),\mathrm{rnk}(y_{jl})\right\}\leq np} \tag{2.29}$$

for n independent observations $\left(y_{i1},y_{j1}\right),\ldots,\left(y_{in},y_{jn}\right)$ of $\mathbf{Y}=\left(Y_i,Y_j\right)$. The function $\mathrm{rnk}(\cdot)$ denotes the rank of the observation relative to its peers, and $n\cdot p$ is the fraction of the observations below the threshold u.

Combing both (2.25) and (2.27), as well as (2.26) and (2.28) yields, for $0<p<1$,

$$\chi\left(p\right)=2-\frac{\log\mathcal{C}\left(p,p\right)}{\log p} \quad \text{and} \quad \bar{\chi}\left(p\right)=\frac{2\log\left(1-p\right)}{\log\bar{\mathcal{C}}\left(p,p\right)}.$$

The estimator $\hat{\chi}\coloneqq\lim_{p\to1}\hat{\chi}(p)$ and $\hat{\bar{\chi}}\coloneqq\lim_{p\to1}\hat{\bar{\chi}}\left(p\right)$ follow from approximating $\mathcal{C}$ by $\hat{\mathcal{C}}$, which yields

$$\hat{\chi}(p)=2-\frac{\log\hat{\mathcal{C}}\left(p,p\right)}{\log p} \quad \text{and} \quad \hat{\bar{\chi}}(p)=\frac{2\log\left(1-p\right)}{\log\hat{\bar{\mathcal{C}}}\left(p,p\right)}-1.$$

Tawn and Ledford (1996) derive what the true parameters that characterize extremal dependence ought to be, for different parametric distributions. Two additional concepts that characterize extremal dependence are introduced in order to link the results presented by Tawn and Ledford (1996) to χ and $\bar{\chi}$. For two random variables with Fréchet marginal distributions, $T_{\mathrm{F}}\left(Y_1\right)$ and $T_{\mathrm{F}}\left(Y_2\right)$, under broad conditions, the joint probability

$$\Pr\left\{T_{\mathrm{F}}\left(Y_1\right)>y,T_{\mathrm{F}}\left(Y_2\right)>y\right\}\sim\mathcal{L}\left(y\right)\Pr\left\{T_{\mathrm{F}}\left(Y_1\right)>y\right\}^{1/\eta}, \quad \text{as} \quad y\to\infty.$$

The parameter $\eta\in(0,1]$ is called the *coefficient of tail dependence*. By definition, $\bar{\chi}\coloneqq2\eta-1$. The function $\mathcal{L}$ can be any arbitrary function that is slowly varying at infinity. A function is said to be *slowly varying at infinity* if, for any t fixed, $\lim_{y\to\infty}\mathcal{L}\left(ty\right)/\mathcal{L}\left(y\right)=1$. For asymptotically dependent random variables, i.e. $\eta=1$, as the function $\mathcal{L}\left(y\right)$ converges to a constant c as $y\to\infty$, it follows that

$$\chi=\lim_{p\to1}\frac{\bar{\mathcal{C}}\left(p,p\right)}{1-p}=c.$$

The true value for χ and $\bar{\chi}$ can be derived for random variables with continuous cumulative distributions functions. A derivation of χ and $\bar{\chi}$ for the multivariate Gaussian distribution is provided in Property 2.2.3. It follows that the multivariate Gaussian distribution is asymptotically independent, since $\chi=0$ and $\bar{\chi}=\rho$, where ρ denotes the correlation between Y_i and Y_j.

Property 2.2.3 (χ and $\bar{\chi}$ for a bivariate Gaussian random variable)**.**
If $\mathbf{Y}\sim\mathcal{N}\left(\mathbf{0},\Sigma\right)$ *where* $\sigma_i=\sigma_j=1$ *and* $\rho\neq0$*, then* $\chi=0$ *and* $\bar{\chi}=\rho$*.*

Proof. As shown by Tawn and Ledford (1996), the coefficient of tail dependence is given by $\eta = (1+\rho)/2$. It follows directly that $\bar{\chi} = \rho$. As $\mathcal{L}(y) = c_\rho (\log y)^{-\frac{\rho}{1+\rho}}$ tends to 0 as $y \to \infty$, it follows that $\chi = 0$. □

The generalized extreme value distribution with symmetric logistic dependence function yields asymptotically dependent data as $\chi = \rho$ and $\bar{\chi} = 1$, see Property 2.2.4.

Property 2.2.4 (χ and $\bar{\chi}$ for the bivariate generalized extreme value distribution with symmetric logistic dependence function[3] with parameter $0 < \rho < 1$)**.**

Proof. For asymptotic dependence, $\eta = 1$ and hence $\bar{\chi} = 1$. Since $\mathcal{L}(y) = 2 - 2^\rho$, it follows that $\chi = 2 - 2^\rho$. □

Since the exact values of χ and $\bar{\chi}$ are known for the aforementioned distributions, the performance of $\hat{\chi}$ and $\hat{\bar{\chi}}$ can be assessed. Diagnostic plots for the χ and $\bar{\chi}$ are shown in Figure 2.6. These figures suggest that the estimators $\hat{\chi}$ and $\hat{\bar{\chi}}$ behave as expected. There is a certain degree of bias in $\hat{\bar{\chi}}$, as shown by Figure 2.6(B) and Figure 2.6(D).

The concepts of asymptotic dependence and asymptotic independence can be extended to higher dimensions. A formal definition of extremal independence is presented in Theorem 2.2.5.

Definition 2.2.5 (Pairwise asymptotic independence (De Haan and Ferreira 2006))**.**
Let $F_{\mathbf{Y}} : \mathbb{R}^d \to \mathbb{R}_+$ be a probability distribution function. Suppose that its marginal distribution functions $F_i : \mathbb{R} \to \mathbb{R}_+$ satisfy

$$\lim_{n\to\infty} F_i^n(b_n y + a_n) = \exp\left\{-(1+\xi_i y)^{-1/\xi_i}\right\}, \quad \forall i \in I,$$

and for all y such that $1 + \xi_i y > 0$. Where

$$\mathbf{a}_n \coloneqq \{a_{1n}, \ldots, a_{dn}\} \in \mathbb{R}^d \quad \textit{and} \quad \mathbf{b}_n \coloneqq \{b_{1n}, \ldots, b_{dn}\} > \mathbf{0}.$$

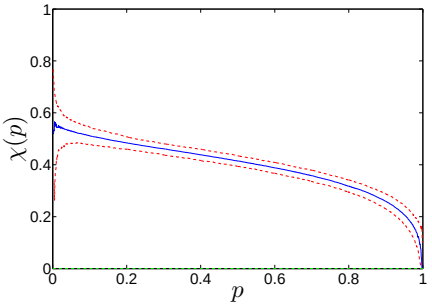
Let $\mathbf{Y} = (Y_1, \ldots, Y_d)$ be an random vector with distribution function $F_{\mathbf{Y}}$. If

$$\lim_{t\to\infty} \frac{\Pr\{Y_i > U_i(t), Y_j > U_j(t)\}}{\Pr\{Y_i > U_i(t)\}} = 0$$

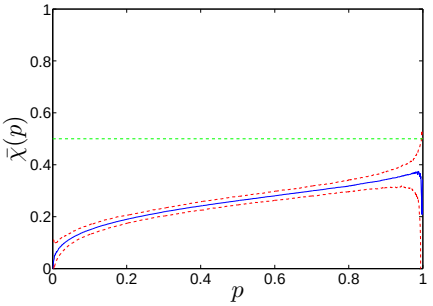
for all $1 \le i < j \le d$, then

$$\lim_{t\to\infty} F_{\mathbf{Y}}^n(\mathbf{b}_n \boldsymbol{y} + \mathbf{a}_n) = \exp\left\{-\sum_{i=1}^{d} (1+\xi_i y_i)^{-1/\xi_i}\right\}$$

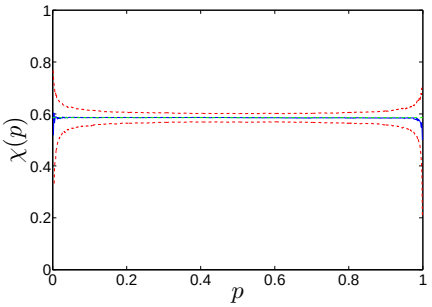
[3]This distribution is used extensively in this thesis, and is formally introduced in Section 3.2.1.



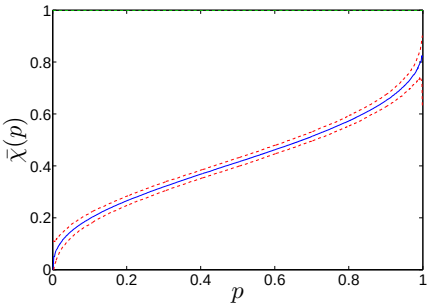
(A) χ for Gaussian data.



(B) $\bar{\chi}$ for Gaussian data.



(C) χ for GEV data with symmetric logistic dependence function data.



(D) $\bar{\chi}$ for GEV data with symmetric logistic dependence function data.

FIGURE 2.6: The sampling distribution of the $\hat{\chi}$ and $\hat{\bar{\chi}}$ estimators (—) as a function of the non-exceedance probability p. Computation of the estimators is repeated $n_B = 10^3$ times, each time with a new sample of $n = 10^4$ realizations from the bivariate Gaussian distribution and the generalized extreme value distribution with symmetric logistic dependence function, both with dependence parameter $\rho = 0.5$. The 95% confidence intervals (- - -) are based on the 2.5% and 97.5% empirical quantiles of the obtained sample of $\hat{\chi}$ and $\hat{\bar{\chi}}$ estimates. The true value (- - -) is obtained through Property 2.2.3 and 2.2.4.

for $1 + \xi_i y_i > 0$ *and* $i \in I$. *Hence the components of the random vector* $(Y_1, \ldots, Y_d)$ *are* asymptotically independent.

Definition 2.2.5 extents the bivariate definitions of χ and $\bar{\chi}$ to higher dimensions in a pairwise sense. However, it would be preferred to jointly characterize tail dependence for multiple components of a random vector. For $\mathbf{Y} \in \mathbb{R}^d$ and any $i \in I$, Wadsworth and Tawn (2013) define the measure of *d-dimensional joint tail dependence* by

$$\boldsymbol{\chi} \coloneqq \lim_{p \to 1} \Pr\left\{F_j(Y_j) > p \colon \forall j \in I, j \neq i \mid F_i(Y_i) > p\right\}, \tag{2.30}$$

If $\boldsymbol{\chi} > \mathbf{0}$, the vector $\mathbf{X}$ is said to exhibit *strong joint tail dependence* and if $\boldsymbol{\chi} = \mathbf{0}$, $\mathbf{X}$ is said to exhibits *weak joint tail dependence*. See Wadsworth and Tawn (2013) for further reference.

Chapter 3

Heffernan and Tawn Model

> Essentially, all models are wrong, but some are useful.
>
> — George Box

The Heffernan and Tawn (2004) model was proposed to address the deficiencies of the multivariate extreme value models introduced in Section 2.2. First of all, these models are appropriate for either asymptotically dependent– or asymptotically independent random variables, but not both. There is no explicit all-embracing parametric model for the two different classes of extremal dependence structures. Secondly, statistical inference for these models is subject to the curse of dimensionality. Fitting models to high dimensional problems is very challenging when data is sparse, and can lead to severe model misspecification.

A description of the Heffernan and Tawn model is provided in Section 3.1, which follows Heffernan and Tawn (2004) closely. Statistical inference for the Heffernan and Tawn model is demonstrated based on a simulation study, the results of which are presented in Section 3.2. Retrieving the true parameter values by minimizing the negative log-likelihood function has proven to be very challenging. This motivated contemplation of the model to identify the leading sources of bias and variance in the maximum likelihood estimator. In particular, the influence of changing the sample size, non-exceedance probability and dependence in the data sample is studied.

3.1 An introduction to the Heffernan and Tawn model

The model proposed by Heffernan and Tawn (2004) is introduced in this section. The mathematical framework is introduced in Section 3.1.1. The Heffernan and Tawn model is introduced in Section 3.1.2. The first step towards an operational model is to make and explicit assumption on the normalizing functions, see Section 3.1.3. Constraints on these normalizing functions that ensure stochastic ordering of the conditional quantiles were proposed by Keef et al. (2013), and are presented in Section 3.1.3. The second step to fully characterize the Heffernan and Tawn model is to assume a parametric limit distribution, see Section 3.1.5. Finally, a fundamental issue of the Heffernan and Tawn model is briefly introduced in Section 3.1.6.

3.1.1 Mathematical framework

The rationale behind the Heffernan and Tawn model is deceptively simple. Define a d-dimensional random variable by $\mathbf{Y} \in \mathbb{R}^d$ alongside the mathematical framework introduced in Section 2.2.1. For an extreme set A, see Definition 2.2.1, the objective is to estimate the probability $\Pr(\mathbf{Y} \in A)$. Start with the observation that $\Pr(\mathbf{Y} \in A)$ can be expressed as

$$\begin{aligned}\Pr(\mathbf{Y} \in A) &= \sum_{i=1}^{d} \Pr(\mathbf{Y} \in A_i), \\ &= \sum_{i=1}^{d} \Pr(\mathbf{Y} \in A_i \mid Y_i > v_i) \Pr(Y_i > v_i),\end{aligned}$$

for an arbitrary high quantile v_i associated to Y_i and A_i. The high quantile v_i is not to be confused with the threshold u_i defined in Section 2.1.3. Without loss of generality, define $v_i \coloneqq \inf\{y_i : \boldsymbol{y} \in A_i\}$. For each $i \in I$, the problem now reduces to estimating both the marginal probability $\Pr(Y_i > v_i)$ and the conditional probability $\Pr(\mathbf{Y} \in A_i \mid Y_i > v_i)$. However, estimating the conditional probability $\Pr(\mathbf{Y} \in A_i \mid Y_i > v_i)$ is not trivial. By definition,

$$\Pr(\mathbf{Y} \in A_i \mid Y_i > v_i) = \int_{v_i}^{y^{\mathrm{F}_i}} \Pr(\mathbf{Y} \in A_i \mid Y_i = y) \frac{\mathrm{d}F_i(y)}{1 - F_i(v_i)}. \tag{3.1}$$

Under the assumption that $F_i \in \mathcal{D}(G_\xi)$, estimating the marginal probability is straightforward and, as Heffernan and Tawn (2004) state, "the derivative of $F_i(y)/\{1 - F_i(v_i)\}$ in (3.1) is the generalized Pareto density function". The problem is now reduced to defining a model for $\Pr(\mathbf{Y} \in A_i \mid Y_i = y)$, where Y_i is referred to as the *conditioning variable*.

The integrand in (3.1) is usually not evaluated on the original scale, but rather on the Gumbel scale (Heffernan and Tawn 2004) or the Laplace scale (Keef et al. 2013; Lugrin et al. 2016) to ensure that the marginal distributions of the transformed random variable $T(\mathbf{Y})$ are equivalent. Here and throughout, the Laplace scale is adopted because of the similar parameterization for both positive– and negative dependent variables.

3.1.2 Model description

In similar fashion as the normalization of component wise maxima discussed in Section 2.1.2, assume the existence of normalizing functions $\mathbf{a}_{|i}(y) \in \mathbb{R}^{d-1}$ and $\mathbf{b}_{|i}(y) \in (0,\infty]^{d-1}$, where

$$\mathbf{a}_{|i}(y) \coloneqq \left\{a_{|1}(y), \ldots, a_{|d}(y)\right\} \quad \text{and} \quad \mathbf{b}_{|i}(y) \coloneqq \left\{b_{|1}(y), \ldots, b_{|d}(y)\right\}.$$

Define the *residual* $Z_{j|i}$ as the normalization of $T_{\mathrm{L}}(Y_j)$ conditional on $T_{\mathrm{L}}(Y_i) = y$ by normalizing functions $a_{j|i}(y)$ and $b_{j|i}(y)$. If $v_i > u_i$, then $y > u_i$ as well. The indexing $j \mid i$ refers to the j-th component of a particular vector, when i is the index of the conditioning variable Y_i. Throughout this thesis report, bold symbols or numbers refer to vectors of appropriate dimension. For example, $\mathbf{a}_{|i}(y) = \mathbf{1}$ refers to a $d-1$ dimensional vector of ones. In vector notation, conditional on $T_{\mathrm{L}}(Y_i) = y$, the residual $\mathbf{Z}_{|i}$ is defined by

$$\mathbf{Z}_{|i} \coloneqq \frac{T_{\mathrm{L}}(\mathbf{Y}_{-i}) - \mathbf{a}_{|i}(y)}{\mathbf{b}_{|i}(y)}, \tag{3.2}$$

where for all $i \in I$,

$$T_{\mathrm{L}}(\mathbf{Y}_{-i}) \coloneqq T_{\mathrm{L}}(\mathbf{Y}) \setminus T_{\mathrm{L}}(Y_i) = \left\{T_{\mathrm{L}}(Y_1), \ldots, T_{\mathrm{L}}(Y_{i-1}), T_{\mathrm{L}}(Y_{i+1}), \ldots, T_{\mathrm{L}}(Y_d)\right\}.$$

Let $G_{|i}$ — which is not to be confused with G_ξ and G_u — denote the limiting distribution of the residuals $Z_{|i}$. The marginal distributions of $G_{|i}$ are non-degenerate as long as the marginal distributions of $F_{\mathbf{Y}}$ are non-degenerate. In addition, the marginal transformation T_{L} ensures that the marginal distributions of the limiting distribution $G_{|i}$ are the same. Different marginal distributions would complicate the formulation of a generic model for the conditional distribution of $T_{\mathrm{L}}(\mathbf{Y}_{-i})$ conditional on the conditioning variable $T_{\mathrm{L}}(Y_i) > u_i$. Define the limit distribution $G_{|i}$, for $u_i \to y^{\mathrm{F}_i}$, by

$$\Pr\left\{\frac{T_{\mathrm{L}}(\mathbf{Y}_{-i}) - \mathbf{a}_{|i}(y)}{\mathbf{b}_{|i}(y)} \leq \mathbf{z}_{|i} \,\middle|\, T_{\mathrm{L}}(Y_i) > u_i\right\} \xrightarrow{d} G_{|i}\left(\mathbf{z}_{|i}\right). \tag{3.3}$$

Under the assumption that $F_i \in \mathcal{D}(G_\xi)$, the re-scaled threshold exceedances of the conditioning variable $(Y_i - u_i)/f(u_i)$ are generalized Pareto distributed. As Heffernan and Tawn (2004) point out, that implies that "the re-scaled conditioning variable is *asymptotically conditionally independent* of the residual $\mathbf{Z}_{|i}$, given that $Y_i > u_i$, as $u_i \to y^{F_i}$". Lugrin et al. (2016) provide an alternative formulation of the limiting argument presented by Heffernan and Tawn (2004), which states

$$\begin{aligned}
&\Pr\left(\mathbf{Z}_{|i} \leq \mathbf{z}_{|i}, \frac{T_{\mathrm{L}}(Y_i) - u_i}{f(u_i)} > y_i \,\middle|\, T_{\mathrm{L}}(Y_i) > u_i\right) \\
&\quad = \Pr\left(\mathbf{Z}_{|i} \leq \mathbf{z}_{|i} \,\middle|\, \frac{T_{\mathrm{L}}(Y_i) - u_i}{f(u_i)} > y_i\right) \Pr\left(\frac{T_{\mathrm{L}}(Y_i) - u_i}{f(u_i)} > y_i \,\middle|\, T_{\mathrm{L}}(Y_i) > u_i\right) \\
&\quad \xrightarrow{d} G_{|i}\left(\mathbf{z}_{|i}\right)\left\{1 - G_{u_i}(y)\right\}, \quad \text{as} \quad u_i \to y^{F_i}.
\end{aligned} \tag{3.4}$$

Two properties of $G_{|i}$ are discussed before appropriate candidate limit distributions are considered.

First of all, if $d \geq 3$, the limit distribution $G_{|i}$ is itself a joint distribution. That means that the problem of estimating $\Pr(\mathbf{Y} \in A)$ is reduced from fitting the a model fro the joint tail distribution of a d dimensional random variable, to fitting d different joint distributions for $d-1$ dimensional residuals defined by (3.2). To simplify statistical inference for the Heffernan and Tawn model, the marginal distributions of $G_{|i}$ are assumed to be asymptotically conditionally independent. The residual $\mathbf{Z}_{|i}$ is said to be *mutually asymptotically conditionally independent* of the conditioning variable Y_i if, conditional on $Y_i = y$,

$$G_{|i}\left(\mathbf{z}_{|i}\right) = \prod_{\substack{j=1 \\ j \neq i}}^{d} G_{j|i}\left(z_{j|i}\right). \tag{3.5}$$

Very few distributions exhibit this property. Among several distributions considered in the original paper by Heffernan and Tawn (2004), only the inverted multivariate extreme value distribution with symmetric logistic dependence function is asymptotically conditionally independent.

Secondly, the resulting limit distribution $G_{|i}$ in (3.3) is *unique up to type*. As Heffernan and Tawn (2004) explain, that implies that "if the normalizing functions $\mathbf{a}_{|i}(y)$ and $\mathbf{b}_{|i}(y)$ give a non-degenerate limit distribution $G_{|i}\left(\mathbf{z}_{|i}\right)$, then for vector constants $\mathbf{c}_a \in \mathbb{R}^{d-1}$ and $\mathbf{c}_b \in (0, \infty]^{d-1}$, the normalizing functions

$$\mathbf{a}^*_{|i}(y) \coloneqq \mathbf{a}_{|i}(y) + \mathbf{c}_a \mathbf{b}_{|i}(y) \quad \text{and} \quad \mathbf{b}^*_{|i} \coloneqq \mathbf{c}_b \mathbf{b}_{|i}(y) \tag{3.6}$$

provide a non-degenerate limit distribution $G_{|i}\left(\mathbf{c}_b \mathbf{z}_{|i} + \mathbf{c}_a\right)$". Hence the limit distribution $G_{|i}\left(\mathbf{z}_{|i}\right)$ and $G_{|i}\left(\mathbf{c}_b \mathbf{z}_{|i} + \mathbf{c}_a\right)$ are *unique up to type*, for all $\mathbf{z}_{|i} \in \mathbb{R}^{d-1}$

and "the normalizing functions $\mathbf{a}_{|i}(y)$ and $\mathbf{b}_{|i}(y)$ can only be identified up to constants $\mathbf{c}_a$ and $\mathbf{c}_b$" in (3.6).

Now that the residual $\mathbf{Z}_{|i}$ and the limit distribution $G_{|i}$ are introduced, conditional on $T_{\mathrm{L}}(Y_i) = y$, the extremal dependence model proposed by Heffernan and Tawn (2004) reduces to the semi-parametric regression model:

$$T_{\mathrm{L}}(\mathbf{Y}_{-i}) = \mathbf{a}_{|i}(y) + \mathbf{b}_{|i}(y)\,\mathbf{Z}_{|i}. \tag{3.7}$$

Implementation of the model requires two explicit choices:

1. Functions for the normalizing functions $\mathbf{a}_{|i}$ and $\mathbf{b}_{|i}$, see Section 3.1.3, and,
2. A probability distribution for the limit distribution $G_{|i}$, see Section 3.1.5.

3.1.3 Explicit expressions for the normalizing functions

The growth of $T_{\mathrm{L}}(\mathbf{Y}_{-i}) \mid T_{\mathrm{L}}(Y_i) = y$ is governed by the dependence between Y_j and Y_i for each $j \in I \setminus \{i\}$. Perfect positive dependence and perfect negative dependence provide bounds on the normalizing functions $\mathbf{a}_{|i}(y)$ and $\mathbf{b}_{|i}(y)$, albeit the degeneracy of the limiting distribution $G_{|i}$ in those cases.

Under perfect dependence, the normalizing functions are given by $\mathbf{b}_{|i}(y) = \mathbf{1}$, as well as $\mathbf{a}_{|i}(y) = \boldsymbol{y}$ for positive perfect dependence and $\mathbf{a}_{|i}(y) = -\boldsymbol{y}$ for negative perfect dependence. For independent random variables, $\mathbf{a}_{|i}(y) = \mathbf{0}$ and $\mathbf{b}_{|i}(y) = \mathbf{1}$. Heffernan and Tawn (2004) derive the true normalizing functions for different probability distributions, some of which are shown in Table 3.1.

The examples provided in Table 3.1 suggest that the parametric family

$$\mathbf{a}_{|i}(y) = \boldsymbol{\alpha}_{|i} y \quad \text{and} \quad \mathbf{b}_{|i}(y) = y^{\boldsymbol{\beta}_{|i}}, \tag{3.8}$$

summarizes the normalizing functions. If (3.8) is considered, conditional on $T_{\mathrm{L}}(Y_i) = y$, the residual $\mathbf{Z}_{|i}$ is defined by

$$\mathbf{Z}_{|i} = \frac{T_{\mathrm{L}}(\mathbf{Y}_{-i}) - \boldsymbol{\alpha}_{|i} y}{y^{\boldsymbol{\beta}_{|i}}}. \tag{3.9}$$

Based on (3.6), it is apparent that parameterization (3.8) is unique, i.e. $\mathbf{a}^*_{|i}(y) = \mathbf{a}_{|i}(y)$ and $\mathbf{b}^*_{|i}(y) = \mathbf{b}_{|i}(y)$ if and only if $\mathbf{c}_a = \mathbf{0}$ and $\mathbf{c}_b = \mathbf{1}$.

If the marginal distributions of $\mathbf{Y}$ are transformed to the Laplace scale, the parameter space for the parameters of the Heffernan and Tawn model is given by

$$\Omega_{\boldsymbol{\alpha}_{|i} \times \boldsymbol{\beta}_{|i}} \subseteq [-1, 1]^{d-1} \times (-\infty, 1]^{d-1}. \tag{3.10}$$

The boundary of the parameter space is denoted by $\partial\Omega_{\boldsymbol{\alpha}_{|i} \times \boldsymbol{\beta}_{|i}}$.

TABLE 3.1: Normalizing functions $\mathbf{a}_{|i}(y)$ and $\mathbf{b}_{|i}(y)$ for different probability distributions of $\mathbf{Y}$ and their true limiting distribution $G_{|i}$. Whether or not $G_{|i}$ is asymptotically conditionally independent (abbreviated to ACI), is also indicated. Source: Heffernan and Tawn (2004, Table 1).

Distribution	Gumbel Scale		Laplace Scale		$G_{\|i}$	ACI
	$\mathbf{a}_{\|i}(y)$	$\mathbf{b}_{\|i}(y)$	$\mathbf{a}_{\|i}(y)$	$\mathbf{b}_{\|i}(y)$		
Perfect positive dependence	$\boldsymbol{y}$	$\mathbf{1}$	$\boldsymbol{y}$	$\mathbf{1}$	Degenerate	NA
Independence	$\mathbf{0}$	$\mathbf{1}$	$\mathbf{0}$	$\mathbf{1}$	Marginal	Yes
Multivariate Gaussian ($\boldsymbol{\rho} > \mathbf{0}$)	$\boldsymbol{\rho}^2 y$	$y^{\mathbf{1/2}}$	$\text{sign}(\boldsymbol{\rho})\,\boldsymbol{\rho}^2 y$	$y^{\mathbf{1/2}}$	Gaussian	No
Multivariate Gaussian ($\boldsymbol{\rho} < 0$)	$-\log(\boldsymbol{\rho}^2 y)$	$y^{-\mathbf{1/2}}$	$\text{sign}(\boldsymbol{\rho})\,\boldsymbol{\rho}^2 y$	$y^{\mathbf{1/2}}$	Gaussian	No
Multivariate extreme value distribution	$\boldsymbol{y}$	$\mathbf{1}$	$\boldsymbol{y}$	$\mathbf{1}$	§	No
Perfect negative dependence	$-\log \boldsymbol{y}$	$\mathbf{1}$	$-\boldsymbol{y}$	$\mathbf{1}$	Degenerate	NA

§: $G_{|i}$ depends on the dependence function.

The relationship between the (in)dependence properties of $T_{\text{L}}(\mathbf{Y}_{-i})$ and the normalizing functions $\mathbf{a}_{|i}(y)$ and $\mathbf{b}_{|i}(y)$ naturally extends to the Heffernan and Tawn parameters $\boldsymbol{\alpha}_{|i}$ and $\boldsymbol{\beta}_{|i}$, as shown in Table 3.2.

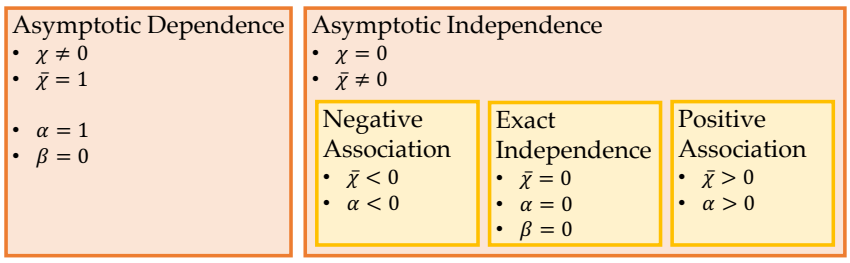


TABLE 3.2: Relationship between different classes of extremal dependence with respect to the parameters of the Heffernan and Tawn model.

The residual defined in (3.9) defines the semi-parametric non-linear regression model

$$T_{\text{L}}(\mathbf{Y}_{-i}) = \boldsymbol{\alpha}_{|i} y + y^{\boldsymbol{\beta}_{|i}} \mathbf{Z}_{|i}. \tag{3.11}$$

The first term on the right hand side of (3.11), is basically a linear trend line with

slope $\boldsymbol{\alpha}_{|i}$, analogous to linear regression. The $\boldsymbol{\beta}_{|i}$ parameter in (3.11) introduces non-linear scaling to the residual $\mathbf{Z}_{|i}$. The $\boldsymbol{\beta}_{|i}$ parameter governs whether the conditional quantiles of the different components of $\mathbf{Y}_{-i}$ either converge or diverge. As Lugrin et al. (2016) point out, when the scale parameter $\beta_{j|i} < 0$, "all the conditional quantiles for Y_j converge to the same value as Y_i increases, which is unlikely in most environmental contexts". A schematic overview of how changing $\boldsymbol{\alpha}_{|i}$ and $\boldsymbol{\beta}_{|i}$ affects (3.11) is shown in Figure 3.1 for a bivariate setting. The distribution of the residual $\mathbf{Z}_{|i}$ is left unspecified at the moment. See Section 3.1.5 for the discussion on appropriate distributions for $\mathbf{Z}_{|i}$.

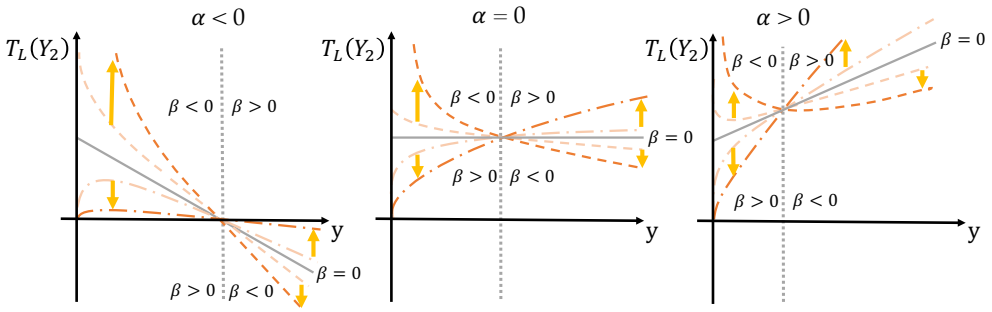


FIGURE 3.1: Overview of how the parameters α and β of the Heffernan and Tawn model in a bivariate context affect the semi-parametric model $Y_2 \mid Y_1 = y \sim \alpha y + y^{\beta}$. This is a deterministic equivalent of (3.11). The special case when $\beta = 0$ yields a linear curve with slope α. For $\beta < 0$ (- - -) and $\beta > 0$ (- · -), the yellow arrows indicate how the lines shift as β decreases for $\beta < 0$ and increases for $\beta > 0$.

3.1.4 Constrained Heffernan and Tawn model

A flaw in the Heffernan and Tawn model parameterization was identified by Keef et al. (2013). It turns out that by adopting (3.8), the joint probability of two events estimated under the Heffernan and Tawn model can be higher than the marginal probability of both events. Mathematically speaking, for a bivariate random variable $\mathbf{Y} = (Y_1, Y_2)$, and return periods $t_1, t_2 > 1$, it is possible that under the Heffernan and Tawn model the estimated joint probability can exceed the marginal probabilities, i.e.

$$\hat{\Pr}\left\{Y_1 > U_1\left(t_1\right), Y_2 > U_2\left(t_2\right)\right\} > \left\{\min\left(t_1, t_2\right)\right\}^{-1}. \tag{3.12}$$

In order to resolve this issue, Keef et al. (2013) propose constraints on the parameter space $\Omega_{\alpha\times\beta}$ to ensure that "conditional quantiles for any form of asymptotic independence cannot be larger than under asymptotic positive dependence, nor can they be smaller than under asymptotic negative dependence". The proposed conditions are presented in Theorem 3.1.1. Imposing

the constraints on (3.8) is referred to as the *constrained Heffernan and Tawn model*. The constrained parameter space is denoted by $\Omega^{\text{KEEF}}_{\boldsymbol{\alpha}_{|i}\times\boldsymbol{\beta}_{|i}}$.

The following definitions enhance the formulation of Theorem 3.1.1. The empirical quantile function of the residuals defined by (3.9) is denoted by $\tilde{U}$. In addition, define the residuals

$$Z^{-}_{j|i} \coloneqq T_{\text{L}}\left(Y_j\right) + T_{\text{L}}\left(Y_i\right) \quad \text{and} \quad Z^{+}_{j|i} \coloneqq T_{\text{L}}\left(Y_j\right) - T_{\text{L}}\left(Y_i\right).$$

For $t \in (1,\infty)$, the empirical quantile functions associated to $Z^{-}_{j|i}$ and $Z^{+}_{j|i}$ are

$$\tilde{U}^{-}\left(t\right) \coloneqq \tilde{F}^{\leftarrow}_{\text{Z}^{-}_{j|i}}\left(1-\frac{1}{t}\right) \quad \text{and} \quad \tilde{U}^{+}\left(t\right) \coloneqq \tilde{F}^{\leftarrow}_{\text{Z}^{+}_{j|i}}\left(1-\frac{1}{t}\right).$$

The derivation of Theorem 3.1.1 considers a bivariate random variable. For higher dimensional random variables, the conditions should be satisfied for each $j \in I\backslash\{i\}$ when Y_i is the conditioning variable. The impact of imposing the constraints presented in Theorem 3.1.1 on statistical inference for the Heffernan and Tawn model is discussed in Section 3.2 and 4.3.

Theorem 3.1.1 (Keef et al. (2013))**.**
For $\upsilon_i > u_i$ — where υ_i is not to be confused with v_i — conditional on $Y_i = y$, the stochastic ordering *constraint*

$$-y + \tilde{U}^{-}\left(t\right) \leq \tilde{U}\left(t\right) \leq y + \tilde{U}^{+}\left(t\right)$$

holds for all $y > \upsilon_i$ and for all $t \in (1,\infty)$, if and only if, both Case I and Case II hold.

- ***Case I**: Either*

$$\alpha_{j|i} \leq \min\left\{1, 1-\beta_{j|i}\tilde{U}\left(t\right)\upsilon_i^{\beta_{j|i}-1}, 1-\upsilon_i^{\beta_{j|i}}\tilde{U}\left(t\right)+\frac{1}{\upsilon_i}\tilde{U}^{+}\left(t\right)\right\}$$

  or, $1-\beta_{j|i}\tilde{U}\left(t\right)\upsilon_i^{\beta_{j|i}-1} < \alpha_{j|i} \leq 1$ *and*

$$\tilde{U}^{+}\left(t\right)+\left(1-\frac{1}{\beta_{j|i}}\right)\left\{\frac{\beta_{j|i}\tilde{U}\left(t\right)}{\left(1-\alpha_{j|i}\right)^{\beta_{j|i}}}\right\}^{1/\left(1-\beta_{j|i}\right)} > 0.$$

- ***Case II**: Either*

$$-\alpha_{j|i} \leq \min\left\{1, 1+\beta_{j|i}\tilde{U}\left(t\right)\upsilon_i^{\beta_{j|i}-1}, 1+\upsilon_i^{\beta_{j|i}}\tilde{U}\left(t\right)-\frac{1}{\upsilon_i}\tilde{U}^{-}\left(t\right)\right\},$$

or, $1 + \beta_{j|i}\tilde{U}(t)\, \upsilon_i^{\beta_{j|i}-1} < -\alpha_{j|i} \leq 1$ *and*

$$-\tilde{U}^{-}(t) + \left(1 - \frac{1}{\beta_{j|i}}\right)\left\{-\frac{\beta_{j|i}\tilde{U}(t)}{\left(1+\alpha_{j|i}\right)^{\beta_{j|i}}}\right\}^{1/\left(1-\beta_{j|i}\right)} > 0.$$

Apart from α, β and the data, the constraints defined by 3.1.1 are a function of both υ_i and t. As Keef et al. (2013) point out, it suffices to evaluate the constraints for $t \to 1$ and $t \to \infty$, as a simulation study performed by the authors showed that the constraints are satisfied for all $t \in (1, \infty)$ if they are satisfied for these two limiting cases.

Choosing υ_i is not trivial. Keef et al. (2013) only point out that "to give the greatest flexibility to the fits, the constraints are only imposed on extrapolations [beyond the observed data]". Thus υ_i is assigned an arbitrary value "above the maximum observed value of Y_i". Throughout this thesis, $\upsilon_i = \sup\{y : y = y_{i1}, \ldots, y_{in}\} + \epsilon$ and $\epsilon = 1$. The constraints show little sensitivity to the choice of ϵ. However, as Keef et al. (2013) point out, "in contrast, when υ_i was smaller [than the maximum observed value of Y_i] the fit of the resulting constrained model was poor".

Throughout this thesis, the constraints presented in Theorem 3.1.1 are imposed implicitly. For any proposal $\boldsymbol{\theta}^* \in \Omega_{\boldsymbol{\theta}}$, the constraints evaluated. An arbitrary large number is returned for the negative log-likelihood if the constraints are not satisfied.

3.1.5 Explicit choice on the limit distribution

The aim of this section is to choose a suitable distribution for the residuals $\mathbf{Z}_{|i}$ to define the semi-parametric regression model defined by (3.11). The Gaussian distribution and the parsimonious empirical distribution are briefly introduced. Although any parametric distribution would suffice, the Gaussian distribution is adopted to define the likelihood function for the Heffernan and Tawn model. As Heffernan and Tawn (2004) point out, "we have considered a range of parametric distributions for the marginals of $\mathbf{Z}_{|i}$ and selected the Gaussian distribution for its simplicity and superior performance in a simulation study". A recent work by Lugrin et al. (2016) considers a Dirichlet process, which the authors show to yield greater flexibility. This is a promising development as the two alternatives discussed in this section each have their deficiencies.

Gaussian distribution

If the limiting distribution $G_{|i}$ is assumed to be Gaussian with parameters $\boldsymbol{\mu}_{|i} \in \mathbb{R}^{d-1}$ and $\boldsymbol{\psi}_{|i}^2 > \mathbf{0}$, the Heffernan and Tawn model is given by

$$T_{\mathrm{L}}\left(\mathbf{Y}_{-i}\right) \mid T_{\mathrm{L}}\left(Y_i\right) = y \sim \mathcal{N}\left(\boldsymbol{\alpha}_{|i} y + y^{\boldsymbol{\beta}_{|i}} \boldsymbol{\mu}_{|i}, y^{2\boldsymbol{\beta}_{|i}} \boldsymbol{\psi}_{|i}^2\right), \quad \forall i \in I. \tag{3.13}$$

The parameters $\boldsymbol{\mu}_{|i}$ and $\boldsymbol{\psi}_{|i}^2$ are referred to as nuisance parameters because these parameters govern the random noise in the regression model formulation (3.7). The Gaussian distribution is shown to be true true limit distribution if the data under scrutiny is itself Gaussian distributed. For simulating purposes, practitioners often turn to the empirical distribution as residuals typically show a poor fit to the Gaussian distribution. The parameter space for the nuisance parameters $\boldsymbol{\mu}_{|i}$ and $\boldsymbol{\psi}_{|i}^2$ is given by

$$\Omega_{\boldsymbol{\mu}_{|i} \times \boldsymbol{\psi}_{|i}^2} \subseteq \mathbb{R}^{d-1} \times [0, \infty)^{d-1}. \tag{3.14}$$

Define $\boldsymbol{\theta} \coloneqq \left\{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\psi}^2\right\}$, such that $\Omega_{\boldsymbol{\theta}} \coloneqq \Omega_{\boldsymbol{\alpha} \times \boldsymbol{\beta}} \times \Omega_{\boldsymbol{\mu} \times \boldsymbol{\psi}^2}$.

Empirical distribution

Once the parameters of the Heffernan and Tawn model have been estimated, plugging these estimates in (3.2) yields a vector of residuals $\hat{\mathbf{Z}}_{|i}$. Rather than simulating residuals from a parametric distribution, instead sample from $\hat{Z}_{|i}$, which is defined by

$$\hat{\mathbf{Z}}_{|i} \coloneqq \frac{T_{\mathrm{L}}\left(\mathbf{Y}_{-i}\right) - \hat{\boldsymbol{\alpha}}_{|i} y}{y^{\hat{\boldsymbol{\beta}}_{|i}}}. \tag{3.15}$$

This pragmatic approach address possible non-Gaussian features in the data that would have been ignored if the residuals would have been sampled from a Gaussian distribution with parameters $\hat{\boldsymbol{\mu}}_{|i}$ and $\hat{\boldsymbol{\psi}}^2$. In this sense, sampling from the empirical distribution reduces the risk of model misspecification due to a poor choice of the limit distribution.

However, if the data sample to which the model is fitted is small, the number of residuals is also small. In that case, simulations under the Heffernan and Tawn model based on the empirical residuals $\hat{\mathbf{Z}}_{|i}$ will yield ray-like results, since each residual is sampled very often. It stands to reason that these simulations are not truly random and are possibly severely biased. Albeit this issue, the empirical distribution is widely used in practice.

3.1.6 Exchangeability and self consistency

By assumption asymptotic conditional independence, there is no guarantee that the different conditional distributions estimated under the Heffernan and Tawn model agree on the joint distribution. The problem is briefly discussed in this section to raise awareness for the issue.

A d-dimensional multivariate random variable $\mathbf{Y}$ is said to be *pairwise exchangeable* if for any $i, j = 1, \ldots, d$ such that $i \neq j$, the dependence of Y_i on Y_j is equivalent to the dependence of Y_j on Y_i. To be more specific, following the definitions presented by Heffernan and Tawn (2004), "the random variables Y_i and Y_j exhibit *weak pairwise extremal exchangebility* if $\theta_{j|i} = \theta_{i|j}$ and *strong pairwise extremal exchangebility* if in addition $G_{i|j} = G_{j|i}$.

In theory, the joint distribution $F_{\mathbf{Y}}$ of $\mathbf{Y}$ governs each of the different conditional distributions $T_{\mathrm{L}}\left(\mathbf{Y}_{-i}\right) \mid T_{\mathrm{L}}\left(Y_i\right) = y$, for $i = 1, \ldots, d$. The conditional distributions are said to be *self consistent* if they agree on the joint distribution. Mathematically speaking, for a bivariate random variable, the self consistency property holds if and only if

$$\frac{\mathrm{d}}{\mathrm{d}y_j}\Pr\left\{T_{\mathrm{L}}\left(Y_j\right) \leq y_j \mid T_{\mathrm{L}}\left(Y_i\right) = y_i\right\} f_i\left(y_i\right) = \frac{\mathrm{d}}{\mathrm{d}y_i}\Pr\left\{T_{\mathrm{L}}\left(Y_i\right) \leq y_i \mid T_{\mathrm{L}}\left(Y_j\right) = y_j\right\} f_j\left(y_j\right). \quad (3.16)$$

Imposing (3.16) on asymptotically independent random variables is too complex in practice. For asymptotically dependent data on the other hand, if either $\alpha_{j|i} = 1$ and/or $\alpha_{i|j} = 1$, (3.16) is trivially satisfied.

See Heffernan and Tawn (2004) and in particular Liu and Tawn (2014) for a more thorough description of the self consistency problem. In practice, the different models for each of the conditional distributions $\mathbf{Y}_{-i} \mid Y_i = y$ are assumed to be independent. This assumption justifies neglecting the constraints related to exchangeability and self consistency and ensures that the model is practically relevant.

The possibility of imposing constraints such that self consistency holds in a subspace of the joint tail region is considered by Liu and Tawn (2014). The authors derive alternative expressions for the residual distribution such that self consistency holds. Although their results extend to higher dimensional problems, in practice the constraints can only be imposed for bivariate problems. The exchangeability and self consistency issue is still an active field of research.

3.2 Statistical inference

The pragmatic assumption that the limit distribution $G_{|i}$ in (3.4) is Gaussian defines a likelihood function for the Heffernan and Tawn model. The goal of this section is to discuss statistical inference based on negative log-likelihood minimization. In general, statisticians distinguish Bayesian inference and frequentist methods.

Bayesians start with the observation that as the data is observed, it is inherently certain. Parameters, on the other hand, are assumed to be random objects with a probability distribution that characterizes them. See Chapter 4 for an elaborate introduction to Bayesian statistics and Bayesian inference for the Heffernan and Tawn model parameters.

Frequentist methods are based on the paradigm that unknown model parameters are deterministic, while the data is random. Maximizing the likelihood of observing the data over the parameter space $\Omega_{\boldsymbol{\theta}}$ is a widely used method for statistical inference. A description of the methodology, as well as maximum likelihood estimates for the Heffernan and Tawn model parameters are presented in this section.

Two different data sets, referred to as Case 1 and Case 2 data, are considered in a simulation study. See Section 3.2.1 for an introduction to the data. The likelihood function for the Heffernan and Tawn model is derived in Section 3.2.2. The gradient and Hessian matrix of the likelihood function and the expected Fisher information matrix are derived in Section 3.2.3. Identifiability of the Heffernan and Tawn model parameters is questioned in Section 3.2.4, and it is shown that the observed– and expected Fisher information matrix are non-invertible unless $\mu = 0$. Bias and variance of the maximum likelihood estimator for the Heffernan and Tawn model parameters is discussed in Section 3.2.6. Bootstrapping provides a pragmatic means of quantifying uncertainty when asymptotic normality of the maximum likelihood estimator does not hold. It is briefly discussed in Section 3.2.7.

3.2.1 Data for simulation study

Two different data samples, referred to as Case 1 and Case 2, are introduced. The main objective of this section is to emphasize key features of the data. Thorough knowledge of the data helps to correctly interpret results of the simulation studies presented in subsequent sections and Chapter 4.

Attention is restricted to bivarate distributions as it is a natural starting point and communicating results is straightforward. Symmetric distributions are considered such that it suffices to present the results for Y_2 given $Y_1 > u$,

and the exchangeability issue discussed in Section 3.1.6 is avoided. The two different cases that are considered are:

- **Case 1**: bivariate Gaussian distribution, see Section 3.2.1 and,
- **Case 2**: bivariate generalized extreme value distribution with symmetric logistic dependence function, see Section 3.2.1.

These particular distributions are chosen as they cover both asymptotic independence and asymptotic dependence, as shown in Section 2.2.6.

As discussed in Section 2.1.3, extreme value analysis requires defining a threshold. The 95%-quantile of the marginal distributions is assumed to be an appropriate choice for the threshold u. This claim is substantiated by Figure 3.2 and 3.3 for Case 1 and Case 2 respectively. A high threshold is preferred to reduce bias in parameter estimates. The total sample size $n_{\mathrm{T}} = 3 \cdot 10^4$ is chosen such that the extreme set A will contain $n = 1500$ threshold exceedances. This sample size is assumed to be sufficiently large to mitigate bias and variance in the maximum likelihood estimator related to the sampling error.

Independent and identically distributed samples are considered in this chapter. Data is transformed such that each of the marginal distributions is equal to the unit Laplace scale, defined by (2.18).

Case 1: the Gaussian distribution

Let the random variable $\mathbf{Y} = (Y_1, Y_2)$ be bivariate Gaussian distributed with zero mean and unit variance, i.e.

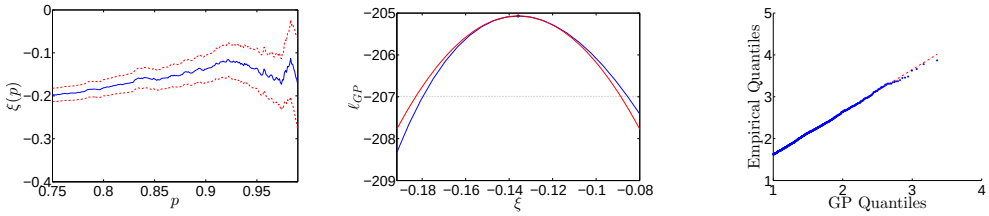
$$\mathbf{Y} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}), \quad \text{where} \quad \mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

Let $\rho = 1/2$ such that Y_1 and Y_2 are positively dependent. Table 3.1 shows that for Gaussian data $\alpha_{\mathrm{T}} = \rho^2$ and $\beta_{\mathrm{T}} = 1/2$. In order to compare the performance of different inferential methods for the Heffernan and Tawn model, take $\sigma_1 = \sigma_2 = 1$, such that $\hat{\alpha}_{\mathrm{MLE}}$ is expected to be equal to $\rho^2 = 1/4$. Simulation from the multivariate Gaussian distribution is straightforward.

Figure 3.2 provides weak evidence that the 95% quantile is an appropriate choice for the threshold u. From the threshold u onward, the generalized Pareto shape parameter $\hat{\xi}$ should show stable behavior. Although a straight line could be drawn trough the 95% confidence interval from the 95% quantile onward, the median of the sampling distribution of $\hat{\xi}_{\mathrm{MLE}}$ does certainly not show stability, as shown in Figure 3.2(A). Uncertainty regarding the maximum likelihood estimate for $p = 0.95$ is visualized in Figure 3.2(B). This plot can be interpreted as a cross section of the sampling distribution shown in Figure 3.2(A) at $p = 0.95$. As reported in Table 2.1, the true value of the shape parameter ξ_{T}

for the Gaussian distribution is equal to 0. Figure 3.2(B) shows the maximum likelihood estimate $\hat{\xi}_{\text{MLE}}$ is significantly different from 0. Both figures confirm that convergence of $\hat{\xi}$ to ξ_{T} is slow, in accordance with the statement in Section 2.1.2.

The quantile-quantile plot against theoretical quantiles of the generalized Pareto distribution, shown in Figure 3.2(C), suggest that when the tail of distribution is assumed to start at the 95% quantile, it is well approximated by the generalized Pareto distribution with parameters $\hat{\xi}_{\text{MLE}} = -0.14_{[-0.19,-0.09]}$ and $\hat{\sigma}_{u,\text{MLE}} = 0.48_{[0.45,0.52]}$.



(A) $\hat{\xi}(p)$. (B) Profile likelihood ξ. (C) Quantile-Quantile plot.

FIGURE 3.2: Diagnostic plots which support the claim that the 95% quantile is an appropriate choice for the threshold u for Case 1 data. The sampling distribution of $\hat{\xi}_{\text{MLE}}$ is shown in Figure 3.2(A) and is summarized by its median (—) and 95% symmetric confidence interval (- · -), based on the estimated variance of the estimator. The intersection between the exact profile likelihood (—) and (· · ·) shown in Figure 3.2(B), as well as the intersection of (· · ·) and the Taylor series expansion around the maximum likelihood estimate for the scale parameter as a function of shape parameter (—) yield two different 95% confidence intervals for $\hat{\xi}_{\text{MLE}}$. A quantile-quantile plot against theoretical quantiles of the generalized Pareto distribution is shown in Figure 3.3(C).

Case 2: the GEV distribution with symmetric logistic dependence function

The generalized extreme value distribution with symmetric logistic dependence function was proposed by Coles et al. (1991). In a bivariate context, the exponent measure for the symmetric logistic dependence function is given by

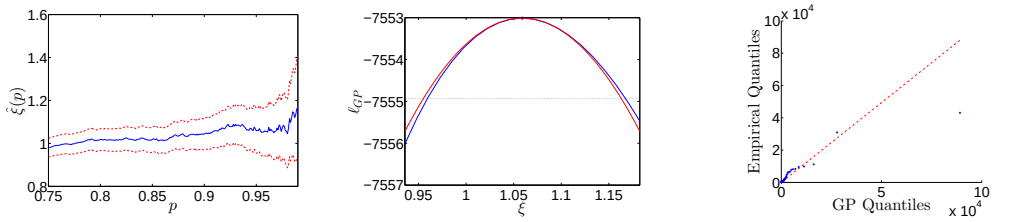
$$V(y_1, y_2) = \left(y_1^{-1/\rho} + y_2^{-1/\rho}\right)^{\rho}.$$

The spectral distribution function $H(\mathrm{d}w)$ defined by (2.23), for the symmetric logistic dependence function is differentiable, such that for $0 < w < 1$ and $0 < \rho < 1$, $h(w)$ is given by

$$h(w) := \frac{\rho^{-1} - 1}{2} \{w(1-w)\}^{-\frac{1+\rho}{\rho}} \left\{w^{-1/\rho} + (1-w)^{-1/\rho}\right\}^{\rho-2}.$$

This model yields independence and perfect dependence between Y_1 and Y_2 when $\rho \uparrow 1$ and $\rho \downarrow 0$ respectively. Let $\rho = 1/2$ such that Y_1 and Y_2 are positive asymptotically dependent. Recall from Table 3.1 that for the multivariate extreme value distribution with symmetric logistic dependence function, $\alpha_{\mathrm{T}} = 1$ and $\beta_{\mathrm{T}} = 0$. The methodology proposed by Stephenson (2003) is used to simulate the data.

Figure 3.3 provides diagnostic plots in favor of the claim that the 95% quantile is an appropriate choice for the high threshold u for Case 2 data. Figure 3.3(A) shows that $\hat{\xi}_{\mathrm{MLE}}$ is stable at the 95% quantile. The profile likelihood shown in Figure 3.3(B) suggests that $\hat{\xi}_{\mathrm{MLE}}$ is close to $\xi_{\mathrm{T}} = 1$. The maximum likelihood estimates and associated confidence intervals are given by $\hat{\xi}_{\mathrm{MLE}} = 1.06_{[0.96,1.16]}$ and $\hat{\sigma_u}_{\mathrm{MLE}} = 19.6_{[17.7,21.7]}$. These arguments provide sufficient evidence to legitimize the assumption that the 95% marginal quantile is a decent choice for the threshold u. The quantile-quantile plot shown in Figure 3.3(C) shows the generalized Pareto distribution does not fit the data particularly very well.



(A) $\hat{\xi}(p)$. (B) Profile likelihood ξ. (C) Quantile-Quantile plot.

FIGURE 3.3: Diagnostic plots which support the claim that the 95% quantile is an appropriate choice for the threshold u for Case 2 data. The sampling distribution of $\hat{\xi}_{\mathrm{MLE}}$ is shown in Figure 3.3(A) and is summarized by its median (—) and 95% symmetric confidence interval (- · -), based on the estimated variance of the estimator. The intersection between the exact profile likelihood (—) and (· · ·) shown in Figure 3.3(B), as well as the intersection of (· · ·) and the Taylor series expansion around the maximum likelihood estimate for the scale parameter as a function of shape parameter (—) yield two different 95% confidence intervals for $\hat{\xi}_{\mathrm{MLE}}$. A quantile-quantile plot against theoretical quantiles of the generalized Pareto distribution is shown in Figure 3.3(C).

3.2.2 Likelihood function for the Heffernan and Tawn model

The purpose of maximum likelihood methods is to estimate model parameters given the observed data. The likelihood function for the Heffernan and Tawn model with Gaussian residual distribution is derived. Maximum likelihood estimates for the parameters of the Heffernan and Tawn model for both

Case 1 and Case 2 are presented. Peculiarities of the likelihood function and maximum likelihood estimator are discussed.

It is natural to compare the performance of an estimator relative to the true value. Given the distribution of the data under scrutiny, the true parameter values $\boldsymbol{\theta}_{\mathrm{T}}$ are known from Table 3.1. A schematic overview is shown in Figure 3.4. For the Heffernan and Tawn model being specified by (3.13), the probabil-

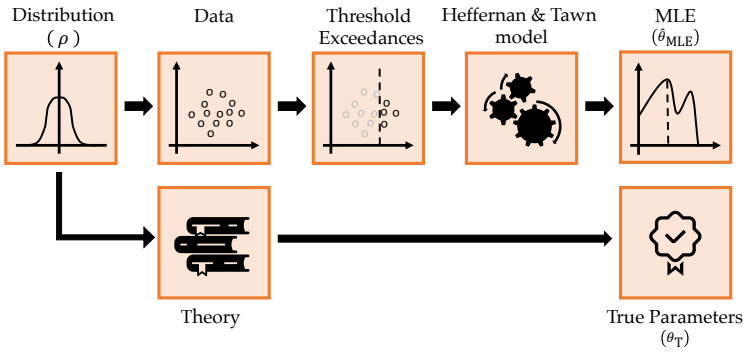


FIGURE 3.4: Different steps in the inference methodology, and how the true parameters $\boldsymbol{\theta}_{\mathrm{T}}$ can be compared to the maximum likelihood estimates $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}$.

ity density function of Y_2 conditional on $Y_1 = y$, where $y > u$ to ensure that Y_1 is extreme, is given by

$$f_{Y_2|Y_1=y_1}\left(y_2 \mid \boldsymbol{\theta}, y_1\right) = \frac{1}{\sqrt{2\pi y_1^{2\beta}\psi^2}} \exp\left\{-\frac{1}{2}\frac{\left(y_2 - \alpha y_1 - y_1^{\beta}\mu\right)^2}{y_1^{2\beta}\psi^2}\right\}.$$

The subscripts of the parameters are left out as these are clear from the context, i.e. $\boldsymbol{\theta} := \boldsymbol{\theta}_{2|1}$. The likelihood function $L\colon \boldsymbol{\theta} \to [0, \infty)$ is a measure of the likelihood of observing a sample $\boldsymbol{y}$ of $\mathbf{Y}$, given the parameters $\boldsymbol{\theta} = \{\alpha, \beta, \mu, \psi^2\}$. The likelihood function is defined by

$$L_{\mathrm{HT}}\left(\boldsymbol{\theta} \mid \boldsymbol{y}\right) := \prod_{l=1}^{n} f_{Y_2|Y_1=y_1}\left(y_{2l} \mid \boldsymbol{\theta}_l, y_{1l}\right) \tag{3.17}$$

The negative log-likelihood function of the Heffernan and Tawn model specified by (3.13) is given by

$$\begin{aligned}\bar{\ell}_{\text{HT}}(\boldsymbol{\theta} \mid \boldsymbol{y}) &= -\sum_{l=1}^{n} \log f_{Y_2|Y_1=y_1}\left(y_{2l} \mid \theta_l, y_{1l}\right), \\ &= \frac{n}{2}\log(2\pi) + \frac{n}{2}\log\left(\psi^2\right) + \beta\sum_{l=1}^{n}\log\left(y_{1l}\right) \\ &\quad + \frac{1}{2}\sum_{l=1}^{n}\frac{\left(y_{2l} - \alpha y_{1l} - y_{1l}^{\beta}\mu\right)^2}{y_{1l}^{2\beta}\psi^2}.\end{aligned} \tag{3.18}$$

Minimizing the negative log-likelihood (3.18) is equivalent to maximizing (3.17) as the logarithm is a monotonic transformation. The former method is preferred as it is superior in terms of numerical stability. Let $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ denote the maximum likelihood estimate of the parameters $\boldsymbol{\theta}$, which is defined by

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} \coloneqq \arg\max_{\boldsymbol{\theta}} L_{\text{HT}}\left(\boldsymbol{\theta} \mid \boldsymbol{y}\right) = \arg\min_{\boldsymbol{\theta}} \bar{\ell}_{\text{HT}}\left(\boldsymbol{\theta} \mid \boldsymbol{y}\right).$$

The maximum likelihood estimator is known to exhibit certain asymptotic properties. First of all, the maximum likelihood estimator is *consistent*, i.e. $\hat{\boldsymbol{\theta}}_{\text{MLE}} \xrightarrow{p} \boldsymbol{\theta}_{\text{T}}$, where $\boldsymbol{\theta}_{\text{T}}$ denotes the true parameter values. Secondly, the *asymptotic normality* property states that the maximum likelihood estimator converges to the true parameter value, i.e.

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_{\text{MLE}} - \boldsymbol{\theta}_{\text{T}}\right) \xrightarrow{d} \mathcal{N}\left\{0, \mathcal{I}\left(\boldsymbol{\theta}_{\text{T}}\right)\right\}, \quad \text{as} \quad n \to \infty, \tag{3.19}$$

where $\mathcal{I}\left(\boldsymbol{\theta}\right)$ denotes the expected Fisher information matrix[1]. These properties are discussed in more detail in Section 3.2.4 and 3.2.6.

Thorough understanding of the likelihood function is important to understand to what extend the aforementioned properties apply to the likelihood function of the Heffernan and Tawn model. Visual exploration of the four dimensional parameter space $\Omega_{\boldsymbol{\theta}}$ is not straightforward. Pairwise profile likelihood contours are shown in Figure 3.5 and 3.6. These contours are obtained by fixing two parameters to their respective maximum likelihood estimate, while varying the other variables. This approach provides a rough idea of the geometry of the parameter space and the likelihood function in the vicinity of the maximum likelihood estimates. On the contrary, if the parameters being fixed are not equal their maximum likelihood estimates, the geometry of the parameter space can deviate significantly. See Figure 3.5 and 3.6, left of the diagonal, for the profile negative log-likelihood contours for Case 1 and Case 2

[1]The expected Fisher information matrix is formally introduced in Section 3.2.3

respectively.

The impact of imposing the constraints which are introduced in Section 3.1.4, is shown by the plots right of the diagonal. The colored surfaces indicate for what subspace of $\Omega_{\boldsymbol{\theta}}$ the profile likelihood function is defined when the constraints are imposed. Several features of the likelihood function that stand out in Figure 3.5 and 3.6 are briefly summarized.

1. The range of the likelihood function shown on the vertical axis of the plots on the diagonal, reveals that changes in $\hat{\alpha}$ and $\hat{\beta}$ have an order of magnitude bigger impact on the negative log-likelihood function than similar changes in $\hat{\mu}$ or $\hat{\psi}^2$.
2. The likelihood is virtually flat along a ridge in the unrestricted likelihood contour plots for $\hat{\alpha}$ – $\hat{\mu}$ and $\hat{\beta}$ – $\hat{\psi}^2$.
3. The color of the contour lines and the distance between likelihood contours suggest the *curvature* of the parameter space $\Omega_{\boldsymbol{\theta}}$ is non-constant.
4. The geometry of the profile likelihood surfaces for Case 1 and Case 2 are similar, up to a shift towards the boundary of the parameter space for the latter case.
5. Maximum likelihood estimates for asymptotically independent data are in the interior of the parameter space $\Omega_{\boldsymbol{\theta}}$. For asymptotically dependent data, the maximum likelihood estimates are on the boundary of the parameter space.

The observations for the unconstrained profile likelihood contours also apply to the plots shown right of the diagonal. Some additional properties for the constrained likelihood surfaces are listed below.

1. Maximum likelihood estimates are not affected by imposing the Keef constraints for asymptotically independent data.
2. The constrained maximum likelihood estimates are on — or very close to — the boundary of the feasible parameter space for both asymptotically dependent data and asymptotically independent data.
3. The curvature is unaffected.
4. The nuisance parameter space $\Omega_{\mu\times\psi^2}$ is unaffected by the constraints.
5. The scale of the vertical axis in the $\hat{\mu}$ – $\hat{\beta}$ and $\hat{\psi}^2$ – $\hat{\beta}$ plots is incredibly small, which is related to the maximum likelihood estimate being in a corner of the parameter space, as shown in the $\hat{\beta}$ – $\hat{\alpha}$ plot. By fixing $\hat{\alpha} = \hat{\alpha}_{\text{MLE}}$, any change in $\hat{\beta}$ will move $\hat{\beta}$ outside the feasible parameter space.

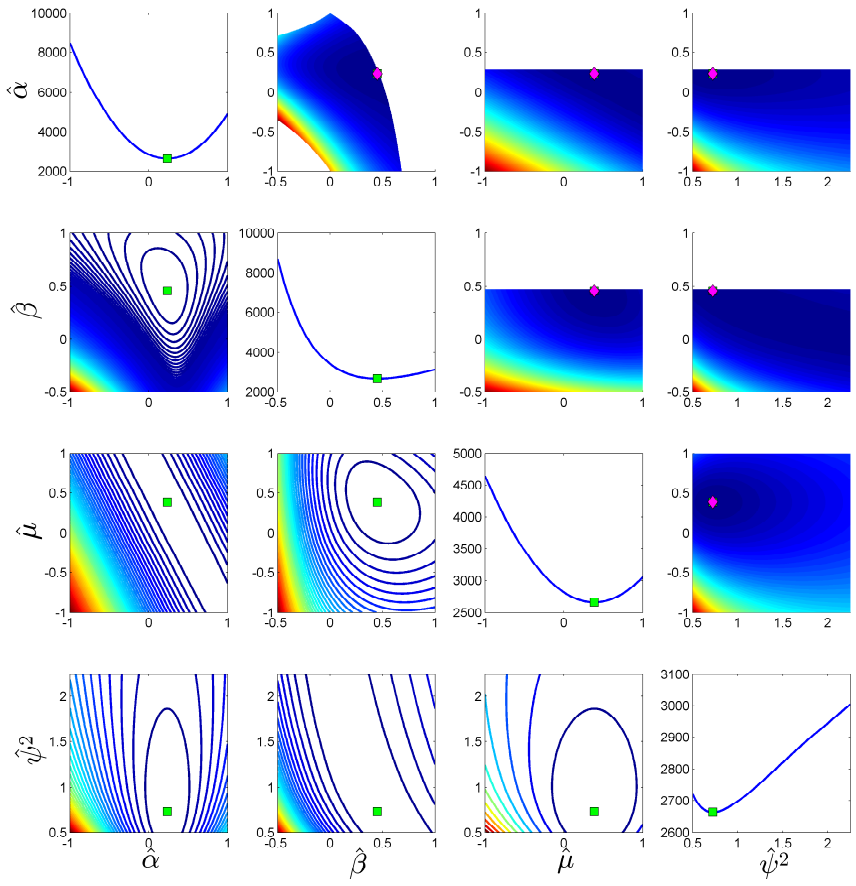


FIGURE 3.5: Profile negative log-likelihood contours around the maximum likelihood estimates (■) for the parameters of the Heffernan and Tawn model for Case 1. The unconstrained likelihood contours (shown left of the diagonal) are spaced such that each contour marks a 250 unit increase in negative log likelihood. The profile negative log-likelihood surfaces on the right of the diagonal show the impact of imposing the conditions proposed by Keef et al. (2013) on the parameter space as well as the feasible maximum likelihood estimates (♦).

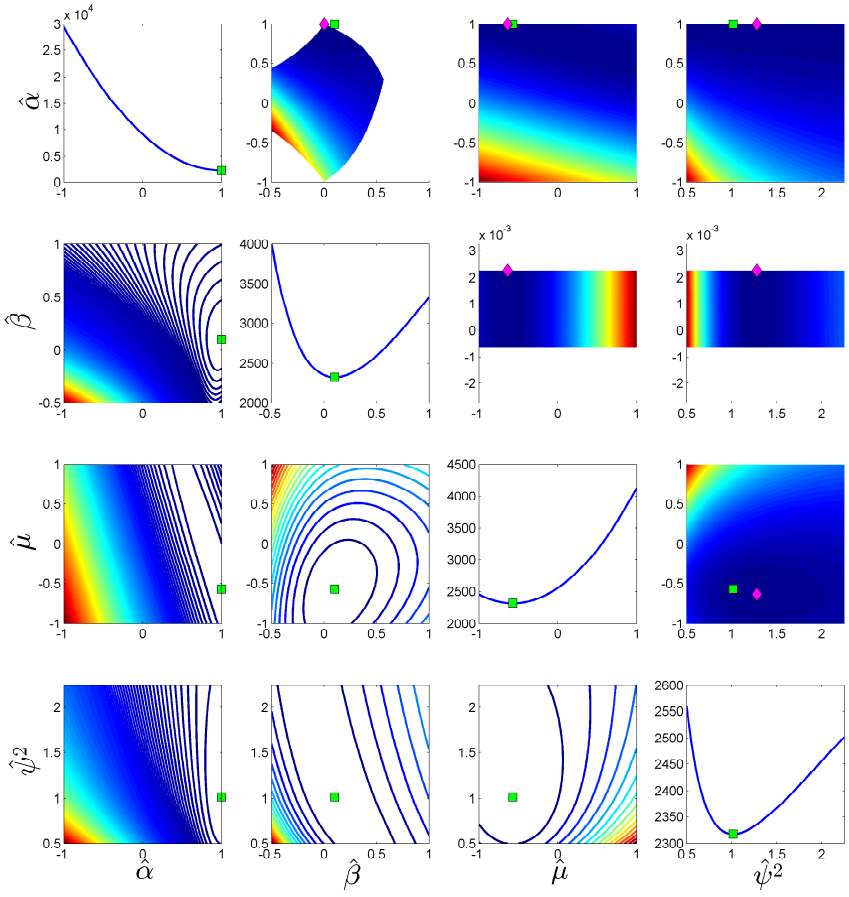


FIGURE 3.6: Profile negative log-likelihood contours around the maximum likelihood estimates (■) for the parameters of the Heffernan and Tawn model for Case 2. The unconstrained likelihood contours (shown left of the diagonal) are spaced such that each contour marks a 250 unit increase in negative log likelihood. The profile negative log-likelihood surfaces on the right of the diagonal show the impact of imposing the conditions proposed by Keef et al. (2013) on the parameter space and the maximum likelihood estimates (♦).

3.2.3 Curvature of the likelihood surface

The *curvature* of the negative log-likelihood function exhibits several interesting features. The *gradient* of (3.18) is defined by

$$\nabla_{\theta}\bar{\ell}_{\text{HT}} \coloneqq \left(\frac{\partial\bar{\ell}_{\text{HT}}}{\partial\alpha} \quad \frac{\partial\bar{\ell}_{\text{HT}}}{\partial\beta} \quad \frac{\partial\bar{\ell}_{\text{HT}}}{\partial\mu} \quad \frac{\partial\bar{\ell}_{\text{HT}}}{\partial\psi^2}\right)^{\intercal}, \tag{3.20}$$

where $\bar{\ell}_{\text{HT}} \coloneqq \bar{\ell}_{\text{HT}}\left(\boldsymbol{\theta} \mid \boldsymbol{y}\right)$. The *Hessian matrix* of (3.18) defines the *observed Fisher information matrix*

$$\mathcal{J}\left(\boldsymbol{\theta}\right) \coloneqq \nabla\nabla^{\intercal}\bar{\ell}_{\text{HT}} = \begin{pmatrix} \frac{\partial^2\bar{\ell}_{\text{HT}}}{\partial\alpha^2} & \frac{\partial^2\bar{\ell}_{\text{HT}}}{\partial\alpha\partial\beta} & \frac{\partial^2\bar{\ell}_{\text{HT}}}{\partial\alpha\partial\mu} & \frac{\partial^2\bar{\ell}_{\text{HT}}}{\partial\alpha\partial\psi^2} \\ \frac{\partial^2\bar{\ell}_{\text{HT}}}{\partial\alpha\partial\beta} & \frac{\partial^2\bar{\ell}_{\text{HT}}}{\partial\beta^2} & \frac{\partial^2\bar{\ell}_{\text{HT}}}{\partial\beta\partial\mu} & \frac{\partial^2\bar{\ell}_{\text{HT}}}{\partial\beta\partial\psi^2} \\ \frac{\partial^2\bar{\ell}_{\text{HT}}}{\partial\alpha\partial\mu} & \frac{\partial^2\bar{\ell}_{\text{HT}}}{\partial\beta\partial\beta} & \frac{\partial^2\bar{\ell}_{\text{HT}}}{\partial\mu^2} & \frac{\partial^2\bar{\ell}_{\text{HT}}}{\partial\mu\partial\psi^2} \\ \frac{\partial^2\bar{\ell}_{\text{HT}}}{\partial\alpha\partial\psi^2} & \frac{\partial^2\bar{\ell}_{\text{HT}}}{\partial\beta\partial\psi^2} & \frac{\partial^2\bar{\ell}_{\text{HT}}}{\partial\mu\partial\psi^2} & \frac{\partial^2\bar{\ell}_{\text{HT}}}{\partial\psi^2\partial\psi^2} \end{pmatrix}. \tag{3.21}$$

The *expected Fisher information* is defined as $\mathcal{I}\left(\boldsymbol{\theta}\right) \coloneqq \mathrm{E}\left\{\mathcal{J}\left(\boldsymbol{\theta}\right)\right\}$. For the likelihood function of Y_2 conditional on $Y_1 = y$, given by (3.18), the entries of the expected Fisher information matrix are given by

$$\mathcal{I}\left(\boldsymbol{\theta}\right)_{ij} = \mathrm{E}\left[\left(\frac{\partial}{\partial\theta_i}\bar{\ell}_{\text{HT}}\right)\left(\frac{\partial}{\partial\theta_j}\bar{\ell}_{\text{HT}}\right)\middle|\boldsymbol{\theta}, y\right]. \tag{3.22}$$

The explicit expressions for the gradient, observed– and expected Fisher information matrix are presented in Appendix A.4 and A.5. If the expected Fisher information is available, it is preferred over the observed Fisher information as it is more stable. In addition, Cao (2013) shows that "the inverse of the expected Fisher information evaluated at $\hat{\boldsymbol{\theta}}_{\text{MLE}}$, outperforms the inverse of the observed Fisher information matrix under a mean squared error criterion".

The gradient and expected Fisher information matrix for the Heffernan and Tawn model likelihood function exhibit special properties when evaluated at the maximum likelihood estimate $\hat{\boldsymbol{\theta}}_{\text{MLE}}$. A necessary condition for $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ to be a proper maximum likelihood estimate states that if $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ lies in the interior of $\Omega_{\boldsymbol{\theta}}$, then

- $\nabla_{\boldsymbol{\theta}}\bar{\ell}_{\text{HT}}\left(\hat{\boldsymbol{\theta}}_{\text{MLE}}\middle|\boldsymbol{y}\right) = \mathbf{0}$, and,
- $\mathcal{I}\left(\hat{\boldsymbol{\theta}}_{\text{MLE}}\right)$ is a positive semi-definite matrix.

The former property enables the highly efficient Newton-Raphson method to find the roots of (3.18). Semi-positive definiteness is particularly useful because it ensures invertibility of the expected Fisher information matrix. The asymptotic normality property then states that the inverse of the expected Fisher information matrix defines an appropriate approximation to a covariance matrix

that summarizes the uncertainty regarding the maximum likelihood estimator $\boldsymbol{\theta}_{\text{MLE}}$.

If, on the contrary, $\hat{\boldsymbol{\theta}}_{\text{MLE}} \in \partial\Omega_{\boldsymbol{\theta}}$, which is the case for asymptotically dependent random variables as $\alpha = 1$, or when the constraints proposed by Keef et al. (2013) are imposed, as shown in Figure 3.5 and 3.6. As long as the unconstrained maximum likelihood estimate is feasible under the constraints, the gradient of the likelihood surface will still be equal to zero when evaluated at $\hat{\boldsymbol{\theta}}_{\text{MLE}}$. However, as shown in Figure 3.6, this is not the case for asymptotically dependent data.

The expected Fisher information matrix can be evaluated on the boundary of the parameter space, as long as the second derivatives of the likelihood function are defined on the boundary of the parameter space. However, symmetric confidence intervals based on inverting the expected Fisher information matrix are clearly not an appropriate way of quantifying uncertainty. Least of all because half the confidence interval will fall outside the parameter space.

Minimizing functions near the boundary of the parameter space is also tricky from a numerical perspective. Algorithms typically fail to identify a maximum or minimum exactly on the boundary of the parameter space because of convergence criteria. This will introduce bias in the maximum likelihood estimates, although for well-behaved functions this bias will typically be small in magnitude. The implications of maximum likelihood estimates being on the boundary of the parameter space on negative log-likelihood minimization are discussed by Self and Liang (1987) and Feng and McCulloch (1992). The authors propose slack constraints and reparameterizations to address the issue. In Chapter 4 the issue will be revisited when encountered in Bayesian inference.

3.2.4 Identifiability of the model parameters

The generic definition of parameter identifiability states that the parameters of the Heffernan and Tawn model are *identifiable*, if for all $\boldsymbol{\theta} \in \Omega_{\boldsymbol{\theta}}$,

$$\boldsymbol{\theta} \neq \hat{\boldsymbol{\theta}}_{\text{MLE}} \quad \Leftrightarrow \quad f_{Y_2|Y_1=y}\left(y_2 \mid \boldsymbol{\theta}, y_1\right) \neq f_{Y_2|Y_1=y}\left(y_2 \mid \hat{\boldsymbol{\theta}}_{\text{MLE}}, y_1\right). \tag{3.23}$$

Issues regarding the identifiability of β have been raised by Cheng et al. (2014), who postulate that β can not be identified from μ or ψ^2.

Identifiability of the parameters of the Heffernan and Tawn model is related to the parameter interactions in (3.18). These interactions are induced by the normalizing functions (3.8). Under the assumption that the residual $Z_{2|1}$ defined by (3.9) is Gaussian, the first two moments of Y_2 given $Y_1 = y$, which

define the Gaussian distribution of the residuals, are given by

$$\mathrm{E}\left\{T_{\mathrm{L}}\left(Y_2\right) \mid T_{\mathrm{L}}\left(Y_1\right) = y\right\} = \alpha y + y^{\beta}\mu \quad \text{and} \quad \mathrm{var}\left\{T_{\mathrm{L}}\left(Y_2\right) \mid T_{\mathrm{L}}\left(Y_1\right) = y\right\} = y^{2\beta}\psi^2. \tag{3.24}$$

It is clear from (3.24) that if $\mu = 0$, the interaction between α and β as well as α and μ is broken. On the other hand, if $\beta \approx 1$, then α can not be distinguished from μ.

The strong interaction between α and μ, as well as β and ψ^2, results in a diagonal ridge in the profile likelihood contours shown in Figure 3.5 and 3.6. The likelihood function is approximately constant along this ridge.

This suggests non-identifiability or *parameter redundancy* — i.e. the problem can be fully parameterized by a subset of $\boldsymbol{\theta}$ — of the Heffernan and Tawn model parameters. If the rank of the expected Fisher information matrix is less than the number of model parameters, then the model is parameter redundant. Catchpole and Morgan (1997) show that "if a model is parameter redundant, then it is not locally identifiable", and Rothenberg (1971) proves that under mild conditions, a model is locally identifiable in a neighborhood of $\boldsymbol{\theta}$, if and only if the expected Fisher information matrix $\mathcal{I}\left(\boldsymbol{\theta}\right)$ is invertible. Although the converse of this statement does not hold, noninvertibility of the expected Fisher information indicates some defect in the parameterization of the Heffernan and Tawn model. Noninvertibility of the expected Fisher information matrix is discussed in Section 3.2.5. It is apparent from (3.24) that the only case when the model is truly parameter redundant is when $\beta = 1$. This case will not be encountered in practice as this might only happen for perfect dependence for which the limit distribution $G_{|i}$ is degenerate anyway. It is concluded that the Heffernan and Tawn model is not parameter redundant.

3.2.5 Noninvertibility of the Fisher information matrix

The expected Fisher information matrix has many useful applications when it is invertible and positive definite. In practice, two different problems can arise. First of all, the matrix can be *singular*, which means that its inverse does not exist. Secondly, the matrix can be non-positive definite, which means that although the inverse of a matrix may exist, it does not define a proper covariance matrix. A matrix that is positive definite is non-singular, but the converse is not necessarily true. Hence in this context, it suffices to study whether the observed– and expected Fisher information matrix is non-positive definite.

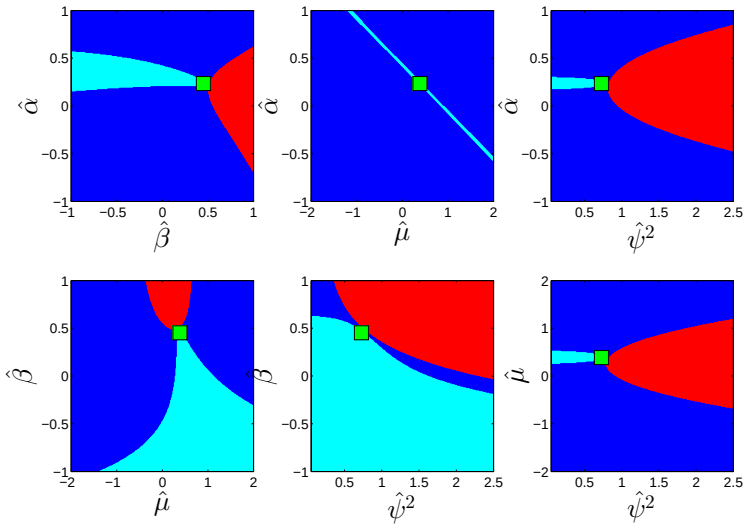
A $d \times d$ matrix $\mathbf{G}$ is said to *positive semi-definite* if $\mathbf{x}^{\intercal}\mathbf{G}\mathbf{x} \geq 0$, for all $\mathbf{x} \in \mathbb{R}^d$. There are several properties of positive semi-definite matrices that can be verified in order to establish whether the expected Fisher information matrix is positive definite. A matrix $\mathbf{G}$ is said to be positive semi-definite, if and only if:

1. All eigenvalues of the matrix $\mathbf{G}$ are non-negative,
2. All its *leading principal minors* are non-negative, where the *n-leading principal minor* is defined as the determinant of the upper left $n \times n$ sub-matrix, and,
3. There exists a unique *Cholesky decomposition* of the form $\mathbf{G} = \mathbf{L}\mathbf{L}^{\intercal}$, where L is a lower-triangular matrix.

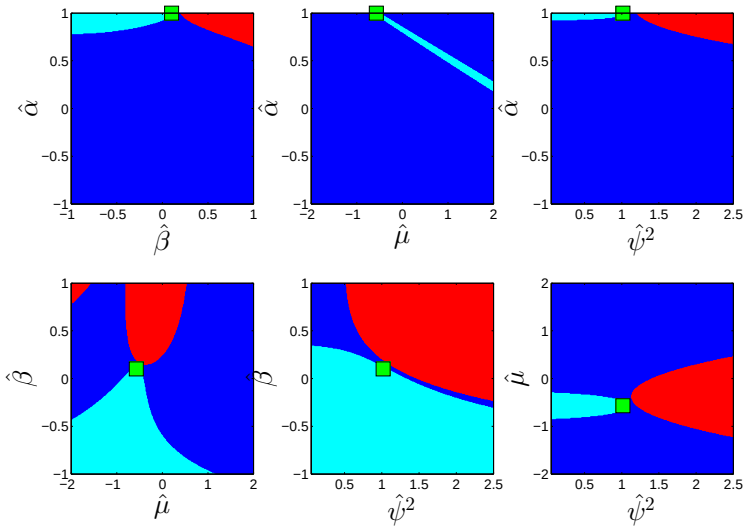
The first property can be easily tested. Figure 3.7 shows what values of $\boldsymbol{\theta}$ yield negative eigenvalues for the observed Fisher information matrix for Case 1 and Case 2. The subspace of $\Omega_{\boldsymbol{\theta}}$ that yields non-negative eigenvalues is marked by the cyan colored surfaces. A similar plot for the expected Fisher information matrix is included in Appendix B in Figure B.2.

The narrow ridges in the $\alpha - \mu$ plots in both Figure 3.7(A)-(B) resemble the ridge in the $\alpha - \mu$ plot in Figure 3.5, and indicate that all eigenvalues are positive along this ridge. For negative eigenvalues, moving along an eigenvector associated to a negative eigenvalue should further decrease the negative log-likelihood. Conversely, as all eigenvalues are positive along the narrow cyan colored ridge, this supports the observation in Section 3.2.2 that the negative log-likelihood is flat along the ridge in the $\alpha - \mu$ plot in Figure 3.7(A). As a consequence, because negative eigenvalues in the observed Fisher information matrix imply non-positive semi-definiteness, which in turn implies non-invertibility, only the confined subspace of $\Omega_{\boldsymbol{\theta}}$ indicated by the cyan colored ridge in Figure 3.7 yields an invertible observed Fisher information matrix. The issue is even more profound for the expected Fisher information matrix, as shown in Figure B.2.

Negative eigenvalues for the observed Fisher information matrix are related to the determinant being negative. Positive definiteness can be forced upon the observed Fisher information matrix by forcing off-diagonal elements to zero, such that each of the leading principal minors are non-negative. In order to preserve the strong interaction between the pairs α– μ and $\beta - \psi^2$, set all non-diagonal elements in the observed Fisher information matrix to zero, except for the entries associated to these pairs. The resulting matrix is referred to as the restrained observed Fisher information matrix and is denoted by $\mathcal{J}^{\mathrm{R}}(\boldsymbol{\theta})$. The union of the blue and cyan colored areas indicates what subspace of $\Omega_{\boldsymbol{\theta}}$ yields a positive semi-definite restrained observed Fisher information matrix. It is remarkable that for certain values of $\boldsymbol{\theta}$, neither the full– nor the restrained observed Fisher information matrix is positive semi-definite. As shown in Figure B.2 the restrained expected Fisher information matrix is stable for each $\boldsymbol{\theta} \in \Omega_{\boldsymbol{\theta}}$.



(A) Case 1: Asymptotically independent data.



(B) Case 2: Asymptotically dependent data.

FIGURE 3.7: Subspace of $\Omega_{\boldsymbol{\theta}}$ that yields non-negative eigenvalues for the observed Fisher information matrix $\mathcal{J}(\boldsymbol{\theta})$ (cyan) and restrained observed Fisher information matrix $\mathcal{J}^{\mathrm{R}}(\boldsymbol{\theta})$ (blue+cyan). The area where neither matrix is semi-positive definite (red) and the maximum likelihood estimates (■) are also indicated.

3.2.6 Bias and variance of the maximum likelihood estimator

Thorough understanding of the factors that influence bias and uncertainty of the maximum likelihood estimator is import for meaningful statistical inference. Three parameters which govern the bias and variance of the maximum likelihood estimator are identified.

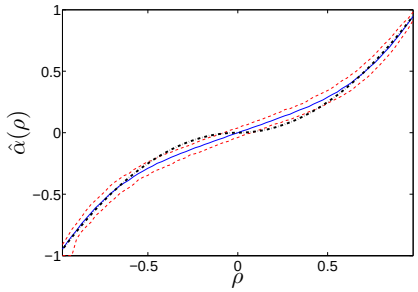
1. Sample size n_{T},
2. Dependence in the data sample, determined by ρ, and,
3. Non-exceedance probability p.

A simulation study is performed to establish a relationship between these parameters and the sampling distribution of the maximum likelihood estimator for the Heffernan and Tawn model parameters.

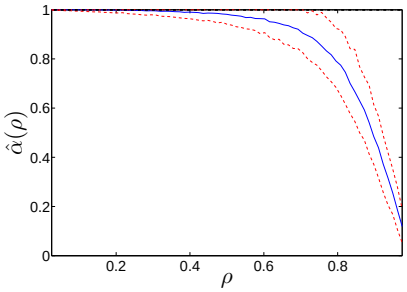
First of all, the influence of the sample size on the maximum likelihood estimator is studied. This relationship is governed by the asymptotic normality property of the maximum likelihood estimator given by (3.19). As long as the maximum likelihood estimate lies in the interior of the parameters space, the variance deflates at a rate $1/\sqrt{n_{\mathrm{T}}}$. In theory, all else being equal, increasing the total sample size deflates the variance of the maximum likelihood estimator and does not affect the bias.

Whether reality sticks to the truth is shown by the results presented in Appendix B in Figure B.3. These results confirm that for Case 1 data, the variance of the maximum likelihood estimator decreases at a rate $1/\sqrt{n_{\mathrm{T}}}$. The estimator is biased as the median of the sampling distribution is approximately constant and does not converge to the true parameter values. This confirms that increasing the sample size leads to a decreases the uncertainty and does not affect the bias. For Case 2 data, $\boldsymbol{\theta}_{\mathrm{T}} \in \partial\Omega_{\boldsymbol{\theta}}$ as $\alpha_{\mathrm{T}} = 1$. Although the median of the sampling distribution approaches 1 as $n \to \infty$ it will not truly converge to 1. Hence $\boldsymbol{\theta}_{\mathrm{T}} \in \partial\Omega_{\boldsymbol{\theta}}$ for Case 2 data is identified as a source of bias in the maximum likelihood estimator.

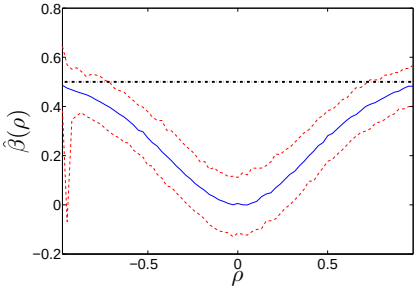
Secondly, the relationship between the strength of dependence in the data sample and the sampling distribution of the maximum likelihood estimator is studied. In Section 2.2.6 and Table 3.1, a relationship between regular dependence in the original data sample, extremal dependence among threshold exceedances and the Heffernan and Tawn model parameters was established. The results are shown in Figure 3.8. All else being equal, i.e. the total sample size ($n_{\mathrm{T}} = 10^5$) and non-exceedance probability ($p = 0.95$) are fixed, the relationship between the sampling distribution and ρ is non-trivial. In the limit $n \to \infty$ and $p \to 1$, the sampling distribution of $\hat{\alpha}_{\mathrm{MLE}}$ and $\hat{\beta}_{\mathrm{MLE}}$ converges to their true values and the variance will deflate. However, finite data samples should be expected to exhibit features similar to those shown in Figure 3.8.



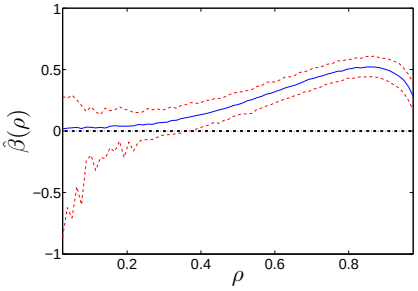
(A) Case 1: $\hat{\alpha}$.



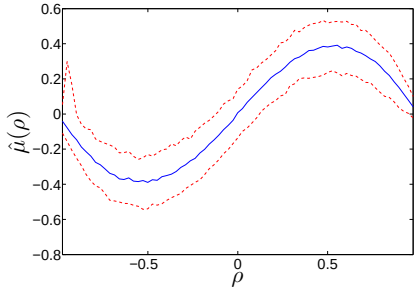
(B) Case 2: $\hat{\alpha}$.



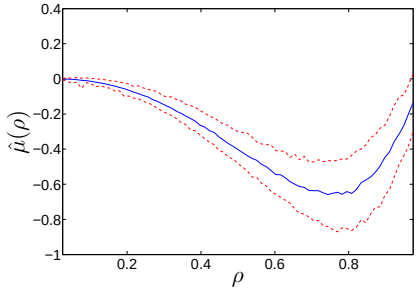
(C) Case 1: $\hat{\beta}$.



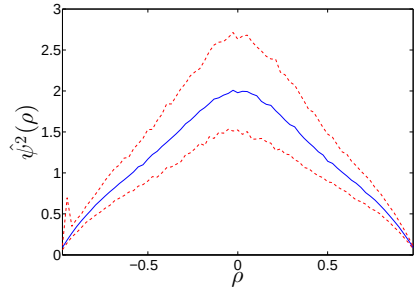
(D) Case 2: $\hat{\beta}$.



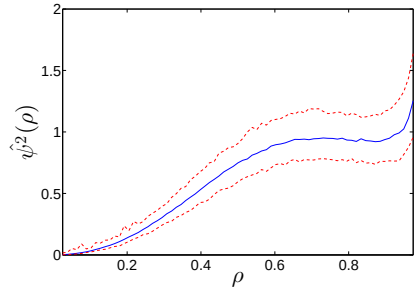
(E) Case 1: $\hat{\mu}$.



(F) Case 2: $\hat{\mu}$.



(G) Case 1: $\hat{\psi}^2$.



(H) Case 2: $\hat{\psi}^2$.

FIGURE 3.8: Influence of the strength of dependence ρ in the data on the sampling distribution of the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ for parameters of the Heffernan and Tawn model. The sample size is 10^5 and the non-exceedance probability is 0.95. The median (—), 2.5% and 97.5% empirical quantile (- - -) and true values α_{T} and β_{T} according to Table 3.1 ($\cdots$) are presented.

Directly observing bias is often not possible. However, it is implicitly defined by the mean squared error and the sample variance. The *mean squared error* is defined as

$$\mathrm{MSE}\left(\hat{\theta}_{\mathrm{MLE}}\right) \coloneqq \mathrm{E}\left[\left(\hat{\theta}_{\mathrm{MLE}} - \theta_{\mathrm{T}}\right)^2\right] = \mathrm{var}\left(\hat{\theta}_{\mathrm{MLE}}\right) + \mathrm{Bias}^2\left(\hat{\theta}_{\mathrm{MLE}}, \theta_{\mathrm{T}}\right).$$

As the true parameter θ_{T} must be known to calculate the mean squared error, the bias can only be determined for α and β. The sample variance of a sample of maximum likelihood estimates can be calculated, and the mean squared error is computed by

$$\mathrm{MSE}\left(\hat{\theta}_{\mathrm{MLE}}\right) = \sum_{l=1}^{n_{\mathrm{B}}} \left(\hat{\theta}^{(l)}_{\mathrm{MLE}} - \theta_{\mathrm{T}}\right)^2.$$

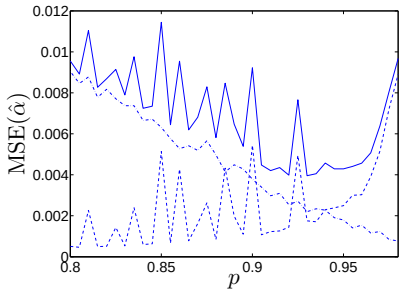
For Case 1 data, variance of the maximum likelihood estimators $\hat{\alpha}_{\mathrm{MLE}}, \hat{\beta}_{\mathrm{MLE}}$ and $\hat{\mu}_{\mathrm{MLE}}$ is approximately constant. A small bias in $\hat{\alpha}_{\mathrm{MLE}}$ is observed when $-1/2 < \rho < 1/2$. This is shown explicitly by the mean squared error $\hat{\alpha}_{\mathrm{MLE}}$ and squared bias in Appendix B in Figure B.5(A). For β on the contrary, the maximum likelihood estimator $\hat{\beta}_{\mathrm{MLE}}$ is severely biased, and the bias is non-constant as a function of ρ, see Figure B.5(C) in Appendix B. Comparing Figure 3.8(C) to Figure 3.8(G) suggests the identifiability issue between β and ψ^2 manifests itself through a significant bias in $\hat{\beta}_{\mathrm{MLE}}$. Features of the sampling distribution are symmetric around $\rho = 0$ as the Gaussian distribution used for simulating Case 1 data is symmetric.

For case 2 data, the sampling distribution behaves as expected for $0 < \rho < 1/2$ which corresponds to strong dependence. As $\rho \to 1$, there is increasingly less evidence in favor of asymptotic dependence. So rather than jumping from $\alpha = 1$ to $\alpha = 0$ when ρ is equal to one, there is a gradual decrease in $\hat{\alpha}_{\mathrm{MLE}}$ as shown in Figure 3.9(B). This also affects $\hat{\mu}_{\mathrm{MLE}}$ because of the strong interaction between these parameters.

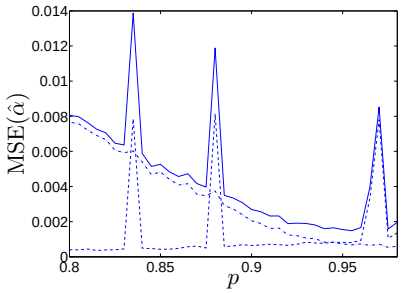
The maximum likelihood estimator for either independence and perfect dependence in Case 1 and Case 2 data should agree with each other. Independence arises for Case 1 data if $\rho = 0$, and for Case 2 data if $\rho = 1$. The median of the sampling distribution for both cases agrees on $\hat{\alpha}_{\mathrm{MLE}} = \hat{\beta}_{\mathrm{MLE}} = \hat{\mu}_{\mathrm{MLE}} = 0$ and $\hat{\psi}^2_{\mathrm{MLE}} \approx 2$. Similarly, for perfect dependence, i.e. $\rho = 1$ for Case 1 data and $\rho = 0$ for Case 2 data, $\hat{\alpha}_{\mathrm{MLE}} = 1$ and $\hat{\beta}_{\mathrm{MLE}} = \hat{\mu}_{\mathrm{MLE}} = \hat{\psi}^2_{\mathrm{MLE}} = 0$ as expected.

Summarizing, certain values of ρ lead to significant bias in the maximum likelihood estimator for β for Case 1 data, and both α and β for Case 2 data. Bias in $\hat{\alpha}_{\mathrm{MLE}}$ for Case 1 data is small in magnitude but not negligible. Variance of the maximum likelihood estimator is constant with respect to ρ for most parameters, except for $\hat{\psi}^2_{\mathrm{MLE}}$ for Case 1 data and $\hat{\mu}_{\mathrm{MLE}}$ for Case 2 data.

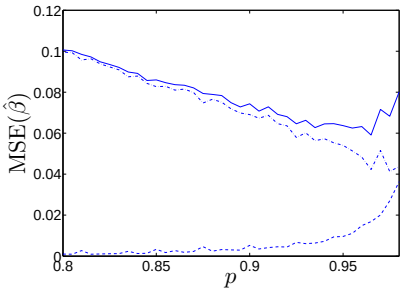
Thirdly, the influence of the non-exceedance probability p on the maximum likelihood estimator is assessed. In theory, bias should decrease and the variance should increase as $p \to 1$. A decomposition of the mean squared error in bias and variance for the maximum likelihood estimator of α and β is shown in Figure 3.9, as a function of p. As expected, bias decreases and the variance increases as $p \to 1$. The presented statistics are not robust, which causes the spikes in the sample variance shown in Figure 3.9(A)-(B). Remarkably, the mean squared error of the maximum likelihood estimator is minimized for $p \approx 0.95$.



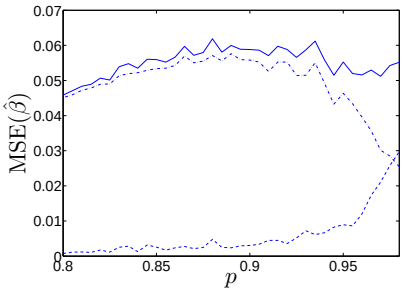
(A) Case 1: MSE $(\hat{\alpha})$.



(B) Case 2: MSE $(\hat{\alpha})$.



(C) Case 1: MSE $\left(\hat{\beta}\right)$.



(D) Case 2: MSE $\left(\hat{\beta}\right)$.

FIGURE 3.9: Mean squared error of the maximum likelihood estimator for the Heffernan and Tawn model parameters. For each value of $p \in [0.8, 0.98]$, a sample of 10^3 maximum likelihood estimates is obtained by repeatedly generating 10^5 observations from either the Gaussian distribution (Case 1) or the generalized extreme value distribution with symmetric logistic dependence function (Case 2) and fitting the Heffernan and Tawn model. Variance (- - -), squared bias (- · -) and the mean squared error (—) are shown.

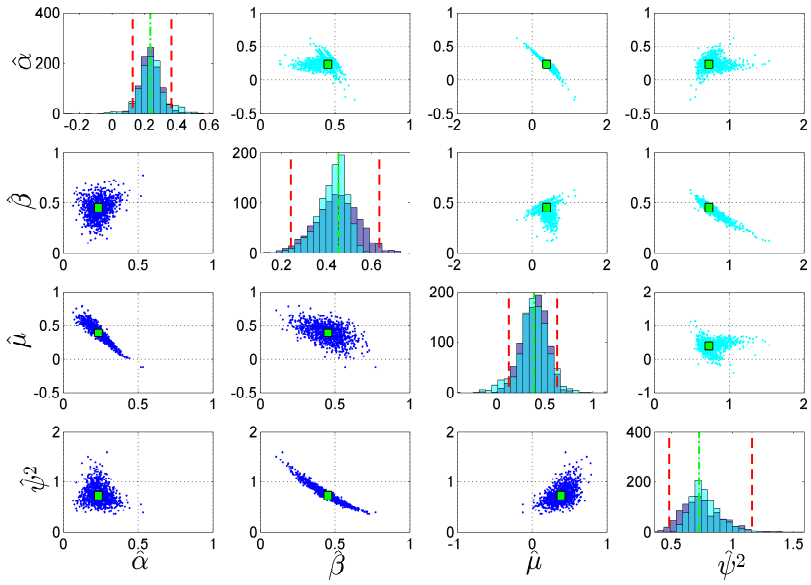
3.2.7 Bootstrapping the maximum likelihood estimator

Rather than relying on the asymptotic properties of the maximum likelihood estimator, *bootstrapping* offers an alternative approach to quantify uncertainty regarding parameter estimates. A bootstrap sample is obtained by repeatedly sampling with replacement from observed data and recomputing the maximum likelihood estimate $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ for each sample. The empirical 2.5% and 97.5% quantile of the bootstrap sample define a 95% confidence interval for $\hat{\boldsymbol{\theta}}_{\text{MLE}}$. This approach is referred to as percentile bootstrap, and is adopted in this section. There are more advanced bootstrap methods available, such as the bias-corrected and accelerated bootstrap proposed by Efron (1987). These methods can also address bias related to the sampling error or non-Gaussian features in the sampling distribution.

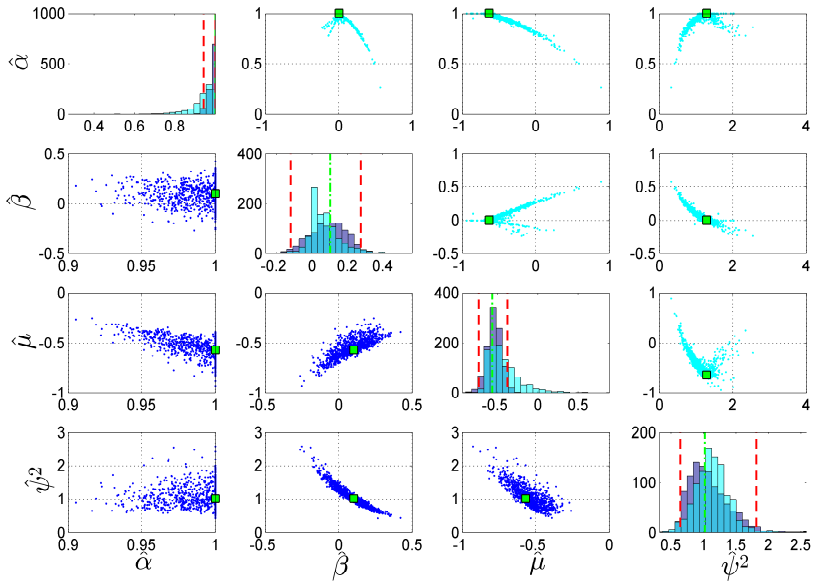
The aim of this section is to demonstrate how to quantify uncertainty regarding a parameter estimate based on bootstrapping maximum likelihood estimates. Both Case 1 and Case 2 data are considered, both with– and without the constraints proposed by Keef et al. (2013) being imposed. The results are presented in Figure 3.10.

First of all, the results for Case 1 shown in Figure 3.10(A) are briefly discussed. The highly correlated bootstrap samples for $\hat{\alpha} - \hat{\mu}$ and $\beta - \psi^2$ stand out, in accordance with the strong interaction between these variables discussed in Section 3.2.4. The histograms shown on the diagonal of Figure 3.10(A) resemble a Gaussian distribution. This is expected as the asymptotic normality property holds for Case 1 data. When the constrained Heffernan and Tawn model is considered, the histogram for β and ψ^2 shows skewness. As the maximum likelihood estimates that are feasible under the constraints are close to the boundary of the constrained parameter space, as shown in Figure 3.5, the asymptotic normality assumption is no longer appropriate.

Secondly, consider the results for Case 2 data, shown in Figure 3.10(B). Notice the strong negative correlation between $\hat{\alpha} - \hat{\mu}$, $\hat{\beta} - \hat{\psi}^2$ and $\hat{\mu} - \hat{\psi}^2$, as well as the strong positive correlation in the $\hat{\beta} - \hat{\mu}$ bootstrap samples. The constrained samples show even stronger correlation due to the maximum likelihood estimate being pushed into a corner of the parameter space, as shown in Figure 3.6.



(A) Case 1: Asymptotically independent data.



(B) Case 2: Asymptotically dependent data.

FIGURE 3.10: Scatterplot matrix for $n_B = 1000$ bootstrapped maximum likelihood estimates of the Heffernan and Tawn model parameters, with (cyan) and without (blue) the constraints proposed by Keef et al. (2013) being imposed. The (feasible) maximum likelihood estimate for the original data sample (■) is also shown.

There is a fundamental conflict between the constraints proposed by Keef et al. (2013) and quantifying uncertainty through bootstrapped maximum likelihood estimates when the maximum likelihood estimate is close to boundary of the parameter space. Since the feasible parameter space under the Keef et al. (2013) conditions depends on the data, resampling will affect the parameter space. It is not trivial how to interpret bootstrap samples that are not feasible under the constraints for the original data sample. The issue is raised in the discussion in Chapter 5. The influence of this discrepancy on higher-level statistics, such as return levels, is expected to be small.

The mean squared error for a bootstrapped sample of maximum likelihood estimates is compared to the mean squared error obtained by generating an entirely new data sample at each iteration, see Figure B.6 in Appendix B. For large non-exceedance probabilities, $p > 0.95$, the variance dominates the mean squared error, and bootstrapping and regenerating data perform similar. Surprisingly, for small non-exceedance probabilities, $p < 0.9$, it differs between Case 1 and Case 2, and between the parameter α and β which method outperforms the other in a mean squared error sense.

All in all, bootstrap methods provide a pragmatic approach to quantify uncertainty when the asymptotic properties of the maximum likelihood estimator fail to provide valid confidence intervals. However, bootstrapping a small sample can lead to severe under- or overestimation of uncertainty regarding the parameter estimates. Furthermore, as bootstrap methods require a reasonably large number of bootstrap samples, these methods become impractical when fitting a computationally expensive model or when are very large number of random variables is considered.

Chapter 4

Bayesian inference on the Heffernan and Tawn models

> Statisticians, like artists, have the bad habit of falling in love with their own models.
>
> — George Box

Conventional methods for statistical inference rely on the assumption that data is identically distributed. Allowing the parameters of a model to be non-constant yields greater flexibility and can enhance the goodness of fit. As Eastoe and Tawn (2009) and Jonathan et al. (2014) show, incorporating covariate effects through non-constant parameterizations can resolve issues such as inefficiency or bias of estimators. Under the assumption of weak identicality, the parameters of the Heffernan and Tawn model model are assumed to be smooth functions with respect to a directional covariate. The aim of this chapter is to present the generalized Heffernan and Tawn model proposed by Jonathan et al. (2014) and demonstrate the proposed Bayesian inference framework.

Bayesian inference is introduced and demonstrated in Section 4.1. Following the roadmap shown in Figure 1.2, the proposed Bayesian inference framework is demonstrated for the following cases in consecutive order,

1. Constant Heffernan and Tawn model, see Section 4.2,
2. Constrained Heffernan and Tawn model, see Section 4.3,
3. Generalized Heffernan and Tawn model, see Section 4.4.

Rather than jumping to Bayesian inference for the generalized Heffernan and Tawn model immediately, treating these models separately allows fundamental issues to be identified as soon as they manifest themselves. Inference on the generalized constrained Heffernan and Tawn model is omitted because additional research regarding the aforementioned models is required.

Discussion of the results for the different models and types of data relies heavily on visuals such as trace- and diagnostic plots. The majority of these figures are moved to the appendix to prevent congestion of the report. See Appendix C for the results of the constant Heffernan and Tawn model. A reparameterization is introduced in Appendix D to map the parameters space $\Omega_{\boldsymbol{\theta}} \rightarrow \mathbb{R}^4$. Inference for the reparameterized constant Heffernan and Tawn model is demonstrated, but because inference on the reparameterized model has some fundamental issues it is disregarded throughout this chapter. Results regarding Bayesian inference for the constrained Heffernan and Tawn model are enclosed in Appendix E. Finally, results for the Bayesian implementation of the generalized Heffernan and Tawn model are presented in Appendix F.

4.1 Bayesian statistics: an introduction

A Bayesian approach to statistical inference for the generalized Heffernan and Tawn model is introduced. The deficiencies of negative log-likelihood minimization, discussed in Section 3.2.6 and 3.2.7, become prohibitive when applied to the generalized Heffernan and Tawn model. Bayesian inference addresses these issues and provides a natural framework to communicate uncertainty regarding the parameter estimates.

Basic concepts of Bayesian statistics are introduced in Section 4.1.1. Sampling algorithms and proposal mechanisms are discussed in Section 4.1.2 and 4.1.3. Summary statistics to assess convergence and mixing of posterior samples are introduced in Section 4.1.4.

4.1.1 Mathematical framework

The first step in Bayesian inference is to specify *prior distributions* for each of the model parameters. Let $\boldsymbol{\theta} = \{\alpha, \beta, \mu, \psi^2\}$ denote the set of model parameters, and n_θ the number of model parameters to be estimated. The first step in setting up a Bayesian model is to specify prior distributions $f_\Theta(\boldsymbol{\theta} \mid \boldsymbol{\eta})$ for each of the model parameters. Parameters that define a prior distribution are referred to as *hyper parameters* and the set of hyper-parameters is denoted by $\boldsymbol{\eta}$. It is important that the support of the prior distributions is equivalent to the parameter space $\Omega_{\boldsymbol{\theta}}$.

The second stage revisits the likelihood function $L(\boldsymbol{\theta} \mid \mathbf{Y}) = f_{\mathbf{Y}|\Theta}(\mathbf{Y} \mid \boldsymbol{\theta})$ as defined in (3.17). The likelihood measures how likely it is to observe a particular sample, given the model and its parameters $\boldsymbol{\theta}$.

Finally, the *posterior distribution* refers to the distribution of $\boldsymbol{\theta}$ when the data is taken into account. The posterior distribution is defined through Bayes' rule,

as the conditional probability

$$f_{\Theta|\mathbf{Y}}\left(\boldsymbol{\theta} \mid \mathbf{Y}, \boldsymbol{\eta}\right) = \frac{f_{\mathbf{Y}|\Theta}\left(\boldsymbol{y} \mid \theta\right) f_{\Theta}\left(\boldsymbol{\theta} \mid \boldsymbol{\eta}\right)}{f_{\mathbf{Y}}\left(\boldsymbol{y} \mid \boldsymbol{\eta}\right)} \propto f_{\mathbf{Y}|\Theta}\left(\boldsymbol{y} \mid \boldsymbol{\theta}\right) f_{\Theta}\left(\boldsymbol{\theta} \mid \boldsymbol{\eta}\right). \tag{4.1}$$

In practice, it suffices to determine the posterior distribution up to proportionality. This is crucial to the practical applicability of Bayesian methods, since the marginal likelihood

$$f_{\mathbf{Y}}\left(\boldsymbol{y} \mid \boldsymbol{\eta}\right) = \int_{\Omega_{\boldsymbol{\theta}}} f_{\mathbf{Y}|\boldsymbol{\theta}}\left(\boldsymbol{y} \mid \boldsymbol{\theta}\right) f_{\Theta}\left(\boldsymbol{\theta} \mid \boldsymbol{\eta}\right) \mathrm{d}\boldsymbol{\theta}$$

is intractable in most applications. The *maximum a posteriori probability estimate* (MAP) is the Bayesian equivalent of the maximum likelihood estimator for negative log-likelihood minimization. It is defined as the mode of the posterior distribution, i.e.

$$\hat{\boldsymbol{\theta}}_{\mathrm{MAP}} \coloneqq \arg\max_{\boldsymbol{\theta}} f_{\Theta|\mathbf{Y}}\left(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{\eta}\right). \tag{4.2}$$

If the prior distribution is uniform or has an approximately flat probability density function, then by substituting (4.1) in (4.2), it follows that $\hat{\boldsymbol{\theta}}_{\mathrm{MAP}} = \hat{\boldsymbol{\theta}}_{\mathrm{MLE}}$.

4.1.2 Sampling algorithms

Markov chain Monte Carlo methods, often abbreviated to MCMC, allow sampling from the posterior distribution $f_{\Theta|\mathbf{Y}}\left(\boldsymbol{\theta} \mid \mathbf{Y}, \boldsymbol{\eta}\right)$. These methods rely on constructing Markov chains such that the *ergodic theorem* is satisfied. Under certain conditions, the ergodic theorem — see Theorem 4.1.1 — guarantees that a Markov chain converges in distribution to a stationary limit distribution. See Gelman et al. (2014) for a full description of the mathematical foundation of Markov chain Monte Carlo methods.

Theorem 4.1.1. *Ergodic theorem*
Let $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(n)}$ be n realizations from an MCMC sampler, that constitute a Markov chain that is aperiodic, irreducable *and* positive recurrent. *Then, for an arbitrary function g such that $\mathrm{E}\left\{g\left(\boldsymbol{\theta}\right)\right\} < \infty$ holds, as $n \to \infty$,*

$$\frac{1}{n}\sum_{l=1}^{n} g\left\{\boldsymbol{\theta}^{(l)}\right\} \overset{a.s.}{\to} \int_{\Omega_{\boldsymbol{\theta}}} g\left(\boldsymbol{\theta}\right) \Pr\left(\boldsymbol{\theta} \mid \boldsymbol{\eta}\right) \mathrm{d}\boldsymbol{\theta}.$$

The first Markov chain Monte Carlo algorithm was proposed by Metropolis et al. (1953), and later improved by Hastings (1970). The Metropolis-Hastings algorithm is as elegant as it is powerful and is deemed by many to be the most influential mathematical invention of the twentieth century. The pseudo code is provided in Algorithm 1. A pragmatic way of choosing feasible starting values is provided in Appendix A.8 where Algorithm 5 is introduced. The

Algorithm 1 Metropolis-Hastings algorithm

Provide a feasible starting value $\boldsymbol{\theta}^{(0)}$ based on Algorithm 5.
for $l = 0$ **to** l_{MAX} **do**
  $\boldsymbol{\theta}^* \sim q\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(l)}\right)$
  $u \sim \mathcal{U}_{[0,1]}$
  if $u \leq \min\left\{1, \frac{f_\Theta(\boldsymbol{\theta}^*)}{f_\Theta(\boldsymbol{\theta}^{(l)})} \frac{q(\boldsymbol{\theta}^{(l)}|\boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(l)})}\right\}$ **then**
    $\boldsymbol{\theta}^{(l+1)} = \boldsymbol{\theta}^*$
  else
    $\boldsymbol{\theta}^{(l+1)} = \boldsymbol{\theta}^{(l)}$
  end if
end for

original Metropolis-Hasting algorithm relies on a multi-dimensional random walk to explore the parameter space $\Omega_{\boldsymbol{\theta}}$. The *transition kernel* $q\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(l)}\right)$ in that case is a multivariate Gaussian distribution. See Section 4.1.3 for an introduction to different proposal mechanisms. The proposed parameters $\boldsymbol{\theta}^*$ are evaluated under the specified prior distributions. If the transition kernel is symmetric, and the ratio $f_\Theta\left(\boldsymbol{\theta}^*\right) / f_\Theta\left(\boldsymbol{\theta}^{(l)}\right) \geq 1$, the proposal $\boldsymbol{\theta}^*$ will be accepted. If $f_\Theta\left(\boldsymbol{\theta}^*\right) / f_\Theta\left(\boldsymbol{\theta}^{(l)}\right) < 1$, the proposal $\boldsymbol{\theta}^*$ can be either accepted or rejected, depending on whether the ratio is greater than some quantity u which is uniformly distributed on $[0, 1]$.

When full conditional distributions for the each of the parameters are known explicitly, *Gibbs sampling* offers an alternative to the rejection sampler defined by Algorithm 1. As Gibbs samplers sample directly from the posterior distribution, they do not require an accept or reject mechanism. Gibbs sampling is particularly useful in hierarchical Bayesian models that are defined by conditional probability distributions, or when the posterior distribution is explicitly known. On the downside, if the model is ill specified, convergence to a stationary limit distribution can be very slow. See Algorithm 2 for the pseudo code of a Gibbs sampler.

Algorithm 2 Gibbs Sampling algorithm

Provide a feasible starting value $\boldsymbol{\theta}^{(0)}$ based on Algorithm 5.
for $l = 0$ **to** l_{MAX} **do**
  $\theta_1^{(l+1)} \sim q\left(\theta_1 \middle| \theta_2^{(l)}, \theta_3^{(l)}, \ldots, \theta_{n_\theta}^{(l)}\right)$
  $\theta_2^{(l+1)} \sim q\left(\theta_2 \middle| \theta_1^{(l)}, \theta_3^{(l)}, \ldots, \theta_{n_\theta}^{(l)}\right)$
  $\vdots$
  $\theta_{n_\theta}^{(l+1)} \sim q\left(\theta_d \mid \theta_1^{(l)}, \theta_2^{(l)}, \ldots, \theta_{n_\theta - 1}^{(l)}\right)$
end for

4.1.3 Transition kernels for the Metropolis-Hastings algorithm

Proposal mechanisms are at the heart of the Metropolis-Hastings algorithm. They should ensure that the likelihood surface is decently explored and the Markov chain mixes well. Each of the transition kernels presented in this section yield proper samples from the posterior distribution according to Theorem 4.1.1. However, when the number of parameters to be estimated is large, the geometry of the negative log-likelihood surface is confined or when the model is misspecified, ignorant proposals lead to very low acceptance rates, poor mixing or high autocorrelation in posterior samples.

Four different proposal mechanisms are introduced in this section. A transition kernel based on a random walk is the default case. The manifold Metropolis adjusted Langevin algorithm proposed by Girolami and Calderhead (2011) and its simplified siblings MALA and smMALA are introduced. The authors provide an elaborate description and comparison of the different algorithms. The most important features of the different transition kernels are briefly summarized in this section.

Random walk (R-W)

The original Metropolis-Hastings algorithm considers a random walk to explore the parameter space. The transition kernel for a Markov chain based on a random walk is given by

$$q\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(l)}\right) \sim \mathcal{N}\left\{\boldsymbol{\nu}\left(\boldsymbol{\theta}^{(l)}, \varepsilon\right), \varepsilon^2 \mathrm{I}_{n_\theta \times n_\theta}\right\},$$

where

$$\boldsymbol{\nu}\left(\boldsymbol{\theta}^{(l)}, \varepsilon\right) = \boldsymbol{\theta}^{(l)}. \tag{4.3}$$

The matrix $\mathrm{I}_{n_\theta \times n_\theta}$ denotes a $n_\theta \times n_\theta$ identity matrix. This transition kernel leads to the $q\left(\boldsymbol{\theta}^{(l)} \mid \boldsymbol{\theta}^*\right) / q\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(l)}\right)$ term in Algorithm 1 being equal to 1 as it is trivially *reversible*, i.e. $q\left(\boldsymbol{\theta}^{(l)} \mid \boldsymbol{\theta}^*\right) = q\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(l)}\right)$.

The continuous time equivalent of this proposal mechanism is a multivariate Wiener-process $\mathbf{W}_t$ for $t \geq 0$, which is defined by the stochastic differential equation $\mathrm{d}\boldsymbol{\theta}(t) = \mathrm{d}\mathbf{W}(t)$.

Simplicity of the random walk transition kernel is both a blessing and a curse. On the one hand, it is very fast to evaluate, but on the other hand, for irregular– or high-dimensional parameter spaces, convergence of the Markov chains generated by Algorithm 1 can be prohibitively slow due to significant autocorrelation.

Manifold Metropolis adjusted Langevin algorithm (mMALA)

More advanced transition kernels for the Metropolis-Hasting algorithm have been proposed by Girolami and Calderhead (2011). As the authors point out, "the parameter space of a statistical model is a Riemann manifold. Therefore, the natural geometric structure of [the parameter space] is defined by the Riemann manifold and associated metric tensor". Exploiting this geometric structure in the proposal mechanism in Algorithm 1 yields faster convergence of the Markov chains.

Rather than a standard Wiener process, Girolami and Calderhead (2011) consider a Langevin diffusion and use the drift term to ensure faster convergence and a reduction in autocorrelation in the posterior samples. The stochastic differential equation for Langevin diffusion is given by

$$\mathrm{d}\boldsymbol{\theta}\left(t\right) = -\tfrac{1}{2}\mathbf{G}^{-1}\left\{\boldsymbol{\theta}\left(t\right), \boldsymbol{y}\right\} \nabla_{\boldsymbol{\theta}} \bar{\ell}_{\mathrm{HT}}\left\{\boldsymbol{\theta}\left(t\right) \mid \boldsymbol{y}\right\} \mathrm{d}t + \mathrm{d}\tilde{\mathbf{W}}\left(t\right). \tag{4.4}$$

In order to exploit the geometry of the parameter space, the preconditioning matrix $\mathbf{G}\left\{\boldsymbol{\theta}\left(t\right), \boldsymbol{y}\right\}$ should define a *metric tensor* on the Riemann manifold associated to the parameter space $\Omega_{\boldsymbol{\theta}}$. More information on $\mathbf{G}\left\{\boldsymbol{\theta}\left(t\right), \boldsymbol{y}\right\}$ is provided below.

First, consider the relationship between the increment $\mathrm{d}\tilde{\mathbf{W}}\left(t\right)$ on the Riemann manifold and its counterpart $\mathrm{d}\mathbf{W}\left(t\right)$ defined on a standard Euclidean space, which for $i = 1, \ldots, n_\theta$ is given by

$$\mathrm{d}\tilde{\mathbf{W}}_i\left(t\right) \coloneqq \frac{1}{\sqrt{\det\left|\mathbf{G}\left\{\boldsymbol{\theta}\left(t\right), \boldsymbol{y}\right\}\right|}} \sum_{j=1}^{n_\theta} \frac{\partial}{\partial \theta_j} \left[\mathbf{G}_{ij}^{-1}\left\{\boldsymbol{\theta}\left(t\right), \boldsymbol{y}\right\} \sqrt{\det\left|\mathbf{G}\left\{\boldsymbol{\theta}\left(t\right), \boldsymbol{y}\right\}\right|}\right] \mathrm{d}t + \sqrt{\mathbf{G}_i^{-1}\left\{\boldsymbol{\theta}\left(t\right), \boldsymbol{y}\right\}} \mathrm{d}\mathbf{W}_i\left(t\right). \tag{4.5}$$

Girolami and Calderhead (2011) explain that "the first term on the right hand side of (4.5) relates to changes in local curvature of the manifold and reduces to 0 if curvature is everywhere constant. The second term provides a position-specific axis alignment of the Wiener process based on the local metric, by transformation of the independent Wiener process $\mathbf{W}\left(t\right)$".

As Girolami and Calderhead (2011) show, substituting (4.5) in (4.4) and applying the Euler-Maruyama discretization yields the transition kernel for the manifold Metropolis adjusted Langevin algorithm given by

$$q\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(l)}\right) \sim \mathcal{N}\left\{\boldsymbol{\nu}\left(\boldsymbol{\theta}^{(l)}, \varepsilon\right), \varepsilon^2 \mathbf{G}^{-1}\left(\boldsymbol{\theta}^{(l)}, \boldsymbol{y}\right)\right\}, \tag{4.6}$$

where

$$\begin{aligned}\boldsymbol{\nu}\left(\boldsymbol{\theta}_i^{(l)}, \varepsilon\right) = \boldsymbol{\theta}_i^{(l)} &- \frac{\varepsilon^2}{2}\left[\mathbf{G}^{-1}\left(\boldsymbol{\theta}^{(l)}, \boldsymbol{y}\right)\nabla_{\boldsymbol{\theta}}\bar{\ell}_{\mathrm{HT}}\left(\boldsymbol{\theta}^{(l)} \mid \boldsymbol{y}\right)\right]_i \\ &- \varepsilon^2 \sum_{i=1}^{n_\theta}\left[\mathbf{G}^{-1}\left(\boldsymbol{\theta}^{(l)}, \boldsymbol{y}\right)\frac{\partial \mathbf{G}\left(\boldsymbol{\theta}^{(l)}, \boldsymbol{y}\right)}{\partial \theta_i}\mathbf{G}^{-1}\left(\boldsymbol{\theta}^{(l)}, \boldsymbol{y}\right)\right]_{ij} \\ &+ \frac{\varepsilon^2}{2}\sum_{i=1}^{n_\theta}\left[\mathbf{G}_{ij}^{-1}\left(\boldsymbol{\theta}^{(l)}, \boldsymbol{y}\right)\operatorname{tr}\left\{\mathbf{G}^{-1}\left(\boldsymbol{\theta}^{(l)}, \boldsymbol{y}\right)\frac{\partial \mathbf{G}^{-1}\left(\boldsymbol{\theta}^{(l)}, \boldsymbol{y}\right)}{\partial \theta_i}\right\}\right]. \quad (4.7)\end{aligned}$$

At this point an explicit definition for the matrix $\mathbf{G}\left(\boldsymbol{\theta}^{(l)}, \boldsymbol{y}\right)$ is required. It was shown by Rao (1945) that the expected Fisher information matrix defined by (3.22) provides a measure of distance between two parameterized probability density functions. Without going into too much detail, this implies the expected Fisher information endows an appropriate *metric tensor* on the Riemann manifold of the parameter space. Intuitively, proposals target regions of high probability density and the step size is scaled such that steps are small when the current state is in the vicinity of the maximum likelihood estimate, and steps are large when the maximum likelihood estimate is still relatively far away from the current state. The metric tensor proposed by Girolami and Calderhead (2011) is given by

$$\begin{aligned}\mathbf{G}\left(\boldsymbol{\theta}, \boldsymbol{y}\right) &= -\mathrm{E}\left[\frac{\partial^2}{\partial \boldsymbol{\theta}^2}\log f_{\mathbf{Y},\Theta}\left(\boldsymbol{\theta}, \boldsymbol{y}\right)\right], \\ &= \mathcal{I}\left(\boldsymbol{\theta}\right) + \mathbf{H}_\Theta\left(\boldsymbol{\theta}\right),\end{aligned}$$

where $\mathcal{I}\left(\boldsymbol{\theta}\right)$ denotes the expected Fisher information matrix and $\mathbf{H}_\Theta\left(\boldsymbol{\theta}\right)$ denotes the Hessian of the log-prior, provided in Appendix A.7 for the Gaussian– and Gamma distribution. If uninformative prior distributions are adopted, then $\mathbf{H}_\Theta\left(\boldsymbol{\theta}\right) = 0$. Hence from here onward, assume $\mathbf{G}\left\{\boldsymbol{\theta}^{(l)}, \boldsymbol{y}\right\} = \mathcal{I}^{\mathrm{R}}\left\{\boldsymbol{\theta}^{(l)}, \boldsymbol{y}\right\}$, where the restrained rather than the full expected Fisher information matrix is adopted to address non-invertibility of the full expected Fisher information, discussed in Section 3.2.5.

In addition to the full mMALA, two additional proposal mechanisms can be defined by strong assumptions on $\mathbf{G}\left\{\boldsymbol{\theta}^{(l)}, \boldsymbol{y}\right\}$, see Figure 4.1 for an overview. These proposal mechanism are formally introduced in Section 4.1.3 and 4.1.3.

Metropolis adjusted Langevin algorithm (MALA)

Under the simplifying assumption $\mathbf{G}\left\{\boldsymbol{\theta}^{(l)}, \boldsymbol{y}\right\} = \mathrm{I}_{n_\theta \times n_\theta}$ mMALA reduces to MALA. Evaluating MALA is computationally less expensive and can be preferred when $\mathbf{G}\left\{\boldsymbol{\theta}^{(l)}, \boldsymbol{y}\right\}$ is large and not sparse, such that its inversion takes a lot of time.

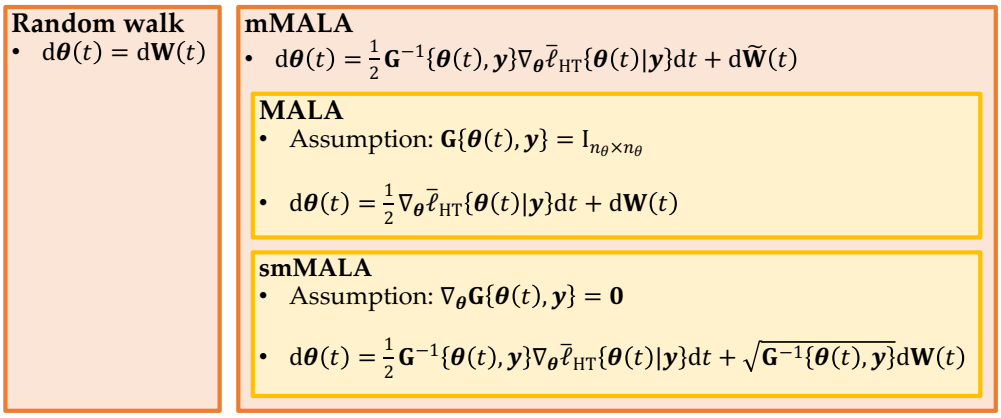


FIGURE 4.1: Relationship between the different proposal mechanisms considered in this chapter.

The transition kernel for MALA is given by

$$q\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(l)}\right) \sim \mathcal{N}\left\{\boldsymbol{\nu}\left(\boldsymbol{\theta}^{(l)}, \varepsilon\right), \varepsilon^2 \mathrm{I}_{n_\theta \times n_\theta}\right\}. \tag{4.8}$$

Where,

$$\boldsymbol{\nu}\left(\boldsymbol{\theta}^{(l)}, \varepsilon\right) = \boldsymbol{\theta}^{(l)} - \frac{\varepsilon^2}{2}\nabla_{\boldsymbol{\theta}}\bar{\ell}_{\mathrm{HT}}\left(\boldsymbol{\theta}^{(l)} \mid \boldsymbol{y}\right). \tag{4.9}$$

Comparing (4.9) to (4.3) shows the only difference between the random walk proposal mechanism and MALA is the gradient term on the right hand side of (4.9), which defines the direction in which proposals are made.

The $q\left(\boldsymbol{\theta}^{(l)} \mid \boldsymbol{\theta}^*\right)/q\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(l)}\right)$ term in Algorithm 1 can no longer be ignored as Markov chains based on MALA are no longer reversible. An appropriate expression for $q\left(\boldsymbol{\theta}^{(l)} \mid \boldsymbol{\theta}^*\right)$ follows from interchanging $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}^{(l)}$ in (4.8).

Simplified manifold Metropolis adjusted Langevin algorithm (smMALA)

As Girolami and Calderhead (2011) point out, "the MALA can be inefficient for highly correlated variables with widely differing variances forcing the step size to accommodate the smallest variance". As discussed in detail in Section 3.2, the interaction between α and μ, as well as the β and ψ^2, is apparent and will lead to highly correlated samples from the posterior distribution.

The simplified manifold Metrolopolis adjusted Langevin algorithm (smMALA) arises from (4.4) by assuming $\nabla_{\boldsymbol{\theta}}\mathcal{I}^{\mathrm{R}}\left\{\boldsymbol{\theta}^{(l)}, \boldsymbol{y}\right\} = \mathbf{0}$. This is a weak assumption as even many of the elements of $\nabla_{\boldsymbol{\theta}}\mathcal{I}\left\{\boldsymbol{\theta}^{(l)}, \boldsymbol{y}\right\}$ are zero anyway, as shown in Appendix A.6. Moreover, several of the non-zero entries will be close to zero if $\mu \approx 0$. Even if "the curvature of the manifold is not constant" and $\nabla_{\boldsymbol{\theta}}\mathcal{I}^{\mathrm{R}}\left\{\boldsymbol{\theta}^{(l)}, \boldsymbol{y}\right\} = \mathbf{0}$ does not hold, "the above simplified proposal mechanism, used in conjunction with the acceptance probability, will still define a

correct MCMC method that converges to the target measure", see Girolami and Calderhead (2011) for the proof of this statement.

The transition kernel for the simplified manifold Metropolis adjusted Langevin algorithm is given by

$$q\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(l)}\right) \sim \mathcal{N}\left\{\boldsymbol{\nu}\left(\boldsymbol{\theta}^{(l)}, \varepsilon\right), \varepsilon^2 \mathcal{I}^{\mathrm{R}}\left\{\boldsymbol{\theta}^{(l)}, \boldsymbol{y}\right\}^{-1}\right\}, \tag{4.10}$$

where

$$\boldsymbol{\nu}\left(\boldsymbol{\theta}^{(l)}, \varepsilon\right) = \boldsymbol{\theta}^{(l)} - \frac{\varepsilon^2}{2} \mathcal{I}^{\mathrm{R}}\left\{\boldsymbol{\theta}^{(l)}, \boldsymbol{y}\right\}^{-1} \nabla_{\boldsymbol{\theta}} \bar{\ell}_{\mathrm{HT}}\left(\boldsymbol{\theta}^{(l)} \mid \mathbf{y}\right). \tag{4.11}$$

4.1.4 Convergence diagnostics and statistics

Theorem 4.1.1 guarantees that a properly sampled Markov chain will converge in distribution to the true posterior distribution. However, it does not guarantee that the Markov chain will converge in a finite number of iterations. Assessing whether a Markov chain has converged is a crucial but non-trivial step in Bayesian inference. Different heuristics and statistics are introduced to asses convergence of the posterior samples.

The step size ε is reported and unless explicitly stated otherwise, a common step size is adopted for each of the parameters of the Heffernan and Tawn model. Performance of Algorithm 1 might be improved if a separate step size is considered for each parameter. However, as tuning four different step sizes is a tedious job and the simplified mMALA addresses the issue by appropriately scaling proposals, the issues is put aside for now. One of the recommendations presented in Chapter 5 address this issue.

The *acceptance rate* (AR) is defined as the fraction of post-burnin samples that is accepted. An acceptance rate of approximately 40% is regarded to be optimal. An acceptance rate that is too low indicates that the chain does not mix well and will fail to effectively explore the parameter space. On the contrary, an acceptance rate that is too high might indicate that the step size is too small and the chain traverses the parameter space very slowly.

Posterior samples are summarized by the median (MED) and a 95% confidence interval ($\mathrm{CI}_{95\%}$). The confidence intervals are based on the 2.5% and 97.5% empirical quantiles. Properly converged posterior samples for parameters that are not on the boundary of their support are expected to be symmetric and resemble a Gaussian distribution.

Consider different posterior samples generated in parallel, the *initial positive sequence estimator* for the *effective sample size* (ESS) as proposed by Geyer (1992), is defined by

$$\mathrm{ESS} := \frac{n_{\mathrm{S}}}{1 + 2\sum_{l=1}^{n_{\mathrm{S}}} \varrho_l \mathbb{1}_{\varrho_l + \varrho_{l-1} < 0}}, \tag{4.12}$$

where ϱ_l denotes the autocorrelation at lag l. The summation in (4.12) must be truncated, because as Hassani (2010) points out, the sample auto correlation will sum up to $-1/2$. The effective sample size is self-explaining, as it provides a measure for the number of samples that is effectively sampled from the posterior distribution.

Advanced transition kernels, such as MALA or simplified mMALA, are expected to yield higher effective sample sizes. It is not fair to judge the performance of a particular proposal mechanism solely on the effective sample size as this does not penalize the additional computational burden. The *effective sample size per second* (ESS/s) is reported to address this issue. Even though the MALA or smMALA might have a higher effective sample size, the random walk transition kernel might still be preferred if it outperforms the other proposal mechanism in terms of effective sample size per second.

If several Markov chains are considered, the Gelman-Rubin statistic measures whether the within-chain variance is significantly different from the between chain covariance. Let n_{C} denote the number of different Markov chains and n_{S} the posterior sample size. Let the within chain variance $\hat{\mathrm{W}}$ is be defined by

$$\hat{\mathrm{W}} \coloneqq \frac{1}{n_{\mathrm{C}}} \sum_{i=1}^{n_{\mathrm{C}}} s_i^2, \quad \text{where} \quad s_i^2 \coloneqq \frac{1}{n_{\mathrm{S}} - 1} \sum_{l=1}^{n_{\mathrm{S}}} \left(\hat{\theta}_{il} - \hat{\theta}_i^{\mathrm{SM}} \right),$$

where $\hat{\theta}_i^{\mathrm{SM}}$ denotes the sample mean of the posterior sample for the parameter θ. The between chain variance $\hat{\mathrm{B}}$ is defined by

$$\hat{\mathrm{B}} \coloneqq \frac{1}{n_{\mathrm{C}}} \sum_{i=1}^{n_{\mathrm{C}}} \left(\hat{\theta}_i^{\mathrm{SM}} - \hat{\theta}^{\mathrm{ASM}} \right)^2, \quad \text{where} \quad \hat{\theta}^{\mathrm{ASM}} \coloneqq \frac{1}{n_{\mathrm{C}}} \sum_{i=1}^{n_{\mathrm{C}}} \hat{\theta}_i^{\mathrm{SM}}.$$

The scale reduction factor, also referred to as the Gelman-Rubin statistic, is defined as

$$\hat{\mathrm{R}} \coloneqq \sqrt{\frac{\left(1 - \frac{1}{n}\right) \mathrm{W} + \frac{1}{n} \mathrm{B}}{\mathrm{W}}}.$$

If the scale reduction factor is close to 1, then this indicates that the different Markov chains have converged in distribution to the same stationary posterior distribution. On the contrary, if $\hat{\mathrm{R}}$ is larger than 1.1–1.2, this indicates the burn-in period was insufficient to ensure convergence.

In addition to the figures and tables presented in this chapter, traceplots of the posterior samples and several other diagnostic plots for the sample likelihood are presented in Appendix C-F. These diagnostic plots are introduced at the start of each appendix.

4.2 Bayesian inference for the constant Heffernan and Tawn model

Bayesian inference for the model proposed by Heffernan and Tawn (2004) has received little attention in the literature. Only the works by Cheng et al. (2014) and Lugrin et al. (2016) consider Bayesian inference on the constant Heffernan and Tawn model. However, as Lugrin et al. (2016) point out, the approach proposed by Cheng et al. (2014) relies on "changes in the structure of the model and adding a noise term in the likelihood function, thereby allowing the likelihood term to be split appropriately". In addition, the model requires strong prior information, which is obtained by negative log-likelihood minimization. The Bayesian inference framework proposed in this section addresses both deficiencies of the methodology proposed by Cheng et al. (2014).

Matrix inversion — as required for the MALA and simplified mMALA — is computationally expensive and can lead to numerical instabilities. Backward substitution rather than matrix inversion is adopted. Secondly, for a positive definite matrix, matrix inversion can be avoided since

$$\log\left(\det\left|\varepsilon^2\mathbf{G}^{-1}\right|\right) = 2n_\theta\log\left(\varepsilon\right) - \log\left(\det\left|\mathbf{G}\right|\right)$$

Furthermore, for a positive definite matrix $\mathbf{G}$ with Cholesky decomposition $\mathbf{G} = \mathbf{L}\mathbf{L}^\intercal$, adopting

$$\log\left\{\det\left|\mathbf{G}\right|\right\} = 2\mathrm{tr}\left\{\log\left(\mathbf{L}\right)\right\},$$

improves the numerical stability. These adjustments to the standard Metropolis-Hastings algorithm are incorporated in Algorithm 3.

Algorithm 3 Metropolis-Hastings algorithm for the Heffernan and Tawn model

Initialize $\boldsymbol{\theta}^{(0)}$ based on Algorithm 5.
for $l = -l_{\mathrm{B}}$ **to** l_{MAX} **do**
  $\boldsymbol{\theta}^* \sim q\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(l)}\right)$
  $u \sim \log\left(\mathcal{U}_{[0,1]}\right)$
  $L^*_{\mathrm{TOT}} = -\bar{\ell}_{\mathrm{HT}}\left(\boldsymbol{\theta}^*\right) + \log\left\{f_\Theta\left(\boldsymbol{\theta}^* \mid \boldsymbol{\eta}\right)\right\} + \log\left\{q\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(l)}\right)\right\}$
  $L^{(l)}_{\mathrm{TOT}} = -\bar{\ell}_{\mathrm{HT}}\left(\boldsymbol{\theta}^{(l)}\right) + \log\left\{f_\Theta\left(\boldsymbol{\theta}^{(l)} \mid \boldsymbol{\eta}\right)\right\} + \log\left\{q\left(\boldsymbol{\theta}^{(l)} \mid \boldsymbol{\theta}^*\right)\right\}$
  if $u \leq \min\left\{0, L^*_{\mathrm{TOT}} - L^{(l)}_{\mathrm{TOT}}\right\}$ **then**
    $\boldsymbol{\theta}^{(l+1)} = \boldsymbol{\theta}^*$
  else
    $\boldsymbol{\theta}^{(l+1)} = \boldsymbol{\theta}^{(l)}$
  end if
end for

The results presented in this section show that adopting MALA or simplified mMALA will significantly improve performance of Bayesian inference on the constant Heffernan and Tawn model. Case 1 and Case 2 data introduced in Section 3.2.1 is considered for the simulation study.

This section is structured as follows. Prior distributions are discussed in Section 4.2.1. Inference on the parameters of the constant Heffernan and Tawn model is demonstrated for Case 1 data and Case 2 data in Section 4.2.2 and 4.2.2 respectively. In both cases, the performance of Algorithm 3 is demonstrated by estimating only the α and β parameter while fixing $\mu^{(l)} = \hat{\mu}_{\text{MLE}}$ and $\psi^{2(l)} = \hat{\psi}^2_{\text{MLE}}$ for all $l = 1, \ldots, n_{\text{S}}$, before addressing joint estimation of all four parameters of the Heffernan and Tawn model. In the two parameter estimation case, the first $n_{\text{B}} = 10^2$ samples are regarded as burn-in, while in the four parameters estimation case, the first $n_{\text{B}} = 10^3$ samples are regarded as burn-in. In both cases, the $n_{\text{S}} = 10^4$ samples following the burn-in are regarded as proper samples from the posterior distribution. Throughout this section, the constraints proposed by Keef et al. (2013) are not imposed, as these are treated separately in Section 4.3.

4.2.1 Prior distributions

A Bayesian framework requires assumptions on the prior distributions of the model parameters. Although there is a lot of flexibility in choosing appropriate prior distributions, there are certain constraints that need to be taken into account.

First of all, the support of the prior distributions should be in accordance with the parameter space. Rather than choosing prior distributions with appropriate support, the support of the Heffernan and Tawn model parameters is incorporated implicitly by assigning an arbitrary large number to the negative log-likelihood function if a proposal does not comply with the parameter support.

Secondly, in order to establish a generic framework that accommodates both asymptotic independent data, as well as asymptotic dependent data, it is desirable to assume uninformative prior distributions. A uniform prior distribution $\mathcal{U}_{[-1,1]}$ for α is considered, and an improper uniform prior $\mathcal{U}_{(-\infty,1]}$ for β. Assume a Gaussian prior distribution $\mathcal{N}\left(\eta^{\mu}_{\mu}, \eta^{\sigma^2}_{\mu}\right)$ for the nuisance parameter μ, and let $\eta^{\mu}_{\mu} = 0$ and $\eta^{\sigma^2}_{\mu} = 100$ such that the density is approximately flat. A Gamma prior $\mathcal{G}(\eta_a, \eta_b)$ is assumed for ψ^2, which — as required — is defined only on the positive real line. Choose the shape parameter $\eta^{a}_{\psi^2} = 10^{-4}$ and the scale $\eta^{b}_{\psi^2} = 10^4$. Practitioners often adopt this prior distribution as the density function is approximately flat for all values on the positive real line away from 0 and is hence uninformative.

4.2.2 Results for Case 1

The aim of this section is to show that the proposed Bayesian framework is an appropriate way to estimate the parameters of the Heffernan and Tawn model and quantify the associated uncertainty. The three different transition kernels for the Metropolis-Hasting algorithm introduced in Section refsec:propmech are considered. Convergence and mixing of the obtained Markov chains is discussed based on the statistics presented in this section and diagnostic plots shown in Appendix C.

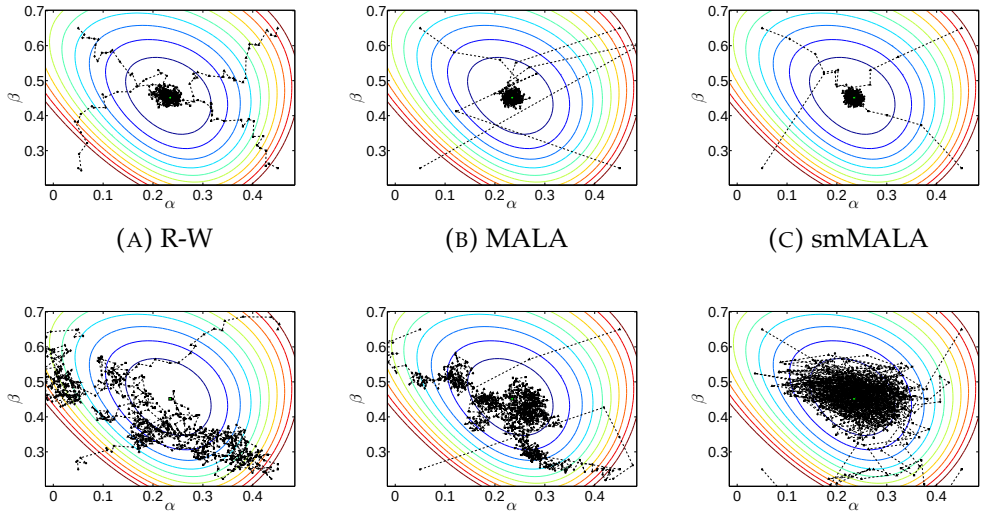
Maximum likelihood estimates for the parameters of the Heffernan and Tawn model for Case 1 data, based on minimizing the negative log-likelihood function, are given by

$$\hat{\alpha}_{\text{MLE}} = 0.23, \quad \hat{\beta}_{\text{MLE}} = 0.45, \quad \hat{\mu}_{\text{MLE}} = 0.39 \quad \text{and} \quad \hat{\psi}^2_{\text{MLE}} = 0.73. \tag{4.13}$$

As uninformative priors are adopted, the posterior samples for each of the parameters are supposed to converge to the associated maximum likelihood estimate, rather than the true values reported in Table 3.1 because of the different sources of bias discussed in Section 3.2.6. First consider the challenge of estimating the α and β parameter of the Heffernan and Tawn model while fixing $\mu^{(l)} = \hat{\mu}_{\text{MLE}}$ and $\psi^{(l)} = \hat{\psi}^2_{\text{MLE}}$ for all $l = 1, \ldots, n_{\text{B}} + n_{\text{S}}$ which provides a proof of concept.

Four different chains are started from dispersed starting values. Each of the chains, for each of the transition kernels, converges to the minimum of the negative log-likelihood surface within the first 100 iterations as shown in Figure 4.2. The rate at which the chains converge increases as the sophistication of the transition kernels increases. Markov chains based on a random walk move around the parameter space relatively slowly, as shown in Figure 4.2(A). For the MALA, the chains converge in only two or three steps, but there is a significant overshoot as the steps are not scaled appropriately. The simplified mMALA restrains the proposals by scaling the step size appropriately and shows rapid convergence.

Several statistics presented in Table 4.1 provide reassurance that the Markov chains have converged. The median and 95% confidence intervals of the posterior samples are nearly identical for any of the transition kernels, and the Gelman-Rubin statistics provide strong evidence in favor of convergence. Additional diagnostic plots are shown in Appendix C in Figure C.1. Traceplots of the posterior sample for α and β, see Figure C.1(A)-(F), show convergence to the maximum likelihood estimates. In addition, the running mean of the sample likelihood for each of the four different chains converges to a common



(A) R-W (B) MALA (C) smMALA

(D) R-W (E) MALA (F) smMALA

FIGURE 4.2: The first 250 burn-in samples for α and β for Case 1 data, based on three different transition kernels: random walk, Metropolis adjusted Langevin algorithm and simplified manifold Metropolis adjusted Langevin algorithm. The top row shows results when only α and β are estimated and $\mu = \hat{\mu}_{\text{MLE}}$ and $\psi = \hat{\psi}^2_{\text{MLE}}$, while all four parameters of the Heffernan and Tawn model are estimated jointly for the figures in the bottom row.

negative log-likelihood level, as shown in Figure C.1(M)-(M), which also indicates convergence. These results are deemed to be sufficient evidence in favor of convergence of the Markov chains.

In addition to convergence, the Markov chains should mix well. The traceplots shown in Figure C.1(A)-(F) suggest the chain mixes well. Markov chains based on the MALA and simplified mMALA show little autocorrelation. This is confirmed by the difference in effective sample size shown in Table 4.1. However, the random walk transition kernel outperforms MALA and smMALA in terms of effective sample size per second.

Now that the Bayesian inference methodology has been demonstrated for the two parameter case, consider joint estimation of all four parameters. Burn-in samples are shown in Figure 4.2(D)-(F). Comparing Figure 4.2(D)-(E) to 4.2(F) suggests that the Markov chains based on the random walk transition kernel and MALA struggle to converge to the minimum of the negative log-likelihood surface.

Poor convergence of the first burn-in samples does not imply poor convergence of the entire posterior sample. At the start of the posterior sample, the Markov chains for each of the parameters and for each of the different transition kernels moves around in the vicinity of the maximum likelihood estimates, as shown in Figure C.2. The median of the posterior sample has converged to

TABLE 4.1: Summary statistics for the posterior samples of the Heffernan and Tawn model parameter estimates for Case 1 data. Different proposal mechanisms are considered: random walk, Metropolis adjusted Langevin algorithm and the simplified manifold Metropolis adjusted Langevin algorithm. Statistics are based on $n_{\text{S}} = 10^4$ posterior samples. The maximum likelihood estimates are given by: $\hat{\alpha}_{\text{MLE}} = 0.23$, $\hat{\beta}_{\text{MLE}} = 0.45$, $\hat{\mu}_{\text{MLE}} = 0.39$ and $\hat{\psi}^2_{\text{MLE}} = 0.73$.

		Two parameter estimation			**Four parameter estimation**		
		R-W	MALA	smMALA	R-W	MALA	smMALA
ε		0.0175	0.0175	1	0.02	0.015	0.8
AR		0.45	0.42	0.52	0.35	0.43	0.46
$\hat{\alpha}$	MED	0.23	0.23	0.23	0.23	0.21	0.24
	$\text{CI}_{95\%}$	[0.21, 0.26]	[0.22, 0.25]	[0.22, 0.25]	[0.03, 0.40]	[0.14, 0.30]	[0.14, 0.33]
	ESS	750	2640	1690	6	15	740
	ESS/s	35	18	12	0.3	0.1	5.1
	$\hat{\text{R}}$	1	1	1	1.04	1.04	1
$\hat{\beta}$	MED	0.45	0.45	0.45	0.46	0.47	0.46
	$\text{CI}_{95\%}$	[0.42, 0.48]	[0.43, 0.47]	[0.43, 0.47]	[0.33, 0.61]	[0.40, 0.55]	[0.38, 0.55]
	ESS	542	1760	1660	12	25	240
	ESS/s	26	12	12	0.5	0.2	1.6
	$\hat{\text{R}}$	1	1	1	1.01	1.04	1
$\hat{\mu}$	MED				0.40	0.42	0.38
	$\text{CI}_{95\%}$				[0.07, 0.67]	[0.26, 0.55]	[0.20, 0.56]
	ESS				7	15	1080
	ESS/s				0.3	0.1	7.4
	$\hat{\text{R}}$				1.04	1.03	1
$\hat{\psi}^2$	MED				0.72	0.70	0.70
	$\text{CI}_{95\%}$				[0.51, 0.98]	[0.59, 0.81]	[0.57, 0.85]
	ESS				11	24	250
	ESS/s				0.5	0.2	1.7
	$\hat{\text{R}}$				1.01	1.04	1.05

the maximum likelihood estimates given by (4.13) for each proposal mechanism. Confidence intervals for the MALA and smMALA appear to have converged faster than those based on random walk proposals. Which indicates the $n_{\text{B}} = 10^3$ burn-in sample was insufficient for the random walk proposal mechanism to yield properly converged posterior samples.

The trace-plots and autocorrelation function plots shown in Appendix C in Figure C.3, show significant and persistent autocorrelation in the Markov chains based on the random walk and MALA proposal mechanisms. This is also reflected in the effective sample size reported in Table 4.1, which is prohibitively small for the random walk and MALA transition kernel. The simplified mMALA outperforms its peers in terms of effective sample per second by a factor 3-5 for the β and ψ^2 parameter, and even a factor 15-75 for the α and μ parameter. This asymmetry suggests that adopting a separate stepsize for α, μ and β, ψ^2 might further improve mixing.

Summarizing, any of the transition kernels discussed in Section 4.1.3 yields Markov chains that properly converge in distribution to the true posterior distribution. The simplified mMALA outperforms the other two proposal mechanisms in terms of autocorrelation and mixing, even when the additional computational burden is taken into account.

4.2.3 Results for Case 2

Bayesian inference is demonstrated for Case 2 data in this section. The results will be presented in similar fashion as those for Case 1 in Section 4.2.2. The data considered in Case 2 is asymptotically dependent, and hence $\alpha_{\mathrm{T}} = 1$, which implies $\boldsymbol{\theta}_{\mathrm{T}} \in \partial\Omega_{\boldsymbol{\theta}}$. Maximum likelihood estimates for the parameters of the Heffernan and Tawn model for Case 2 data, based on minimizing the negative log-likelihood function, are given by

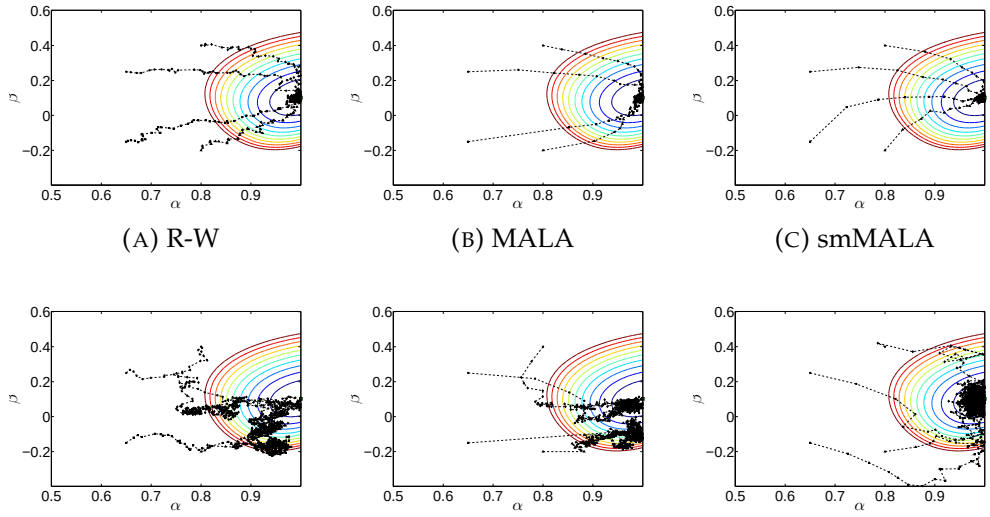
$$\hat{\alpha}_{\mathrm{MLE}} = 1^{-}, \quad \hat{\beta}_{\mathrm{MLE}} = 0.10, \quad \hat{\mu}_{\mathrm{MLE}} = \text{-}0.57 \quad \text{and} \quad \hat{\psi}^2_{\mathrm{MLE}} = 1.02, \tag{4.14}$$

where the superscript in 1^{-} is used to indicate a very small deviation from 1.

The first 200 burn-in samples suggest convergence for each proposal mechanism, as shown in Figure 4.3. However, once the chain approaches the maximum likelihood estimate, the boundary of the parameter space starts to manifest itself.

Summary statistics for the posterior samples are provided in Table 4.2. When the parameters are sampled on the original scale, the step size for the random walk transition kernel and MALA must be very small, in order to approach the boundary of the parameter space. The results presented in Table 4.2 are in line with previous remarks on the performance of the different transition kernels for Case 1 data. The effective sample size increases when the proposals become more sophisticated. When only two parameters are estimated, the additional computational burden outweighs the positive impact on the effective sample size, as the effective sample size per second for the random walk is much better. The Gelman-Rubin statistic indicates convergence for any of the proposal mechanisms.

The results shown in Table 4.2 and Figure 4.3 conceal the fundamental issue of estimating a parameter on the boundary of the parameter space. The trace-plots shown in Appendix C in Figure C.8 reveal the posterior sample for α gets close to the boundary, but fails to put any probability mass on the actual boundary. The proposals literally hit a wall and bounce off. Due to the strong correlation between the posterior samples of α and μ, the bias will also be apparent in the posterior sample for μ. By further reducing the step size, the



(A) R-W (B) MALA (C) smMALA

(D) R-W (E) MALA (F) smMALA

FIGURE 4.3: The first 250 burn-in samples for α and β for Case 2 data, based on three different transition kernels: random walk, Metropolis adjusted Langevin algorithm and simplified manifold Metropolis adjusted Langevin algorithm. The top row shows results when only α and β are estimated and $\mu = \hat{\mu}_{\text{MLE}}$ and $\psi = \hat{\psi}^2_{\text{MLE}}$, while all four parameters of the Heffernan and Tawn model are estimated jointly for the figures in the bottom row.

Markov chains will be able to get even closer to the boundary, but the problem will persist.

The fundamental issues raised for inference on α and β apply equally well to the four parameter estimation problem. Figure 4.3(D)-(F) suggests the random walk has considerable difficulty to reach the maximum likelihood estimate and tends to get stuck. This is supported by the Gelman-Rubin statistic which provides evidence against convergence of the posterior samples of β and ψ^2 for the random walk transition kernal.

Summarizing, as the true value α_{T} for asymptotic dependent data lies on the boundary of the parameter space, conventional inference will result in biased parameter estimates and unreliable confidence intervals. It stands to reason the bias is sufficiently small and does not have a significantly affect simulations from the Heffernan and Tawn model based on these parameters estimates. In addition, the interaction between α and μ counteracts the bias, as a decrease in α leads to an increase in μ and both parameters have the same explanatory effect, to a certain extend. It would be interesting to see whether this hypothesis holds, and it presented as a recommendation for further research in Chapter 5.

TABLE 4.2: Summary statistics for the posterior samples of the Heffernan and Tawn model parameter estimates for Case 2 data. Different proposal mechanisms are considered: random walk, Metropolis adjusted Langevin algorithm and the simplified manifold Metropolis adjusted Langevin algorithm. Statistics are based on $n_{\text{S}} = 10^4$ posterior samples. The maximum likelihood estimates are given by: $\hat{\alpha}_{\text{MLE}} = 1^{-}$, $\hat{\beta}_{\text{MLE}} = 0.10$, $\hat{\mu}_{\text{MLE}} = \text{-}0.57$ and $\hat{\psi}^2_{\text{MLE}} = 1.02$.

		Two parameter estimation			**Four parameter estimation**		
		R-W	MALA	smMALA	R-W	MALA	smMALA
ε		0.01	0.007	0.7	0.013	0.0125	0.6
AR		0.41	0.42	0.44	0.42	0.39	0.46
$\hat{\alpha}$	MED	0.99	1^{-}	1^{-}	0.98	0.99	0.99
	$\text{CI}_{95\%}$	$[0.98, 1]$	$[0.99, 1]$	$[0.99, 1]$	$[0.94, 1^{-}]$	$[0.96, 1^{-}]$	$[0.96, 1^{-}]$
	ESS	600	1090	970	20	65	740
	ESS/s	32	8	7	1	0.5	6
	$\hat{\text{R}}$	1	1	1	1.02	1.01	1
$\hat{\beta}$	MED	0.1	0.1	0.1	0.06	0.08	0.1
	$\text{CI}_{95\%}$	$[0.07, 0.13]$	$[0.08, 0.12]$	$[0.08, 0.12]$	[-0.08, 0.21]	$[0, 0.17]$	$[0.02, 0.18]$
	ESS	210	340	710	5	9	100
	ESS/s	11	3	5	0.2	0.1	0.7
	$\hat{\text{R}}$	1	1	1	1.11	1.04	1
$\hat{\mu}$	MED				-0.53	-0.54	-0.53
	$\text{CI}_{95\%}$				[-0.65, -0.40]	[-0.62, -0.44]	[-0.60, -0.43]
	ESS				10	16	210
	ESS/s				0.4	0.1	1.5
	$\hat{\text{R}}$				1.02	1	1
$\hat{\psi}^2$	MED				1.13	1.08	1.01
	$\text{CI}_{95\%}$				$[0.78, 1.60]$	$[0.87, 1.30]$	$[0.84, 1.23]$
	ESS				5	10	100
	ESS/s				0.2	0.1	0.7
	$\hat{\text{R}}$				1.13	1.04	1

A reparameterization is proposed to resolve the issues set-forth in this section. The idea is that the reparameterization $\boldsymbol{\theta}^*\colon \Omega_{\boldsymbol{\theta}} \to \mathbb{R}^4$ can get arbitrarily close to the boundary of the parameter space on the original scale. This is shown by the traceplots for α when estimated using the reparameterization. The results are presented in Appendix D. Reparameterization (D.1) — in particular the logistic transformation of α — resolve the boundary issue, but shift the true value α^*_{T} to infinity on the reparameterized scale. The simplified mMALA breaks down as the expected Fisher information matrix is singular. In addition, tuning the algorithms is more difficult as the step size for the α Markov chains must be significantly bigger to effectively explore the parameter space under the reparameterization. Albeit the bias in the posterior samples for α and μ when the parameters are evaluated on the original scale, the reparameterization proposed in Appendix D is disregarded as its benefits do not weigh up its disadvantages.

Although the reparameterization provides a sound way to quantify uncertainty for parameter estimates, it would be even better to restrict the Markov chain samplers to a subspace of $\Omega_{\boldsymbol{\theta}}$ by fixing $\alpha = 1$. This can be achieved through a reversible jump algorithm as proposed by Green (1995). An alternative would be to compute the χ and $\bar{\chi}$ statistic introduced in Section 2.2.6 for each random variable, and asses whether the data is asymptotically dependent or not. In case there is sufficient evidence in favor of asymptotic dependence, fix α to 1.

4.3 Bayesian inference for the constrained Heffernan and Tawn model

The aim of this section is to show that the Bayesian inference methodology introduced in Section 4.2 can also be applied to the constrained Heffernan and Tawn model which was introduced in Section 3.1.4. Exploring the confined parameter space poses a challenge to the deployed Markov chain Monte Carlo algorithms. The issues raised in Section 4.2 are expected to manifest themselves in this case as well.

4.3.1 Prior distributions

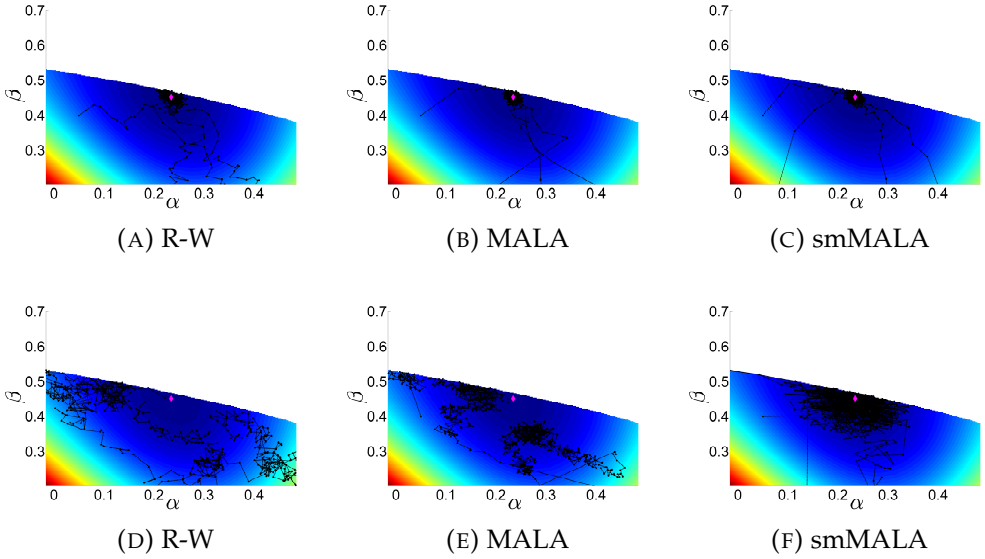
As inference for the constrained Heffernan and Tawn model is inherently similar to inference for the regular Heffernan and Tawn model, the uninformative prior distributions specified in Section 4.2.1 are adopted. However, the impact of the constraints on the support of α and β should be accounted for. Imposing the constraints through defining prior distributions with feasible support is not possible, as a closed form expression for the constraints on either α or β does not exist. Hence the constraints are imposed implicitly through the likelihood function, which returns an arbitrary high number if the constraints are not satisfied. An alternative approach to imposing the constraints is discussed in Chapter 5 as a recommendation for further research.

4.3.2 Results Case 1

For asymptotic independent data, the maximum likelihood estimates (4.13) are still feasible under the constraints proposed by Keef et al. (2013). However, as shown in Figure 3.5, the maximum likelihood estimate are close to the boundary of the feasible parameter space. The same issues as those raised in Section 4.2.3 — concerning uncertainty quantification when a parameter is on the boundary of the parameter space — will apply to statistical inference for the constrained Heffernan and Tawn model as well. Similar to Section 4.2, first the two parameter estimation problem will be discussed, before turning to inference for the full Heffernan and Tawn model.

Burn-in samples obtained when only α and β are estimated are shown in Figure 4.4(A)-(C). All four chains converge for each of the considered transition kernels. Summary statistics for the two parameter estimation case are presented in Table E.1, and additional diagnostic plots are provided in Appendix E in Figure E.1. These are not discussed separately, as results for jointly estimating all four parameters are more informative.

Several trials for different starting values — the results of which are not included — showed the MALA does not converge if both starting values $\alpha_0, \beta_0 < 0$. Although this region of the parameter space is feasible under the constraints, gradients of the negative log-likelihood function are so large the MALA will keep making proposals outside the feasible parameter space. Thus preventing the Markov chains from moving around. Comparing Figure 4.4(D)-(E) to



(A) R-W (B) MALA (C) smMALA

(D) R-W (E) MALA (F) smMALA

FIGURE 4.4: The first 200 burn-in samples for the two parameter estimation problem for Case 1 data, shown in (A)-(C), as well as the first 500 burn-in samples for the four parameter estimation of the constrained Heffernan and Tawn model, shown in (D)-(F). The constrained maximum likelihood estimates (♦) are also indicated.

Figure 4.4(F) suggests the simplified mMALA converges, while the random walk transition kernel and MALA struggle to converge during burn-in. It is remarkable that even for proper starting values, MALA yields poorer results compared to the constant Heffernan and Tawn model even for proper starting values, as shown in Figure 4.2(D).

The trace-plots shown in Appendix E in Figure E.1(A)-(F) for inference on α, β, and Figure E.2 for inference on all four parameters, suggest convergence of the posterior samples for each of the estimated parameters for all three proposal mechanisms. The posterior samples for β shown in Figure E.2(D)-(F) appear to be biased. The scatterplot matrix shown in Figure E.4-E.6 suggests that the bias is introduced because the maximum likelihood estimate $\hat{\beta}_{\text{MLE}}$ is very close to the boundary of the constrained parameter space. The strong correlation in the posterior samples of β and ψ^2 induces the bias in the posterior sample for β onto the posterior sample for ψ^2.

The summary statistics presented in Table E.1 show great resemblance to

the results for the constant Heffernan and Tawn model for Case 1 data, presented in Section 4.2.2. The Gelman-Rubin statistic confirms convergence of the posterior samples. The different proposal mechanisms perform equally well for the two parameter estimation problem, while the simplified mMALA outperforms its peers for the four parameter estimation problem in terms of effective sample size and effective sample size per second.

As the maximum likelihood estimates are close to the boundary and the constraints are imposed implicitly, the proposal mechanisms can not sense the presence of the boundary of the constrained parameter space. Consequently, a considerable number of proposals will be rejected. Hence a smaller step size is required to guarantee acceptable acceptance rates. It is not trivial how to address this issue. As previously stated, incorporating the constraints in the support of the prior distributions explicitly is not possible. An alternative approach would be to incorporate the constraints directly in the proposal distribution, by sampling from a truncated multivariate Gaussian distribution. This possibility is raised as a promising recommendation for further research in Chapter 5.

4.3.3 Results Case 2

Issues related to inference on the constrained Heffernan and Tawn model for asymptotically dependent data have been addressed in Section 4.2.3 and issues related to imposing the constraints have been discussed in Section 4.3.2. Hence it suffices to only briefly discuss the results presented in this section, as many of the previously discussed issues will be apparent in this case as well.

The unconstrained maximum likelihood estimates are not feasible under the constraints. Maximum likelihood estimates obtained by minimizing the negative log-likelihood function subject to the constraints for Case 2 data are given by

$$\hat{\alpha}_{\text{MLE}} = 1^{-}, \quad \hat{\beta}_{\text{MLE}} = 0.00, \quad \hat{\mu}_{\text{MLE}} = \text{-}0.63 \quad \text{and} \quad \hat{\psi}^2_{\text{MLE}} = 1.28. \tag{4.15}$$

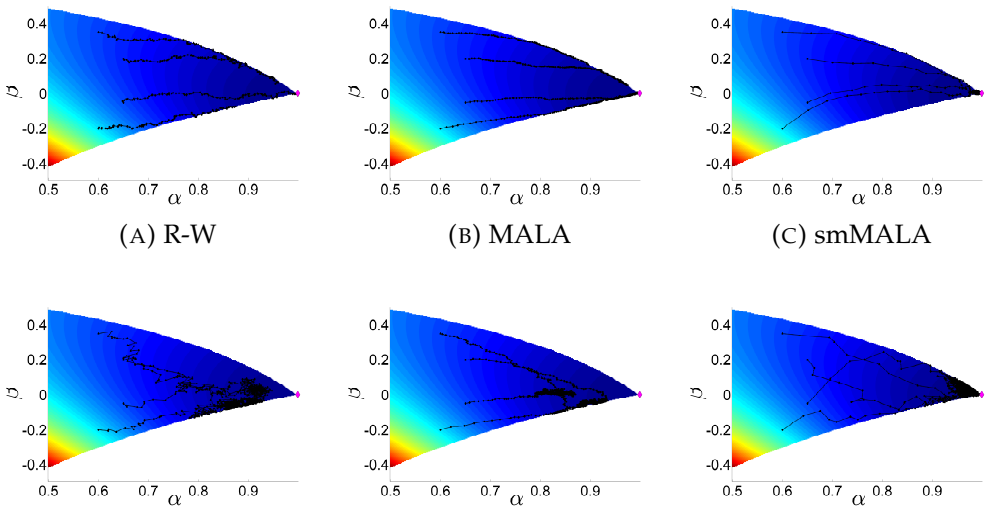
Remarkably, imposing the constraints leads to $\hat{\alpha}_{\text{MLE}} \approx \alpha_{\text{T}}$ and $\hat{\beta}_{\text{MLE}} \approx \beta_{\text{T}}$.

The first thing that stands out in Figure 4.5 is the confined geometry of the parameter space $\Omega_{\alpha\times\beta}$. The maximum likelihood estimates (4.15) are located in the corner of the profile-likelihood surface, as shown in Figure 4.5. Hence forcing the Markov chains to squeeze through the funnel shaped parameter space. Effectively exploring the confined parameter space is challenging and the step size of both the random walk proposal mechanism and MALA need to be reduced to ensure acceptable acceptance rates. As if inference on a parameter on the boundary of the parameter space was not challenging enough, the funnel

shaped parameter space makes it virtually impossible for the Markov chains to mix well.

Albeit the geometry of the parameter space, each of the Markov chains shown in Figure 4.5 appears to converge to the maximum likelihood estimate. However, for the four parameter estimation problem, both MALA and the random walk proposal mechanism struggle considerably to converge during burn-in. The step size for MALA is forced to be small in order to ensure feasible proposals. This leads to severe autocorrelation and very slow convergence. Given the bias associated to $\hat{\alpha}_{\text{MLE}} \in \partial\Omega_{\boldsymbol{\theta}}$, the Gelman-Rubin statistics presented in Table E.2 suggest convergence towards a stationary limit distribution. For the simplified mMALA, the effective sample size is even decent.

The traceplots shown in Figure E.8 reveal significant bias in the parameter estimates. The bias which was already apparent in the results presented in Section 4.2.3 is amplified by the funnel shaped geometry of the parameter space.



(A) R-W (B) MALA (C) smMALA

(D) R-W (E) MALA (F) smMALA

FIGURE 4.5: The first 200 burn-in samples for the two parameter estimation problem for Case 2 data, shown in (A)-(C), as well as the first 500 burn-in samples for the four parameter estimation of the constrained Heffernan and Tawn model, shown in (D)-(F). The constrained maximum likelihood estimates (♦) are also indicated.

4.4 Bayesian inference for the generalized Heffernan and Tawn model

Bayesian inference for a generalization of the Heffernan and Tawn model is the magnum opus of this thesis project. The aim of this section is to describe and demonstrate Bayesian inference for a generalization of the Heffernan and Tawn model proposed by Jonathan et al. (2014). Consider the bivariate Heffernan and Tawn model defined by (3.18), and assume the model parameters to be smooth functions with respect to a directional covariate $X_1 \in [-\pi, \pi]$ which is associated to the conditioning variable. See section 4.4.1 for an introduction to P-spline curves. Under the assumption that the limit distribution $G_{2|1}$ is Gaussian, the generalized Heffernan and Tawn model is given by

$$T_{\mathrm{L}}(Y_2) \mid T_{\mathrm{L}}(Y_1) = y, X_1 = x \sim \mathcal{N}\left\{\alpha(x)\, y + y^{\beta(x)}\mu(x), y^{2\beta(x)}\psi^2(x)\right\}. \quad (4.16)$$

Two different extremal dependence structures — both in the asymptotic independence class — are considered, to demonstrate the flexibility of the proposed model. The data is introduced in Section 4.4.2. Asymptotic dependent data and the constraints proposed by Keef et al. (2013) are disregarded because of the fundamental issues raised in Section 4.2 and 4.3. Although the implemented Bayesian inference framework could treat these cases, it does not make sense to demonstrate inference as long as the fundamental issues raised in Section 4.2.3 and 4.3 have been addressed. Prior distributions in relation to the proposed hierarchical Bayesian model are discussed in Section 4.4.3. By choosing a specific pair of prior distributions, a blend of the Metropolis-Hastings and Gibbs sampler can be adopted, see Section 4.4.4. Results for the two different types of data are presented in Section 4.4.5 and 4.4.6.

The discussion on the results focuses on convergence and mixing of the posterior samples presented in Appendix F. The results provided in Appendix F cover an example of uninformative– and informative prior distributions. The discussion in this section is primarily based on the results for uninformative priors, as it has greater practical significance.

4.4.1 Mathematical framework

The aim of this section is to introduce the generalized parameterization for the Heffernan and Tawn model proposed by Jonathan et al. (2014). The primary objective of the generalized parameterization is to accommodate weakly-identically distributed data and account for directional covariate effects in the extremal dependence structure. Parameters of the Heffernan and Tawn model are assumed to be smooth 2π periodic functions with respect to the directional

covariate X_i, which is related to the conditioning variable Y_i. Although higher dimensional covariates can be considered, the aim of this thesis is to provide a minimum working example based on a single covariate.

Statisticians distinguish parametric and non-parametric models. The former requires prohibitively strong assumptions on an appropriate functional form, while the latter class generally lacks structure. Combining the best of both yields a third class, referred to as semi-parametric models. Let $\theta(x) \in \left\{\alpha(x), \beta(x), \mu(x), \psi^2(x)\right\}$ denote an arbitrary parameter of the Heffernan and Tawn model. Following Jonathan et al. (2014), the semi-parametric model adopted in this thesis is defined by a matrix of *basis functions* $\mathbf{B}_\theta$ and a vector of *weights* $\boldsymbol{\zeta}_\theta$, such that $\theta(x) = \mathbf{B}_\theta(x)\boldsymbol{\zeta}_\theta$. Smoothness of the resulting function is controlled through a *roughness penalty* R. Hence the parameters of the Heffernan and Tawn model can be expressed as

$$\begin{aligned}
\alpha(x) &= \mathbf{B}_\alpha(x)\boldsymbol{\zeta}_\alpha, \\
\beta(x) &= \mathbf{B}_\beta(x)\boldsymbol{\zeta}_\beta, \\
\mu(x) &= \mathbf{B}_\mu(x)\boldsymbol{\zeta}_\mu, \quad \text{and,} \\
\psi^2(x) &= \mathbf{B}_{\psi^2}(x)\boldsymbol{\zeta}_{\psi^2}.
\end{aligned}$$

To explicitly define the model, an appropriate choice for the matrix of basis functions is required. Several suitable semi-parametric methods exist, e.g. Fourier– or wavelet transforms, but following Jonathan et al. (2014) the penalized basis spline (P-spline) curve parameterization proposed by Eilers and Marx (1996) is adopted. A brief introduction to spline curves is provided in Section 4.4.1 and 4.4.1.

B-Spline curves

A brief introduction to the basic concepts of basis spline functions is provided in this section. See the standard work by De Boor (1978) for more in-depth discussion on the subject. The starting point for the definition of spline curves is a sequence of *knots* or *control points*, denoted by $k_0, \ldots, k_{n_\mathrm{K}}$. Partition the covariate space Ω_X in n_K non-overlapping sub-intervals, such that

$$\inf\{x\colon x \in \Omega_X\} \coloneqq k_0 < k_1 < \cdots < k_{n_\mathrm{K}-1} < k_{n_\mathrm{K}} \coloneqq \sup\{x\colon x \in \Omega_X\},$$

Sub-intervals of equal length are usually adopted, but non-equidistant partitions can be considered as well. Given the knot vector $\mathbf{k} \coloneqq \{k_0, \ldots, k_{n_\mathrm{K}}\}$, a

spline basis functions (B-spline) of order r is defined by

$$B_{i,1}(x) \coloneqq \begin{cases} 1 & \text{if} \quad k_i \leq x < k_{i+1} \\ 0 & \text{otherwise} \end{cases},$$

for a polynomial of degree $r = 1$. For $r > 1$, the B-spline function is given by

$$B_{i,r}(x) \coloneqq \begin{cases} \frac{x-k_i}{k_{i+r-1}-k_i} B_{i,r-1}(x) + \frac{k_{i+r}-x}{k_{i+r}-k_{i+1}} B_{i+1,r-1}(x) & \text{if} \quad k_i \leq x < k_{i+r} \\ 0 & \text{otherwise} \end{cases}.$$

In matrix notation, for each $\theta \in \left\{\alpha, \beta, \mu, \psi^2\right\}$, the vector of basis functions is given by

$$\mathbf{B}_\theta(x) \coloneqq \left\{B_{1,r}(x), \ldots, B_{n_\kappa,r}(x)\right\}.$$

A *B-spline curve* S is a piece-wise polynomial function. Given a knot vector $\mathbf{k}$, a B-spline curve is uniquely defined by a linear combination of weights $\boldsymbol{\zeta}$ and B-spline basis functions of degree r, i.e. for $n_\kappa \geq r - 1$,

$$S(x) \coloneqq \sum_{i=0}^{n_\kappa} \zeta_i B_{i,r}, \quad \forall x \in [k_{r-1}, k_{n_\kappa}].$$

A recursion scheme proposed by De Boor et al. (1976) is implemented to generate appropriate spline bases.

P-spline curves

The flexibility of a B-spline curve is both a blessing and a curse. On the one hand, these curves allow arbitrary functions to be well approximated. On the contrary, it is not obvious how to come up with an optimal spacing of the knots and there is a fundamental trade-off between fidelity to the data and smoothness of the spline curve. As Lang and Brezger (2004) state, "a small number of knots may result in a function space which is not flexible enough to capture the variability of the data, while a large number of knots may lead to serious over-fitting". To address these issues, Eilers and Marx (1996) propose to impose a roughness penalty to control the smoothness of the B-spline curve, where they penalize "(higher-order) finite differences of the coefficients of adjacent B-splines". Although higher order *difference matrices* can be considered,

the first order difference matrix $\mathbf{D}$ is adopted and given by

$$\mathbf{D} = \begin{pmatrix} -1 & 1 & & \cdots & 0 \\ & -1 & 1 & & \\ \vdots & & \ddots & \ddots & \vdots \\ 1 & & \cdots & -1 & 1 \end{pmatrix}$$

The lower-left entry being equal to 1 ensures periodicity. Setting it to zero resets this option. The $n_{\mathrm{K}} \times n_{\mathrm{K}}$ *penalty matrix* is then defined as $\mathbf{P} \coloneqq \mathbf{D}^{\intercal}\mathbf{D}$, which yields

$$\mathbf{P} = \begin{pmatrix} 2 & -1 & \cdots & -1 & 1 \\ -1 & 2 & -1 & & \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ -1 & & & & -1 \\ 1 & \cdots & & -1 & 2 \end{pmatrix}.$$

Penalized B-spline curves are also referred to as P-spline curves. Smoothness of the P-spline curve is controlled through the roughness coefficient λ_θ. The *roughness penalty* R_θ is defined as

$$R_\theta \coloneqq \tfrac{1}{2}\lambda_\theta \boldsymbol{\zeta}_\theta^{\intercal}\mathbf{P}\boldsymbol{\zeta}_\theta.$$

4.4.2 Data for simulation study

Bayesian inference for the generalized Heffernan and Tawn is demonstrated through a simulation study for two different types of data, referred to as Case 1.1 and Case 1.2. Consider a bivariate Gaussian random variable $T_{\mathrm{L}}(\mathbf{Y})$ with mean $\mathbf{0}$ and covariance matrix defined in Section 4.4.2 and 4.4.2. Define $T_{\mathrm{L}}(Y_1)$ to be the conditioning variable. A sample $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ is considered, where the sample size is chosen differently for both cases.

Without loss of generality, let the covariate $X_1 \sim \mathcal{U}_{[-\pi,\pi]}$. The results presented here are based on a one dimensional covariate, as visualizing the results in this case is more natural and it suffices as a minimum working example. The two different cases are briefly introduced.

Compute the 95% empirical quantile $u_1 = \breve{F}^{\leftarrow}_{T_{\mathrm{L}}(Y_1)}(0.95)$ on the Laplace scale. As symmetric distributions are considered it suffices to demonstrate Bayesian inference on the observations $\{\boldsymbol{y} \colon y_1 > u_1\}$. Recall from Section 2.1.3 that a homogeneous Poisson process governs the rate at which threshold exceedances for the conditioning variable arrive. This guarantees that the distribution of the covariate values associated to the threshold exceedances is still uniform.

The assumption of homogeneity of the Poisson process seldomly holds in practice. Hence also affecting the distribution of the covariate values associated to the threshold exceedances. The issue is discussed by Jonathan et al. (2014) who adopt non-crossing quantile regression "for transformation of non-stationary marginal distributions to standard stationary form". Ensuring that the transfromed marginal distributions are stationary is important to guarantee that the limit distribution in (3.4) has equivalent marginal distributions. This will not be a problem for the two dimensional random variable and one dimensional covariate considered in this section, but should be taken into account in higher dimensional applications.

Dependence structure for Case 1.1

The first case concerns an extremal dependence structure that is itself a smooth and periodic function of the covariate X_i. Data is sampled from a multivariate Gaussian distribution with zero mean and unit variance. The correlation is defined by

$$\rho^2(x) = 0.5 + 0.2\sin(x) \in [0.3, 0.7], \quad \forall x \in [-\pi, \pi]. \tag{4.17}$$

Draw a sample $x_1, \ldots, x_n$ of $X \sim \mathcal{U}_{[-\pi,\pi]}$ with sample size $n_{\mathrm{T}} = 3 \cdot 10^4$ and compute $\rho^2(x)$. For each $x_1, \ldots, x_n$, generate a bivariate Gaussian random number where the covariance matrix is defined by (4.17).

Maximum likelihood estimates for the spline curves based on minimizing the negative log-likelihood functions and cross-validating the model to determine the optimal roughness coefficient, are not provided due to time constraints. See Jonathan et al. (2014) for related results. Given that (4.17) governs the extremal dependence structure, convergence of the fitted spline curve is assessed by comparing the results to to the true value $\alpha_{\mathrm{T}}(x) = \rho^2(x)$, which is defined by (4.17), and $\beta_{\mathrm{T}}(x) = 1/2$.

Dependence structure for Case 1.2

Following the simulation study presented by Jonathan et al. (2014), a mixture of Gaussian distributions is considered. This is interesting because approximating a piece-wise constant function by a smooth spline curve is challenging.

Consider six equidistantly space sectors over the domain $[-\pi, \pi]$. Generate a sample, with sample size $n_{\mathrm{T}}/6 = 3 \cdot 10^4$ for each sector from a bivariate Gaussian distribution with zero mean, unit variance and correlation for each sector defined by

$$\rho(x) \in \left\{\sqrt{0.6}, \sqrt{0.8}, \sqrt{0.2}, -\sqrt{0.7}, -\sqrt{0.3}, \sqrt{0.4}\right\} \tag{4.18}$$

Considering a mixture of both positively– and negatively correlated samples demonstrates that the proposed Bayesian inference framework is able to accommodate both, thanks to the unified parameterization arising from the Laplace marginal transformation. This can be regarded as a further generalization with respect to original model proposed by Heffernan and Tawn (2004) and the generalization proposed by Jonathan et al. (2014).

As for Case 1.1, the maximum likelihood estimate of the spline curve is not provided for aforementioned reasons. However, as the data for Case 1.2 is a mixture of Gaussian distributions with different extremal dependence structures governed by (4.18), the maximum likelihood estimates and true values for each different sector are shown when the results are reported in Section 4.4.6.

4.4.3 Prior distributions

Rather than directly estimating the parameters of the Heffernan and Tawn model, the weights $\boldsymbol{\zeta}_\theta$ and roughness coefficient λ_θ are the parameters to be estimated. Let $\boldsymbol{\eta}$ the set of hyper-prior parameters. The joint posterior distribution is given by

$$f_{\Theta|\mathbf{Y}}\left(\zeta_\theta, \lambda_\theta : \theta\left(x\right) \in \boldsymbol{\theta}\left(x\right) \mid \boldsymbol{y}, x\right) \propto f_{\mathbf{Y}|\Theta}\left(y_2 \mid y_1, x, \boldsymbol{\theta}\left(x\right)\right) \prod_{\theta \in \boldsymbol{\theta}} f_\Theta\left(\zeta_\theta \mid \lambda_\theta\right) f_\Theta\left(\lambda_\theta \mid \boldsymbol{\eta}_\theta\right).$$

The Bayesian inference scheme for the generalized Heffernan and Tawn model can be represented as a directed acyclical graph, shown in Figure 4.6. Assume

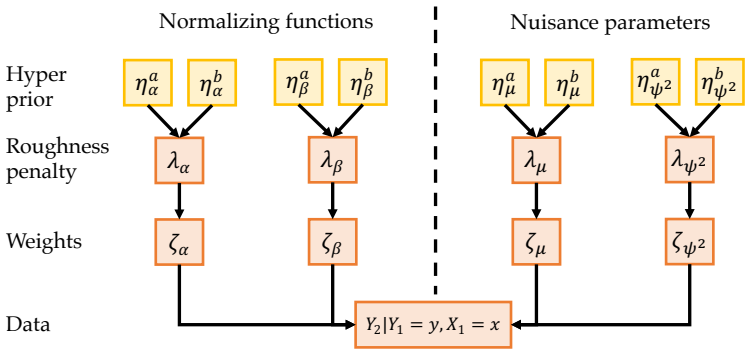


FIGURE 4.6: Hierarchical Bayesian model for the generalized Heffernan and Tawn model.

a Gaussian prior for the distribution of $\boldsymbol{\zeta}_\theta \mid \lambda_\theta$,

$$f_\Theta\left(\boldsymbol{\zeta}_\theta \mid \lambda_\theta\right) \propto \lambda_\theta^{1/2} \exp\left(-\frac{1}{2}\lambda_\theta \boldsymbol{\zeta}_\theta^\intercal \mathbf{P} \boldsymbol{\zeta}_\theta\right).$$

In this set-up, the roughness coefficient λ_θ can be interpreted as the *parameter precision* for the weight vector $\boldsymbol{\zeta}_\theta$. A gamma prior distribution is adopted for $\lambda_\theta \sim \mathcal{G}\left(\eta_\theta^a, \eta_\theta^b\right)$, which is a common choice for parameters that serve as parameter precision. As the Gamma distribution is conjugate to a Gaussian likelihood, the full conditional distributions for $\lambda_\theta \mid \boldsymbol{\zeta}_\theta$ are explicitly known, i.e.

$$\lambda_\theta \mid \boldsymbol{\zeta}_\theta \sim \mathcal{G}\left(\eta_\theta^a + \frac{n_{\mathrm{K}}}{2}, \eta_\theta^b + \frac{1}{2}\boldsymbol{\zeta}_\theta^{\intercal}\mathbf{P}\boldsymbol{\zeta}_\theta\right).$$

Both uninformative– and informative priors are considered to show how prior knowledge can affect the results. For each $\theta \in \boldsymbol{\theta}$, uninformative prior distributions are obtained by $\eta_\theta^a = 10^{-4}$ and $\eta_\theta^b = 10^4$. The informative priors are chosen as

$$\begin{aligned} \eta_\alpha^a &= 10^{-4}, & \eta_\alpha^b &= 10^4, \\ \eta_\beta^a &= 10, & \eta_\beta^b &= 20, \\ \eta_\mu^a &= 3, & \eta_\mu^b &= 4, \\ \eta_{\psi^2}^a &= 7, & \eta_{\psi^2}^b &= 15. \end{aligned}$$

4.4.4 Gibbs within Metropolis-Hastings algorithm

A sampling algorithm for the generalized Heffernan and Tawn model is introduced in this section. As the full posterior distribution of $\boldsymbol{\zeta}_\theta$ and $\lambda)\theta$ is analytically intractable, a Metropolis-Hastings sampler rather than a Gibbs sampler is adopted. However, simultaneous updating of both weights and roughness coefficients will affect the effectiveness of the Metropolis-Hastings sampler.

Luckily, conjugacy of the gamma prior distribution with a Gaussian likelihood permits sampling directly from the conditional distribution of $\lambda_\theta \mid \boldsymbol{\zeta}_\theta$ by Gibbs sampling through Algorithm 2. The resulting Gibbs within Metropolis-Hastings algorithm is presented in Algorithm 4.

Choosing appropriate starting values for $\boldsymbol{\zeta}_{\boldsymbol{\theta}}^{(0)}$ has proven to be non-trivial, as the resulting spline curves $\boldsymbol{\theta}^{(l)} = \mathbf{B}_{\boldsymbol{\theta}}\boldsymbol{\zeta}_{\boldsymbol{\theta}}^{(0)}$ will have to agree with the parameter space $\Omega_{\boldsymbol{\theta}}$. The current implementation proposes $\boldsymbol{\zeta}_{\boldsymbol{\theta}}^{(0)}$ at random untill these constraints are satisfied.

4.4.5 Results for Case 1.1

Results for the Case 1.1 data are presented and discussed in this section. All four parameters of the Heffernan and Tawn model are jointly estimated. The number of spline knots is fixed to 10 for each of the four parameters of the Heffernan and Tawn model. A third degree spline function is considered. The

Algorithm 4 Gibbs within Metropolis-Hastings algorithm for the generalized Heffernan and Tawn model

Initialize $\boldsymbol{\zeta}_{\boldsymbol{\theta}}^{(0)}$ and compute $\boldsymbol{\theta}^{(l)} = \mathbf{B}_{\boldsymbol{\theta}}\boldsymbol{\zeta}_{\boldsymbol{\theta}}^{(0)}$.
for $l = -l_{\text{B}}$ **to** l_{MAX} **do**
  for $\theta \in \boldsymbol{\theta}$ **do**
    $\lambda_\theta \mid \boldsymbol{\zeta}_\theta \sim \mathcal{G}\left(\eta_\theta^a + \frac{n_{\text{K}}}{2}, \eta_\theta^b + \frac{1}{2}\boldsymbol{\zeta}_\theta^{\intercal}\mathbf{P}\boldsymbol{\zeta}_\theta\right)$.
    $\boldsymbol{\zeta}^* \sim q\left(\boldsymbol{\zeta}^* \mid \boldsymbol{\zeta}^{(l)}\right)$
    $R_\theta^* = \frac{1}{2}\lambda_\theta \boldsymbol{\zeta}_\theta^{*\intercal}\mathbf{P}\boldsymbol{\zeta}_\theta^*$
    $R_\theta^{(l)} = \frac{1}{2}\lambda_\theta \boldsymbol{\zeta}_\theta^{(l)\intercal}\mathbf{P}\boldsymbol{\zeta}_\theta^{(l)}$
    $\theta^*(\mathbf{x}) = \mathbf{B}_\theta \boldsymbol{\zeta}_\theta^*$.
  end for
  $u \sim \log\left(\mathcal{U}_{[0,1]}\right)$
  $L_{\text{TOT}}^* = -\bar{\ell}_{\text{HT}}\left(\boldsymbol{\theta}^*(\mathbf{x})|\boldsymbol{y}, \mathbf{x}\right) + \log\left\{f_\Theta\left(\boldsymbol{\theta}^*(\mathbf{x}) \mid \mathbf{x}, \boldsymbol{\eta}\right)\right\}$
    $+ \log\left\{q\left(\boldsymbol{\theta}^*(\mathbf{x}) \mid \boldsymbol{\theta}^{(l)}(\mathbf{x})\right)\right\} + R_{\boldsymbol{\theta}}^*$
  $L_{\text{TOT}}^{(l)} = -\bar{\ell}_{\text{HT}}\left(\boldsymbol{\theta}^{(l)}(\mathbf{x})\middle|\boldsymbol{y}, \mathbf{x}\right) + \log\left\{f_\Theta\left(\boldsymbol{\theta}^{(l)}(\mathbf{x}) \mid \mathbf{x}, \boldsymbol{\eta}\right)\right\}$
    $+ \log\left\{q\left(\boldsymbol{\theta}^{(l)}(\mathbf{x}) \mid \boldsymbol{\theta}^*(\mathbf{x})\right)\right\} + R_{\boldsymbol{\theta}}^{(l)}$
  if $u \leq \min\left\{0, L_{\text{TOT}}^* - L_{\text{TOT}}^{(l)}\right\}$ **then**
    $\boldsymbol{\theta}^{(l+1)} = \boldsymbol{\theta}^*$
  else
    $\boldsymbol{\theta}^{(l+1)} = \boldsymbol{\theta}^{(l)}$
  end if
end for

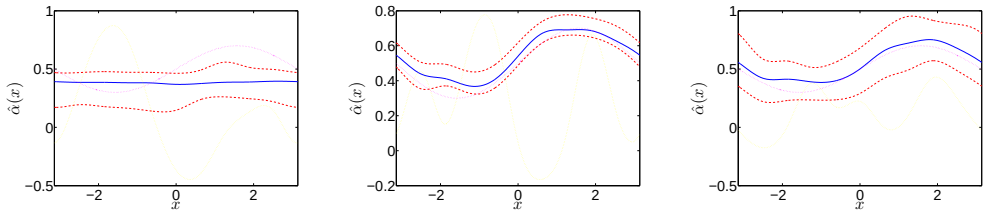
first $n_{\text{B}} = 2 \cdot 10^4$ are regarded as burn-in, and the next $n_{\text{S}} = 10^4$ realizations are assumed to be proper samples from the posterior distribution.

Presenting summary statistics and diagnostic plots for the generalized Heffernan and Tawn model is challenging as the total number of model parameters is large; $4 \cdot 10$ weights and 4 roughness coefficients in this case. Statistics and diagnostics plots concerning the Markov chains for the weights, are presented in Appendix F. Traceplots of the posterior samples for the weights, as well as the Gelman-Rubin statistic and effective sample size, are an average of the statistics computed for individual weights $\boldsymbol{\zeta}_{\boldsymbol{\theta}}$. Medians and confidence intervals based on the posterior sample of a single weight coefficient are not provided as their scale has no trivial interpretation. Traceplots and other diagnostic plots for the roughness coefficients are also presented in Appendix F. See the introduction to Appendix F for a more elaborate introduction to the presented results.

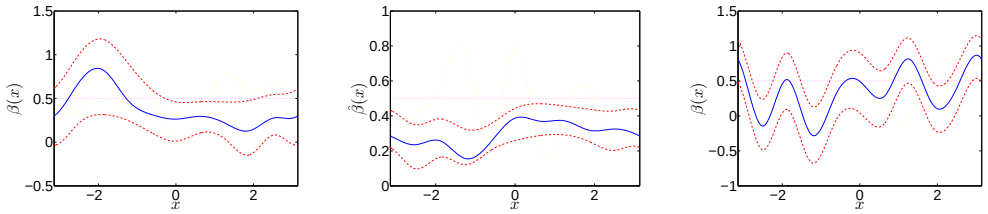
The median and 95% confidence interval for the posterior spline curves for α and β are shown in Figure 4.7. These summary statistics are computed by taking the median and 2.5% and 97.5% quantile of the posterior sample for each weight coefficient, and multiplying these results by the spline basis matrix $\mathbf{B}_\theta$. As shown in Figure 4.7(B)-(C), the median and 95% confidence interval for the α spline curve successfully converge to their true value. Figure 4.7(F) suggests

over-fitting in the spline curve for β based on the simplified mMALA, while for MALA, the curve seems biased, as shown in Figure 4.7(E).

The results for the random walk look suspicious and suggest the chains might not have converged yet. This statement is confirmed by the running mean of the likelihood for four different chains started at different starting values shown in Appendix F in Figure F.2(D). It is immediately clear that the Markov chains based on a random walk transitional kernel do not move around. Although the likelihood and autocorrelation for a single chain looks decent and, as shown in Figure F.2(A)-(G), the results for the random walk transition kernel are indeed not to be trusted. It is concluded that Bayesian inference for the generalized Heffernan and Tawn model based on a random walk transition kernel is not viable, as convergence of the chains is prohibitively slow. The diagnostic plots for the random walk will be provided for completeness, but a discussion of the results is omitted.



(A) R-W: $\alpha(x)$ (B) MALA: $\alpha(x)$ (C) smMALA: $\alpha(x)$



(D) R-W: $\beta(x)$ (E) MALA: $\beta(x)$ (F) smMALA: $\beta(x)$

FIGURE 4.7: Summary statistics of the spline curves based on the posterior samples for the weight coefficients $\boldsymbol{\zeta}_\alpha$ and $\boldsymbol{\zeta}_\beta$ for Case 1.1 data. The median and 2.5% and 97.5% quantiles are computed for the posterior samples of each single weight coefficient. The resulting smooth spline curves are obtained by multiplying these statistics with the basis matrix $\mathbf{B}_\theta$.

The running mean of the likelihood and the trace-plots of posterior samples for selected weights ζ_θ for the MALA and simplified mMALA, shown in Appendix F in Figure F.2(E)-(E) and Figure F.3 respectively, provide sufficient evidence that each of the chains — which are started at different starting values — have converged in distribution to the same stationary target distribution. However, these trace-plots reveal that chains based on the simplified mMALA show much better mixing compared to MALA. Figure F.3(K) even suggests that

certain chains for MALA get stuck for a considerable number of iterations.

If posterior samples are constant for certain ranges, this induces high correlation between these samples. The heatmap of the correlation matrices shown in Figure F.2(K)-(L) shows the pairwise correlation between the posterior samples of different weight coefficients for each of the four Heffernan and Tawn model parameters. For the simplified mMALA, the correlation matrix resembles the structure imposed by the restrained expected Fisher information matrix. For MALA on the contrary, there is significant correlation between certain posterior samples while there is not supposed to be any.

A closer look at the roughness penalties reveals why the β spline curve in Figure 4.7(F) appears to be over-fitting the data. The trace-plots and histograms shown in Figure F.4 and F.5 reveal the roughness coefficients λ_β and λ_{ψ^2} are close to 0 for the simplified mMALA, allowing the spline curves for β and ψ^2 to vary wildly. It indicates there is insufficient information in the data to control the smoothness of the spline curves for these parameters. Assuming a common step-size ε might also contribute to the problem, as the magnitude of the proposals might overrule the scale at which the roughness coefficient tries to enforce smoothness.

Adopting informative prior distributions for the roughness coefficient is an obvious way to address this issue. The spline curves shown in Figure F.6 indeed show greater smoothness. Figure F.7(K) shows a significant reduction in correlation for the off-diagonal entries in the correlation matrix of the posterior samples generated with MALA. The increase in correlation for simplified mMALA in the $\beta - \beta$, $\psi^2 - \psi^2$ and $\beta - \psi^2$ entries, as shown in Figure F.7(K), arises because the Hessian of the log-prior was assumed to be equal to zero, which is clearly no longer true and is not accounted for. It is suspicious that the histogram of the posterior sample for $\lambda_\beta, \lambda_\mu$ and λ_{ψ^2} shows nearly perfect resemblance with the prior densities, as shown in Figure F.9. This reassures that there is little information in the data to force a constant spline curve $\beta(x)$.

Summarizing, the random walk transition kernel is an inappropriate proposal mechanism for the generalized Heffernan and Tawn model. For uninformative prior distributions, both MALA and simplified mMALA yield satisfactory results, up to a certain degree of over-fitting in the spline curves for $\beta(x)$ and $\psi^2(x)$. Adopting informative priors $\boldsymbol{\eta}_\beta$ and $\boldsymbol{\eta}_{\psi^2}$ is a natural way to address this issue.

4.4.6 Results for Case 1.2

Now that Bayesian inference for the generalized Heffernan and Tawn model has been demonstrated for Case 1.1, the more challenging Case 1.2 is considered. Several of the issues raised in Section 4.4.5 are expected to manifest themselves in this case as well. Summary statistics for the spline curves for α and β are presented in Figure 4.8, based on the median and 95% empirical quantiles of the posterior sample for the weight coefficients.

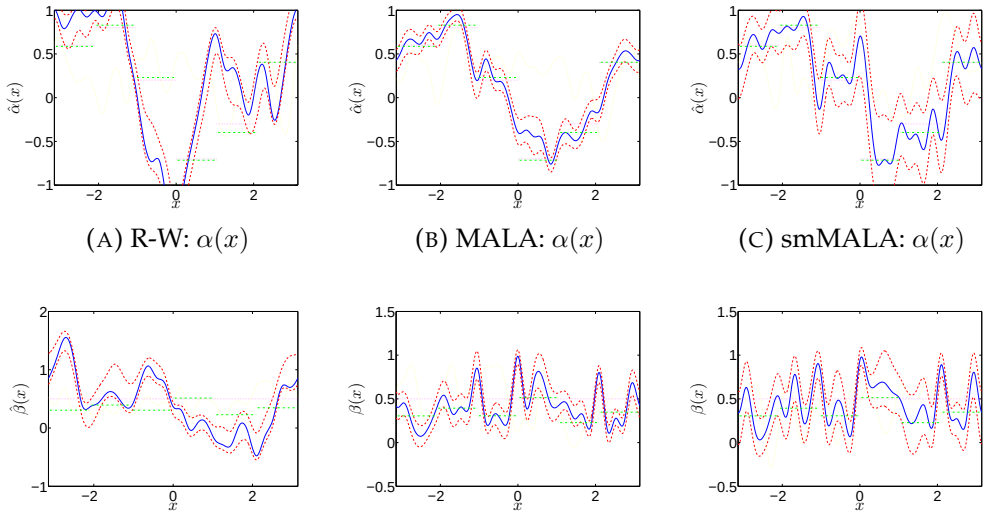
The random walk transition kernel — again — has failed to converge. This is confirmed by the running mean of the likelihood for four different chains, which do not converge to a common negative log-likelihood level, as shown in Figure F.12(D). By increasing the number of burn-in samples and applying *thinning* to the posterior samples, converged and properly mixing posterior samples should be obtained, as Theorem 4.1.1 dictates. However, these results suggest it will require an impractical number of burn-in iterations. Since the MALA and simplified mMALA yield properly converged posterior samples, no further attempts are made to tune the step size, number of iterations and hyper-prior parameters to end up with properly converged posterior samples for the random walk transition kernel. Hence the results, although included, are disregarded from further discussion in this section.

A first impression based on the results shown in Figure 4.8 suggest the model is an appropriate fit to the data. The spline curves retrieve the imposed extremal dependence structure very well. The confidence bounds for the simplified mMALA even cover the true values entirely.

Discontinuities at the boundary between sectors pose a significant challenge for the spline curves, as is clearly shown in Figure 4.8(E)-(F). The model struggles to be piece-wise constant within the sectors, but at the same time cope with the discontinuities at the boundaries between the sectors. Smoothness of the spline curves is controlled by a single roughness coefficient that has to find a balance to cope with these two features in the extremal dependence structure.

The overshoot at the discontinuities persists if informative priors are adopted, as shown by comparing Figure F.11 to F.16 in Appendix F. One way to address this issue is to locally increase the number of knots in order to provide greater control of the spline curve, as proposed by Eck and Hadenfeld (1995). Adding and removing spline knots is deemed to be a promising improvement to the methodology proposed in Section 4.4.

Markov chains based on MALA or simplified mMALA successfully converge in only 100-200 burn-in iterations, as shown in Figure F.12(E)-(F). Traceplots of the likelihood for the posterior sample suggest proper mixing and only a small degree of autocorrelation. However, the autocorrelation function itself shows significant and persistent autocorrelation in the posterior samples.



(A) R-W: $\alpha(x)$ (B) MALA: $\alpha(x)$ (C) smMALA: $\alpha(x)$

(D) R-W: $\beta(x)$ (E) MALA: $\beta(x)$ (F) smMALA: $\beta(x)$

FIGURE 4.8: Summary statistics of the spline curves based on the posterior samples for the weight coefficients $\boldsymbol{\zeta}_\alpha$ and $\boldsymbol{\zeta}_\beta$ for Case 1.2 data. The median and 2.5% and 97.5% quantiles are computed for the posterior samples of each single weight coefficient. The resulting smooth spline curves are obtained by multiplying these statistics with the basis matrix $\mathbf{B}_\theta$.

Trace-plots for a selection of the weight coefficients shows the simplified mMALA yields faster convergence and better mixing of the posterior samples, see Figure F.1. The correlation matrix for MALA reveals significant correlation between many posterior samples, while there is not supposed to be any. This incidates the simplified mMALA is superior to MALA in terms of mixing and autocorrelation. Even if the number of knots is increased, the result of which are not reported, the simplified mMALA proves to be reliable, while posterior samples based on MALA fail to converge or exhibit extensive autocorrelation.

Overfitting related to the discontinuities has already been discussed, but similar to Case 1.1, the splines curves also show overfitting in general. Trace-plots and histograms of the roughness coefficients are shown in Figure F.4 and F.15. The roughness coefficient is very close to zero for each parameter, for both MALA and simplified mMALA. Adopting informative priors to address this issue is surprisingly ineffective. Comparing Figure F.15 to F.20 reveals the posterior distributions of the roughness coefficients have a lot of probability mass close to zero, albeit the informative priors in favor of higher roughness penalties. This suggests there is strong information in the data in favor of small roughness penalties and wildly varying spline curves. Accommodating discontinuities in the spline curve takes a hefty toll on the smoothness.

Summarizing, the MALA and simplified mMALA successfully retrieve the imposed extremal dependence structure. Both algorithms yield convergent

posterior samples, and the simplified mMALA shows superior mixing of the Markov chains. Fitting a smooth spline curve to a piece-wise constant function is shown to be challenging, as the global roughness coefficient has to balance the extremal dependence function being constant within each sector, and discontinuous at the boundary between two sectors. Even strong informative priors are not an effective measure to address this issue.

Chapter 5

Conclusion and Discussion

The aim of this thesis project was to propose a robust methodology to quantify uncertainty regarding the parameters of the generalized Heffernan and Tawn model, which was proposed by Jonathan et al. (2014). This generalization assumes the parameters of the the Heffernan and Tawn model to be smooth functions with respect a covariate, by adopting penalized basis spline functions as introduced by Eilers and Marx (1996). Inference demonstrated by Jonathan et al. (2014), relies on likelihood minimization and cross-validating the entire model to obtain an optimal roughness coefficient. Sampling with replacement from the data and repeating the entire procedure was required to quantify uncertainty regarding the model parameters. This is very time consuming, tedious and puts practical constraints on the number of random variables that can be taken into account.

Bayesian inference arose as a natural candidate to address these issues. A novel Bayesian model was developed and demonstrated for the different models presented in the road map shown in Figure 1.2. By adopting the simplified manifold Metropolis adjusted Langevin algorithm (smMALA) proposed by Girolami and Calderhead (2011), the methodology presented in this thesis is concluded to be successful, easy to fit and robust. Even when a large number of knots is considered and the number of dimensions of the associated parameter space is large. As the main advantages are practical in nature, quantifying or proving superior performance compared to frequentist inference was omitted. The most important issues encountered along the way, as well as recommendations to address these issues, are discussed in this chapter. By adopting the proposed recommendations, the methodology proposed in this thesis can ultimately be adopted for inference on the generalized constrained Heffernan and Tawn model.

First, Bayesian inference on the constant Heffernan and Tawn model was demonstrated. Exploring the parameter space based on a random walk proposal mechanism results in highly correlated posterior samples. This issue was

addressed by adopting the simplified manifold Metropolis adjusted Langevin algorithm (smMALA) proposed by Girolami and Calderhead (2011). As the gradient of the likelihood function and the expected Fisher information matrix characterize the geometry of the parameter space, exploiting this information yields more sophisticated proposals. The smMALA is shown to outperform the other proposal mechanisms, even if the additional computation time is taken into account. Uninformative prior distributions proved to be sufficient to obtain satisfactory results.

For asymptotically dependent data, the maximum likelihood estimates for the parameters the Heffernan and Tawn model lie on the boundary of the the parameter space. Conventional methods fail to quantifying uncertainty in this case as asymptotic normality of the maximum likelihood estimator does not hold. In a Bayesian setting, Markov chain Monte Carlo algorithms will also fail to put probability mass on the boundary of the parameter space. A reparameterization was considered to address this issue. Adopting the reparameterization annihilates the resulting bias in the posterior samples. On the contrary, ensuring convergence of the Markov chains is much more challenging and the simplified mMALA even breaks down as the expected Fisher information matrix is non-invertible. Hence an alternative solution to inference on the Heffernan and Tawn model for asymptotic dependent data is required. One solution would be to fix $\alpha = 1$ when there is strong evidence in favor of the data being asymptotically dependent. Adopting the reversible jump Markov chain Monte Carlo algorithm proposed by Green (1995) can reduce the number of dimensions of the parameter space by fixing $\alpha = 1$. In similar fashion, the case where $\beta = 1$ could be accommodated. Implementing a reversible jump algorithm is recommended as it ensures that both classes of extremal dependence can be accommodated by a single algorithm. This is valuable as accommodating both asymptotic independent– and asymptotic dependent data is a distinguishing feature of the Heffernan and Tawn model, compared to other multivariate extreme value models. This recommendation was also raised by Lugrin et al. (2016).

Secondly, the indispensable constraints on the parameters of the Heffernan and Tawn model proposed by Keef et al. (2013) were taken into consideration. These constraints ensure a stochastic ordering on conditional quantiles under the Heffernan and Tawn model. Inference on the constrained Heffernan and Tawn model is challenging as the parameter space is severely confined, in particular for asymptotic dependent data. Previously raised issues regarding uncertainty quantification for the maximum likelihood estimator for asymptotic

dependent data, now apply to both classes of extremal dependence. In addition, as the constraints are currently imposed implicitly, the proposal mechanism will keep on making unfeasible proposals affecting the efficiency of the sampler. Simulating from a truncated multivariate Gaussian distribution is recommended to address this issue. However, as Botev (2016) points out, "simulation from the truncated multivariate Gaussian distribution in high dimensions is a recurrent problem in statistical computing and is typically only feasible by using approximate Markov chain Monte Carlo sampling". As the truncated multivariate Gaussian density should be determined at each iteration of the MALA or simplified mMALA, imposing the constraints this way is extremely computationally expensive. Botev (2016) proposes a "minimax tilting method for [. . .] generating samples from the truncated multivariate Gaussian distribution" which yields a promising approach to impose the constraints explicitly.

Finally, inference on the generalized Heffernan and Tawn model was demonstrated. When setting up the model, the number of knots and their spacing need to be defined. As it is not obvious beforehand how to choose these parameters, appropriate values are obtained by trial and error. Misspecification of these parameters can lead to serious over– or under-fitting. Moreover, discontinuities and local data sparsity can have a profound impact on the goodness of fit by overruling the roughness penalty that ensures smoothness of the spline curve. A method to add– or remove knots during burn-in of the Markov chains was proposed by Eck and Hadenfeld (1995). Implementing this method will save time when setting up the model and yields greater flexibility, as it is unfeasible to tune the number of knots and the knot spacing for each of the different parameters of the Heffernan and Tawn model.

Bayesian inference for the generalization of the Heffernan and Tawn model proposed by Jonathan et al. (2014) is the magnum opus of this thesis. A simulation study concerning asymptotically independent data for two distinct extremal dependence structures, shows the simplified mMALA yields posterior samples that converge and mix well. The proposed Bayesian inference framework was particularly successful for certain parameters of the generalized Heffernan and Tawn model, while it proved difficult to control overfitting for other parameters. The posterior samples of the roughness coefficient reveal that for other parameters there is little information in the data to control the smoothness of the spline curve. Strong informative priors can address this issue, but choosing appropriate prior distributions is not trivial. Considering higher order difference matrices or higher order spline functions will provide greater control and might improve the goodness of fit of the resulting spline curves. After discussing the results and performance of the algorithm with Philip Jonathan and David Randell, it is concluded that the proposed Bayesian

inference methodology is a superior alternative to the methodology adopted by Jonathan et al. (2014).

In addition to the results for the demonstrated Bayesian inference, several properties of the likelihood function of the Heffernan and Tawn model have been studied. The results of a simulation study suggest that the model parameters are *nearly unidentifiable* along a ridge in the likelihood function. What can be concluded with certainty, is that the parameters are unidentifiable if $\beta = 1$, and they are identifiable if $\mu = 0$. These issues stem from the Gaussianity assumption on the limit distribution for the residuals, which leads to non-linear relationships between certain model parameters. Closely related, the full observed– and expected Fisher information matrix have been shown to be non-positive definite unless the mean of the residual distribution is approximately equal to zero. The restrained expected Fisher information matrix was introduced to guarantee for the entire parameter space. Further research might focus on formalizing the assertions on identifiability of the Heffernan and Tawn model parameters.

The influence of sample size, dependence in the original data sample and the non-exceedance probability on the sampling distribution of the maximum likelihood estimator have been studied. For weakly dependent data, the maximum likelihood estimator was shown to be severely biased. All else being equal, adopting the non-exceedance probability $p = 0.95$ was shown to minimize the mean squared error of the maximum likelihood estimator for the parameters of the Heffernan and Tawn model.

Miscellaneous concluding remarks

The simulation studies performed within the context of this thesis only consider the Gaussian distribution and the generalized extreme value distribution with symmetric logistic dependence function. Throughout this thesis, these two distributions are deemed to be representative for their respective class of extremal dependence. It is possible that by considering other distributions or real world data, remarks regarding the performance of the proposed Bayesian inference methodology for a particular class of extremal dependence no longer holds.

Some of the issues raised in Chapter 3 and 4 might be less profound on an aggregate level. For example, bias in posterior samples for α and μ for asymptotically dependent data is significant, but the impact on return level estimation under the posterior sample might be negligible. As time constraints did not permit to address return level estimation, it would be interesting to perform these analyses.

Large sample sizes were adopted in the simulation studies to reduce the impact of sampling errors on the results. The availability of large samples is a luxury not often encountered in real world applications. Considering small samples introduces a lot of uncertainty regarding the parameter estimates for the constant Heffernan and Tawn model, as shown in Section 3.2.6. For the generalized Heffernan and Tawn model, this issue will be even more profound. In addition, even if the total sample size is large, but data is locally sparse, parameter estimates for the generalized Heffernan and Tawn model have a significant exposure to sampling errors.

The constraints proposed by Keef et al. (2013) are a function of the threshold exceedances of the conditioning variable. This leads to an interesting conjunction where Bayesians and frequentists will part. Consider uncertainty quantification based on resampling data with replacement and computing maximum likelihood estimates for each sample. As the data affects the feasible parameter space, for each sample, the negative log-likelihood function will be minimized over a slightly different parameter space. This leads to strange structure in the bootstrap sample, in particular when the data sample is small. On the contrary, as the Bayesian paradigm assumes the observed data to be fixed, a single constrained parameter space is taken into consideration and hence the boundary of the parameter space is fixed as well.

Delft University of Technology
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft Institute of Applied Mathematics

Appendix

A thesis submitted to the
Delft Institute of Applied Mathematics
in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE
in
APPLIED MATHEMATICS

by

THIJS WILLEMS

Delft, the Netherlands,
November 2016

Appendix A

Derivations and Proofs

A.1 The link between the GEV and GP distribution

Let N be a Poisson random variable with rate parameter $\bar{p}$, and let $Y_1, \ldots, Y_N$ be generalized Pareto distributed. The generalized extreme value distribution arises as the distribution of normalized partial maxima $M_N \coloneqq \max\{Y_1, \ldots, Y_N\}$, since

$$\begin{aligned}
\Pr(M_N \leq x) &= \sum_{n=0}^{\infty} \Pr(N = n) \cdot \Pr\left(Y_1 \leq y, \ldots, Y_n \leq y\right), \\
&= \sum_{n=0}^{\infty} \mathrm{e}^{-\bar{p}} \frac{\bar{p}^n}{n!} \cdot \left\{1 - \left(1 + \xi \frac{y}{\sigma_u}\right)_+^{-1/\xi}\right\}^n, \\
&= \exp(-\bar{p}) \cdot \exp\left[\bar{p}\left\{1 - \left(1 + \xi \frac{y}{\sigma_u}\right)_+^{-1/\xi}\right\}\right], \\
&= \exp\left\{-\left(\frac{\sigma_u}{\sigma}\right)^{-1/\xi} \left(1 + \xi \frac{y}{\sigma_u}\right)_+^{-1/\xi}\right\}, \\
&= \exp\left\{-\left(1 + \xi \frac{y + u - \mu}{\sigma}\right)_+^{-1/\xi}\right\}, \\
&= G_\xi\left(y + u\right).
\end{aligned}$$

A.2 Bivariate Distributions

- **Bilogistic Model**
  By Joe et al. (1992).

$$h(x;a,b) = \frac{1}{2}\frac{(1-a)(1-u)u^{1-a}}{\{a(1-u)+bu\}(1-x)x^2} \tag{A.1}$$

  where $0 < x < 1$ and $0 < a, b < 1$. The scalar $u = u(x,a,b)$ is given by the solution of

$$(1-a)(1-w)(1-u)^b - (1-a)wu^b = 0 \tag{A.2}$$

- **Dirichlet Model**
  By Coles et al. (1991).

$$h(x;a,b) = \frac{ab}{2}\frac{\Gamma(a+b+1)}{\Gamma(a)\Gamma(b)}\frac{(aw)^{a-1}\{b(1-w)\}^{b-1}}{\{aw+b(1-w)\}^{a+b+1}} \tag{A.3}$$

A.3 Deriving the negative log-Likelihood function

The likelihood function is defined by

$$L_{\text{HT}}(\boldsymbol{\theta} \mid \mathbf{y}) = \prod_{l=1}^{n} f_{\text{HT}}\left(y_{2l} \mid \boldsymbol{\theta}, y_{1l}\right). \tag{A.4}$$

In a bivariate setting, the density function for the Heffernan and Tawn model for a Gaussian residual distribution is given by

$$f_{\text{HT}}\left(y_{2l} \mid \boldsymbol{\theta}, y_{1l}\right) \coloneqq \frac{1}{\sqrt{2\pi y_{1l}^{2\beta}\psi^2}} \exp\left\{-\frac{1}{2}\frac{\left(y_{2l} - \alpha y_{1l} - y_{1l}^{\beta}\mu\right)^2}{y_{1l}^{2\beta}\psi^2}\right\}. \tag{A.5}$$

The negative log-likelihood is defined by $\bar{\ell}\left(\theta \mid \mathbf{y}\right) \coloneqq -\log L\left(\theta \mid \mathbf{y}\right)$. After substitution, this yields

$$\begin{aligned}
\bar{\ell}_{\text{HT}}(\boldsymbol{\theta} \mid \mathbf{y}) &= -\log \mathrm{L}_{\text{HT}}\left(\boldsymbol{\theta} \mid \mathbf{y}\right), \\
&= -\log\left\{\prod_{l=1}^{n} f_{\text{HT}}\left(y_{2l} \mid \boldsymbol{\theta}, y_{1j}\right)\right\}, \\
&= -\sum_{l=1}^{n} \log f_{\text{HT}}\left(y_{2l} \mid \boldsymbol{\theta}, y_{1l}\right).
\end{aligned}$$

Which is equivalent to

$$\begin{aligned}
\bar{\ell}_{\text{HT}}(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{n}{2}\log(2\pi) + \frac{n}{2}\log\psi^2 + \beta\frac{1}{2}\sum_{l=1}^{n}\log y_{1j} \\
+ \frac{1}{2}\sum_{l=1}^{n}\frac{\left(y_{2j} - \alpha y_{1j} - y_{1j}^{\beta}\mu\right)^2}{y_{1j}^{2\beta}\psi^2}.
\end{aligned} \tag{A.6}$$

A.4 Derivatives of the likelihood function

As derived in Appendix A.3, the negative log-likelihood for the bivariate Heffernan and Tawn model, when Y_1 is the conditioning variable, is given by (A.6).

First order derivatives

$$\begin{aligned}
\frac{\partial \bar{\ell}}{\partial \alpha} &= -\frac{1}{\psi^2} \sum_{l=1}^{n} \frac{y_{2l} - \alpha y_{1l} - y_{1l}^{\beta}\mu}{y_{1l}^{2\beta-1}}, \\
\frac{\partial \bar{\ell}}{\partial \beta} &= -\frac{1}{\psi^2} \sum_{l=1}^{n} \log\left(y_{1l}\right) \left\{ \left(y_{2l} - \alpha y_{1l}\right) \frac{\left(y_{2l} - \alpha y_{1l} - y_{1l}^{\beta}\mu\right)}{y_{1l}^{2\beta}} + \psi^2 \right\}, \\
\frac{\partial \bar{\ell}}{\partial \mu} &= -\frac{1}{\psi^2} \sum_{l=1}^{n} \frac{y_{2l} - \alpha y_{1l} - y_{1l}^{\beta}\mu}{y_{1l}^{\beta}}, \\
\frac{\partial \bar{\ell}}{\partial \psi^2} &= -\frac{1}{\psi^2} \sum_{l=1}^{n} \frac{1}{2} \frac{\left(y_{2l} - \alpha y_{1l} - y_{1l}^{\beta}\mu\right)^2 - y_{1l}^{2\beta}}{y_{1l}^{2\beta}\psi^2}.
\end{aligned}$$

Second order derivatives

$$\begin{aligned}
\frac{\partial^2 \bar{\ell}}{\partial \alpha^2} &= \frac{1}{\psi^2} \sum_{l=1}^{n} y_{1l}^{2-2\beta}, \\
\frac{\partial^2 \bar{\ell}}{\partial \beta^2} &= \frac{1}{\psi^2} \sum_{l=1}^{n} \log^2\left(y_{1l}\right) \left(y_{2l} - \alpha y_{1l}\right) \frac{2y_{2l} - 2\alpha y_{1l} - y_{1l}^{\beta}\mu}{y_{1l}^{2\beta}}, \\
\frac{\partial^2 \bar{\ell}}{\partial \mu^2} &= \frac{n}{\psi^2}, \\
\frac{\partial^2 \bar{\ell}}{\partial \psi^2 \partial \psi^2} &= \frac{1}{\psi^2} \sum_{l=1}^{n} \frac{1}{\psi^4} \frac{\left(y_{2l} - \alpha y_{1l} - y_{1l}^{\beta}\mu\right)^2 - y_{1l}^{2\beta}}{y_{1l}^{2\beta}\psi^2}.
\end{aligned}$$

Mixed derivatives

$$\frac{\partial^2 \bar{\ell}}{\partial\alpha\partial\beta} = \frac{1}{\psi^2}\sum_{l=1}^{n}\log\left(y_{1l}\right)\frac{2y_{2l}-2\alpha y_{1l}-y_{1l}^{\beta}\mu}{y_{1l}^{2\beta-1}},$$

$$\frac{\partial^2 \bar{\ell}}{\partial\alpha\partial\mu} = \frac{1}{\psi^2}\sum_{l=1}^{n} y_{1l}^{1-\beta},$$

$$\frac{\partial^2 \bar{\ell}}{\partial\alpha\partial\psi^2} = \frac{1}{\psi^4}\sum_{l=1}^{n}\frac{y_{2l}-\alpha y_{1l}-y_{1l}^{\beta}\mu}{y_{1l}^{2\beta-1}},$$

$$\frac{\partial^2 \bar{\ell}}{\partial\beta\partial\mu} = \frac{1}{\psi^2}\sum_{l=1}^{n}\log\left(y_{1l}\right)\frac{y_{2l}-\alpha y_{1l}}{y_{1l}^{\beta}},$$

$$\frac{\partial^2 \bar{\ell}}{\partial\beta\partial\psi^2} = \frac{1}{\psi^4}\sum_{l=1}^{n}\log\left(y_{1l}\right)\left(y_{2l}-\alpha y_{1l}\right)\frac{y_{2l}-\alpha y_{1l}-y_{1l}^{\beta}\mu}{y_{1l}^{2\beta}},$$

$$\frac{\partial^2 \bar{\ell}}{\partial\mu\partial\psi^2} = \frac{1}{\psi^4}\sum_{l=1}^{n}\frac{y_{2l}-\alpha y_{1l}-y_{1l}^{\beta}\mu}{y_{1l}^{\beta}}.$$

A.5 Expected Fisher information matrix for the Heffernan and Tawn model

The expected Fisher information can be determined based on the derivatives presented in Appendix A.4. If the Heffernan and Tawn model with Gaussian residual distribution is specified by

$$T_{\mathrm{L}}\left(Y_2\right) \mid T_{\mathrm{L}}\left(Y_1\right) = y \sim \mathcal{N}\left(\alpha y + y^{\beta}\mu, y^{2\beta}\psi^2\right), \tag{A.7}$$

taking the expectation with respect Y_2 given $Y_1 = y$, for each of the second derivatives, yields the entries of the Fisher information matrix.

First order derivatives

$$\begin{aligned}
\mathrm{E}\left(\frac{\partial \bar{\ell}}{\partial \alpha}\right) &= 0, \\
\mathrm{E}\left(\frac{\partial \bar{\ell}}{\partial \beta}\right) &= 0, \\
\mathrm{E}\left(\frac{\partial \bar{\ell}}{\partial \mu}\right) &= 0, \\
\mathrm{E}\left(\frac{\partial \bar{\ell}}{\partial \psi^2}\right) &= 0.
\end{aligned}$$

Second order derivatives

$$\mathrm{E}\left(\frac{\partial^2 \bar{\ell}}{\partial \alpha^2}\right) = \frac{1}{\psi^2}\sum_{l=1}^{n} y_{1l}^{2-2\beta}, \tag{A.8}$$

$$\mathrm{E}\left(\frac{\partial^2 \bar{\ell}}{\partial \beta^2}\right) = \left(2 + \frac{\mu^2}{\psi^2}\right)\sum_{l=1}^{n} \log^2\left(y_1\right), \tag{A.9}$$

$$\mathrm{E}\left(\frac{\partial^2 \bar{\ell}}{\partial \mu^2}\right) = \frac{n}{\psi^2}, \tag{A.10}$$

$$\mathrm{E}\left(\frac{\partial^2 \bar{\ell}}{\partial \psi^2 \partial \psi^2}\right) = \frac{n}{2\psi^4}. \tag{A.11}$$

Mixed derivatives

$$
\begin{aligned}
\mathrm{E}\left(\frac{\partial^2 \bar{\ell}}{\partial\alpha\partial\beta}\right) &= \frac{1}{\psi^2}\mu\sum_{l=1}^{n}\log\left(y_{1l}\right)y_{1l}^{2-2\beta},\\
\mathrm{E}\left(\frac{\partial^2 \bar{\ell}}{\partial\alpha\partial\mu}\right) &= \frac{1}{\psi^2}\sum_{l=1}^{n}y_{1l}^{1-\beta},\\
\mathrm{E}\left(\frac{\partial^2 \bar{\ell}}{\partial\alpha\partial\psi^2}\right) &= 0,\\
\mathrm{E}\left(\frac{\partial^2 \bar{\ell}}{\partial\beta\partial\mu}\right) &= \frac{1}{\psi^2}\mu\sum_{l=1}^{n}\log\left(y_{1l}\right),\\
\mathrm{E}\left(\frac{\partial^2 \bar{\ell}}{\partial\beta\partial\psi^2}\right) &= \frac{1}{\psi^2}\sum_{l=1}^{n}\log\left(y_{1l}\right),\\
\mathrm{E}\left(\frac{\partial^2 \bar{\ell}}{\partial\mu\partial\psi^2}\right) &= 0.
\end{aligned}
$$

A.6 Derivatives under the the reparameterization

The reparameterization of the Heffernan and Tawn model introduced in Section 4.2 is given by

$$\alpha^* \coloneqq \log\left(\frac{1+\alpha}{1-\alpha}\right), \quad \beta^* \coloneqq -\log\left(1-\beta\right) \quad \text{and} \quad \psi^{2*} \coloneqq \log\left(\psi^2\right).$$

This requires an appropriate scaling of the first and second derivatives.

First order derivatives

The first order derivatives for the likelihood function of the reparameterized Heffernan and Tawn model can be derived through the chain rule. Since, for $\theta \in \{\alpha, \beta, \psi^2\}$, and $\theta^* \in \{\alpha^*, \beta^*, \psi^{2*}\}$, the first derivatives are given by

$$\frac{\partial \bar{\ell}}{\partial \theta^*} = \frac{\mathrm{d}\theta}{\mathrm{d}\theta^*}\frac{\partial \bar{\ell}}{\partial \theta}, \tag{A.12}$$

where each the first derivatives with respect to θ are presented in Appendix A.4. The first order derivatives of functions that the define the reparameterization are given by

$$\frac{\mathrm{d}\alpha}{\mathrm{d}\alpha^*} = \frac{1}{2}\mathrm{sech}\left(\frac{\alpha^*}{2}\right), \tag{A.13}$$

$$\frac{\mathrm{d}\beta}{\mathrm{d}\beta^*} = \exp\left(-\beta^*\right), \tag{A.14}$$

$$\frac{\mathrm{d}\psi^2}{\mathrm{d}\psi^{2*}} = \exp\left(\psi^{2*}\right), \tag{A.15}$$

where sech denotes the hyperbolic secant function. As the mixed partial derivatives are all zero, the Jacobian matrix $\mathbf{J}_{\text{RP}}$ of the reparameterization is given by

$$\mathbf{J}_{\text{RP}} \coloneqq \begin{pmatrix} \frac{1}{2}\mathrm{sech}\left(\frac{\alpha^*}{2}\right) & 0 & 0 & 0 \\ 0 & \exp\left(-\beta^*\right) & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \exp\left(\psi^{2*}\right) \end{pmatrix}. \tag{A.16}$$

Second order derivatives

$$\begin{aligned}\frac{\partial^2 \bar{\ell}}{\partial \theta^*} &= \frac{\partial}{\partial \theta^*}\left(\frac{\mathrm{d}\theta}{\mathrm{d}\theta^*}\frac{\partial \bar{\ell}}{\partial \theta}\right), \\ &= \frac{\partial \bar{\ell}}{\partial \theta}\left(\frac{\mathrm{d}}{\mathrm{d}\theta^*}\frac{\mathrm{d}\theta}{\mathrm{d}\theta^*}\right) + \frac{\mathrm{d}\theta}{\mathrm{d}\theta^*}\left(\frac{\partial}{\partial \theta^*}\frac{\partial \bar{\ell}}{\partial \theta}\right), \\ &= \frac{\partial \bar{\ell}}{\partial \theta}\frac{\mathrm{d}^2\theta}{\mathrm{d}\theta^{*2}} + \left(\frac{\mathrm{d}\theta}{\mathrm{d}\theta^*}\right)^2 \frac{\partial^2 \bar{\ell}}{\partial \theta^2}\end{aligned}$$

The Hessian matrix $\mathbf{H}_{\text{RP}}$ of the reparameterization is given by

$$\mathbf{H}_{\text{RP}} \coloneqq \begin{pmatrix} -4\sinh\left(\frac{\alpha^*}{2}\right)\operatorname{csch}(\alpha^*) & 0 & 0 & 0 \\ 0 & -\exp(-\beta^*) & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \exp\left(\psi^{2*}\right) \end{pmatrix},$$

where $\sinh$ denotes the hyperbolic sinus function and csch denotes the hyperbolic cosecant function.

Expected Fisher information matrix

$$\mathrm{E}\left(\frac{\partial^2 \bar{\ell}}{\partial \theta^*}\right) = \frac{\mathrm{d}^2\theta}{\mathrm{d}\theta^{*2}}\mathrm{E}\left(\frac{\partial \bar{\ell}}{\partial \theta}\right) + \left(\frac{\mathrm{d}\theta}{\mathrm{d}\theta^*}\right)^2 \mathrm{E}\left(\frac{\partial^2 \bar{\ell}}{\partial \theta^2}\right) \tag{A.17}$$

The first term on the right hand side in (A.17) drops out as the expectation of the first order derivative is zero for each $\theta \in \{\alpha, \beta, \psi^2\}$, as shown in Appendix A.5. Hence, the expectation of the second derivatives for the likelihood function of the reparameterized Heffernan and Tawn model, is defined by scaling each expression in (A.8) by the squared first derivatives defined by (A.13).

Because the off diagonal elements of the Jacobian matrix in (A.16) are all zero, the mixed derivatives are directly defined by scaling the mixed derivatives presented in Appendix A.5. For $\theta_1, \theta_2 \in \{\alpha, \beta, \psi^2\}$ such that $\theta_1 \neq \theta_2$, the mixed derivative is defined by

$$\mathrm{E}\left(\frac{\partial^2 \bar{\ell}}{\partial \theta_1^* \partial \theta_2^*}\right) = \frac{\mathrm{d}\theta_1}{\mathrm{d}\theta_1^*}\frac{\mathrm{d}\theta_2}{\mathrm{d}\theta_2^*}\mathrm{E}\left(\frac{\partial^2 \bar{\ell}}{\partial \theta_1 \partial \theta_2}\right).$$

Gradient of the metric tensor

Contrary to the simplified mMALA which assumes constant curvature of the parameter space manifold, full mMALA requires computing the metric tensor that defines the curvature of the manifold. The metric tensor is defined by the Jacobian matrix of the expected Fisher information matrix, who's entries are provided below.

Second Order Derivatives

$$\begin{aligned}
\nabla_\theta \mathrm{E}\left(\frac{\partial^2 \bar{\ell}}{\partial \alpha^2}\right) &= \left\{0 \quad -\frac{2}{\psi^2} y_1^{2-2\beta} \log(y_1) \quad 0 \quad -\frac{1}{\psi^4} y_1^{2-2\beta}\right\}^\intercal, \\
\nabla_\theta \mathrm{E}\left(\frac{\partial^2 \bar{\ell}}{\partial \beta^2}\right) &= \left\{0 \quad 0 \quad \frac{2}{\psi^2}\mu \log^2(y_1) \quad -\frac{1}{\psi^4}\mu^2 \log^2(y_1)\right\}^\intercal, \\
\nabla_\theta \mathrm{E}\left(\frac{\partial^2 \bar{\ell}}{\partial \mu^2}\right) &= \left\{0 \quad 0 \quad 0 \quad -\frac{1}{\psi^4}\right\}^\intercal, \\
\nabla_\theta \mathrm{E}\left(\frac{\partial^2 \bar{\ell}}{\partial \psi^2 \partial \psi^2}\right) &= \left\{0 \quad 0 \quad 0 \quad -\frac{1}{\psi^6}\right\}^\intercal.
\end{aligned}$$

Mixed Derivatives

$$\begin{aligned}
\nabla_\theta \mathrm{E}\left(\frac{\partial^2 \bar{\ell}}{\partial \alpha \partial \beta}\right) &= -\frac{1}{\psi^2}\frac{\log(y_1)}{y_1^{2\beta-2}} \{0 \quad 2\mu \log(y_1) \quad -1 \quad \mu\}^\intercal, \\
\nabla_\theta \mathrm{E}\left(\frac{\partial^2 \bar{\ell}}{\partial \alpha \partial \mu}\right) &= -\frac{1}{\psi^2} y_1^{1-\beta} \left\{0 \quad \log(y_1) \quad 0 \quad \frac{1}{\psi^2}\right\}^\intercal, \\
\nabla_\theta \mathrm{E}\left(\frac{\partial^2 \bar{\ell}}{\partial \alpha \partial \psi^2}\right) &= \{0 \quad 0 \quad 0 \quad 0\}^\intercal, \\
\nabla_\theta \mathrm{E}\left(\frac{\partial^2 \bar{\ell}}{\partial \beta \partial \mu}\right) &= -\frac{1}{\psi^2}\log(y_1) \{0 \quad 0 \quad -1 \quad \mu\}^\intercal, \\
\nabla_\theta \mathrm{E}\left(\frac{\partial^2 \bar{\ell}}{\partial \beta \psi^2}\right) &= -\frac{1}{\psi^2} \left\{0 \quad 0 \quad 0 \quad \frac{1}{\psi^2}\log(y_1)\right\}^\intercal, \\
\nabla_\theta \mathrm{E}\left(\frac{\partial^2 \bar{\ell}}{\partial \mu \partial \psi^2}\right) &= \{0 \quad 0 \quad 0 \quad 0\}^\intercal.
\end{aligned}$$

A.7 Derivatives for the log-prior distributions

The gradient and Hessian for the log-prior distributions used in Chapter 4 are derived in this section. If informative prior distributions are considered the Hessian of the log-prior distributions should be subtracted from the expected Fisher information matrix.

Gaussian prior distribution

First, consider the Gaussian distribution is chosen as a prior distribution for the parameter θ. Let $\boldsymbol{\eta}_\theta = \left(\eta_\theta^\mu, \eta_\theta^{\sigma^2}\right)$ denote the hyper parameters. The probability density function of the Gaussian distribution is given by

$$f_\Theta\left(\theta \mid \eta_\mu, \eta_{\sigma^2}\right) = \frac{1}{\sqrt{2\pi\eta_\theta^{\sigma^2}}} \exp\left\{-\frac{\left(\theta - \eta_\theta^\mu\right)^2}{2\eta_\theta^{\sigma^2}}\right\},$$

such that

$$-\log f_\Theta = \frac{1}{2}\log 2\pi + \frac{1}{2}\log \eta_\theta^{\sigma^2} + \frac{\left(\theta - \eta_\mu\right)^2}{2\eta_{\sigma^2}}$$

The gradient is given by

$$-\nabla_{\boldsymbol{\eta}} \log f_\Theta\left(\theta\right) = \begin{pmatrix} \frac{\theta - \eta_\mu}{\eta_{\sigma^2}} \\ \frac{1}{2\eta_\theta^{\sigma^2}} - \frac{\left(\theta - \eta_\theta^\mu\right)^2}{2\left(\eta_\theta^{\sigma^2}\right)^2} \end{pmatrix}$$

The Hessian matrix is given by

$$\mathbf{H} = -\frac{\partial^2}{\partial \boldsymbol{\eta}} \log f_\Theta\left(\theta\right) = \begin{pmatrix} \frac{1}{\eta_\theta^{\sigma^2}} & \frac{\theta - \eta_\theta^\mu}{\left(\eta_\theta^{\sigma^2}\right)^2} \\ \frac{\theta - \eta_\theta^\mu}{\left(\eta_\theta^{\sigma^2}\right)^2} & \frac{\left(\theta - \eta_\theta^\mu\right)^2}{\left(\eta_\theta^{\sigma^2}\right)^3} - \frac{1}{2\left(\eta_\theta^{\sigma^2}\right)^2} \end{pmatrix}$$

Gamma prior distribution

For the Gamma distribution, let $\boldsymbol{\eta}_\theta = \left(\eta_\theta^a, \eta_\theta^b\right)$ denote the hyper parameters. The probability density function of the Gamma distribution is given by

$$f_\Theta\left(\theta \mid \eta_\theta^a, \eta_\theta^b\right) = \frac{1}{\Gamma\left(\eta_\theta^a\right)\left(\eta_\theta^b\right)^{\eta_\theta^a}} \theta^{\eta_\theta^a - 1} \exp\left(-\frac{\theta}{\eta_\theta^b}\right),$$

such that

$$-\log f_\Theta = -\log\left\{\Gamma\left(\eta_\theta^a\right)\right\} - \eta_\theta^a \log \eta_\theta^b + \left(\eta_\theta^a - 1\right)\log\theta - \frac{\theta}{\eta_\theta^b}.$$

The gradient is given by

$$\nabla_{\boldsymbol{\eta}} - \log f_{\Theta}(\theta) = \begin{pmatrix} di\Gamma^{(0)}(\eta_\theta^a) + \log \eta_\theta^b - \log \theta \\ \frac{\eta_\theta^a \eta_\theta^b - \theta}{\left(\eta_\theta^b\right)^2} \end{pmatrix}$$

where $\mathrm{di}\Gamma^{(k)}$ denotes the k-th derivative of the di-Gamma function. The Hessian matrix is given by

$$\mathbf{H} = \frac{\partial^2}{\partial \boldsymbol{\eta}} - \log f_{\Theta}(\theta) = \begin{pmatrix} di\Gamma^{(1)}(\eta_\theta^a) & \frac{2\theta - \eta_\theta^a \eta_\theta^b}{\left(\eta_\theta^b\right)^3} \\ \frac{2\theta - \eta_\theta^a \eta_\theta^b}{\left(\eta_\theta^b\right)^3} & \frac{1}{\eta_\theta^b} \end{pmatrix}$$

A.8 Feasible starting values for minimization algorithm

Maximum likelihood estimates are obtained by minimizing (3.18) with the simplex search algorithm proposed by Reeds et al. (1998). The algorithm requires feasible starting values in order to converge. Proposing feasible starting values, in particular when the constraints proposed by Keef et al. (2013) are imposed, is not trivial. In particular since the constraints itself depend on the data sample. Algorithm 5 addresses this issues by proposing random starting values in the unconstrained parameter space. The proposed starting value $\boldsymbol{\theta}_0$ is either accepted or rejected based on whether the constraints are satisfied. Once the first feasible starting value is found, an arbitrary but small number of additional feasible starting values are identified. The proposed starting value with the smallest associated negative log-likelihood value is chosen as a relative optimal starting value. The pseudo code for this routine is provided in Algorithm 5. The boundaries of the unconstrained parameter space, and prior knowledge

Algorithm 5 Feasible starting value algorithm

Define i_{MAX} and j_{MAX}. Initialize $i, j = 1$.
while $i \leq i_{\text{MAX}}$ **and** $j \leq j_{\text{MAX}}$ **do**
  Sample $\boldsymbol{\theta}_0^*$ from appropriate prior distributions.
  if $\boldsymbol{\theta}_0$ is feasible under the Keef et al. (2013) conditions. **then**
    $\boldsymbol{\theta}_0^{(i)} \leftarrow \boldsymbol{\theta}_0^*$
    $i \leftarrow i + 1$
  end if
  $j \leftarrow j + 1$.
end while
$i_0 \leftarrow \arg\max_{1 \leq i \leq i_{\text{MAX}}} \bar{\ell}_{\text{HT}} \left(\boldsymbol{\theta}_0^{(i)} \mid \boldsymbol{y} \right)$
return $\boldsymbol{\theta}_0 \leftarrow \boldsymbol{\theta}_0^{(i_0)}$

of the true parameter values for Case 1 and Case 2 can be incorporated in the proposal distributions. For Case 1, let

$$\alpha_0 \sim \mathcal{B}(2,5), \quad \beta_0 \sim 1 - \mathcal{W}(2,2.5), \quad \mu_0 \sim \mathcal{N}(0.5,1) \quad \text{and} \quad \psi_0^2 \sim \mathcal{W}(1,1.5).$$

These distributions reflect strong prior knowledge of starting values that are likely to be feasible. For Case 2, the distribution for μ_0 and ψ_0^2 is similar, but

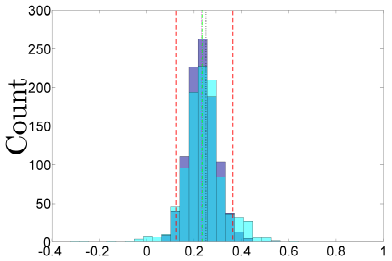
$$\alpha_0 \sim \mathcal{B}(10,2) \quad \text{and} \quad \beta_0 \sim 1 - \mathcal{W}(1,3.5).$$

The proposed methodology is practical yet certainly not optimal. The algorithm may not converge if the prior distributions are misspecified.

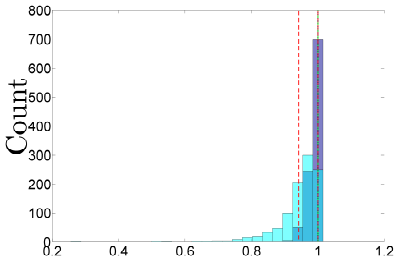
Appendix B

Additional figures Chapter 3

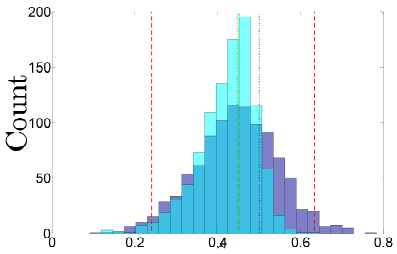
FIGURE B.1: Histogram of the bootstrapped maximum likelihood estimates for the parameters of the Heffernan and Tawn model for both Case 1 and Case 2 data. Maximum likelihood estimates (- · -) and 95% confidence bounds (- - -) obtained by computing the 2.5% and 97.5% quantiles of sample of maximum likelihood estimates. The bootstrap sample is obtained by sampling $n_B = 10^3$ times with replacement from the original data, while estimating the maximum likelihood estimates at each iteration.



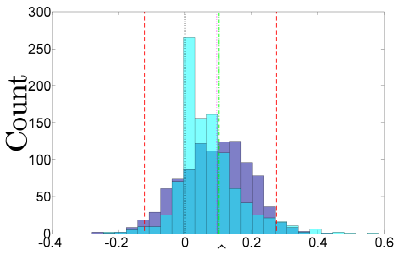
(A) Case 1: $\hat{\alpha}_{2|1}$.



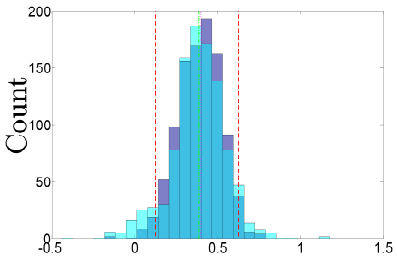
(B) Case 2: $\hat{\alpha}_{2|1}$.



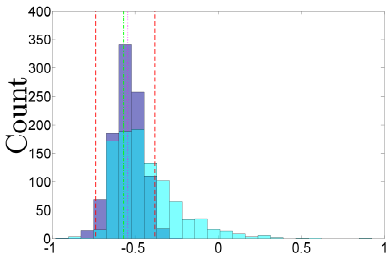
(C) Case 1: $\hat{\beta}_{2|1}$.



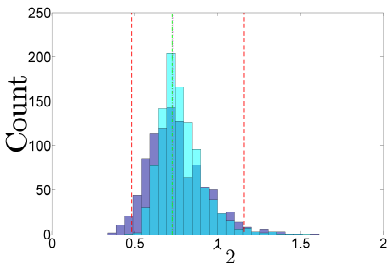
(D) Case 2: $\hat{\beta}_{2|1}$.



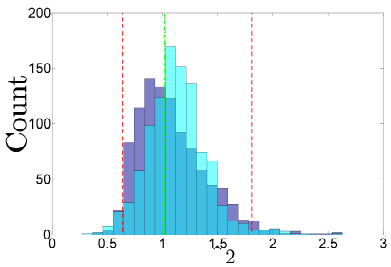
(E) Case 1: $\hat{\mu}$



(F) Case 2: $\hat{\mu}$

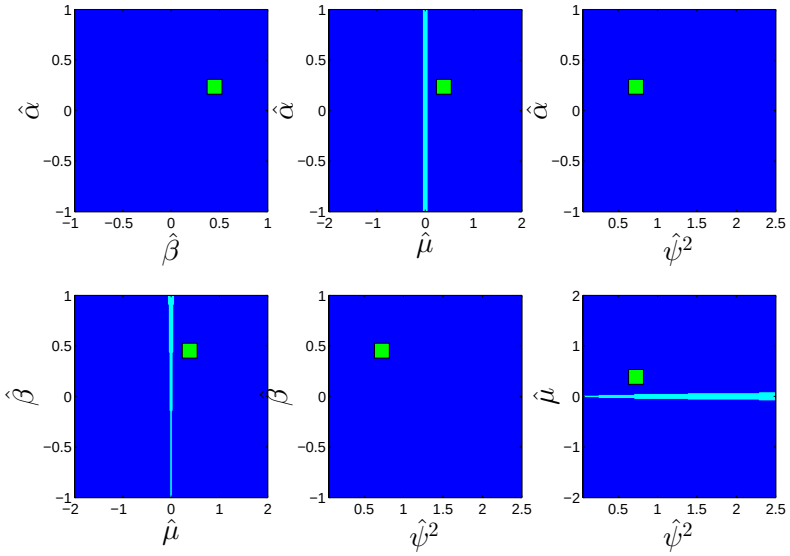


(G) Case 1: $\hat{\psi}^2$

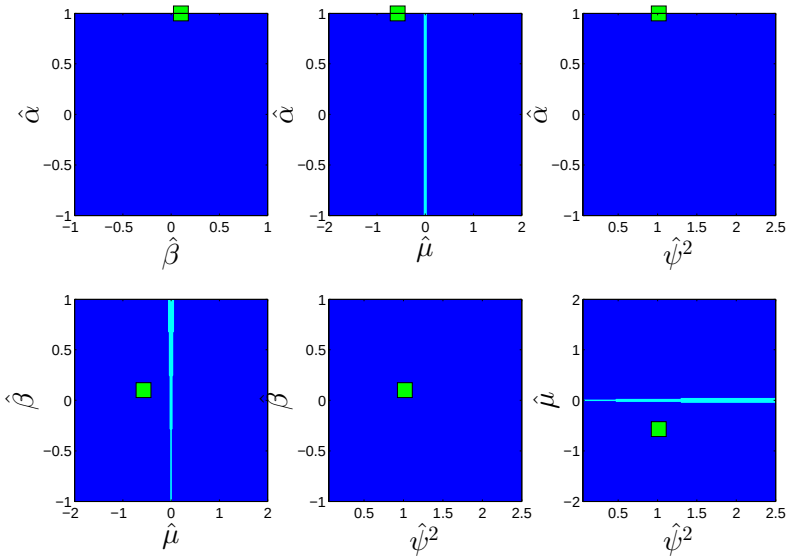


(H) Case 2: $\hat{\psi}^2$

FIGURE B.2: Subspace of $\Omega_{\boldsymbol{\theta}}$ that yields non-negative eigenvalues for the expected Fisher information matrix $\mathcal{I}(\boldsymbol{\theta})$ (cyan)) and restrained expected Fisher information matrix $\mathcal{I}^{\mathrm{R}}(\boldsymbol{\theta})$ (blue+cyan). The maximum likelihood estimates (■) are also indicated.

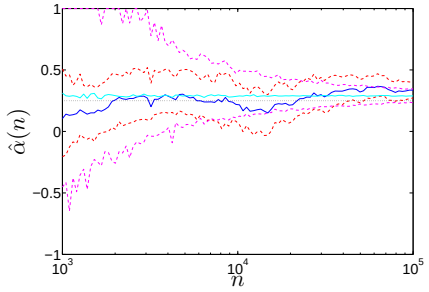


(A) Case 1: Asymptotically independent data.

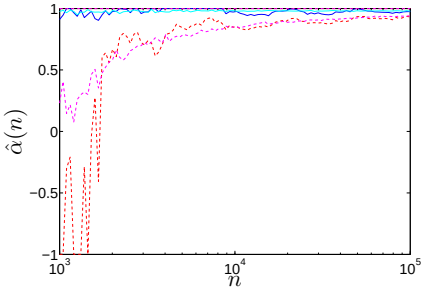


(B) Case 2: Asymptotically dependent data.

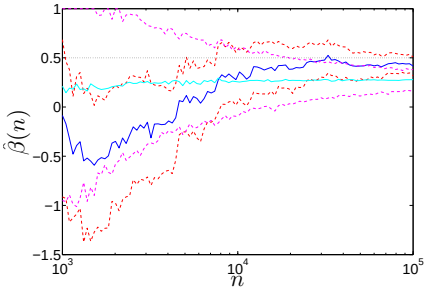
FIGURE B.3: Influence of sample size n_T of the maximum likelihood estimator for the Heffernan and Tawn model parameters. The non-exceedance probability $p = 0.95$. The median (—) and 95% confidence interval (- - -) of the sampling distribution are shown, as well as the median (—) and 95% confidence interval (- - -) of the sample of maximum likelihood estimates obtained by resampling with replacement from the data and estimating the maximum likelihood estimates for each iteration.



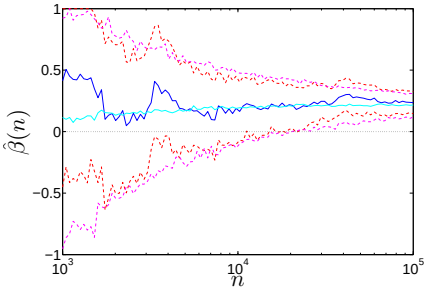
(A) Case 1: $\hat{\alpha}$.



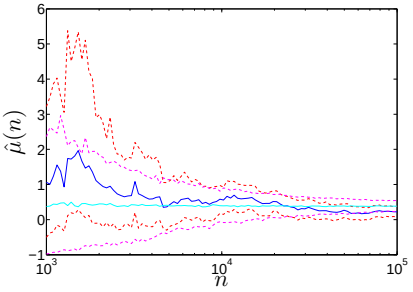
(B) Case 2: $\hat{\alpha}$.



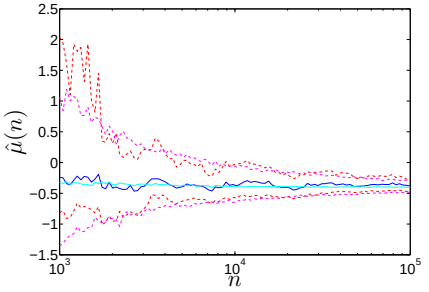
(C) Case 1: $\hat{\beta}$.



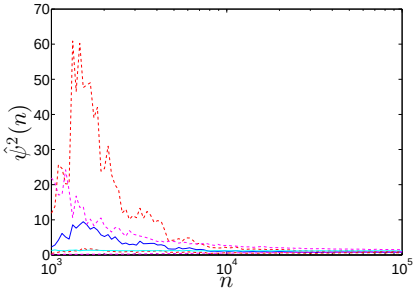
(D) Case 2: $\hat{\beta}$.



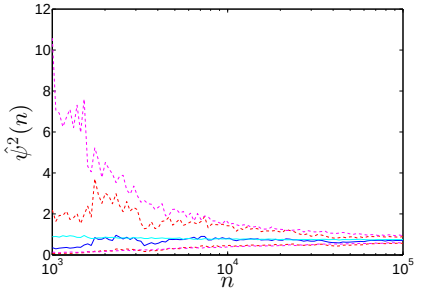
(E) Case 1: $\hat{\mu}$.



(F) Case 2: $\hat{\mu}$.

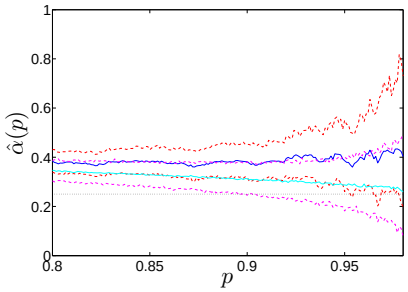


(G) Case 1: $\hat{\psi}^2$.

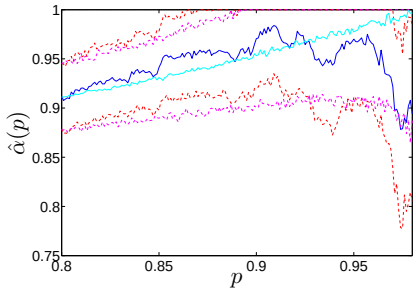


(H) Case 2: $\hat{\psi}^2$.

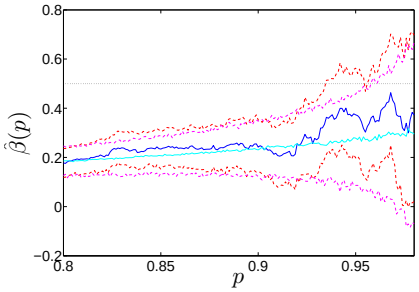
FIGURE B.4: Influence of threshold uncertainty of the maximum likelihood estimator for the Heffernan and Tawn model parameters. The median (—) and 95% confidence interval (- - -) of the sampling distribution are shown, as well as the median (—) and 95% confidence interval (- - -) of bootstrapped maximum likelihood estimates.



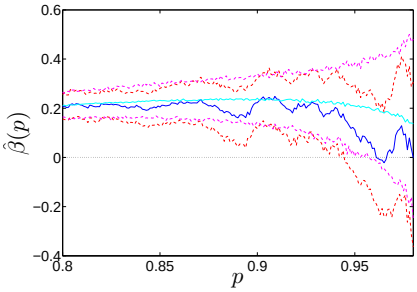
(A) Case 1: $\hat{\alpha}$.



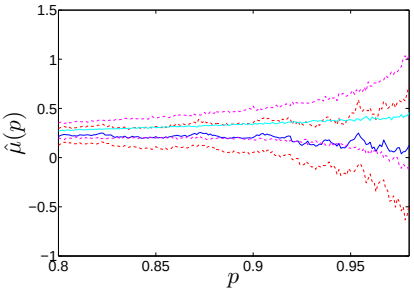
(B) Case 2: $\hat{\alpha}$.



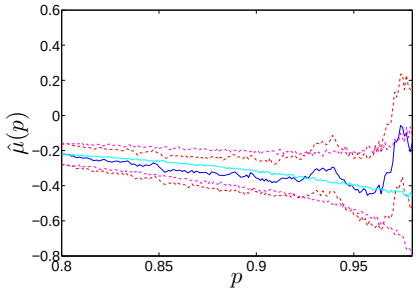
(C) Case 1: $\hat{\beta}$.



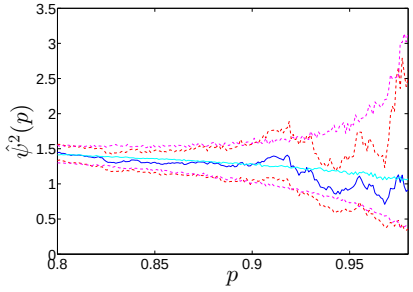
(D) Case 2: $\hat{\beta}$.



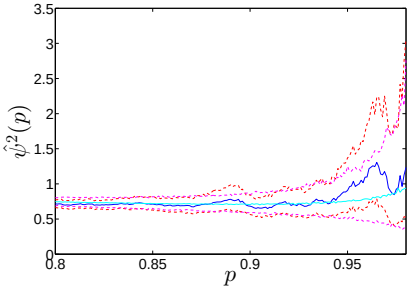
(E) Case 1: $\hat{\mu}$.



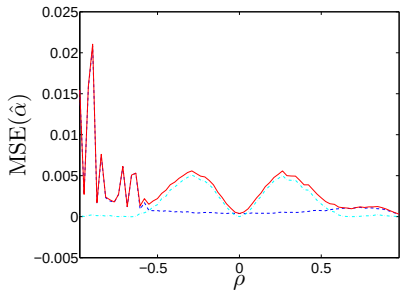
(F) Case 2: $\hat{\mu}$.



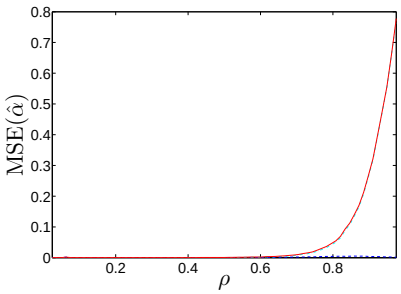
(G) Case 1: $\hat{\psi}^2$.



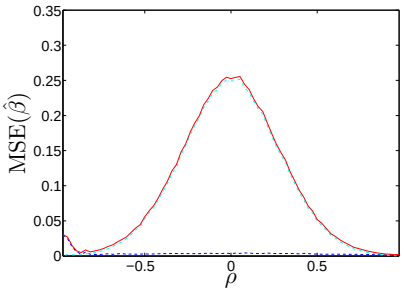
(H) Case 2: $\hat{\psi}^2$.



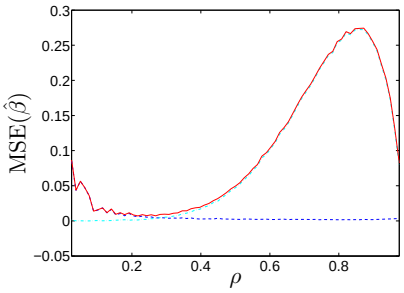
(A) Case 1: MSE $(\hat{\alpha})$.



(B) Case 2: MSE $(\hat{\alpha})$.

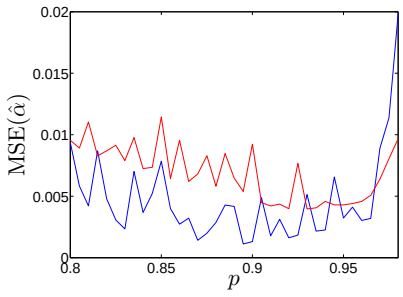


(C) Case 1: MSE $\left(\hat{\beta}\right)$.

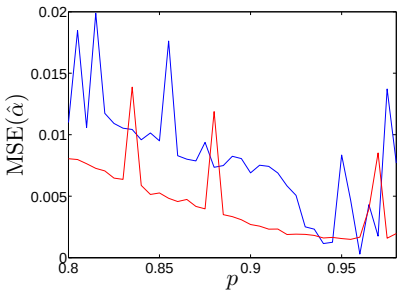


(D) Case 2: MSE $\left(\hat{\beta}\right)$.

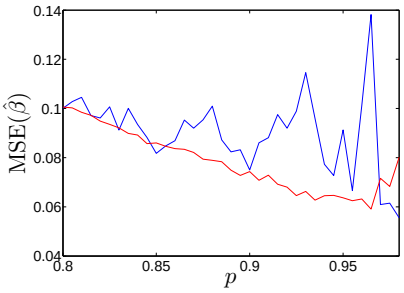
FIGURE B.5: Mean squared error (—) for the maximum likelihood estimator for the Heffernan and Tawn model parameters, as a function of ρ. The squared bias (- · -) and variance (- - -) are also shown. A sample of maximum likelihood estimates $\hat{\theta}_{\text{MLE}}$ is obtained by generating a new data sample ($n_{\text{T}} = 10^5$ and $p = 0.95$) at each iteration, and fitting the Heffernan and Tawn model to that sample.



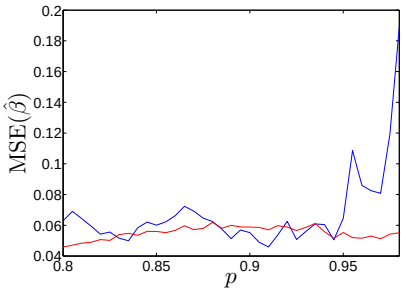
(A) Case 1: MSE $(\hat{\alpha})$.



(B) Case 2: MSE $(\hat{\alpha})$.



(C) Case 1: MSE $\left(\hat{\beta}\right)$.



(D) Case 2: MSE $\left(\hat{\beta}\right)$.

FIGURE B.6: Mean squared error for the maximum likelihood estimator for the Heffernan and Tawn model parameters, as a function of $p \in [0.8, 0.98]$. Samples with maximum likelihood estimates are obtained in two different ways. The first approach (—) relies on resampling with replacement, i.e. bootstrapping, from a single data sample. The second approach (—) approximates the sampling distribution of $\hat{\theta}_{\text{MLE}}$ by generating a new data sample at each iteration, and fitting the Heffernan and Tawn model to that sample.

Appendix C

Diagnostic plots: constant Heffernan and Tawn model

Diagnostic plots for the Bayesian analysis discussed in Section 4.2 are presented in this appendix. Both the asymptotic independent Case 1 data and asymptotic dependent Case 2 data are considered. The results are presented in consecutive order:

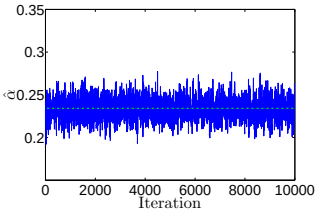
1. Case 1: α, β, see Figure C.1,
2. Case 1: All four parameters $\alpha, \beta, \mu, \psi^2$, see Figure C.2 and C.3,
3. Case 2: α, β, see Figure C.7,
4. Case 2: All four parameters $\alpha, \beta, \mu, \psi^2$, see Figure C.8 and C.9.

For each of the aforementioned cases, the following diagnostic plots are presented).

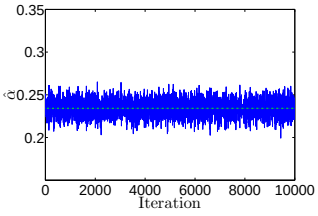
- Trace-plots of the posterior samples of the parameters to be estimated (—) with the maximum likelihood estimate (- · -) as a reference. A converged and properly mixing posterior sample should resemble a white noise process around the maximum likelihood estimate.
- Trace-plot of the likelihood. The absence of a trend and autoregressive features, as well as constant variability, indicate the Markov chain has converged to a stationary limit distribution.
- Autocorrelation function of the sample likelihood. The faster the autocorrelation decreases, the better.
- Running mean of the sample likelihood for four different chains with different starting values. If the running mean of the likelihood for each single chain converges to the same value, this suggests the Markov chains have converged to a global minimum in the negative log-likelihood function.

- The generated Markov chains, when all four parameters are estimated simultaneously, are summarized in scatterplot matrices. Burn-in is shown left of the diagonal, and the posterior samples are shown right of the diagonal with maximum likelihood estimates (■). The histograms on the diagonal show the median (· · ·) and 95% confidence interval (- - -) based on the 2.5% and 97.5% quantile as well as the maximum likelihood estimate (- · -).

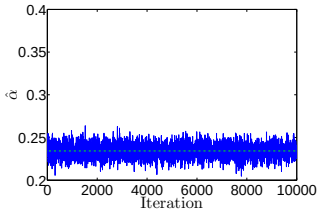
FIGURE C.1: **Case 1** (α, β): Traceplots of the posterior samples and maximum likelihood estimates (- - -), as well as diagnostic plots for the sample likelihood when α and β are estimated simultaneously.



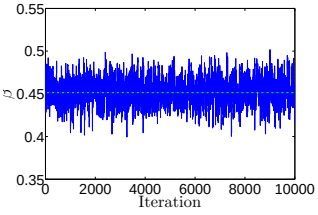
(A) R-W: $\hat{\alpha}$



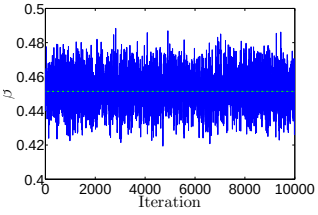
(B) MALA: $\hat{\alpha}$



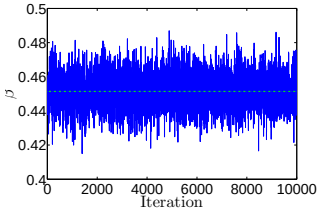
(C) smMALA: $\hat{\alpha}$



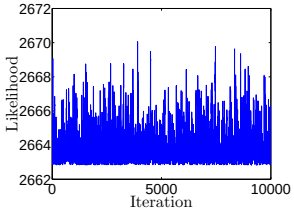
(D) R-W: $\hat{\beta}$



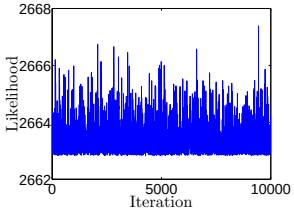
(E) MALA: $\hat{\beta}$



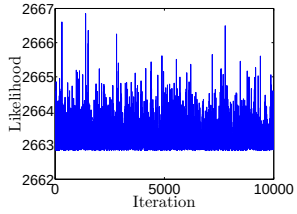
(F) smMALA: $\hat{\beta}$



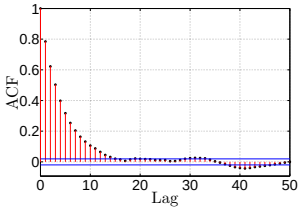
(G) R-W: Likelihood



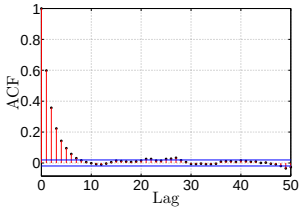
(H) MALA: Likelihood



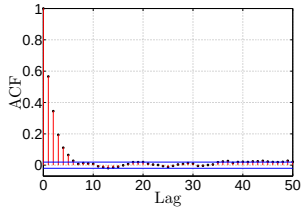
(I) smMALA: Likelihood



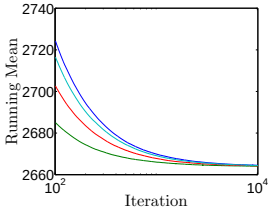
(J) R-W: ACF



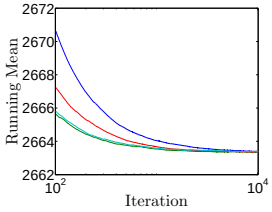
(K) MALA: ACF



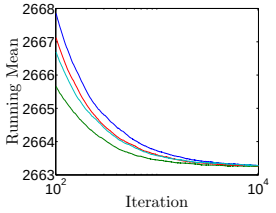
(L) smMALA: ACF



(M) R-W: Running Mean

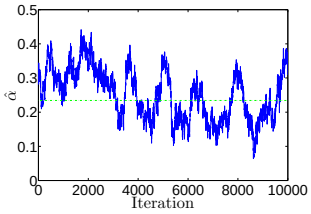


(N) MALA: Running Mean

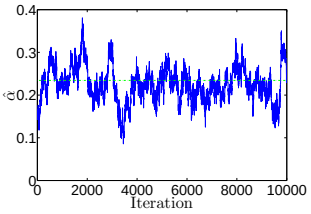


(O) smMALA: Running Mean

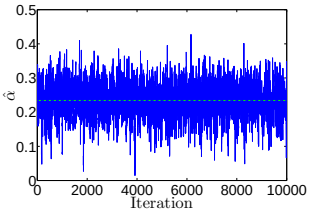
FIGURE C.2: **Case 1** $(\alpha, \beta, \mu, \psi^2)$: Traceplots of the posterior samples and maximum likelihood estimates (- - -) when all four parameters of the Heffernan and Tawn model are estimated simultaneously.



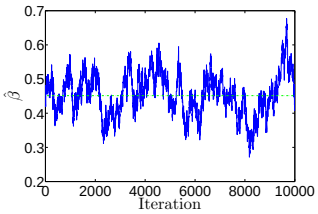
(A) R-W: $\hat{\alpha}$



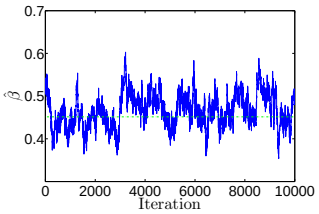
(B) MALA: $\hat{\alpha}$



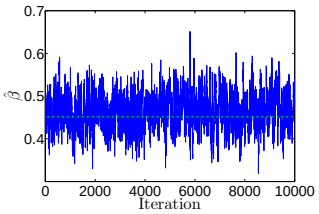
(C) smMALA: $\hat{\alpha}$



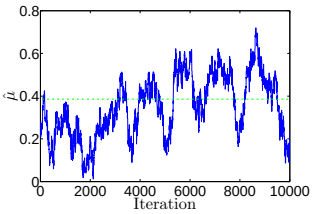
(D) R-W: $\hat{\beta}$



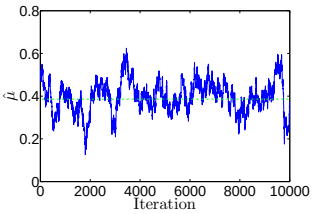
(E) MALA: $\hat{\beta}$



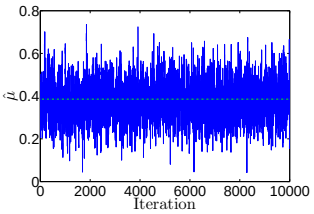
(F) smMALA: $\hat{\beta}$



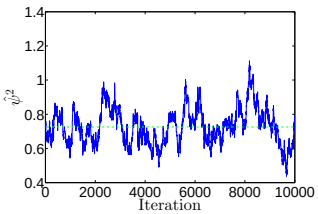
(G) R-W: $\hat{\mu}$



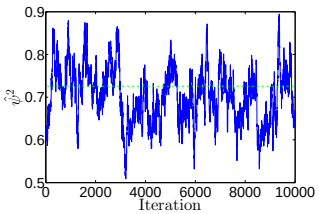
(H) MALA: $\hat{\mu}$



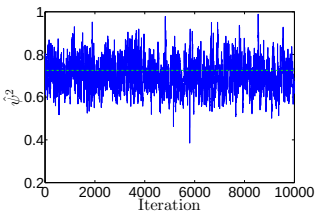
(I) smMALA: $\hat{\mu}$



(J) R-W: $\hat{\psi}^2$

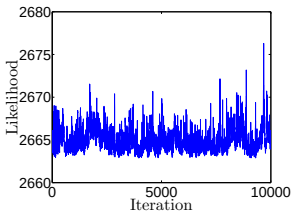


(K) MALA: $\hat{\psi}^2$

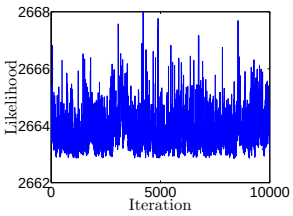


(L) smMALA: $\hat{\psi}^2$

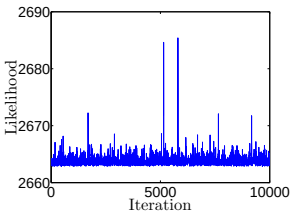
FIGURE C.3: **Case 1** $(\alpha, \beta, \mu, \psi^2)$: Diagnostic plots of the sample likelihood when all four parameters of the Heffernan and Tawn model are estimated simultaneously.



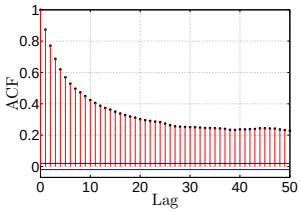
(A) R-W: Likelihood



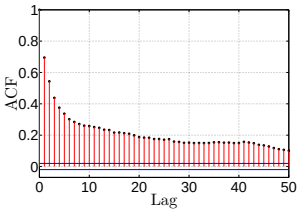
(B) MALA: Likelihood



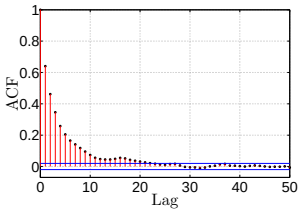
(C) smMALA: Likelihood



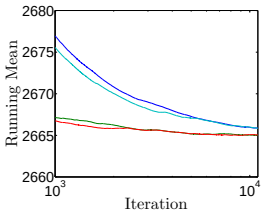
(D) R-W: ACF



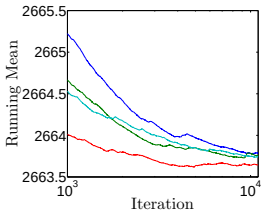
(E) MALA: ACF



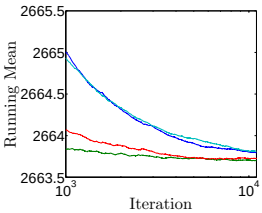
(F) smMALA: ACF



(G) R-W: Running Mean



(H) MALA: Running Mean



(I) smMALA: Running Mean

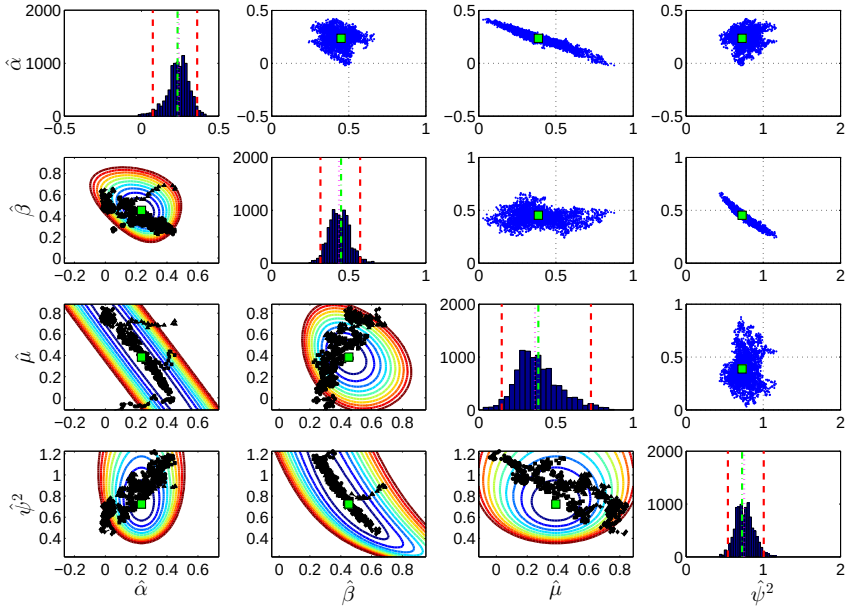


FIGURE C.4: **Case 1** $(\alpha, \beta, \mu, \psi^2)$: Scatterplot matrix of the burn-in sample (left of diagonal) and posterior sample (right of diagonal) based on the random walk transition kernel.

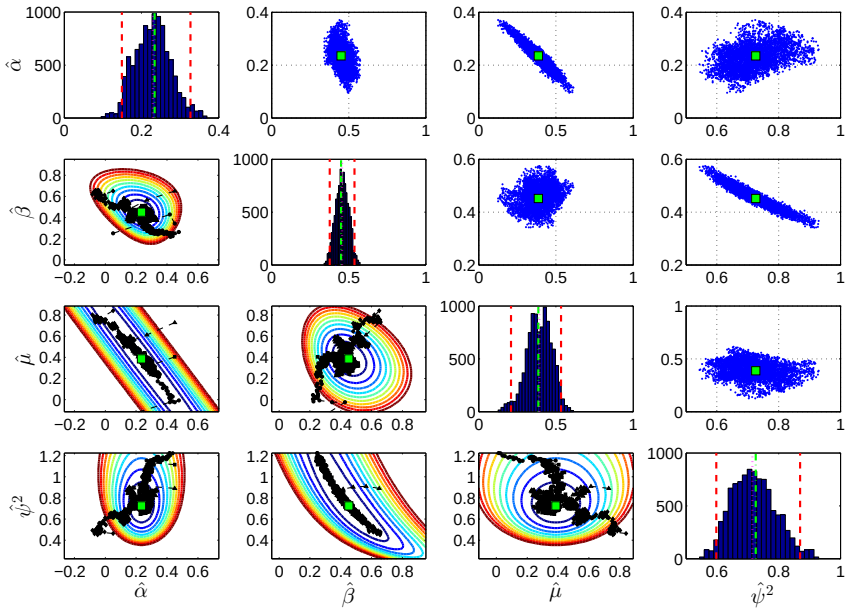


FIGURE C.5: **Case 1** $(\alpha, \beta, \mu, \psi^2)$: Scatterplot matrix of the burn-in (left of diagonal) and posterior sample (right of diagonal) based on the MALA.

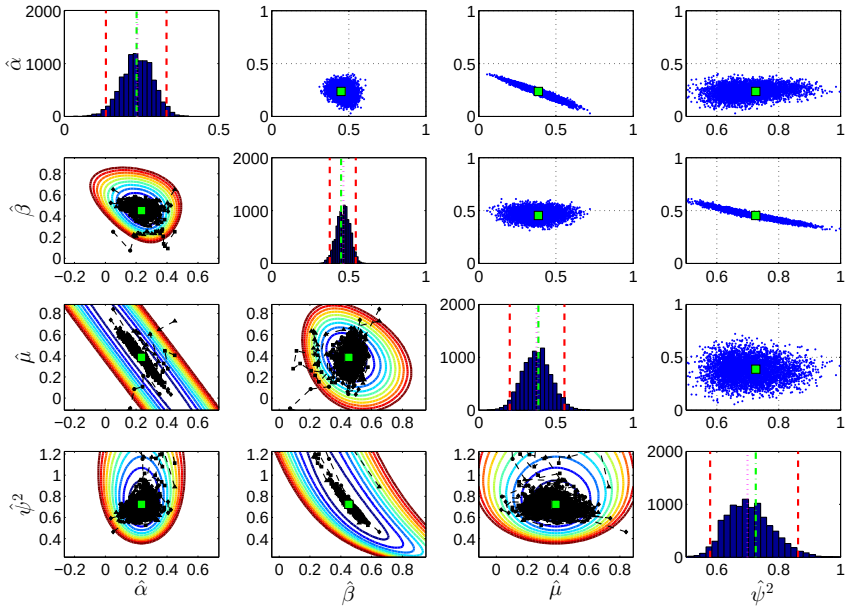
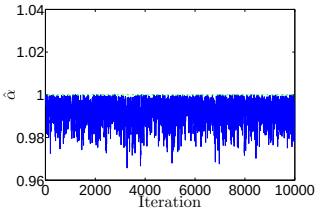
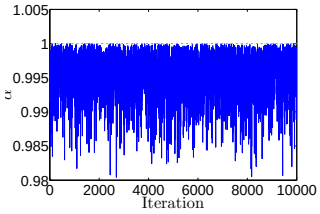


FIGURE C.6: **Case 1** $(\alpha, \beta, \mu, \psi^2)$: Scatterplot matrix of the burn-in (left of diagonal) and posterior sample (right of diagonal) based on the simplified mMALA.

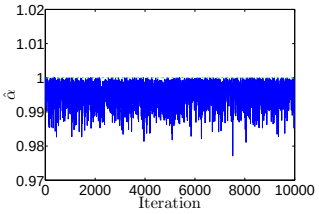
FIGURE C.7: **Case 2** (α, β): Traceplots of the posterior samples and maximum likelihood estimates (- - -), as well as diagnostic plots for the sample likelihood when α and β are estimated simultaneously.



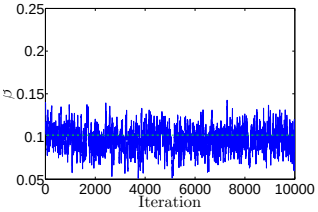
(A) R-W: $\hat{\alpha}$



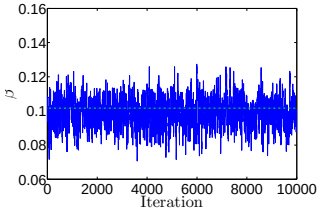
(B) MALA: $\hat{\alpha}$



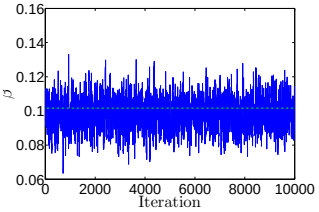
(C) smMALA: $\hat{\alpha}$



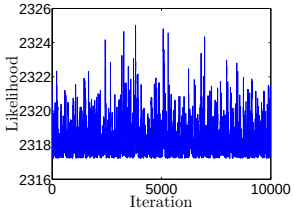
(D) R-W: $\hat{\beta}$



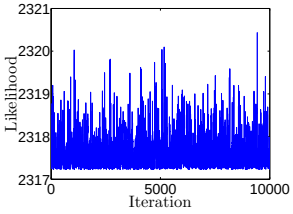
(E) MALA: $\hat{\beta}$



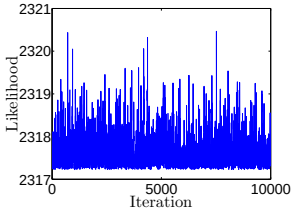
(F) smMALA: $\hat{\beta}$



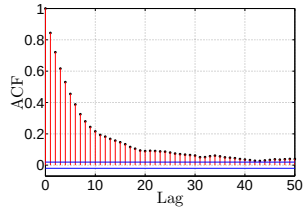
(G) R-W: Likelihood



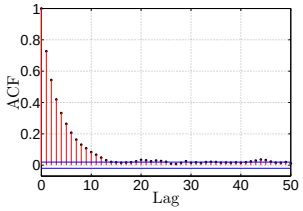
(H) MALA: Likelihood



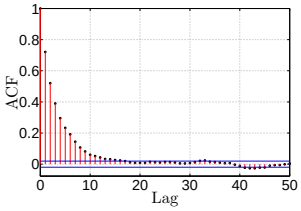
(I) smMALA: Likelihood



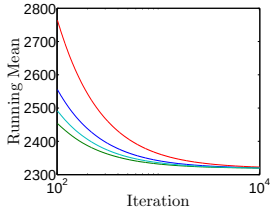
(J) R-W: ACF



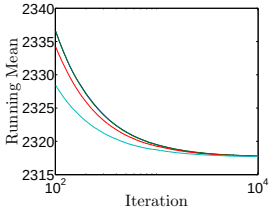
(K) MALA: ACF



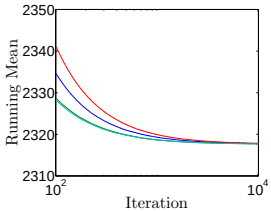
(L) smMALA: ACF



(M) R-W: Running Mean

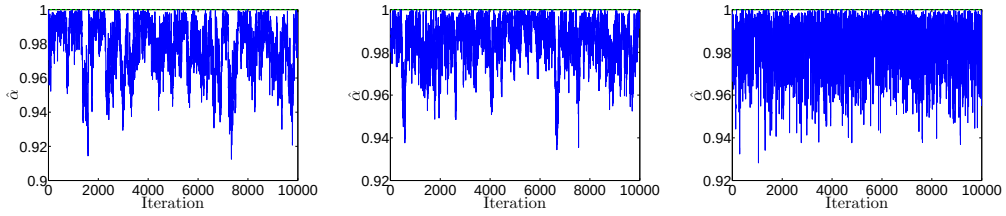


(N) MALA: Running Mean

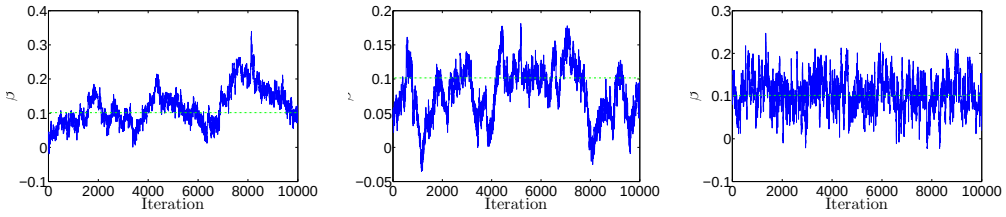


(O) smMALA: Running Mean

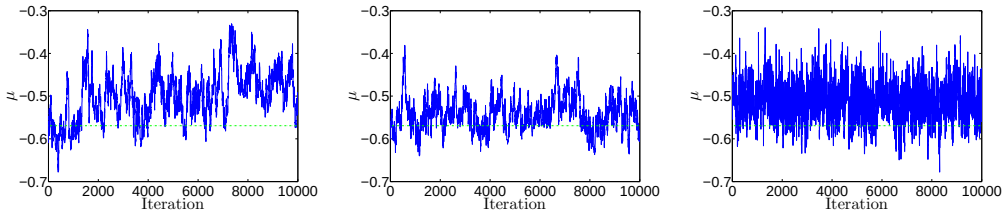
FIGURE C.8: **Case 2** $(\alpha, \beta, \mu, \psi^2)$: Traceplots of the posterior samples and maximum likelihood estimates (- - -) when all four parameters of the Heffernan and Tawn model are estimated simultaneously.



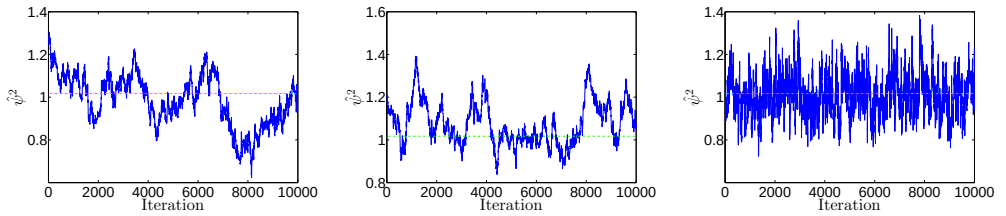
(A) R-W: $\hat{\alpha}$ (B) MALA: $\hat{\alpha}$ (C) smMALA: $\hat{\alpha}$



(D) R-W: $\hat{\beta}$ (E) MALA: $\hat{\beta}$ (F) smMALA: $\hat{\beta}$

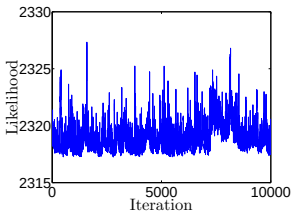


(G) R-W: $\hat{\mu}$ (H) MALA: $\hat{\mu}$ (I) smMALA: $\hat{\mu}$

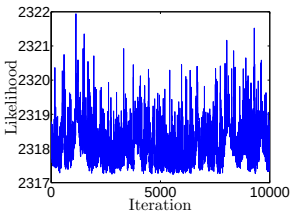


(J) R-W: $\hat{\beta}^2$ (K) MALA: $\hat{\psi}^2$ (L) smMALA: $\hat{\psi}^2$

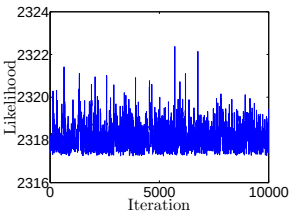
FIGURE C.9: **Case 2** $(\alpha, \beta, \mu, \psi^2)$: Diagnostic plots of the sample likelihood when all four parameters of the Heffernan and Tawn model are estimated simultaneously.



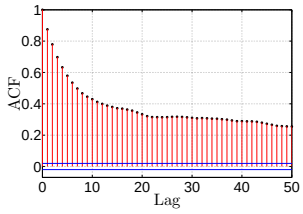
(A) R-W: Likelihood



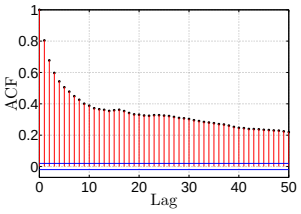
(B) MALA: Likelihood



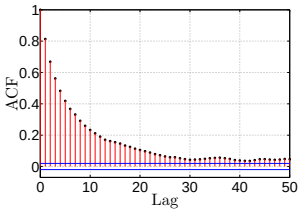
(C) smMALA: Likelihood



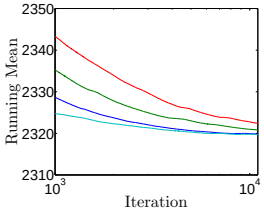
(D) R-W: ACF



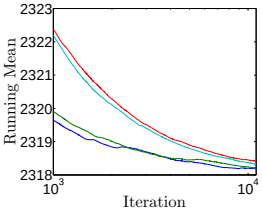
(E) MALA: ACF



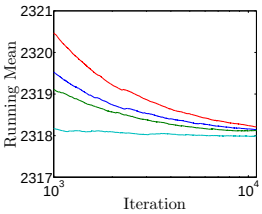
(F) smMALA: ACF



(G) R-W: Running Mean



(H) MALA: Running Mean



(I) smMALA:Running Mean

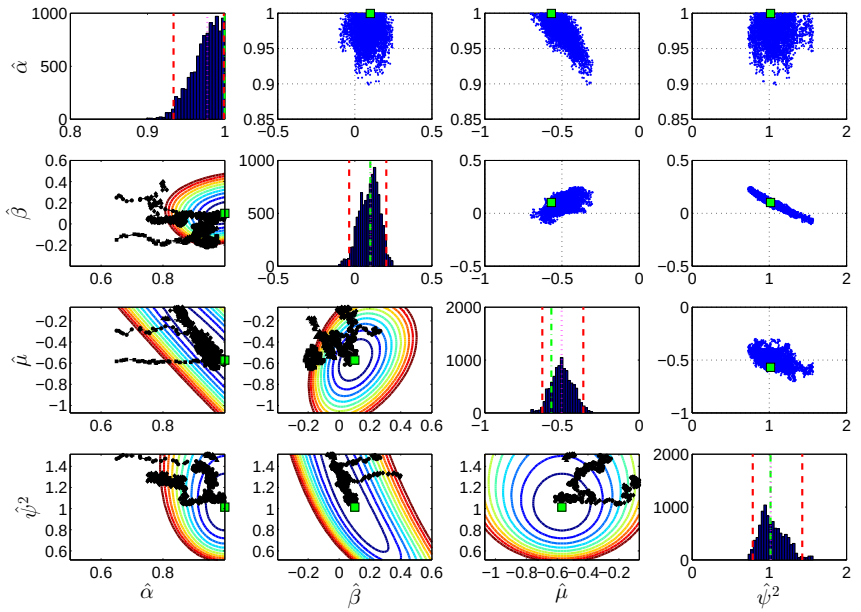


FIGURE C.10: **Case 2** $(\alpha, \beta, \mu, \psi^2)$: Scatterplot matrix of the burn-in sample (left of diagonal) and posterior sample (right of diagonal) based on the random walk transition kernel.

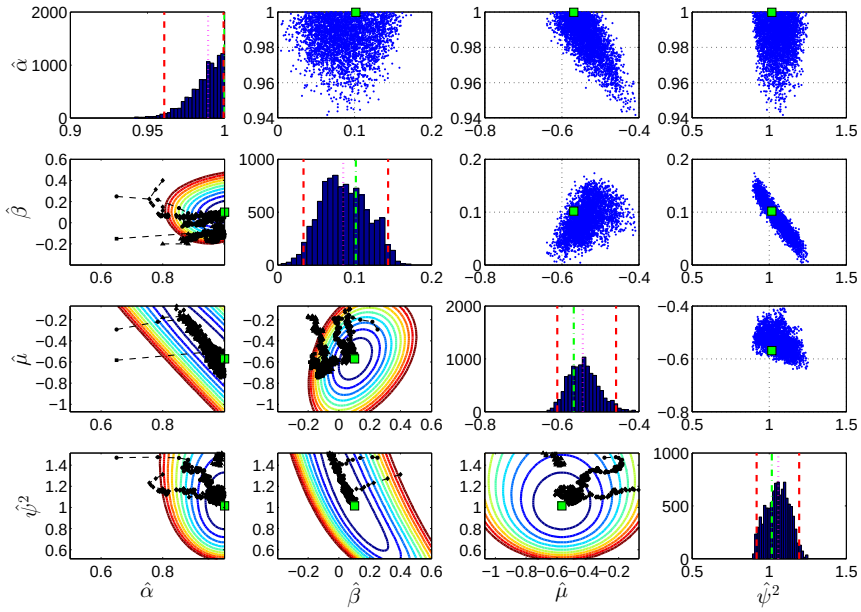


FIGURE C.11: **Case 2** $(\alpha, \beta, \mu, \psi^2)$: Scatterplot matrix of the burn-in (left of diagonal) and posterior sample (right of diagonal) based on the MALA.

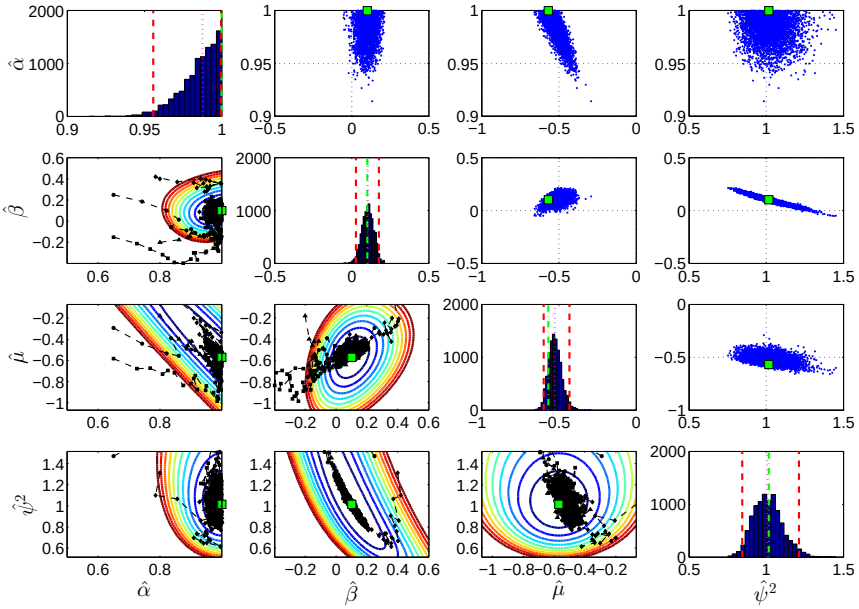


FIGURE C.12: **Case 2** $(\alpha, \beta, \mu, \psi^2)$: Scatterplot matrix of the burn-in (left of diagonal) and posterior sample (right of diagonal) based on the simplified mMALA.

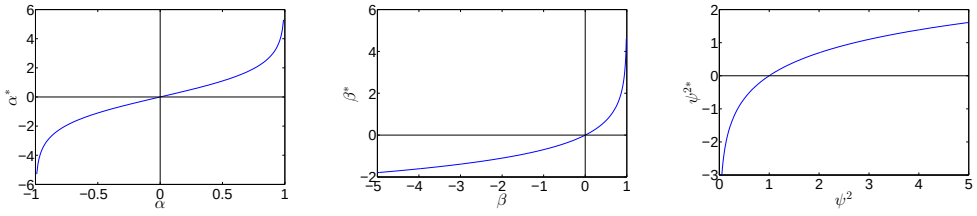
Appendix D

Diagnostic plots: reparameterized Heffernan and Tawn model

A reparameterization is proposed to map $\Omega_{\boldsymbol{\theta}}$ to $\mathbb{R}^4$. Consider

$$\alpha^* \coloneqq \log\left(\frac{1+\alpha}{1-\alpha}\right), \quad \beta^* \coloneqq -\log\left(1-\beta\right) \quad \text{and} \quad \psi^{2*} \coloneqq \log\left(\psi^2\right). \tag{D.1}$$

Although $\alpha^* \to \infty$ as $\alpha \to 1$ for asymptotic dependent data, the transformation will work in practice as $\hat{\alpha}_{\text{MLE}}$ might be arbitrarily close to 1, but will never be exactly equal to 1 for finite data samples. The reparameterization functions in (D.1) are presented in Figure D.1. If α or β is close to zero or when ψ^2 is close to 1, then the mapping $\theta^* : \theta \to \mathbb{R}$ is approximately linear.



(A) $\alpha^* : \alpha \to \mathbb{R}$ (B) $\beta^* : \beta \to \mathbb{R}$ (C) $\psi^{2*} : \psi^2 \to \mathbb{R}$

FIGURE D.1: Reparameterization of the parameters of the Heffernan and Tawn model..

Both the asymptotic independent Case 1 data and asymptotic dependent Case 2 data are considered. The results are presented in consecutive order:

1. Case 1: α^*, β^*, see Figure D.3,
2. Case 1: All four parameters $\alpha^*, \beta^*, \mu, \psi^{2*}$, see Figure D.4 and D.5,
3. Case 2: α^*, β^*, see Figure D.11.

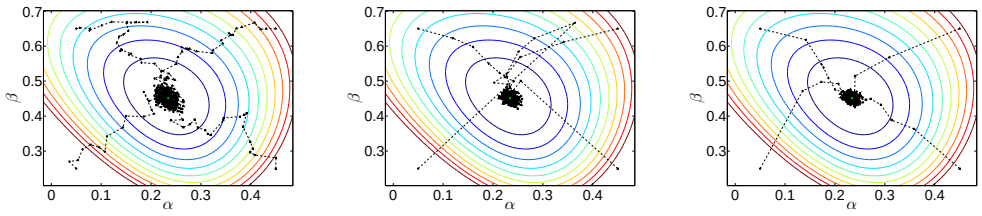
4. Case 2: All four parameters $\alpha^*, \beta^*, \mu, \psi^{2*}$, see Figure D.12 and D.13.

For each of the aforementioned cases, the following diagnostic plots are presented).

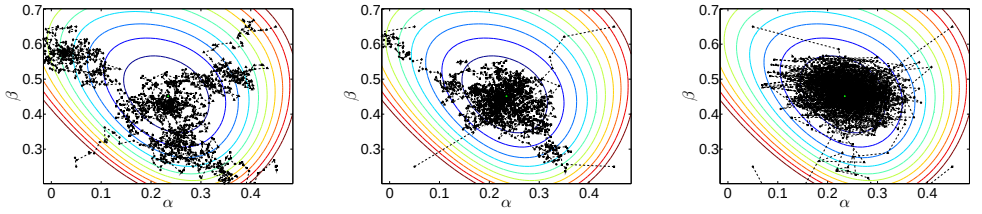
- Trace-plots of the posterior samples of the parameters to be estimated (—) with the maximum likelihood estimate (- · -) as a reference. A converged and properly mixing posterior sample should resemble a white noise process around the maximum likelihood estimate.
- Trace-plot of the likelihood. The absence of a trend and autoregressive features, as well as constant variability, indicate the Markov chain has converged to a stationary limit distribution.
- Autocorrelation function of the sample likelihood. The faster the autocorrelation decreases, the better.
- Running mean of the sample likelihood for four different chains with different starting values. If the running mean of the likelihood for each single chain converges to the same value, this suggests the Markov chains have converged to a global minimum in the negative log-likelihood function.
- The generated Markov chains, when all four parameters are estimated simultaneously, are summarized in scatterplot matrices. Burn-in is shown left of the diagonal, and the posterior samples are shown right of the diagonal with maximum likelihood estimates (■). The histograms on the diagonal show the median (· · ·) and 95% confidence interval (- - -) based on the 2.5% and 97.5% quantile as well as the maximum likelihood estimate (- · -).

Comparing Figure 4.2(A)-(C) with Figure D.2(A)-(C) suggests the reparameterization does not affect the results. This is confirmed by the results presented in Table D.1 and the resemblance between the trace– and diagnostic plots for both cases, shown in Figure C.1 and D.3. The only remarkable difference is that the reparameterization allows the stepsize to be increased to obtain a similar acceptance rate.

FIGURE D.2: The first 250 burn-in samples for α^* and β^* for Case 2 data, based on three different transition kernels: random walk, Metropolis adjusted Langevin algorithm and simplified manifold Metropolis adjusted Langevin algorithm. Results are presented on the original scale. The top row shows results when only α and β are estimated and $\mu = \hat{\mu}_{\text{MLE}}$ and $\psi = \hat{\psi}^2_{\text{MLE}}$, while all four parameters of the Heffernan and Tawn model are estimated jointly for the figures in the bottom row.



(A) R-W (B) MALA (C) smMALA

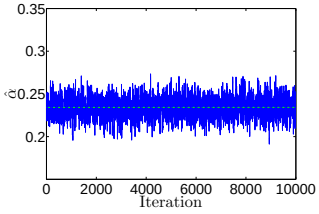


(D) R-W (E) MALA (F) smMALA

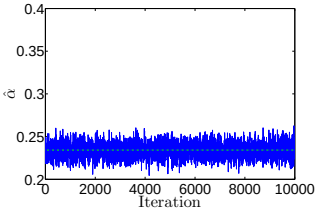
TABLE D.1: Summary statistics for the posterior samples of the Heffernan and Tawn model parameter estimates evaluated on the reparamterized scale, for Case 1 data. Different proposal mechanisms are considered: random walk, Metropolis adjusted Langevin algorithm and the simplified manifold Metropolis adjusted Langevin algorithm.

		Two parameter estimation			**Four parameter estimation**		
		RW	MALA	smMALA	RW	MALA	smMALA
ε		0.035	0.03	1	0.025	0.025	0.8
AR		0.44	0.46	0.52	0.44	0.42	0.45
$\hat{\alpha}$	MED	0.23	0.23	0.23	0.24	0.24	0.23
	$CI_{95\%}$	$[0.21, 0.26]$	$[0.22, 0.25]$	$[0.22, 0.25]$	$[0.11, 0.42]$	$[0.12, 0.32]$	$[0.14, 0.33]$
	ESS	780	1870	1780	4	19	830
	ESS/s	34	13	12	0.2	0.1	5.6
	$\hat{R}$	1	1	1	1.15	1.04	1
$\hat{\beta}$	MED	0.45	0.45	0.45	0.46	0.46	0.45
	$CI_{95\%}$	$[0.42, 0.48]$	$[0.43, 0.47]$	$[0.43, 0.47]$	$[0.32, 0.58]$	$[0.37, 0.54]$	$[0.38, 0.55]$
	ESS	580	1720	1690	10	25	200
	ESS/s	26	12	12	0.4	0.2	1.3
	$\hat{R}$	1	1	1	1.05	1	1
$\hat{\mu}$	MED				0.38	0.37	0.45
	$CI_{95\%}$				$[0.02, 0.62]$	$[0.22, 0.58]$	$[0.20, 0.54]$
	ESS				5	20	1060
	ESS/s				0.2	0.1	7.1
	$\hat{R}$				1.12	1.04	1
$\hat{\psi}^2$	MED				0.72	0.71	0.45
	$CI_{95\%}$				$[0.54, 0.99]$	$[0.59, 0.87]$	$[0.58, 0.86]$
	ESS				11	24	210
	ESS/s				0.4	0.2	1.4
	$\hat{R}$				1	1	1

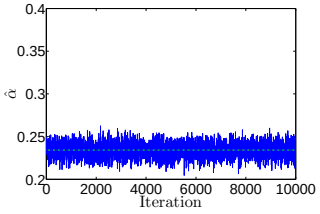
FIGURE D.3: **Case 1** (α^*, β^*): Traceplots of the posterior samples and maximum likelihood estimates (- - -), as well as diagnostic plots for the sample likelihood when α and β are estimated simultaneously. Results are presented on the original scale.



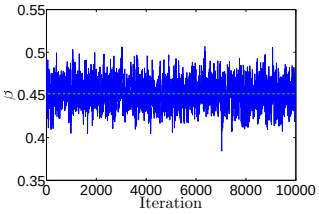
(A) R-W: $\hat{\alpha}$



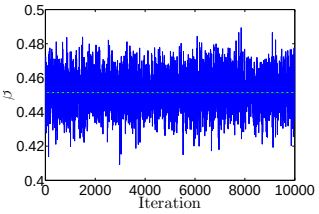
(B) MALA: $\hat{\alpha}$



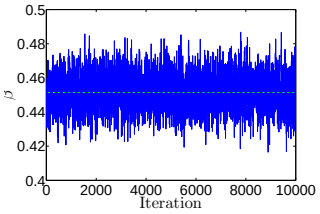
(C) smMALA: $\hat{\alpha}$



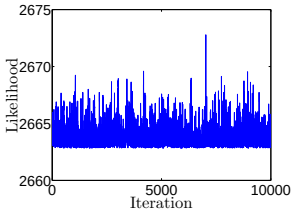
(D) R-W: $\hat{\beta}$



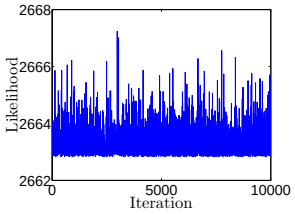
(E) MALA: $\hat{\beta}$



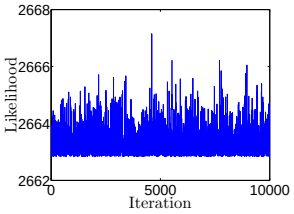
(F) smMALA: $\hat{\beta}$



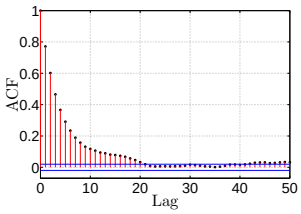
(G) R-W: Likelihood



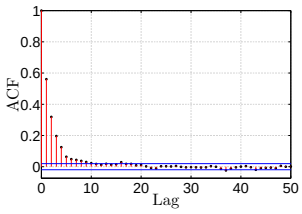
(H) MALA: Likelihood



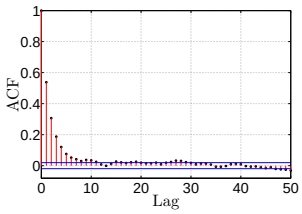
(I) smMALA: Likelihood



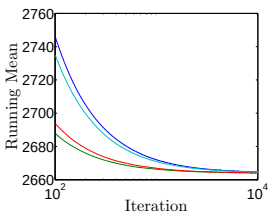
(J) R-W: ACF



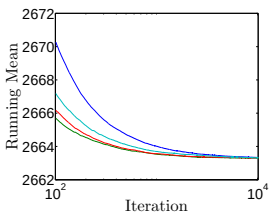
(K) MALA: ACF



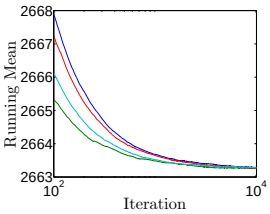
(L) smMALA: ACF



(M) R-W: Running Mean

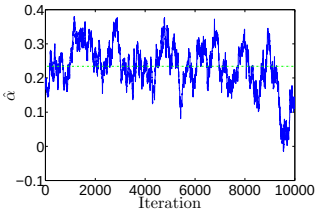


(N) MALA: Running Mean

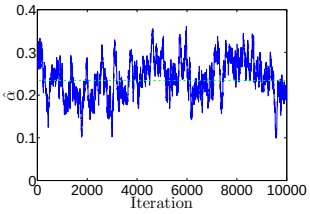


(O) smMALA: Running Mean

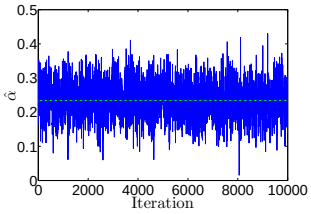
FIGURE D.4: **Case 1** $\left(\alpha^*, \beta^*, \mu, \psi^{2*}\right)$: Traceplots of the posterior samples and maximum likelihood estimates (- - -) when all four parameters of the Heffernan and Tawn model are estimated simultaneously. Results are presented on the original scale.



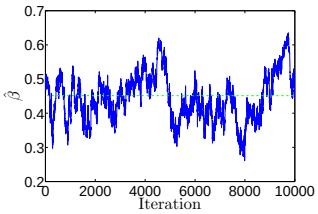
(A) R-W: $\hat{\alpha}$



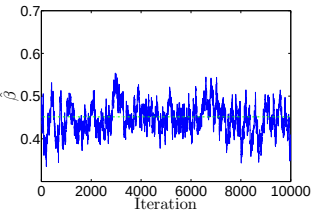
(B) MALA: $\hat{\alpha}$



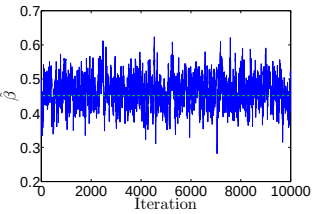
(C) smMALA: $\hat{\alpha}$



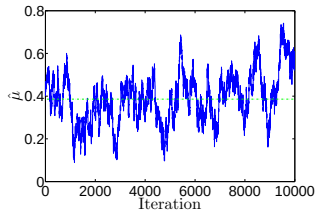
(D) R-W: $\hat{\beta}$



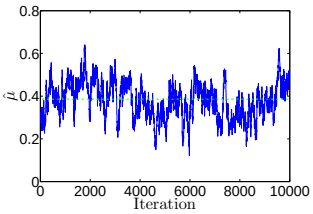
(E) MALA: $\hat{\beta}$



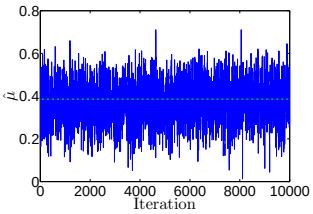
(F) smMALA: $\hat{\beta}$



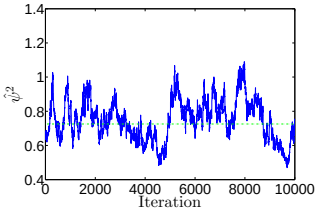
(G) R-W: $\hat{\mu}$



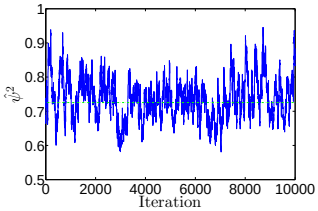
(H) MALA: $\hat{\mu}$



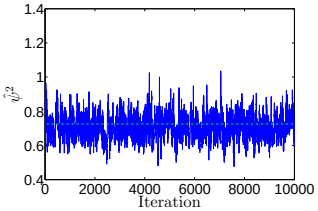
(I) smMALA: $\hat{\mu}$



(J) R-W: $\hat{\psi}^2$

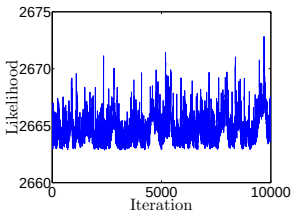


(K) MALA: $\hat{\psi}^2$

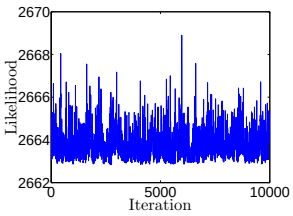


(L) smMALA: $\hat{\psi}^2$

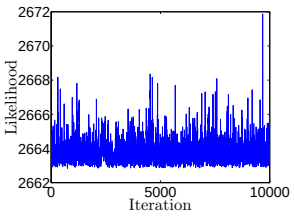
FIGURE D.5: **Case 1** $\left(\alpha^*, \beta^*, \mu, \psi^{2*}\right)$: Diagnostic plots of the sample likelihood when all four parameters of the Heffernan and Tawn model are estimated simultaneously.



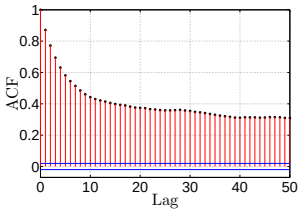
(A) R-W: Likelihood



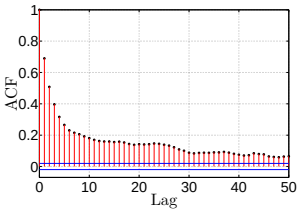
(B) MALA: Likelihood



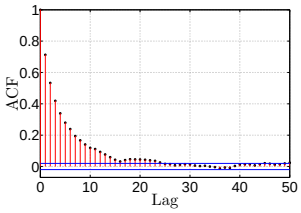
(C) smMALA: Likelihood



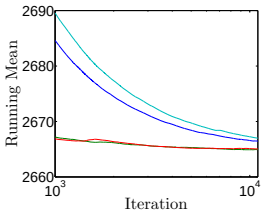
(D) R-W: ACF



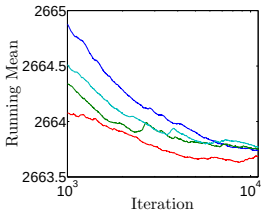
(E) MALA: ACF



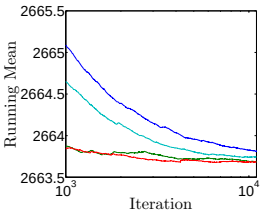
(F) smMALA: ACF



(G) R-W: Running Mean



(H) MALA: Running Mean



(I) smMALA: Running Mean

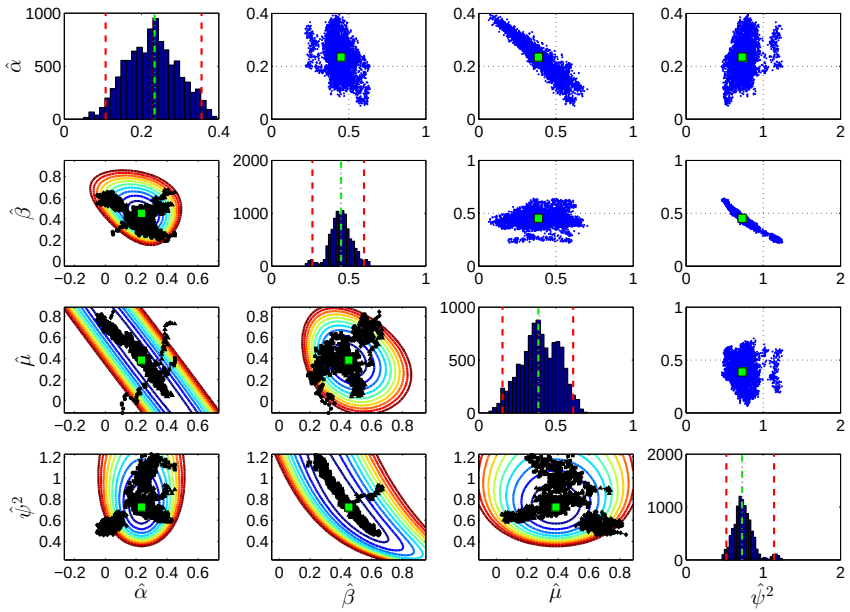


FIGURE D.6: **Case 1** $(\alpha^*, \beta^*, \mu, \psi^{2*})$: Scatterplot matrix of the burn-in sample (left of diagonal) and posterior sample (right of diagonal) based on the random walk transition kernel.

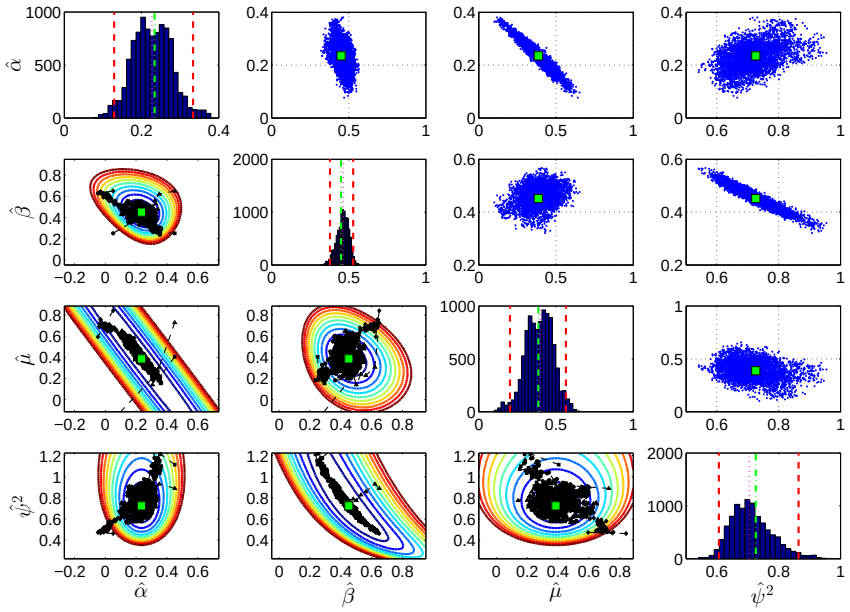


FIGURE D.7: **Case 1** $(\alpha^*, \beta^*, \mu, \psi^{2*})$: Scatterplot matrix of the burn-in sample (left of diagonal) and posterior sample (right of diagonal) based on the MALA.

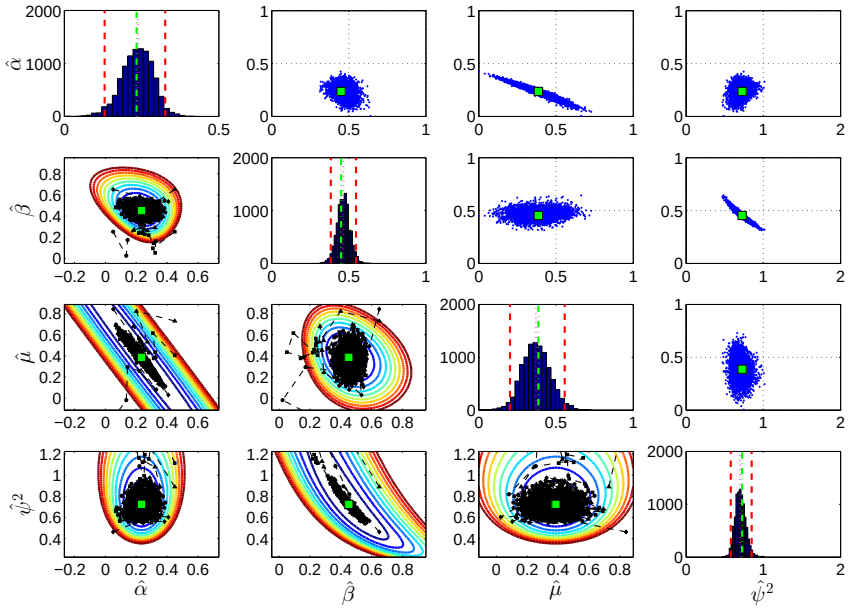
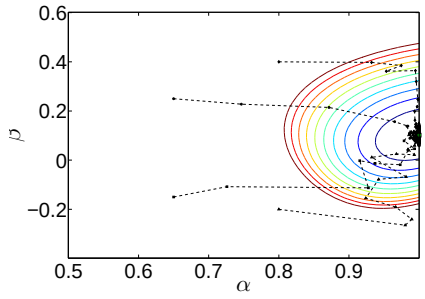


FIGURE D.8: **Case 1** $(\alpha^*, \beta^*, \mu, \psi^{2*})$: Scatterplot matrix of the burn-in sample (left of diagonal) and posterior sample (right of diagonal) based on the smMALA.

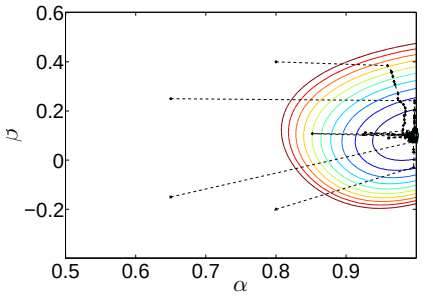
TABLE D.2: Summary statistics for the posterior samples of the Heffernan and Tawn model parameter estimates evaluated on the reparameterized scale, for Case 2 data. Different proposal mechanisms are considered: random walk, Metropolis adjusted Langevin algorithm and the simplified manifold Metropolis adjusted Langevin algorithm.

		Two parameter estimation		**Four parameter estimation**	
		R-W	MALA	R-W	MALA
ε		0.04	0.01	0.02	0.015
ε_α		1	2	1	2
AR		0.42	0.78	0.51	0.58
$\hat{\alpha}$	MED	1	1	1	1
	$CI_{95\%}$	$[1^-, 1]$	$[1^-, 1]$	$[0.99, 1]$	$[1^-, 1]$
	ESS	140	820	60	7580
	ESS/s	7	6	2.6	53.1
	$\hat{R}$	1.04	1	1.03	1
$\hat{\beta}$	MED	0.1	0.1	0.09	0.11
	$CI_{95\%}$	$[0.07, 0.13]$	$[0.08, 0.12]$	[-0.01, 0.19]	$[0.03, 0.19]$
	ESS	1380	1200	11	16
	ESS/s	65	9	0.5	0.1
	$\hat{R}$	1	1	1.02	1
$\hat{\mu}$	MED			-0.57	-0.56
	$CI_{95\%}$			[-0.67, -0.50]	[-0.62, -0.5]
	ESS			13	22
	ESS/s			0.6	0.2
	$\hat{R}$			1.04	1.04
$\hat{\psi}^2$	MED			1.03	0.99
	$CI_{95\%}$			$[0.83, 1.34]$	$[0.82, 1.22]$
	ESS			10	15
	ESS/s			0.5	0.1
	$\hat{R}$			1.02	1

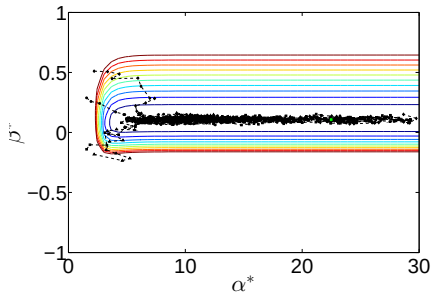
FIGURE D.9: The first 250 burn-in samples for α^* and β^* for Case 2 data presented on the original scale. Only α^* and β^* are estimated and $\mu = \hat{\mu}_{\text{MLE}}$ and $\psi = \hat{\psi}^2_{\text{MLE}}$ are fixed. Three different transition kernels are considered: random walk, Metropolis adjusted Langevin algorithm and simplified manifold Metropolis adjusted Langevin algorithm. Results on the original scale are shown on the top row, while the the results on the reparameterized scale are shown on the bottom row.



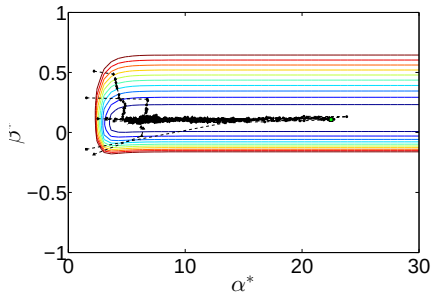
(A) R-W



(B) MALA

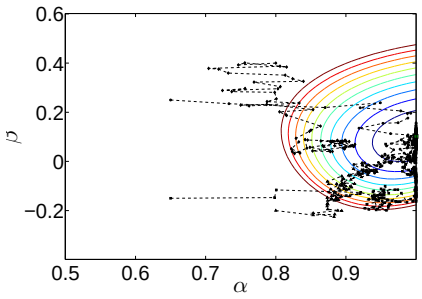


(C) R-W

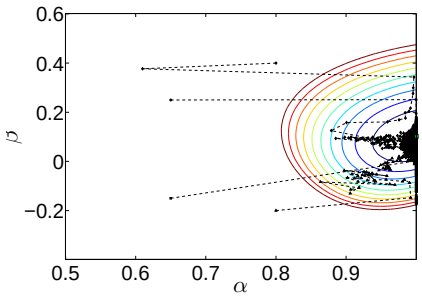


(D) MALA

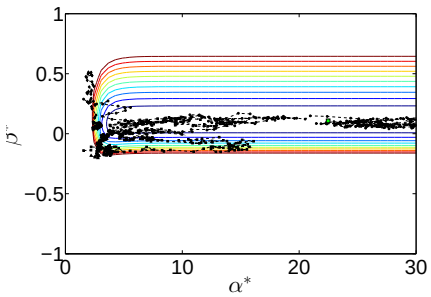
FIGURE D.10: The first 250 burn-in samples for α^* and β^* for Case 2 data presented on the original scale. All four reparameterized parameters are estimated jointly. Three different transition kernels are considered: random walk, Metropolis adjusted Langevin algorithm and simplified manifold Metropolis adjusted Langevin algorithm. Results on the original scale are shown on the top row, while the the results on the reparameterized scale are shown on the bottom row.



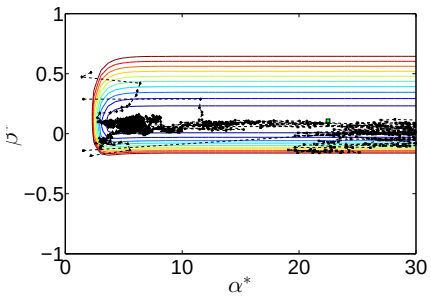
(A) R-W



(B) MALA

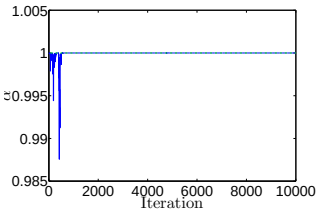


(C) R-W

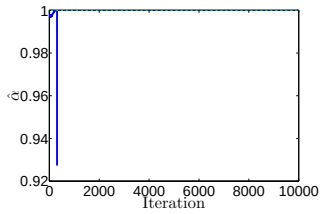


(D) MALA

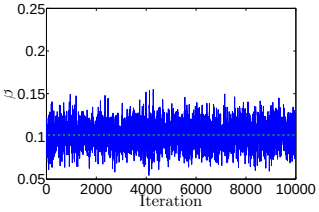
FIGURE D.11: **Case 2** (α^*, β^*): Traceplots of the posterior samples and maximum likelihood estimates (- - -), as well as diagnostic plots for the sample likelihood when α and β are estimated simultaneously. Results are presented on the original scale.



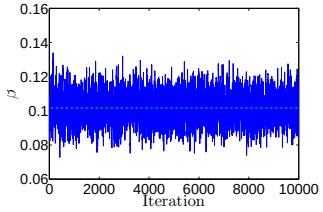
(A) R-W: $\hat{\alpha}$



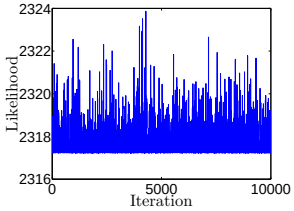
(B) MALA: $\hat{\alpha}$



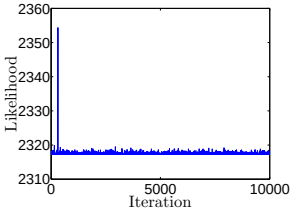
(C) R-W: $\hat{\beta}$



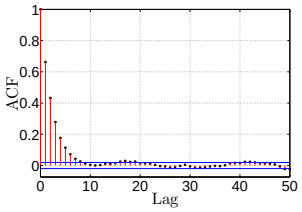
(D) MALA: $\hat{\beta}$



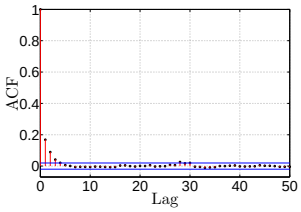
(E) R-W: Likelihood



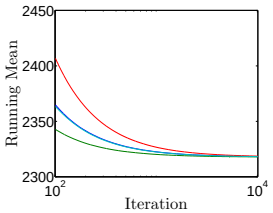
(F) MALA: Likelihood



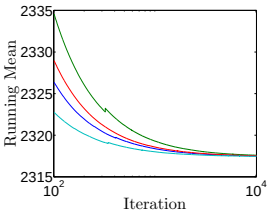
(G) R-W: ACF



(H) MALA: ACF

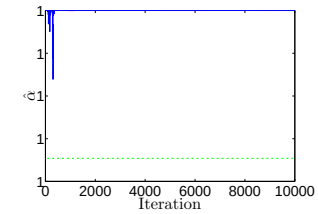


(I) R-W: Running Mean

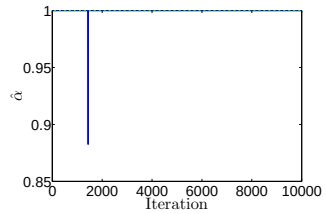


(J) MALA: Running Mean

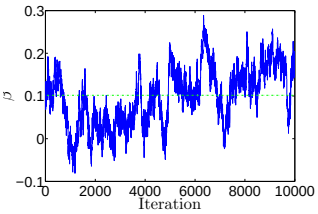
FIGURE D.12: **Case 2** $\left(\alpha^*, \beta^*, \mu, \psi^{2*}\right)$: Traceplots of the posterior samples and maximum likelihood estimates (- - -) when all four parameters of the Heffernan and Tawn model are estimated simultaneously. Results are presented on the original scale.



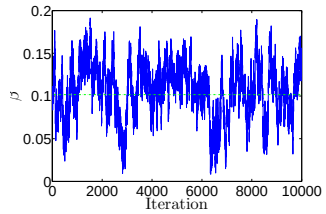
(A) R-W: $\hat{\alpha}$



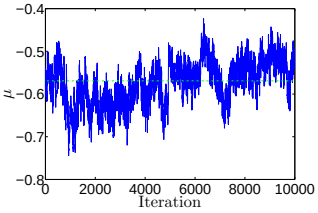
(B) MALA: $\hat{\alpha}$



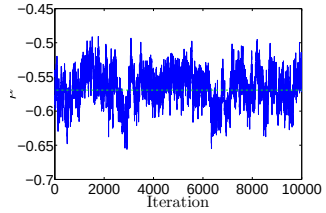
(C) R-W: $\hat{\beta}$



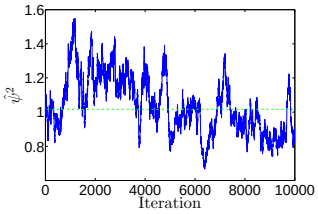
(D) MALA: $\hat{\beta}$



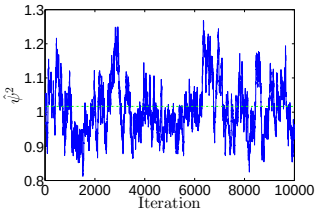
(E) R-W: $\hat{\mu}$



(F) MALA: $\hat{\mu}$

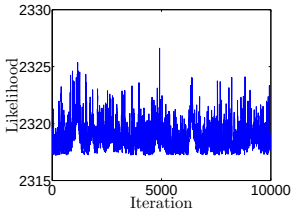


(G) R-W: $\hat{\psi}^2$

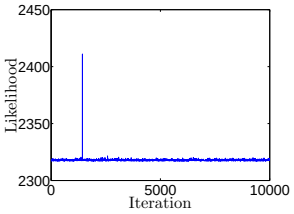


(H) MALA: $\hat{\psi}^2$

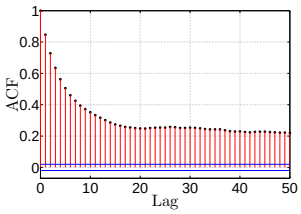
FIGURE D.13: **Case 2** $\left(\alpha^*, \beta^*, \mu, \psi^{2*}\right)$: Diagnostic plots of the likelihood when all four reparameterized parameters of the Heffernan and Tawn model are estimated simultaneously. Results are shown on the original scale.



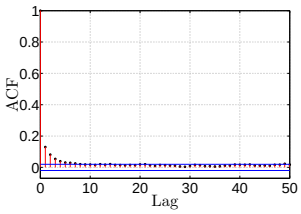
(A) R-W: Likelihood



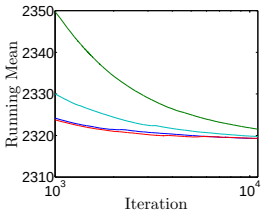
(B) MALA: Likelihood



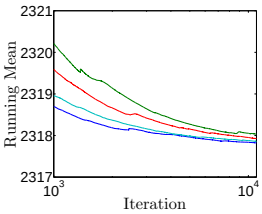
(C) R-W: ACF



(D) MALA: ACF



(E) R-W: Running Mean



(F) MALA: Running Mean

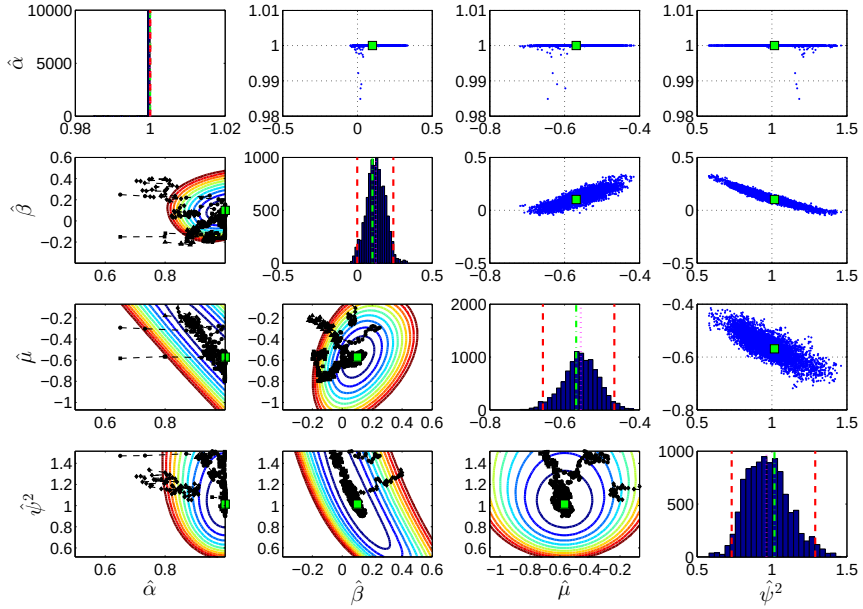


FIGURE D.14: **Case 1** $(\alpha^*, \beta^*, \mu, \psi^{2*})$: Scatterplot matrix of the burn-in sample (left of diagonal) and posterior sample (right of diagonal) based on the random walk transition kernel.

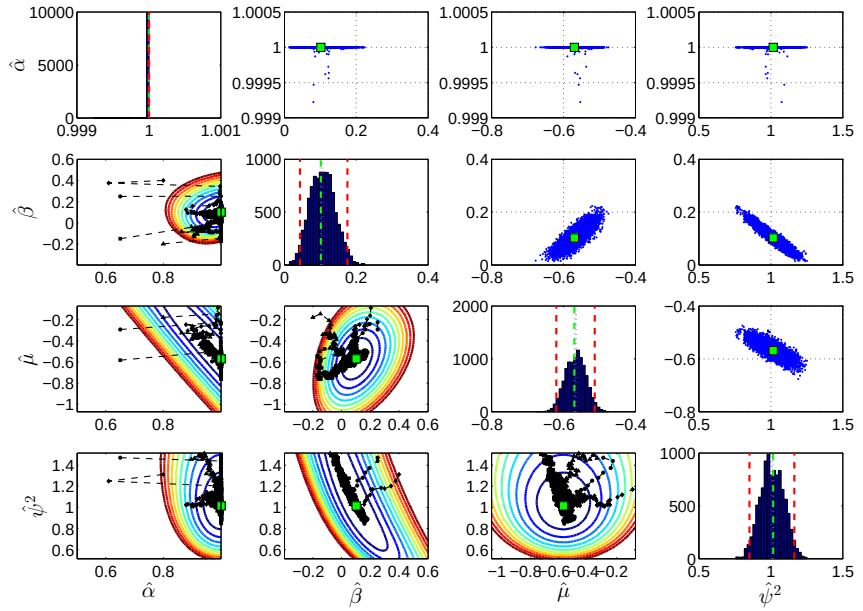


FIGURE D.15: **Case 1** $(\alpha^*, \beta^*, \mu, \psi^{2*})$: Scatterplot matrix of the burn-in sample (left of diagonal) and posterior sample (right of diagonal) based on the MALA.

Appendix E

Diagnostic plots: constrained Heffernan and Tawn model

Diagnostic plots for the Bayesian analysis of the Heffernan and Tawn model subject to the constraints proposed by Keef et al. (2013), discussed in Section 4.3 are presented in this appendix. In consecutive order, the following cases are presented.:

1. Case 1: α, β, see Figure E.1.
2. Case 1: All four parameters $\alpha, \beta, \mu, \psi^2$, see Figure E.2 and E.3.
3. Case 2: α, β, see Figure E.7.
4. Case 2: All four parameters $\alpha, \beta, \mu, \psi^2$, see Figure E.8 and E.9.

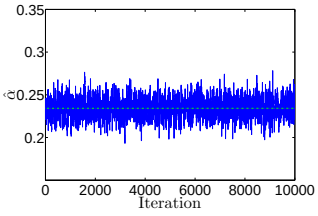
See the introduction to Appendix C for an explanation on how to interpret the diagnostic plots presented in this appendix.

The burn-in samples in the scatterplot matrices are drawn on top of a profile likelihood surface. For the four parameter estimation problem, these surfaces are deceiving and suggest certain samples fall outside the feasible parameter space. This is related to the fact the the profile likelihood surfaces are generated by fixing two parameters to their maximum likelihood estimate. The burn-in samples can attain other values than the maximum likelihood estimates, such that they are still feasible under the constrains but fall outside the profile likelihood surface.

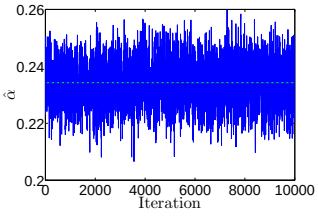
TABLE E.1: Statistics introduced in Section 4.1.4 of the posterior samples for all four parameters of the constrained Heffernan and Tawn model for Case 1 data, obtained with different transition kernels: random walk, Metropolis adjusted Langevin algorithm and the simplified manifold Metropolis adjusted Langevin algorithm.

		Two parameter estimation			Four parameter estimation		
		RW	MALA	smMALA	RW	MALA	smMALA
ε		0.0175	0.01	0.7	0.02	0.012	0.65
AR		0.41	0.59	0.61	0.32	0.48	0.43
$\hat{\alpha}$	MED	0.24	0.23	0.23	0.23	0.25	0.23
	$CI_{95\%}$	$[0.21, 0.26]$	$[0.22, 0.25]$	$[0.22, 0.25]$	$[0.13, 0.32]$	$[0.15, 0.32]$	$[0.14, 0.33]$
	ESS	690	1430	1010	10	15	380
	ESS/s	7.4	6.8	4.7	0.1	0.1	1.7
	$\hat{R}$	1	1	1	1.05	1.01	1
$\hat{\beta}$	MED	0.45	0.45	0.45	0.42	0.43	0.44
	$CI_{95\%}$	$[0.42, 0.47]$	$[0.43, 0.47]$	$[0.43, 0.47]$	$[0.37, 0.48]$	$[0.37, 0.48]$	$[0.38, 0.55]$
	ESS	580	1010	1020	10	20	350
	ESS/s	6.2	4.8	5.3	0.1	0.1	1.1
	$\hat{R}$	1	1	1	1.04	1	1
$\hat{\mu}$	MED				0.42	0.37	0.39
	$CI_{95\%}$				$[0.23, 0.66]$	$[0.23, 0.55]$	$[0.20, 0.54]$
	ESS				10	15	490
	ESS/s				0.1	0.1	2.3
	$\hat{R}$				1.05	1.01	1
$\hat{\psi}^2$	MED				0.79	0.76	0.75
	$CI_{95\%}$				$[0.65, 1.07]$	$[0.67, 0.88]$	$[0.58, 0.86]$
	ESS				10	20	260
	ESS/s				0.1	0.1	1.2
	$\hat{R}$				1.04	1	1

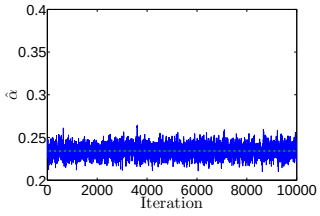
FIGURE E.1: **Case 1** (α, β): Traceplots of the posterior samples and maximum likelihood estimates (- - -), as well as diagnostic plots for the sample likelihood when the constrained parameters α and β are estimated simultaneously.



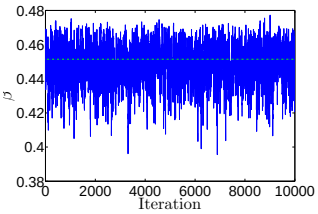
(A) R-W: $\hat{\alpha}$



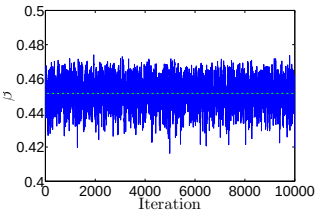
(B) MALA: $\hat{\alpha}$



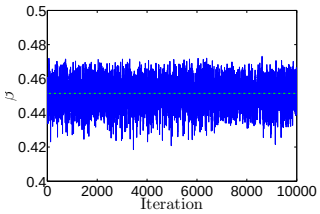
(C) smMALA: $\hat{\alpha}$



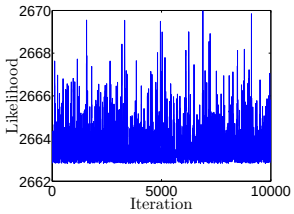
(D) R-W: $\hat{\beta}$



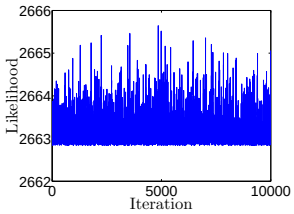
(E) MALA: $\hat{\beta}$



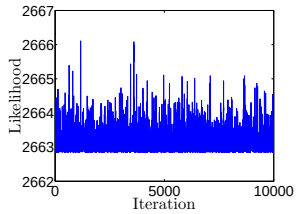
(F) smMALA: $\hat{\beta}$



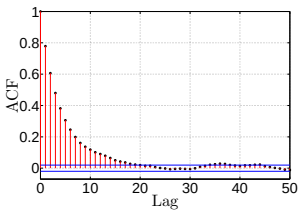
(G) R-W: Likelihood



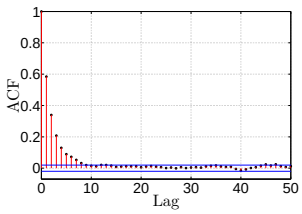
(H) MALA: Likelihood



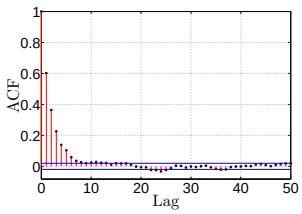
(I) smMALA: Likelihood



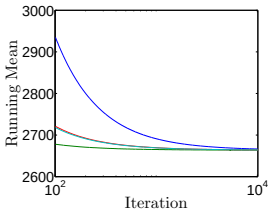
(J) R-W: ACF



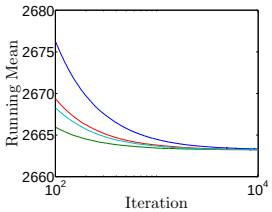
(K) MALA: ACF



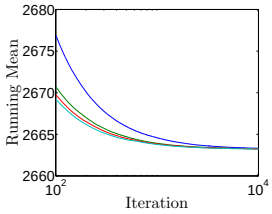
(L) smMALA: ACF



(M) R-W: Running Mean

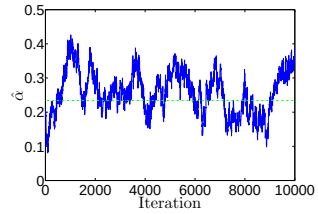


(N) MALA: Running Mean

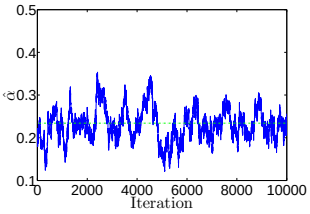


(O) smMALA: Running Mean

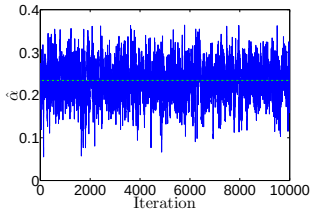
FIGURE E.2: **Case 1** $(\alpha, \beta, \mu, \psi^2)$: Traceplots of the posterior samples and maximum likelihood estimates (- - -) when all four parameters of the constrained Heffernan and Tawn model are estimated simultaneously.



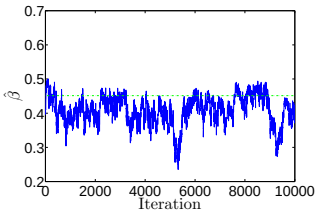
(A) R-W: $\hat{\alpha}$



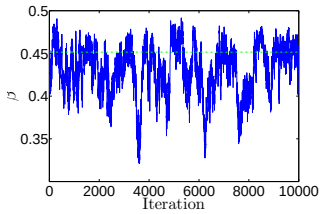
(B) MALA: $\hat{\alpha}$



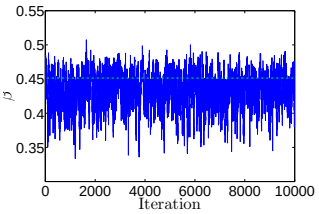
(C) smMALA: $\hat{\alpha}$



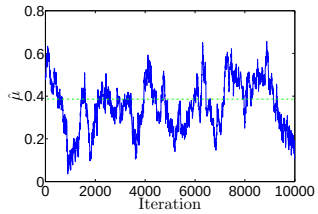
(D) R-W: $\hat{\beta}$



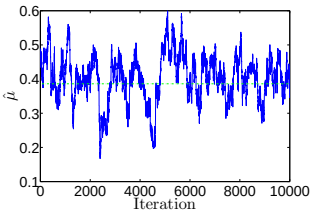
(E) MALA: $\hat{\beta}$



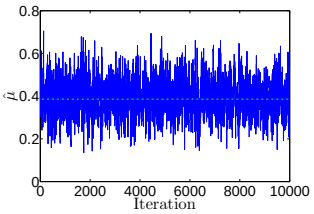
(F) smMALA: $\hat{\beta}$



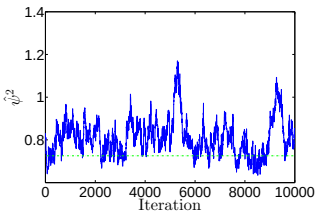
(G) R-W: $\hat{\mu}$



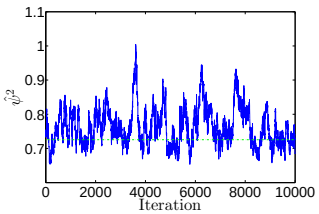
(H) MALA: $\hat{\mu}$



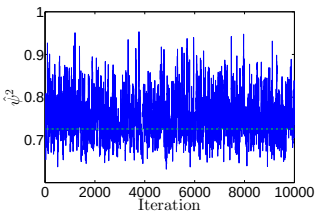
(I) smMALA: $\hat{\mu}$



(J) R-W: $\hat{\psi}^2$

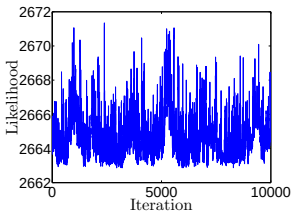


(K) MALA: $\hat{\psi}^2$

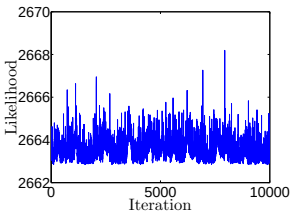


(L) smMALA: $\hat{\psi}^2$

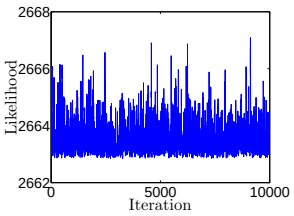
FIGURE E.3: **Case 1** $(\alpha, \beta, \mu, \psi^2)$: Diagnostic plots of the sample likelihood when all four parameters of the constrained Heffernan and Tawn model are estimated simultaneously.



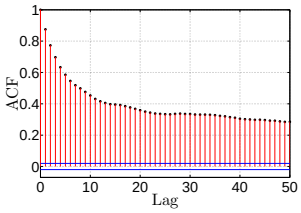
(A) R-W: Likelihood



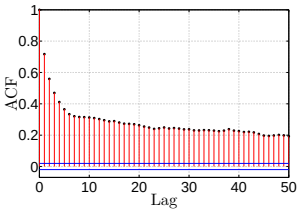
(B) MALA: Likelihood



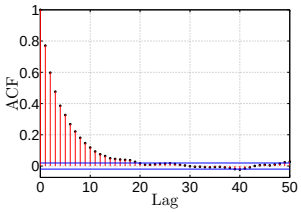
(C) smMALA: Likelihood



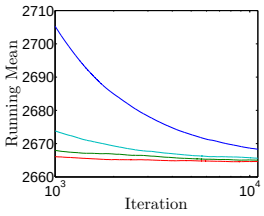
(D) R-W: ACF



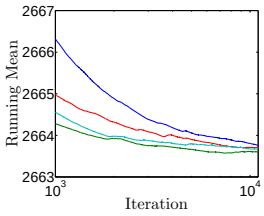
(E) MALA: ACF



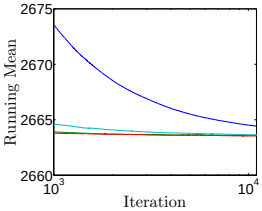
(F) smMALA: ACF



(G) R-W: Running Mean



(H) MALA: Running Mean



(I) smMALA: Running Mean

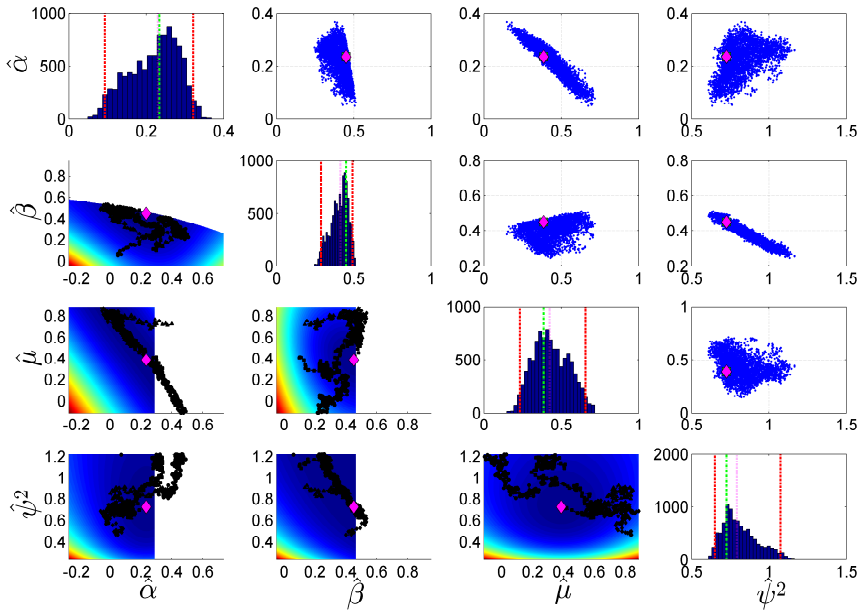


FIGURE E.4: **Case 1** $(\alpha, \beta, \mu, \psi^2)$: Scatterplot matrix of the burn-in sample (left of diagonal) and posterior sample (right of diagonal) based on the random walk transition kernel.

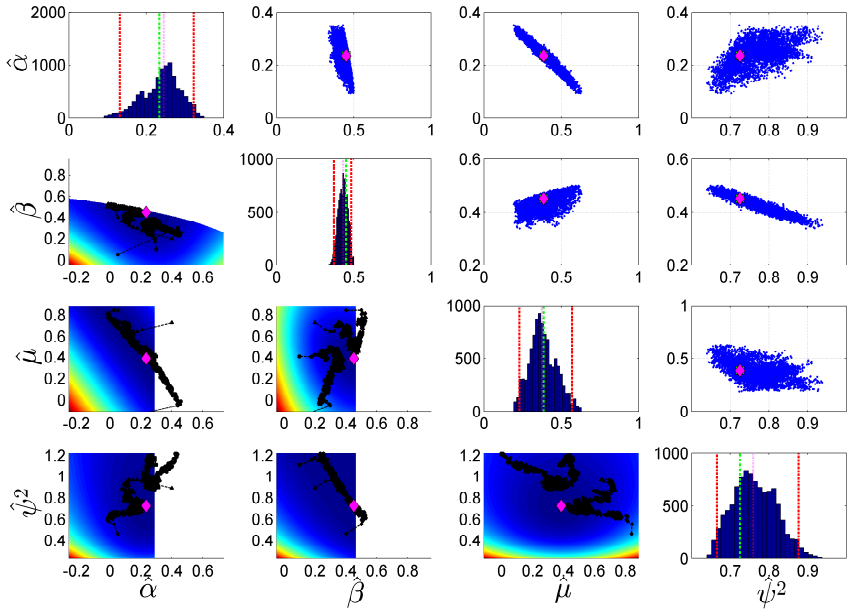


FIGURE E.5: **Case 1** $(\alpha, \beta, \mu, \psi^2)$: Scatterplot matrix of the burn-in sample (left of diagonal) and posterior sample (right of diagonal) based on the MALA.

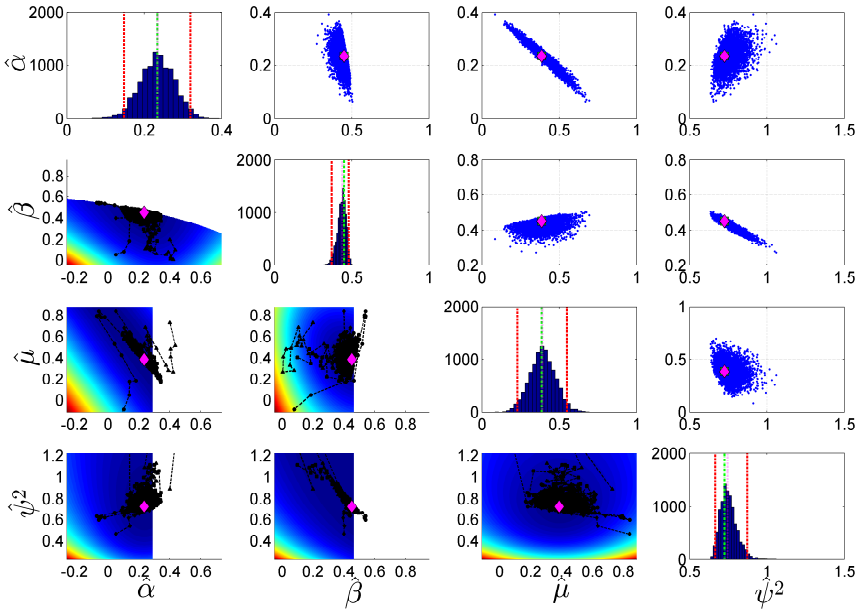
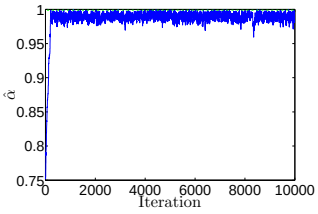


FIGURE E.6: **Case 1** $(\alpha, \beta, \mu, \psi^2)$: Scatterplot matrix of the burn-in sample (left of diagonal) and posterior sample (right of diagonal) based on the smMALA.

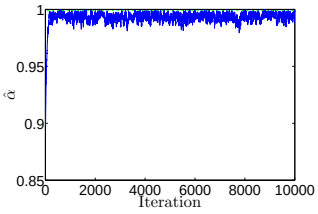
TABLE E.2: Statistics introduced in Section 4.1.4 of the posterior samples for all four parameters of the constrained Heffernan and Tawn model for Case 2 data, obtained with different transition kernels: random walk, Metropolis adjusted Langevin algorithm and the simplified manifold Metropolis adjusted Langevin algorithm.

		Two parameter estimation			**Four parameter estimation**		
		RW	MALA	smMALA	RW	MALA	smMALA
ε		0.005	0.0018	0.5	0.013	0.003	0.55
AR		0.52	0.73	0.39	0.39	0.68	0.35
$\hat{\alpha}$	MED	0.99	0.99	0.99	0.96	0.97	0.97
	$CI_{95\%}$	[0.97, 1-]	[0.97, 1-]	[0.99, 1-]	[0.91, 0.99]	[0.87, 0.99]	[0.93, 0.99]
	ESS	70	85	570	15	5	155
	ESS/s	0.8	0.4	2.8	0.2	0.02	0.7
	$\hat{R}$	1	1.01	1	1.01	1.07	1.01
$\hat{\beta}$	MED	0.00	0.00	0.00	0.04	0.03	0.04
	$CI_{95\%}$	[-0.01, 0.03]	[-0.01, 0.06]	[-0.01, 0.01]	[-0.01, 0.10]	[-0.07, 0.06]	[-0.01, 0.10]
	ESS	85	65	860	10	5	115
	ESS/s	1.0	0.3	4.2	0.1	0.01	0.6
	$\hat{R}$	1.01	1.01	1	1.01	1.24	1.01
$\hat{\mu}$	MED				-0.47	-0.51	-0.50
	$CI_{95\%}$				[-0.62, -0.33]	[-0.59, -0.22]	[-0.60, 0.36]
	ESS				10	5	110
	ESS/s				0.1	0.02	0.5
	$\hat{R}$				1.01	1.10	1.01
$\hat{\psi}^2$	MED				1.19	1.19	1.16
	$CI_{95\%}$				[1.01, 1.36]	[1.11, 1.73]	[1.02, 1.30]
	ESS				10	5	125
	ESS/s				0.1	0.01	0.6
	$\hat{R}$				1.01	1.18	1.01

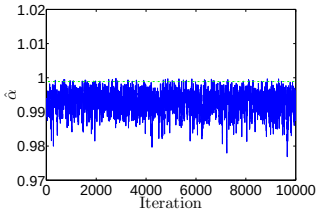
FIGURE E.7: **Case 2** (α, β): Traceplots of the posterior samples and maximum likelihood estimates (- - -), as well as diagnostic plots for the sample likelihood when α and β are estimated simultaneously.



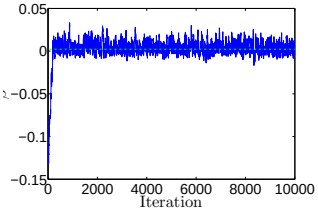
(A) R-W: $\hat{\alpha}$



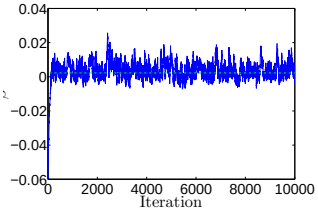
(B) MALA: $\hat{\alpha}$



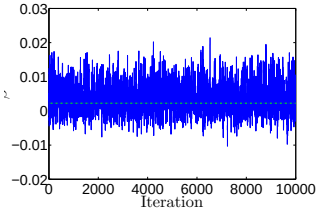
(C) smMALA: $\hat{\alpha}$



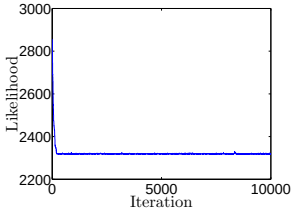
(D) R-W: $\hat{\beta}$



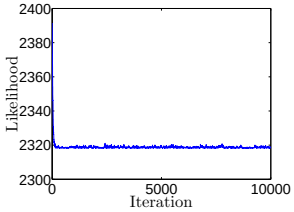
(E) MALA: $\hat{\beta}$



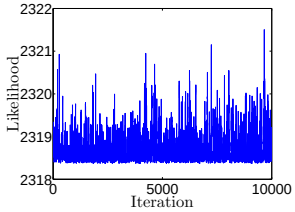
(F) smMALA: $\hat{\beta}$



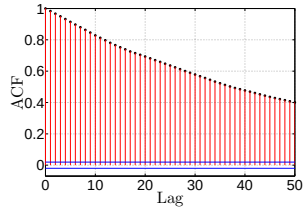
(G) R-W: Likelihood



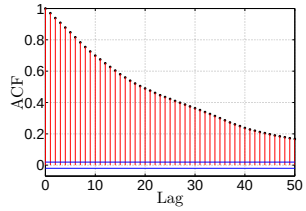
(H) MALA: Likelihood



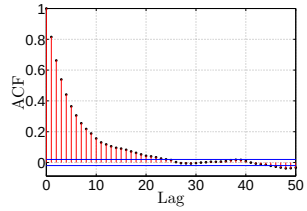
(I) smMALA: Likelihood



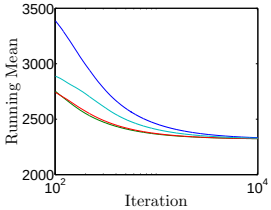
(J) R-W: ACF



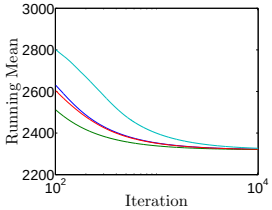
(K) MALA: ACF



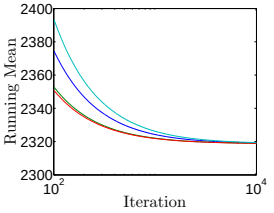
(L) smMALA: ACF



(M) R-W: Running Mean

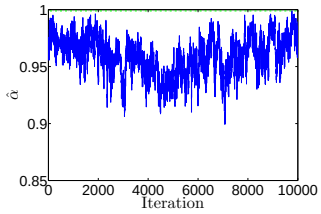


(N) MALA: Running Mean

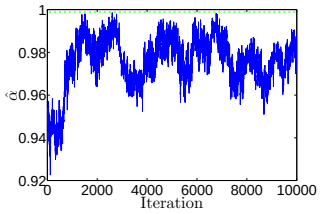


(O) smMALA: Running Mean

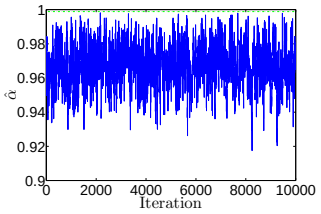
FIGURE E.8: **Case 2** $(\alpha, \beta, \mu, \psi^2)$: Traceplots of the posterior samples and maximum likelihood estimates (- - -) when all four parameters of the Heffernan and Tawn model are estimated simultaneously.



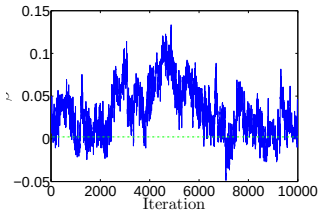
(A) R-W: $\hat{\alpha}$



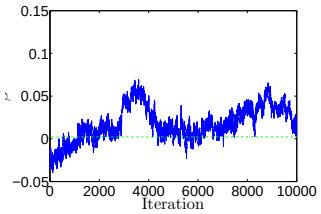
(B) MALA: $\hat{\alpha}$



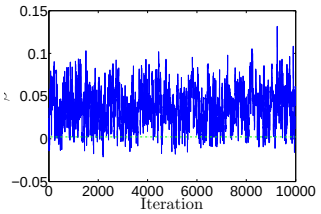
(C) smMALA: $\hat{\alpha}$



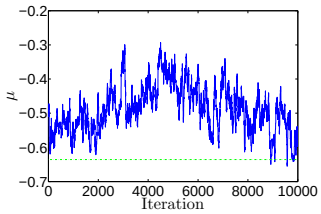
(D) R-W: $\hat{\beta}$



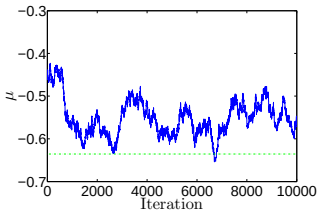
(E) MALA: $\hat{\beta}$



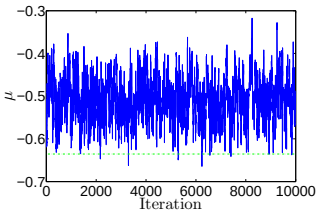
(F) smMALA: $\hat{\beta}$



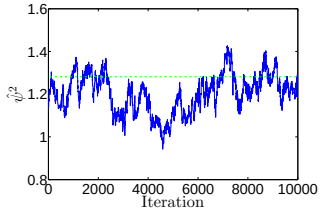
(G) R-W: $\hat{\mu}$



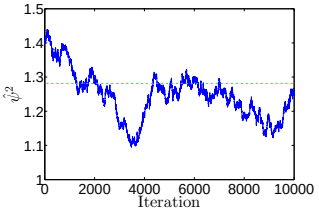
(H) MALA: $\hat{\mu}$



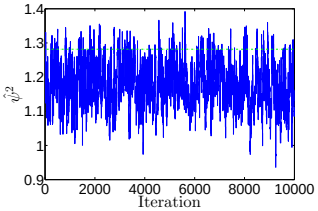
(I) smMALA: $\hat{\mu}$



(J) R-W: $\hat{\psi}^2$

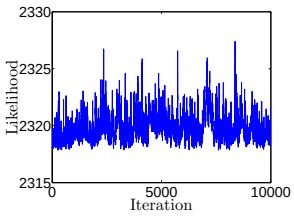


(K) MALA: $\hat{\psi}^2$

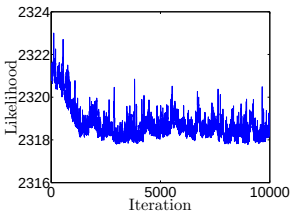


(L) smMALA: $\hat{\psi}^2$

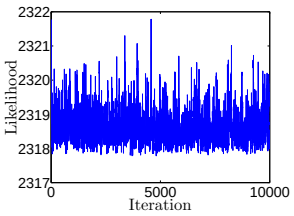
FIGURE E.9: **Case 2** $(\alpha, \beta, \mu, \psi^2)$: Diagnostic plots of the sample likelihood when all four parameters of the Heffernan and Tawn model are estimated simultaneously.



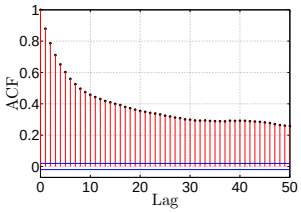
(A) R-W: Likelihood



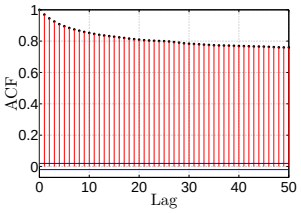
(B) MALA: Likelihood



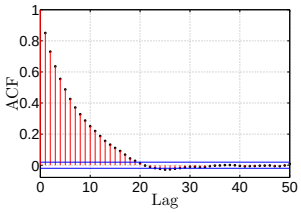
(C) smMALA: Likelihood



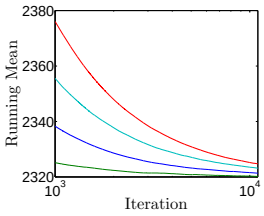
(D) R-W: ACF



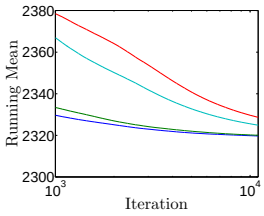
(E) MALA: ACF



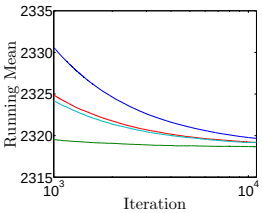
(F) smMALA: ACF



(G) R-W: Running Mean



(H) MALA: Running Mean



(I) smMALA:Running Mean

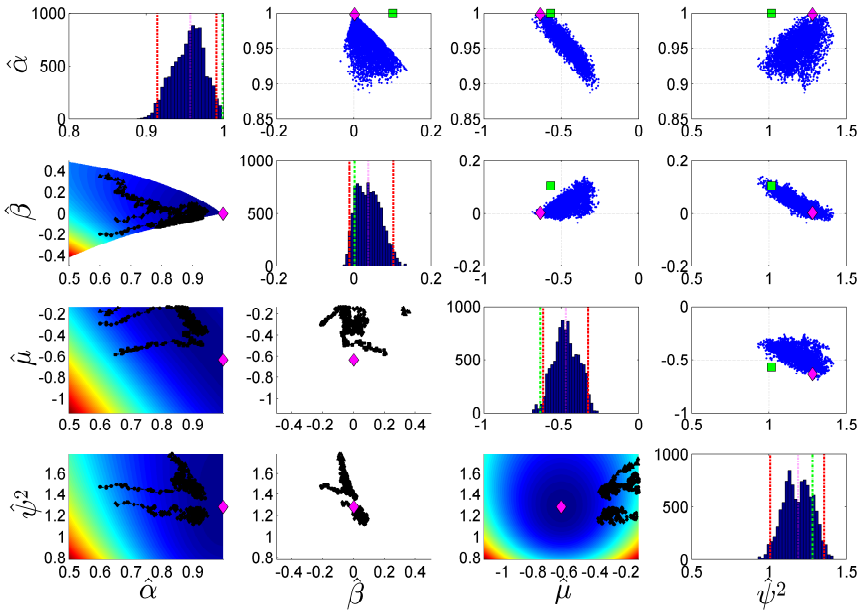


FIGURE E.10: **Case 2** $(\alpha, \beta, \mu, \psi^2)$: Scatterplot matrix of the burn-in sample (left of diagonal) and posterior sample (right of diagonal) based on the random walk transition kernel.

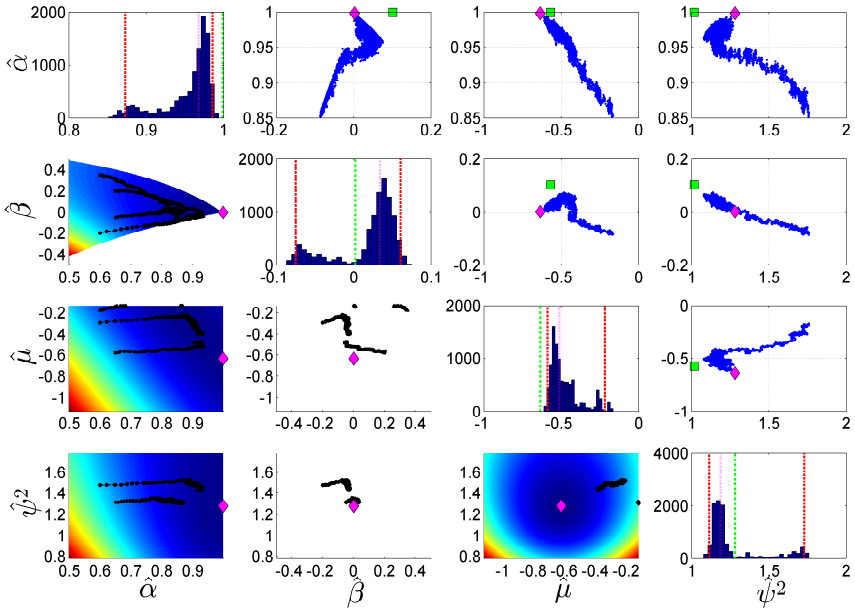


FIGURE E.11: **Case 2** $(\alpha, \beta, \mu, \psi^2)$: Scatterplot matrix of the burn-in sample (left of diagonal) and posterior sample (right of diagonal) based on the MALA.

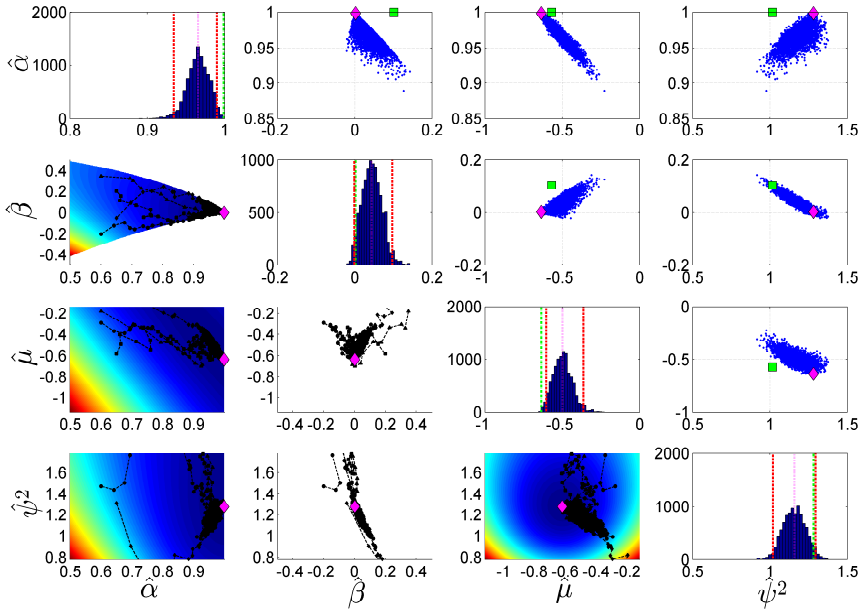


FIGURE E.12: **Case 2** $(\alpha, \beta, \mu, \psi^2)$: Scatterplot matrix of the burn-in sample (left of diagonal) and posterior sample (right of diagonal) based on the smMALA.

Appendix F

Diagnostic plots: generalized Heffernan and Tawn model

Diagnostic plots for the Bayesian analysis of the generalized Heffernan and Tawn model, discussed in Section 4.4. In consecutive order, the following cases are presented:

1. Case 1.1 with uninformative prior distributions, see Figure F.1–F.5.
2. Case 1.1 with informative prior distributions, see Figure F.6–F.10.
3. Case 1.2 with uninformative prior distributions, see Figure F.11–F.15.
4. Case 1.2 with informative prior distributions, see Figure F.16–F.20.

For each of the aforementioned cases, the following diagnostic plots are presented).

- Median (—) and 95% confidence interval (- - -) of the posterior sample of the spline curves for all four parameters of the Heffernan and Tawn model, as a function of the covariate. The true values for $\alpha(x)$ and $\beta(x)$ (· · ·) and maximum likelihood estimates (- · -) are provided if they are available. The starting values (· · ·) are intentionally poorly visible, such that they do not attract to much attention.
- Several diagnostic plots for the sample likelihood. Including a trace-plot of the likelihood for the posterior sample, the likelihood for multiple chains during burn-in and the autocorrelation function of the likelihood of the posterior sample.
- Correlation matrix which shows the correlation between the posterior samples of each of the weight coefficients. There should be no significant correlation other than on the leading diagonals of the correlation matrix.
- Trace-plots for the posterior samples of the weight coefficients ζ_θ. A converged and properly mixing posterior sample should resemble a white noise process around the maximum likelihood estimate.

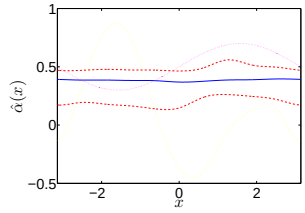
- Trace-plots for the roughness coefficient λ_θ. The absence of a trend and auto-regressive features, as well as constant variability, indicate the Markov chain has converged to a stationary limit distribution.
- Prior density and histogram of the posterior sample of the roughness coefficient λ_θ. If the histogram resembles the prior density line (—) the model is extremely sensitive to the specified prior distributions.

Several summary statistics are presented in Table F.1 and F.2.

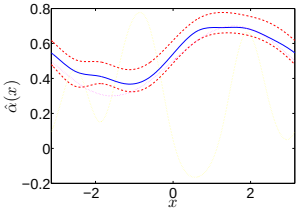
TABLE F.1: Summary statistics for the posterior samples of the weight coefficients. The presented statistics are averages of the the values obtained for individual posterior samples. A burn-in of $n_{\text{B}} = 2 \cdot 10^4$ is considered, and the following $n_{\text{S}} = 10^4$ samples are assumed to be valid observations from the posterior distribution. Three different proposal mechanisms are considered: random walk, Metropolis adjusted Langevin algorithm and simplified manifold Metropolis adjusted Langevin algorithm.

		Uninformative priors			**Informative priors**		
		RW	MALA	smMALA	RW	MALA	smMALA
ε		0.02	0.015	0.275	0.025	0.015	0.275
AR		0.01	0.12	0.46	0.17	0.34	0.36
$\hat{\alpha}$	ESS	5	8	50	6	7	47
	ESS/s	0.4	0.1	0.1	0.4	0.2	0.1
	$\hat{\text{R}}$	7.8	8.7	14.3	37.6	21.7	17.2
$\hat{\beta}$	ESS	8	5	53	5	8	36
	ESS/s	0.6	0.1	0.1	0.3	0.2	0.1
	$\hat{\text{R}}$	18.6	12.2	33.3	66.8	34.6	46.8
$\hat{\mu}$	ESS	3	3	47	4	6	44.4
	ESS/s	0.2	0.1	0.1	0.3	0.1	12
	$\hat{\text{R}}$	126.6	95.2	37.1	34.5	43.2	40.6
$\hat{\psi}$	ESS	3	4	47	4	6	35.7
	ESS/s	0.3	0.1	0.1	0.3	0.1	12
	$\hat{\text{R}}$	36.2	7.7	23.0	64.3	33.7	37.0

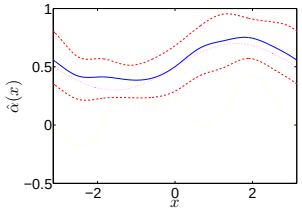
FIGURE F.1: **Case 1.1 with uninformative priors**: Median and 95% confidence interval of the posterior sample of the spline curves.



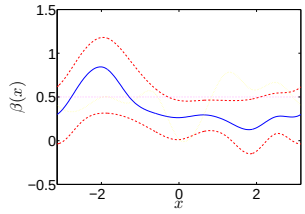
(A) R-W: $\hat{\alpha}(x)$



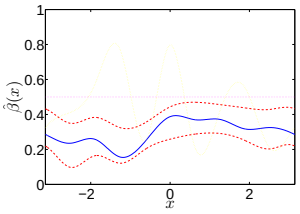
(B) MALA: $\hat{\alpha}(x)$



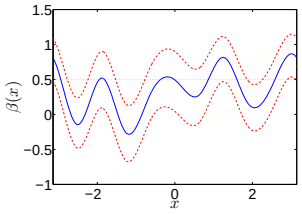
(C) smMALA: $\hat{\alpha}(x)$



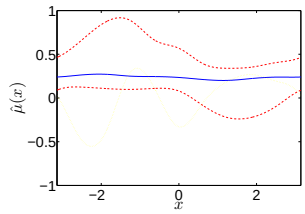
(D) R-W: $\hat{\beta}(x)$



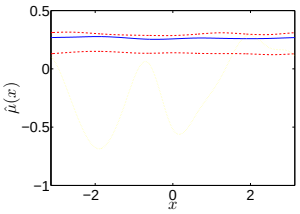
(E) MALA: $\hat{\beta}(x)$



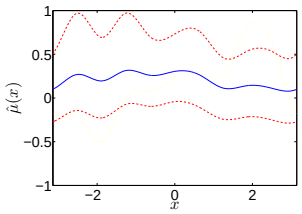
(F) smMALA: $\hat{\beta}(x)$



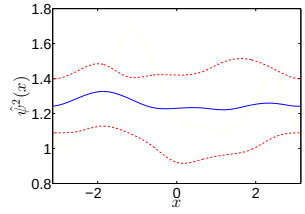
(G) R-W: $\hat{\mu}(x)$



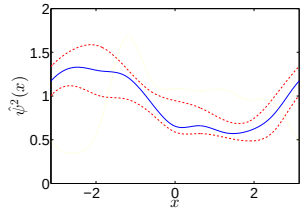
(H) MALA: $\hat{\mu}(x)$



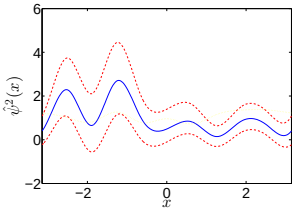
(I) smMALA: $\hat{\mu}(x)$



(J) R-W: $\hat{\psi}^2(x)$

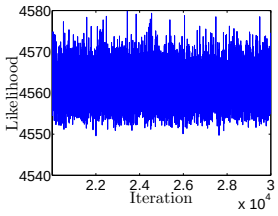


(K) MALA: $\hat{\psi}^2(x)$

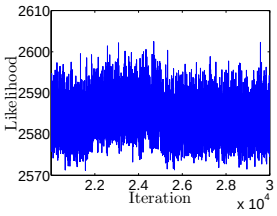


(L) smMALA: $\hat{\psi}^2(x)$

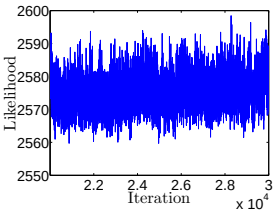
FIGURE F.2: **Case 1.1 with uninformative priors**: Diagnostic plots of the sample likelihood.



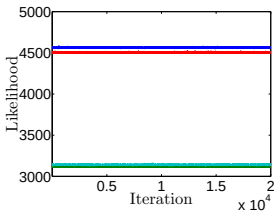
(A) R-W: Likelihood



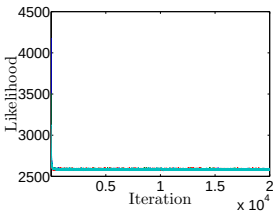
(B) MALA: Likelihood



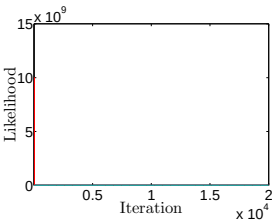
(C) smMALA: Likelihood



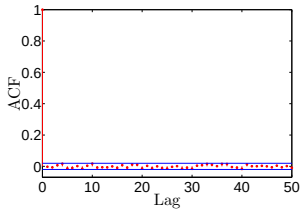
(D) R-W: Likelihood



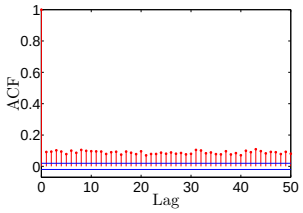
(E) MALA: Likelihood



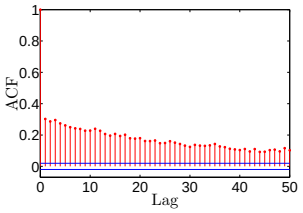
(F) smMALA: Likelihood



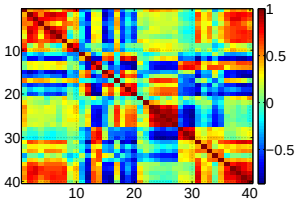
(G) R-W: ACF



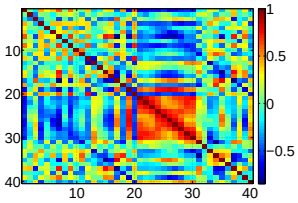
(H) MALA: ACF



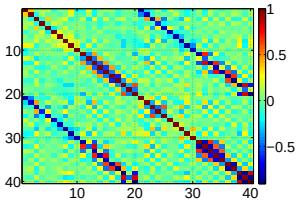
(I) smMALA: ACF



(J) R-W: Correlation

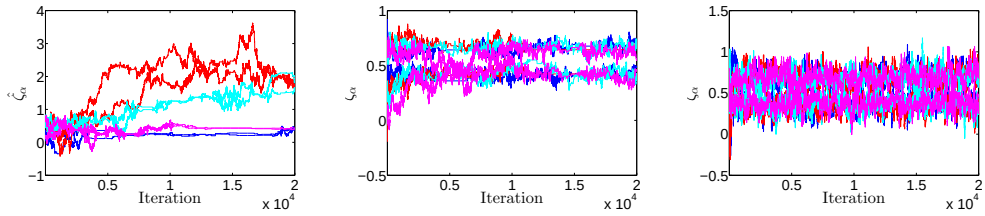


(K) MALA: Correlation

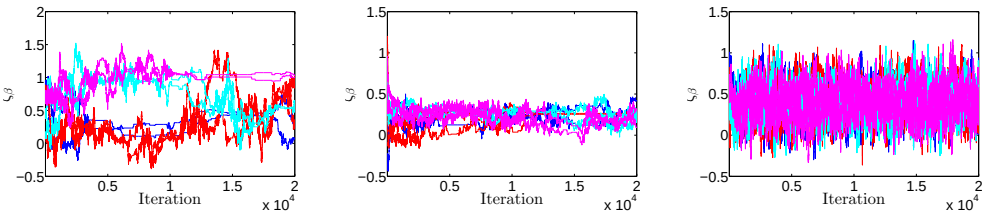


(L) smMALA: Correlation

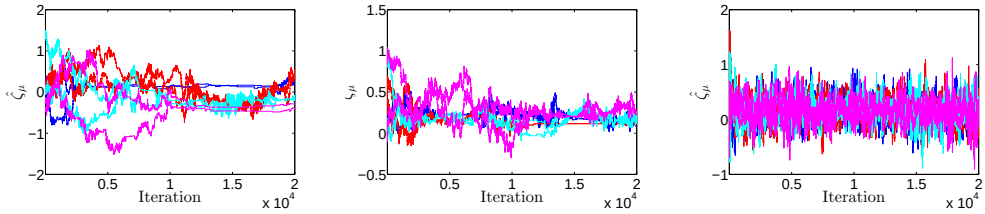
FIGURE F.3: **Case 1.1 with uninformative priors**: Traceplots of the posterior sample of a selection of the weight coefficients.



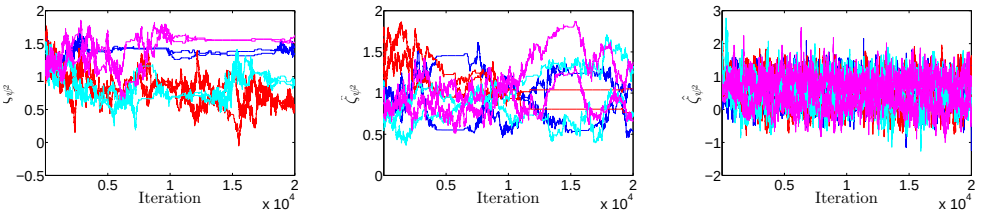
(A) R-W: $\hat{\zeta}_\alpha$ (B) MALA: $\hat{\zeta}_\alpha$ (C) smMALA: $\hat{\zeta}_\alpha$



(D) R-W: $\hat{\zeta}_\beta$ (E) MALA: $\hat{\zeta}_\beta$ (F) smMALA: $\hat{\zeta}_\beta$

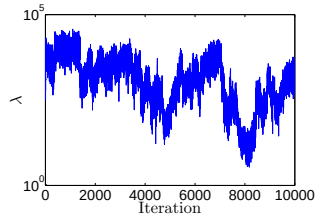


(G) R-W: $\hat{\zeta}_\mu$ (H) MALA: $\hat{\zeta}_\mu$ (I) smMALA: $\hat{\zeta}_\mu$

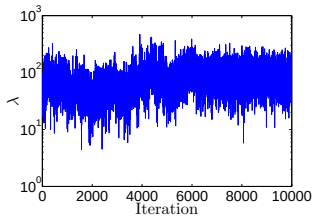


(J) R-W: $\hat{\zeta}_{\psi 2}$ (K) MALA: $\hat{\zeta}_{\psi 2}$ (L) smMALA: $\hat{\zeta}_{\psi 2}$

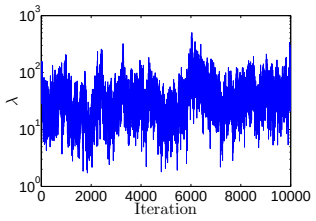
FIGURE F.4: **Case 1.1 with uninformative priors**: Traceplots for the roughness coefficient λ_θ.



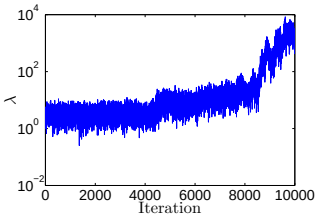
(A) R-W: $\hat{\lambda}_\alpha$



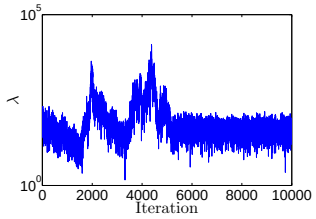
(B) MALA: $\hat{\lambda}_\alpha$



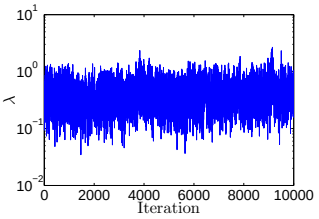
(C) smMALA: $\hat{\lambda}_\alpha$



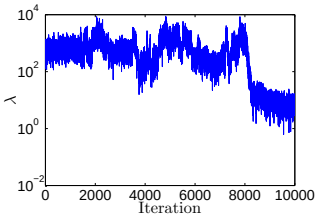
(D) R-W: $\hat{\lambda}_\beta$



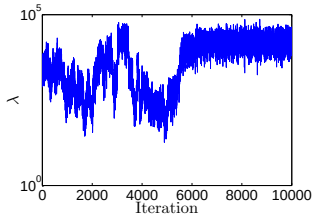
(E) MALA: $\hat{\lambda}_\beta$



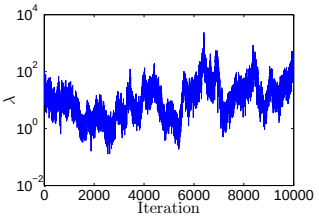
(F) smMALA: $\hat{\lambda}_\beta$



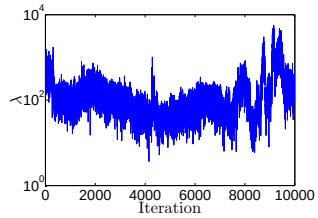
(G) R-W: $\hat{\lambda}_\mu$



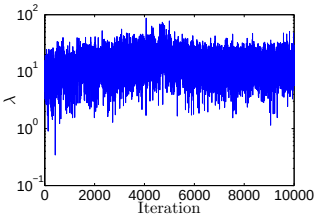
(H) MALA: $\hat{\lambda}_\mu$



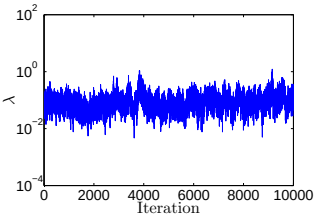
(I) smMALA: $\hat{\lambda}_\mu$



(J) R-W: $\hat{\lambda}_{\psi^2}$

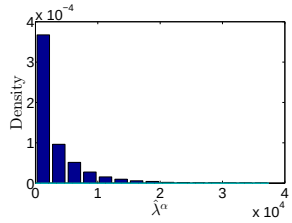


(K) MALA: $\hat{\lambda}_{\psi^2}$

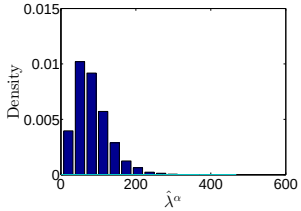


(L) smMALA: $\hat{\lambda}_{\psi^2}$

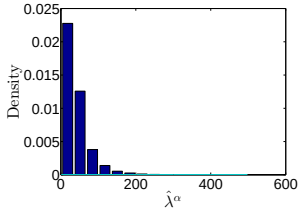
FIGURE F.5: **Case 1.1 with uninformative priors**: Prior density and histogram of the posterior sample for the roughness coefficient λ_θ.



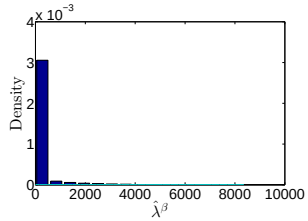
(A) R-W: $\hat{\lambda}_\alpha$



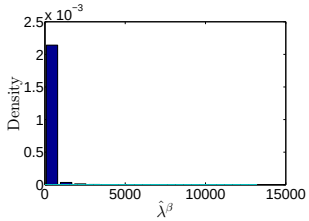
(B) MALA: $\hat{\lambda}_\alpha$



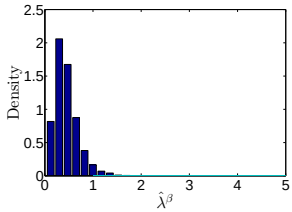
(C) smMALA: $\hat{\lambda}_\alpha$



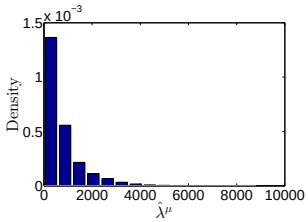
(D) R-W: $\hat{\lambda}_\beta$



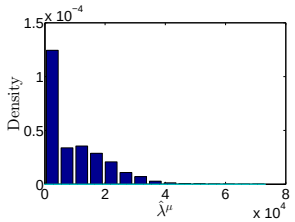
(E) MALA: $\hat{\lambda}_\beta$



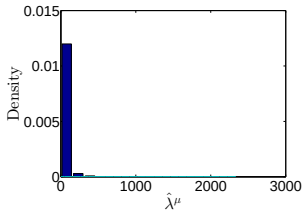
(F) smMALA: $\hat{\lambda}_\beta$



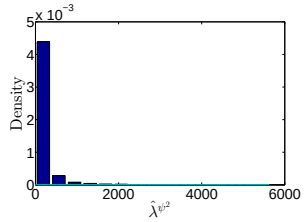
(G) R-W: $\hat{\lambda}_\mu$



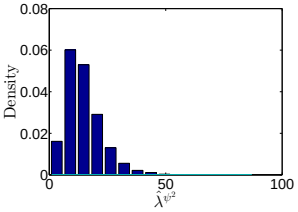
(H) MALA: $\hat{\lambda}_\mu$



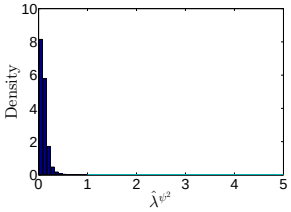
(I) smMALA: $\hat{\lambda}_\mu$



(J) R-W: $\hat{\lambda}_{\psi^2}$

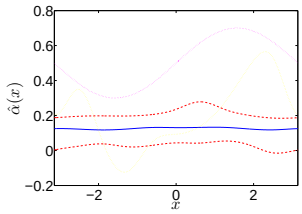


(K) MALA: $\hat{\lambda}_{\psi^2}$

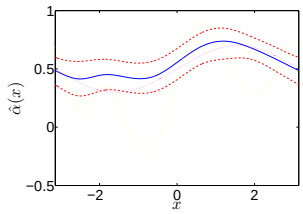


(L) smMALA: $\hat{\lambda}_{\psi^2}$

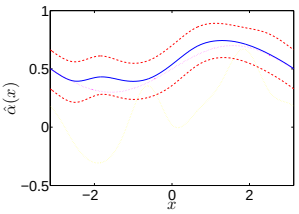
FIGURE F.6: **Case 1.1 with informative priors**: Median and 95% confidence interval of the posterior sample of the spline curves.



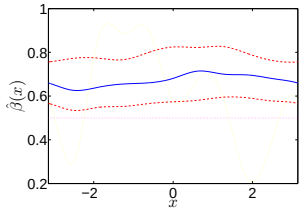
(A) R-W: $\hat{\alpha}(x)$



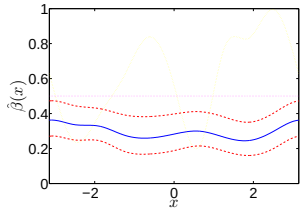
(B) MALA: $\hat{\alpha}(x)$



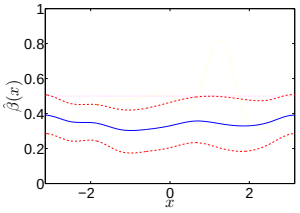
(C) smMALA: $\hat{\alpha}(x)$



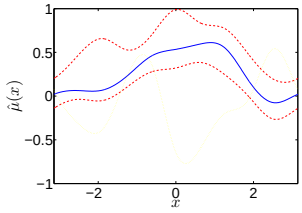
(D) R-W: $\hat{\beta}(x)$



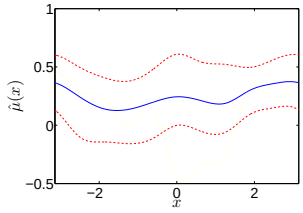
(E) MALA: $\hat{\beta}(x)$



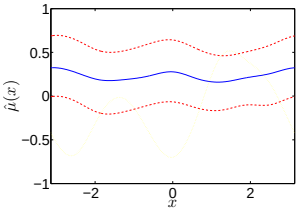
(F) smMALA: $\hat{\beta}(x)$



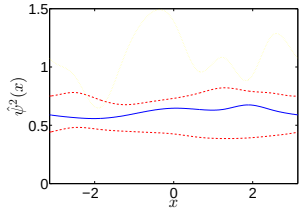
(G) R-W: $\hat{\mu}$



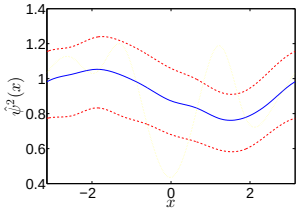
(H) MALA: $\hat{\mu}(x)$



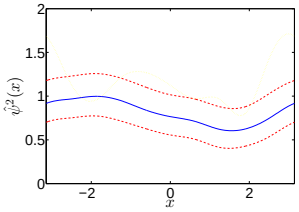
(I) smMALA: $\hat{\mu}(x)$



(J) R-W: $\hat{\psi}^2(x)$

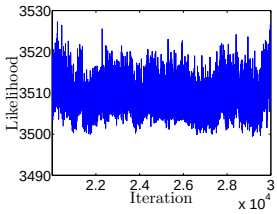


(K) MALA: $\hat{\psi}^2(x)$

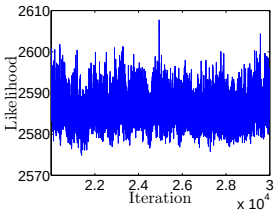


(L) smMALA: $\hat{\psi}^2(x)$

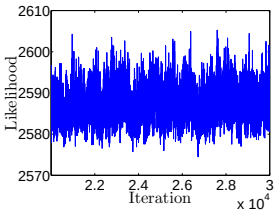
FIGURE F.7: **Case 1.1 with informative priors**: Diagnostic plots of the sample likelihood.



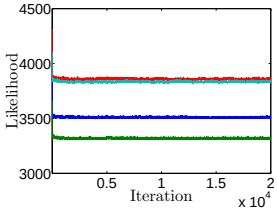
(A) R-W: Posterior



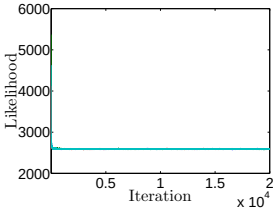
(B) MALA: Posterior



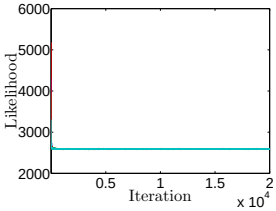
(C) smMALA: Posterior



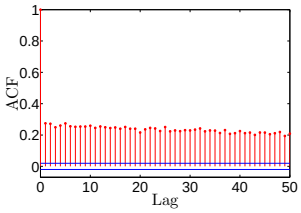
(D) R-W: Burn-in



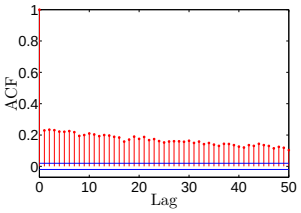
(E) MALA: Burn-in



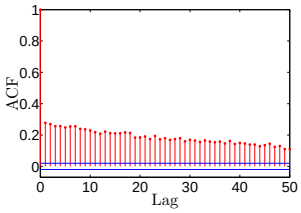
(F) smMALA:Burn-in



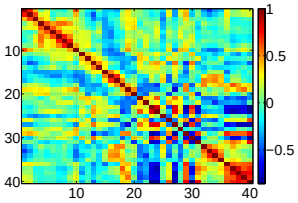
(G) R-W: ACF



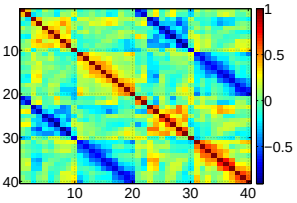
(H) MALA: ACF



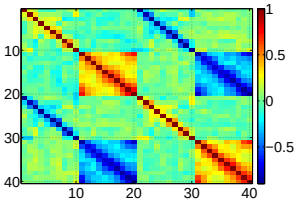
(I) smMALA: ACF



(J) R-W: Correlation

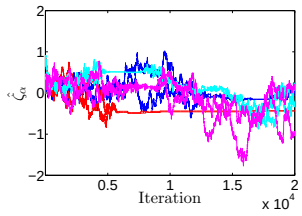


(K) MALA: Correlation

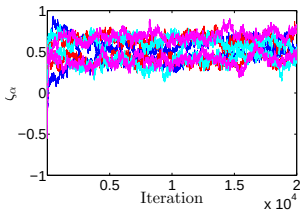


(L) smMALA: Correlation

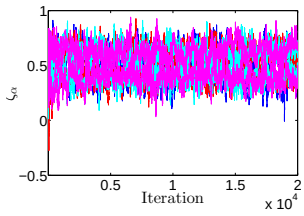
FIGURE F.8: **Case 1.1 with informative priors**: Traceplots of the posterior sample of a selection of the weight coefficients.



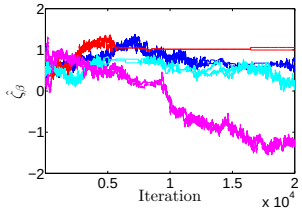
(A) R-W: $\hat{\zeta}_\alpha$



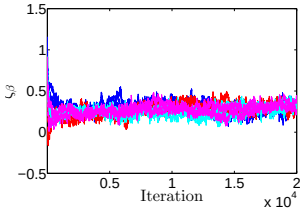
(B) MALA: $\hat{\zeta}_\alpha$



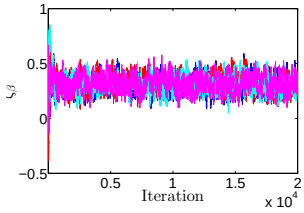
(C) smMALA: $\hat{\zeta}_\alpha$



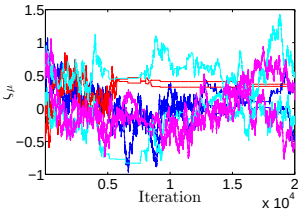
(D) R-W: $\hat{\zeta}_\beta$



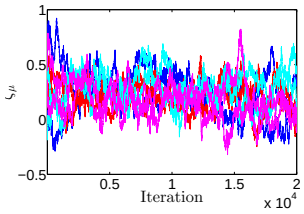
(E) MALA: $\hat{\zeta}_\beta$



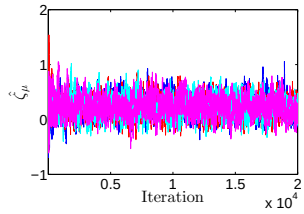
(F) smMALA: $\hat{\zeta}_\beta$



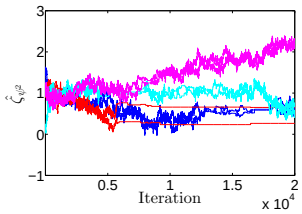
(G) R-W: $\hat{\zeta}_\mu$



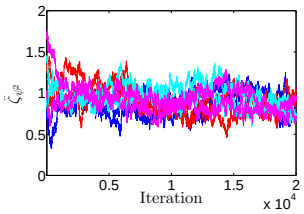
(H) MALA: $\hat{\zeta}_\mu$



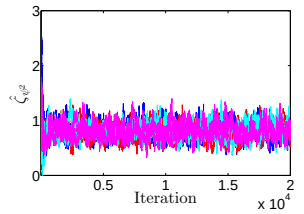
(I) smMALA: $\hat{\zeta}_\mu$



(J) R-W: $\hat{\zeta}_{\psi^2}$

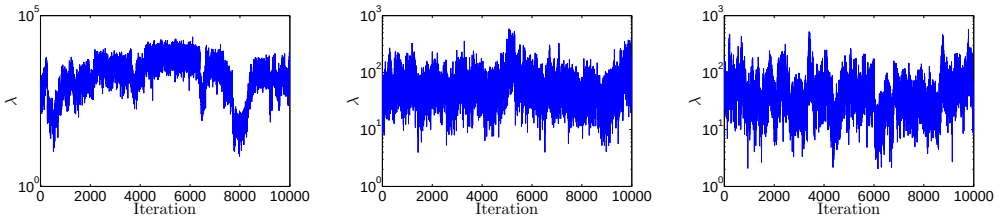


(K) MALA: $\hat{\zeta}_{\psi^2}$

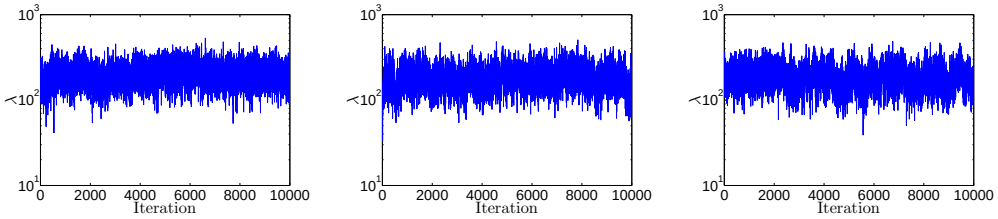


(L) smMALA: $\hat{\zeta}_{\psi^2}$

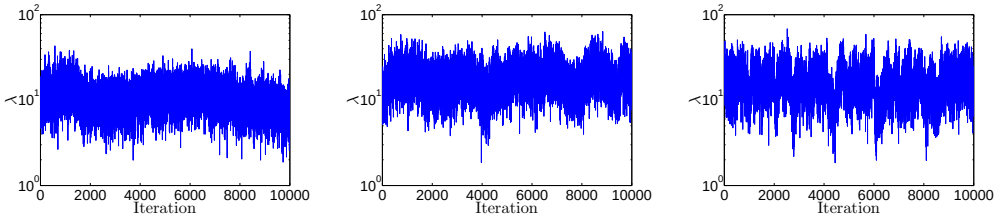
FIGURE F.9: **Case 1.1 with informative priors**: Traceplots for the roughness coefficient λ_θ.



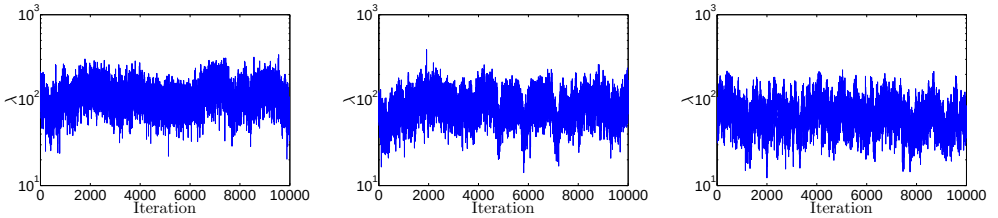
(A) R-W: $\hat{\lambda}_\alpha$ (B) MALA: $\hat{\lambda}_\alpha$ (C) smMALA: $\hat{\lambda}_\alpha$



(D) R-W: $\hat{\lambda}_\beta$ (E) MALA: $\hat{\lambda}_\beta$ (F) smMALA: $\hat{\lambda}_\beta$

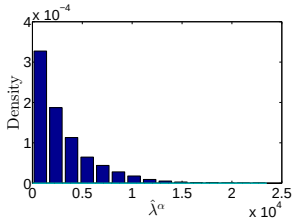


(G) R-W: $\hat{\lambda}_\mu$ (H) MALA: $\hat{\lambda}_\mu$ (I) smMALA: $\hat{\lambda}_\mu$

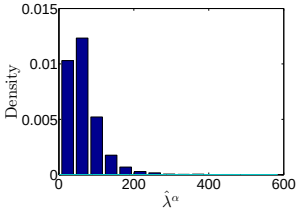


(J) R-W: $\hat{\lambda}_{\psi^2}$ (K) MALA: $\hat{\lambda}_{\psi^2}$ (L) smMALA: $\hat{\lambda}_{\psi^2}$

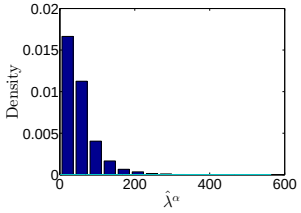
FIGURE F.10: **Case 1.1 with informative priors**: Prior density and histogram of the posterior sample for the roughness coefficient λ_θ.



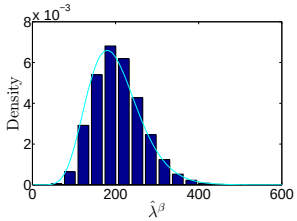
(A) R-W: $\hat{\lambda}_\alpha$



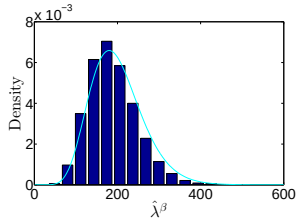
(B) MALA: $\hat{\lambda}_\alpha$



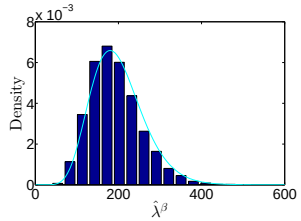
(C) smMALA: $\hat{\lambda}_\alpha$



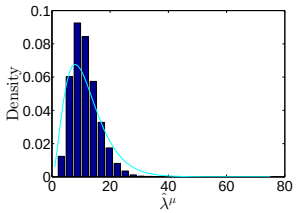
(D) R-W: $\hat{\lambda}_\beta$



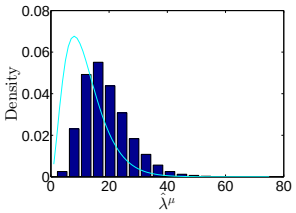
(E) MALA: $\hat{\lambda}_\beta$



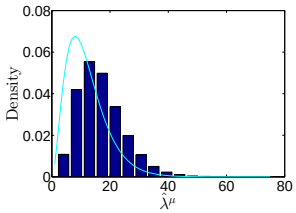
(F) smMALA: $\hat{\lambda}_\beta$



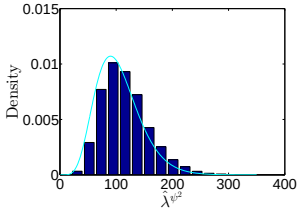
(G) R-W: $\hat{\lambda}_\mu$



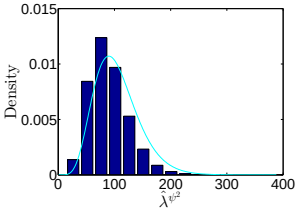
(H) MALA: $\hat{\lambda}_\mu$



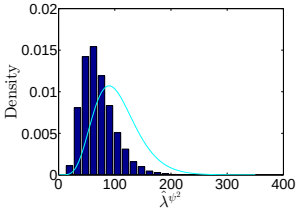
(I) smMALA: $\hat{\lambda}_\mu$



(J) R-W: $\hat{\lambda}_{\psi^2}$



(K) MALA: $\hat{\lambda}_{\psi^2}$

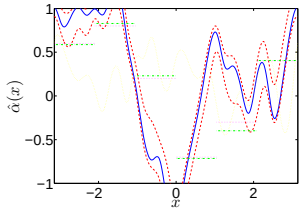


(L) smMALA: $\hat{\lambda}_{\psi^2}$

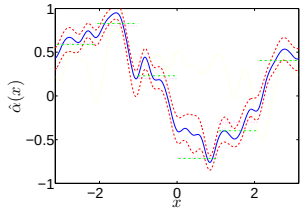
TABLE F.2: Summary statistics for the posterior samples of the weight coefficients. The presented statistics are averages of the the values obtained for individual posterior samples. A burn-in of $n_{\text{B}} = 2 \cdot 10^4$ is considered, and the following $n_{\text{S}} = 10^4$ samples are assumed to be valid observations from the posterior distribution. Three different proposal mechanisms are considered: random walk, Metropolis adjusted Langevin algorithm and simplified manifold Metropolis adjusted Langevin algorithm.

		Uninformative priors			**Informative priors**		
		RW	MALA	smMALA	RW	MALA	smMALA
ε		0.025	0.0075	0.1	0.01	0.01	0.125
AR		0.13	0.26	0.52	0.47	0.13	0.33
$\hat{\alpha}$	ESS	3	4	10	4	5	13
	ESS/s	0.2	0.04	0.03	0.2	0.06	0.04
	$\hat{\text{R}}$	4.9	3.2	4.3	13.1	2.4	5.7
$\hat{\beta}$	ESS	5	4	11	4	8	8
	ESS/s	0.3	0.04	0.04	0.3	0.04	0.03
	$\hat{\text{R}}$	13.6	4.4	5.5	24.9	4.4	5.2
$\hat{\mu}$	ESS	3	3	11	4	6	11
	ESS/s	0.2	0.03	0.04	0.1	0.04	0.04
	$\hat{\text{R}}$	14.5	4.3	5.8	7.1	3.4	6.0
$\hat{\psi}$	ESS	4	2	9	3	6	9
	ESS/s	0.2	0.03	0.03	0.3	0.04	0.03
	$\hat{\text{R}}$	51.7	2.1	7.4	28.3	2.9	6.0

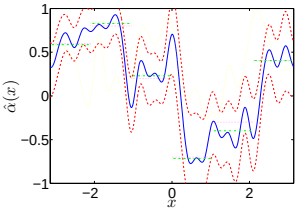
FIGURE F.11: **Case 1.2 with uninformative priors**: Median and 95% confidence interval of the posterior sample of the spline curves.



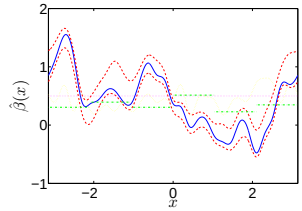
(A) R-W: $\hat{\alpha}(x)$



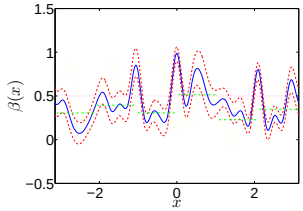
(B) MALA: $\hat{\alpha}(x)$



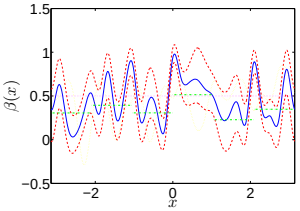
(C) smMALA: $\hat{\alpha}(x)$



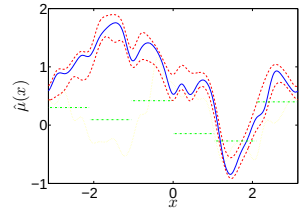
(D) R-W: $\hat{\beta}(x)$



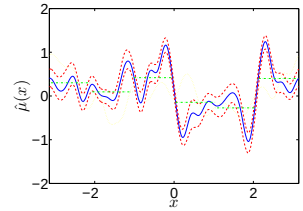
(E) MALA: $\hat{\beta}(x)$



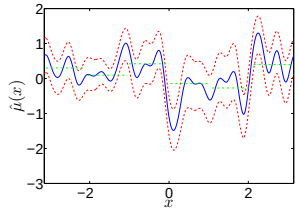
(F) smMALA: $\hat{\beta}(x)$



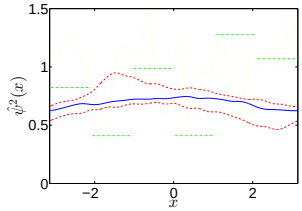
(G) R-W: $\hat{\mu}(x)$



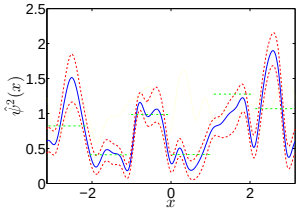
(H) MALA: $\hat{\mu}(x)$



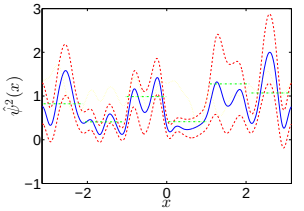
(I) smMALA: $\hat{\mu}(x)$



(J) R-W: $\hat{\psi}^2(x)$

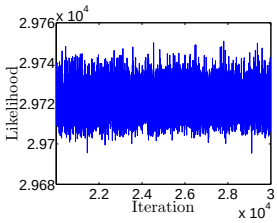


(K) MALA: $\hat{\psi}^2(x)$

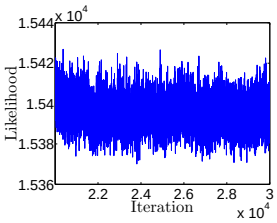


(L) smMALA: $\hat{\psi}^2(x)$

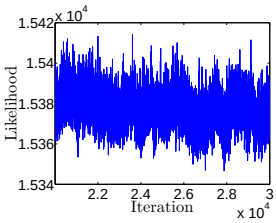
FIGURE F.12: **Case 1.2 with uninformative priors**: Diagnostic plots of the sample likelihood.



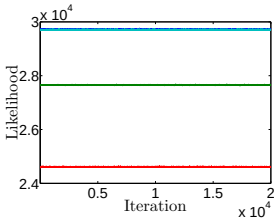
(A) R-W: Posterior



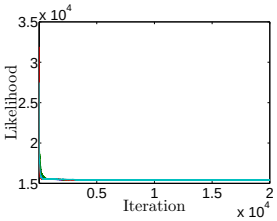
(B) MALA: Posterior



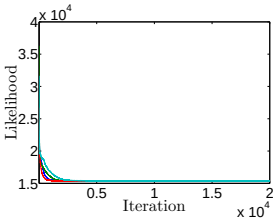
(C) smMALA: Posterior



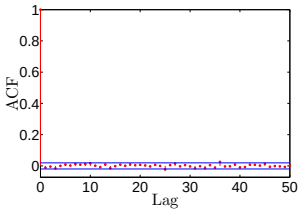
(D) R-W: Burn-in



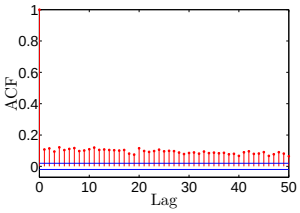
(E) MALA: Burn-in



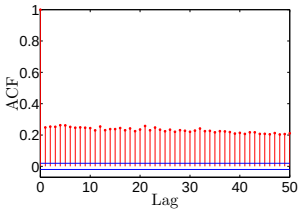
(F) smMALA: Burn-in



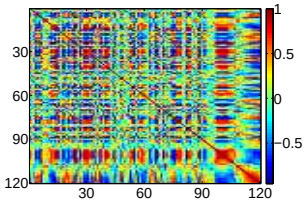
(G) R-W: ACF



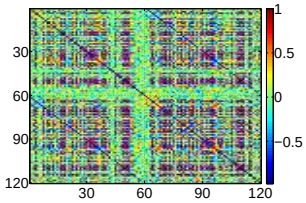
(H) MALA: ACF



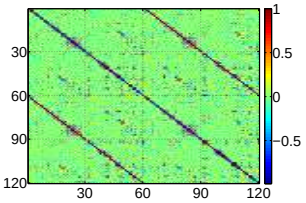
(I) smMALA: ACF



(J) R-W: Correlation

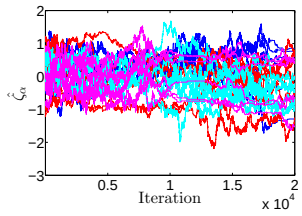


(K) MALA: Correlation

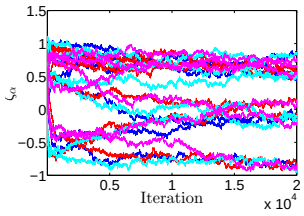


(L) smMALA: Correlation

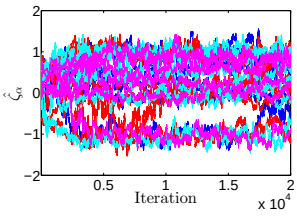
FIGURE F.13: **Case 1.2 with uninformative priors**: Traceplots of the posterior sample of a selection of the weight coefficients.



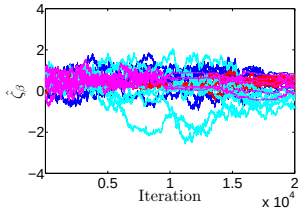
(A) R-W: $\hat{\zeta}_\alpha$



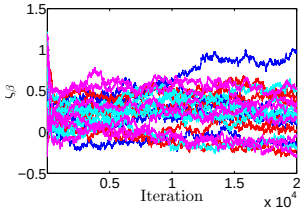
(B) MALA: $\hat{\zeta}_\alpha$



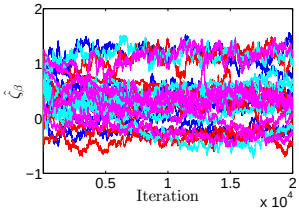
(C) smMALA: $\hat{\zeta}_\alpha$



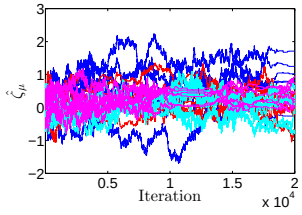
(D) R-W: $\hat{\zeta}_\beta$



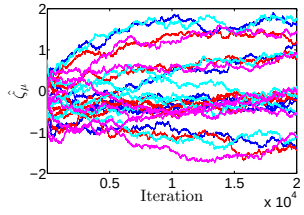
(E) MALA: $\hat{\zeta}_\beta$



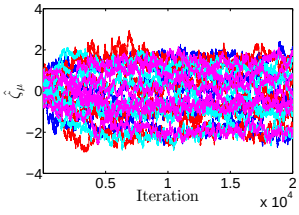
(F) smMALA: $\hat{\zeta}_\beta$



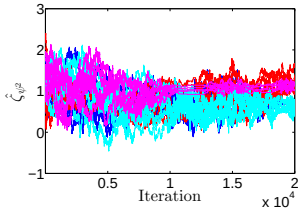
(G) R-W: $\hat{\zeta}_\mu$



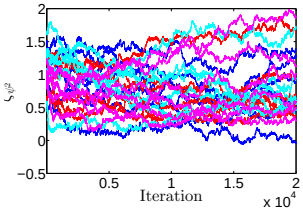
(H) MALA: $\hat{\zeta}_\mu$



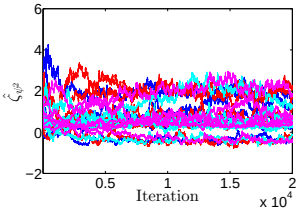
(I) smMALA: $\hat{\zeta}_\mu$



(J) R-W: $\hat{\zeta}_{\psi^2}$

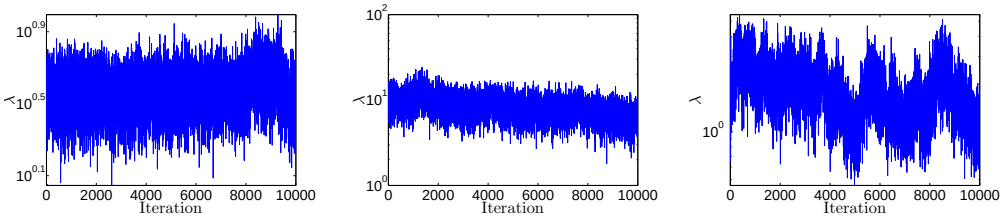


(K) MALA: $\hat{\zeta}_{\psi^2}$

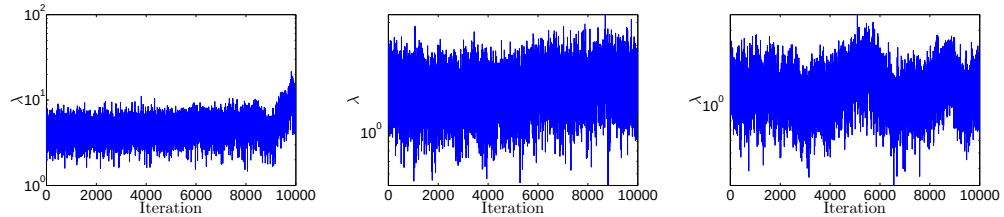


(L) smMALA: $\hat{\zeta}_{\psi^2}$

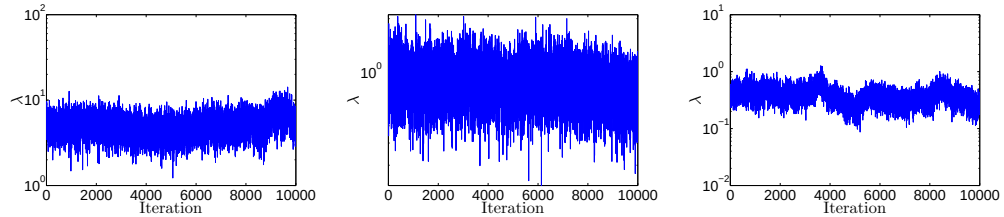
FIGURE F.14: **Case 1.2 with uninformative priors**: Traceplots for the roughness coefficient λ_θ.



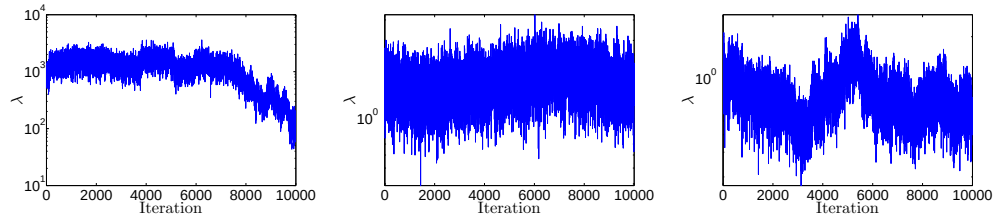
(A) R-W: $\hat{\lambda}_\alpha$ (B) MALA: $\hat{\lambda}_\alpha$ (C) smMALA: $\hat{\lambda}_\alpha$



(D) R-W: $\hat{\lambda}_\beta$ (E) MALA: $\hat{\lambda}_\beta$ (F) smMALA: $\hat{\lambda}_\beta$

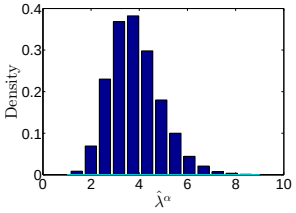


(G) R-W: $\hat{\lambda}_\mu$ (H) MALA: $\hat{\lambda}_\mu$ (I) smMALA: $\hat{\lambda}_\mu$

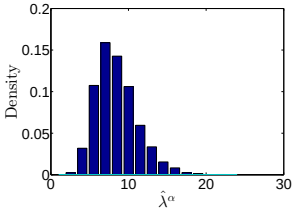


(J) R-W: $\hat{\lambda}_{\psi^2}$ (K) MALA: $\hat{\lambda}_{\psi^2}$ (L) smMALA: $\hat{\lambda}_{\psi^2}$

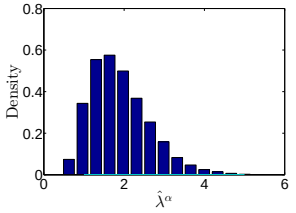
FIGURE F.15: **Case 1.2 with uninformative priors**: Prior density and histogram of the posterior sample for the roughness coefficient λ_θ.



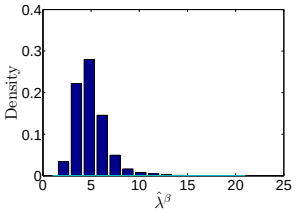
(A) R-W: $\hat{\lambda}_\alpha$



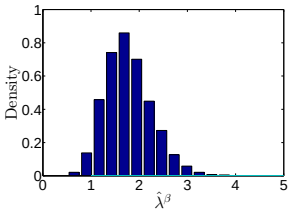
(B) MALA: $\hat{\lambda}_\alpha$



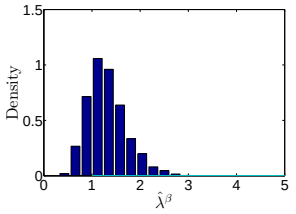
(C) smMALA: $\hat{\lambda}_\alpha$



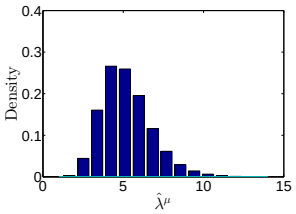
(D) R-W: $\hat{\lambda}_\beta$



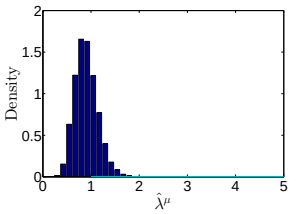
(E) MALA: $\hat{\lambda}_\beta$



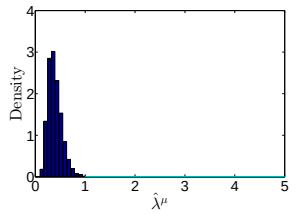
(F) smMALA: $\hat{\lambda}_\beta$



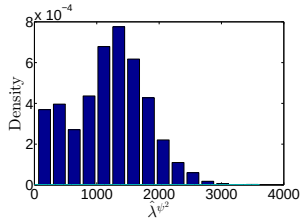
(G) R-W: $\hat{\lambda}_\mu$



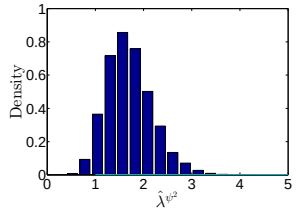
(H) MALA: $\hat{\lambda}_\mu$



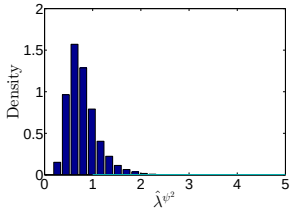
(I) smMALA: $\hat{\lambda}_\mu$



(J) R-W: $\hat{\lambda}_{\psi^2}$

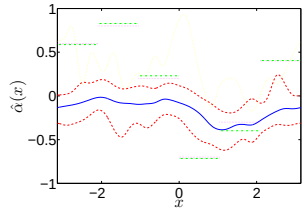


(K) MALA: $\hat{\lambda}_{\psi^2}$

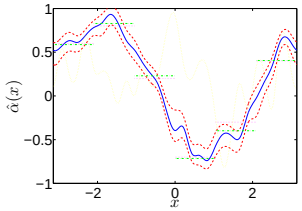


(L) smMALA: $\hat{\lambda}_{\psi^2}$

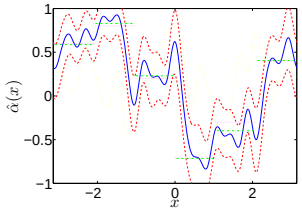
FIGURE F.16: **Case 1.2 with informative priors**: Median and 95% confidence interval of the posterior sample of the spline curves.



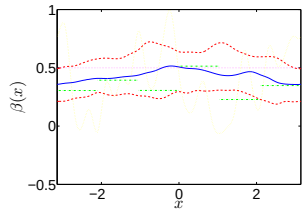
(A) R-W: $\hat{\alpha}\,(x)$



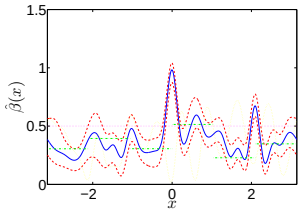
(B) MALA: $\hat{\alpha}\,(x)$



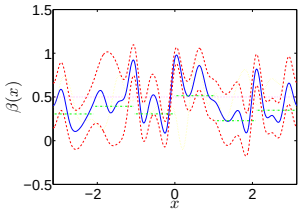
(C) smMALA: $\hat{\alpha}\,(x)$



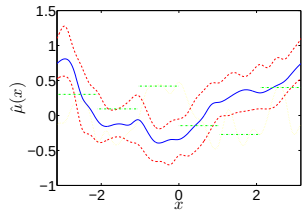
(D) R-W: $\hat{\beta}\,(x)$



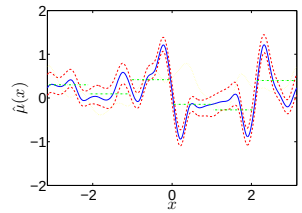
(E) MALA: $\hat{\beta}\,(x)$



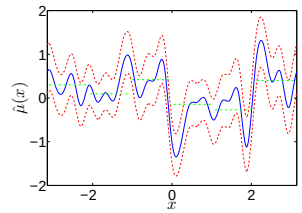
(F) smMALA: $\hat{\beta}\,(x)$



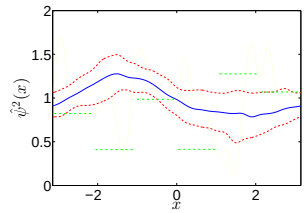
(G) R-W: $\hat{\mu}\,(x)$



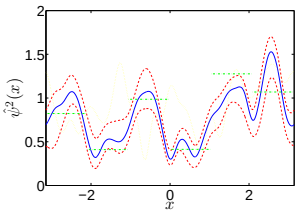
(H) MALA: $\hat{\mu}\,(x)$



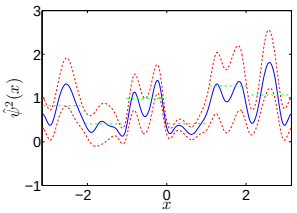
(I) smMALA: $\hat{\mu}\,(x)$



(J) R-W: $\hat{\psi}^2\,(x)$

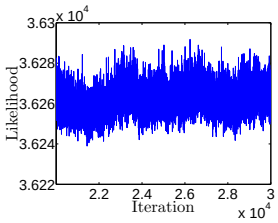


(K) MALA: $\hat{\psi}^2\,(x)$

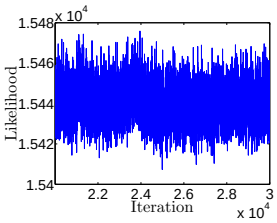


(L) smMALA: $\hat{\psi}^2\,(x)$

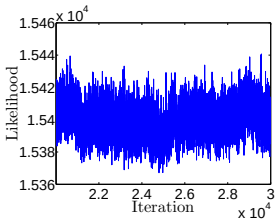
FIGURE F.17: **Case 1.2 with informative priors**: Diagnostic plots of the sample likelihood.



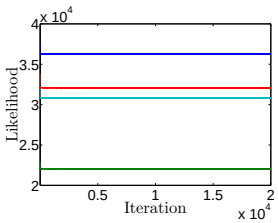
(A) R-W: Posterior



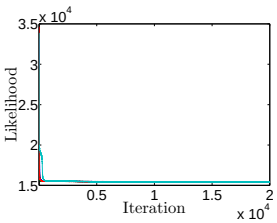
(B) MALA: Posterior



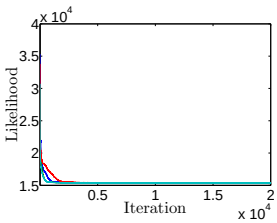
(C) smMALA: Posterior



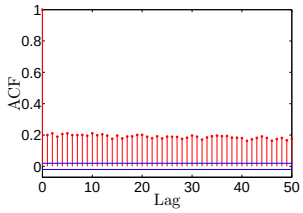
(D) R-W: Burn-in



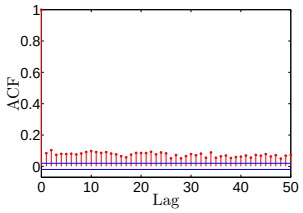
(E) MALA: Burn-in



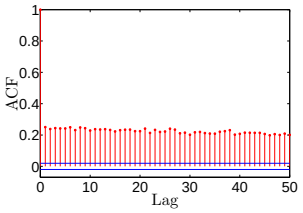
(F) smMALA: Burn-in



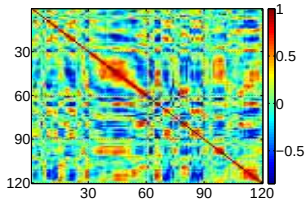
(G) R-W: ACF



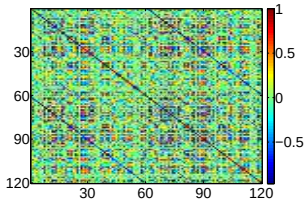
(H) MALA: ACF



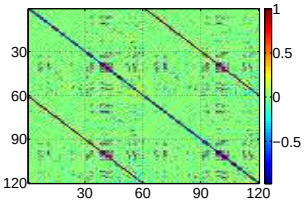
(I) smMALA: ACF



(J) R-W: Correlation

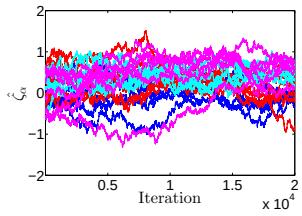


(K) MALA: Correlation

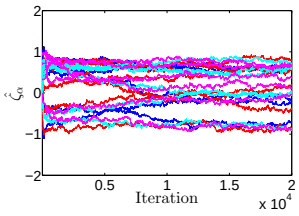


(L) smMALA: Correlation

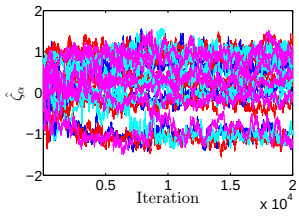
FIGURE F.18: **Case 1.2 with informative priors**: Traceplots of the posterior sample of a selection of the weight coefficients.



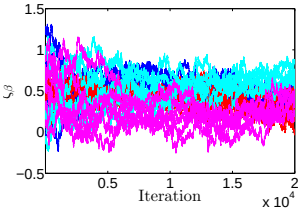
(A) R-W: $\hat{\zeta}_\alpha$



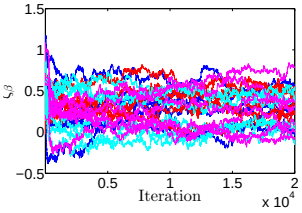
(B) MALA: $\hat{\zeta}_\alpha$



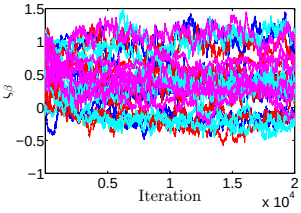
(C) smMALA: $\hat{\zeta}_\alpha$



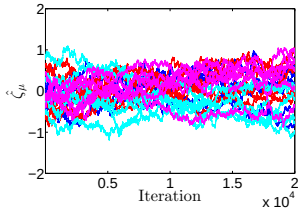
(D) R-W: $\hat{\zeta}_\beta$



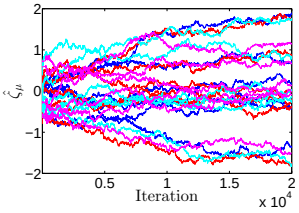
(E) MALA: $\hat{\zeta}_\beta$



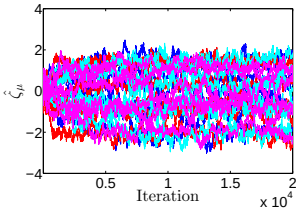
(F) smMALA: $\hat{\zeta}_\beta$



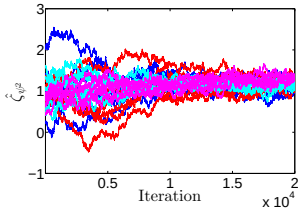
(G) R-W: $\hat{\zeta}_\mu$



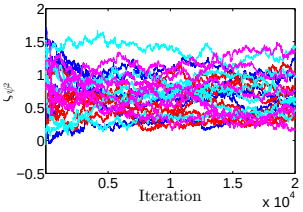
(H) MALA: $\hat{\zeta}_\mu$



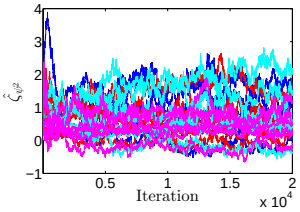
(I) smMALA: $\hat{\zeta}_\mu$



(J) R-W: $\hat{\zeta}_{\psi 2}$

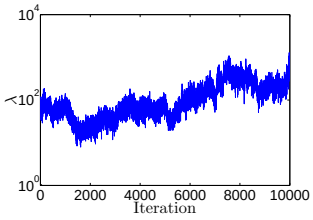


(K) MALA: $\hat{\zeta}_{\psi 2}$

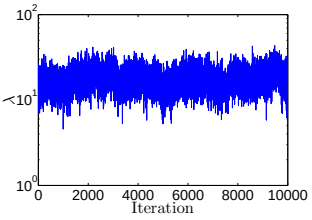


(L) smMALA: $\hat{\zeta}_{\psi 2}$

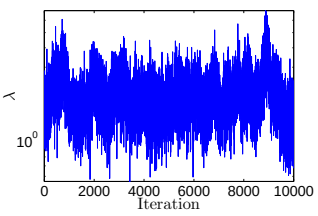
FIGURE F.19: **Case 1.2 with informative priors**: Traceplots for the roughness coefficient λ_θ.



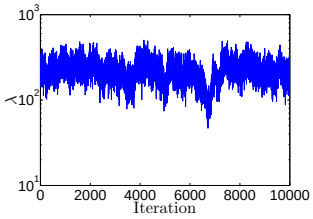
(A) R-W: $\hat{\lambda}_\alpha$



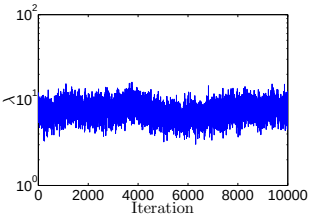
(B) MALA: $\hat{\lambda}_\alpha$



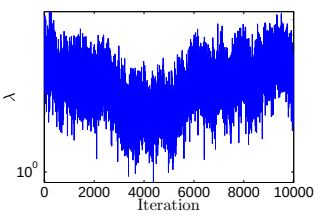
(C) smMALA: $\hat{\lambda}_\alpha$



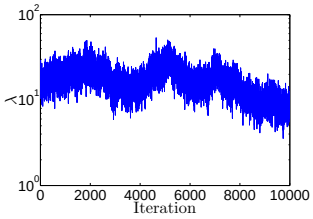
(D) R-W: $\hat{\lambda}_\beta$



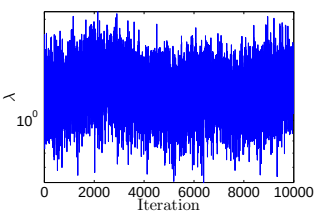
(E) MALA: $\hat{\lambda}_\beta$



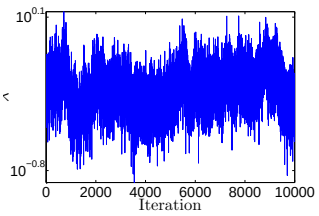
(F) smMALA: $\hat{\lambda}_\beta$



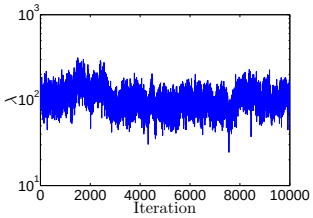
(G) R-W: $\hat{\lambda}_\mu$



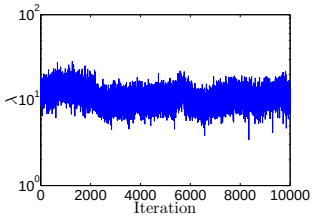
(H) MALA: $\hat{\lambda}_\mu$



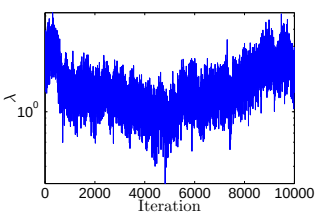
(I) smMALA: $\hat{\lambda}_\mu$



(J) R-W: $\hat{\lambda}_{\psi^2}$

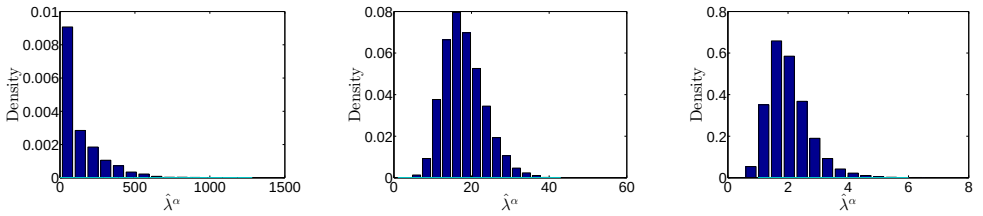


(K) MALA: $\hat{\lambda}_{\psi^2}$

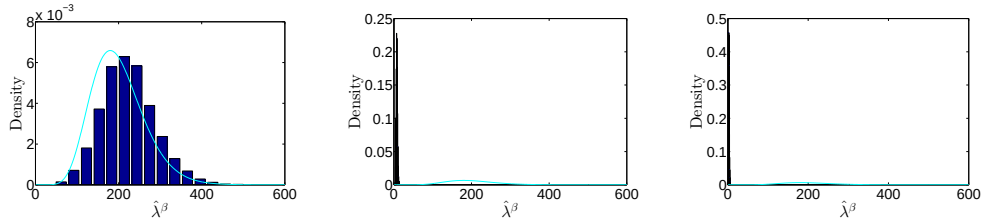


(L) smMALA: $\hat{\lambda}_{\psi^2}$

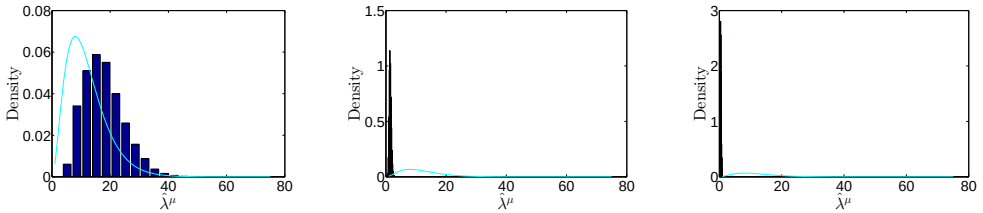
FIGURE F.20: **Case 1.2 with informative priors**: Prior density and histogram of the posterior sample for the roughness coefficient λ_θ.



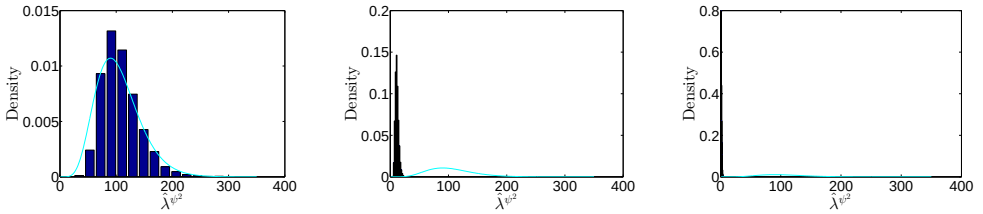
(A) R-W: $\hat{\lambda}_\alpha$ (B) MALA: $\hat{\lambda}_\alpha$ (C) smMALA: $\hat{\lambda}_\alpha$



(D) R-W: $\hat{\lambda}_\beta$ (E) MALA: $\hat{\lambda}_\beta$ (F) smMALA: $\hat{\lambda}_\beta$



(G) R-W: $\hat{\lambda}_\mu$ (H) MALA: $\hat{\lambda}_\mu$ (I) smMALA: $\hat{\lambda}_\mu$



(J) R-W: $\hat{\lambda}_{\psi^2}$ (K) MALA: $\hat{\lambda}_{\psi^2}$ (L) smMALA: $\hat{\lambda}_{\psi^2}$

Bibliography

AghaKouchak, Amir, David Easterling, Kuolin Hsu, Siegfried Schubert, and Soroosh Sorooshian (2013). *Extremes in a changing climate. Detection, Analysis and Uncertainty*, p. 423 (cit. on p. 30).

Bagnoli, Mark and Ted Bergstrom (1989). "Log-Concave Probability and Its Applications by" (cit. on p. 18).

Balkema, A. A. and Laurens De Haan (1974). "Residual Life Time at Great Age". In: *The Annals of Probability* 2.5, pp. 792–804 (cit. on p. 18).

Beirlant, Jan, Yuri Goegebeur, Johan Segers, and Jozef Teugels (2004). *Statistics of Extremes: Theory and Applications*. Ed. by Daniel De Waal and Chris Ferro. Wiley New York (cit. on p. 10).

Botev, Z. I. (2016). "The normal law under linear restrictions: Simulation and estimation via minimax tilting". In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* (cit. on p. 103).

Cao, Xumeng (2013). "Relative Performance of Expected and Observed Fisher Information in Covariance Estimation for Maximum Likelihood Estimators". Dissertation for the degree of Doctor of Philospohy. Johns Hopkins University (cit. on p. 53).

Catchpole, E. A. and B. J. T. Morgan (1997). "Detecting Parameter Redundancy". In: *Biometrika* 84.1, pp. 187–192 (cit. on p. 55).

Chavez-Demoulin, Valerie and Anthony C. Davison (2012). "Modelling time series extremes". In: *Revstat Statistical Journal* 10.1, pp. 109–133 (cit. on p. 1).

Cheng, Linyin, Eric Gilleland, Matthew J. Heaton, and Amir Aghakouchak (2014). "Empirical Bayes estimation for the conditional extreme value model". In: *Stat* 3.1, pp. 391–406 (cit. on pp. 5, 54, 75).

Coles, Stuart G. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer (cit. on pp. 10, 14, 24).

Coles, Stuart G. and Jonathan A. Tawn (1994). "Statistical Methods for Multivariate Extremes: an Application to Structural Design". In: *Journal of the Royal Statistical Society. Series C: Applied Statistics* 43.1, pp. 1–48 (cit. on p. 23).

Coles, Stuart G., Jonathan A. Tawn, and Stuart G. Coles (1991). "Modelling Extreme Multivariate Events". In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 53.2, pp. 377–392 (cit. on pp. 27, 46, 4).

Das, Bikramjit and Sidney I. Resnick (2011). "Conditioning on an extreme component: Model consistency with regular variation on cones". In: *Bernoulli* 17.1, pp. 226–252 (cit. on p. 24).

Davison, Anthony C., S. A. Padoan, and M. Ribatet (2012). "Statistical Modeling of Spatial Extremes". In: *Statistical Science* 27.2, pp. 161–186 (cit. on p. 1).

Davison, Anthony C., Peiman Asadi, and Sebastian Engelke (2015a). "Extremes on river networks". In: *Annals of Applied Statistics* 9.4, pp. 2023–2050 (cit. on pp. 1, 21).

Davison, Anthony C., Jennifer L. Wadsworth, Jonathan A. Tawn, and Daniel Elton (2015b). "Modelling across extremal dependence classes" (cit. on p. 29).

De Boor, C. (1978). *A Practical Guide to Splines*. Springer (cit. on p. 89).

De Boor, C., Tom Lyche, and Larry Schumaker (1976). *Numerische Methoden der Approximationstheorie/Numerical Methods of Approximation Theory*. Springer, pp. 123–146 (cit. on p. 90).

De Haan, Laurens and Ana Ferreira (2006). *Extreme Value Theory: An Introduction*. Ed. by Thomas V. Mikosch, Sidney I. Resnick, and Stephen M. Robinson. Springer (cit. on pp. 10, 17, 26, 31).

Dümbgen, Lutz and Kaspar Rufibach (2009). "Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency". In: *Bernoulli* 15.1, pp. 40–68 (cit. on p. 18).

Eastoe, Emma F. and Jonathan A. Tawn (2009). "Modelling non-stationary extremes with application to surface level ozone". In: *Journal of the Royal Statistical Society. Series C: Applied Statistics* 58.1, pp. 25–45 (cit. on p. 65).

Eck, Matthias and Jan Hadenfeld (1995). "Knot removal for B-spline curves". In: *Computer Aided Geometric Design* 12.3, pp. 259–282 (cit. on pp. 98, 103).

Efron, Bradley (1987). "Better Bootstrap Confidence Intervals". In: *Journal of the American Statistical Association* 82.397, pp. 171–185 (cit. on p. 62).

Eilers, Paul H. C. and Brian D. Marx (1996). "Flexible Smoothing with B-splines and Penalties". In: *Statistical Science* 11.2, pp. 89–102 (cit. on pp. 89, 90, 101).

Feng, Ziding and Charles E. McCulloch (1992). "Statistical inference using maximum likelihood estimation and the generalized likelihood ratio when the true parameter is on the boundary of the parameter space". In: *Statistics and Probability Letters* 13, pp. 325–332 (cit. on p. 54).

Fisher, R. A. and L. H. C. Tippett (1928). "Limiting forms of the frequency distribution of the large sample". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 24.2, pp. 180–190 (cit. on pp. 13, 14).

Fréchet, Maurice René (1927). "A review of mathematical functions for the analysis of growth in poultry". In: *Annales de la Société Polonaise de Mathematique* 6, pp. 93–116 (cit. on p. 14).

Galambos, Janos (1978). *The asymptotic theory of extreme order statistics*. Wiley New York (cit. on p. 10).

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin (2014). *Bayesian data analysis*. 2nd. Taylor and Francis (cit. on p. 67).

Geyer, Charles J. (1992). "Practical Markov Chain Monte Carlo". In: *Statistical Science* 7.4, pp. 473–483 (cit. on p. 73).

Girolami, Mark and Ben Calderhead (2011). "Riemann manifold Langevin and Hamiltonian Monte Carlo methods". In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 73.2, pp. 123–214 (cit. on pp. iii, 69–73, 101, 102).

Gnedenko, Boris (1943). "Sur la distribution limite du terme maximum d'une serie aleatoire". In: *Annals of Mathematics* 44.3, pp. 423–453 (cit. on p. 14).

Green, Peter J. (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination". In: *Biometrika* 82.4, pp. 711–732 (cit. on pp. 83, 102).

Hassani, Hossein (2010). "A note on the sum of the first n primes". In: *Quarterly Journal of Mathematics* 61.1, pp. 109–115 (cit. on p. 74).

Hastings, W K (1970). "Monte Carlo sampling methods using Markov chains and their applications". In: *Biometrika* 57.1, pp. 97–109 (cit. on p. 67).

Heffernan, Janet E. and Sidney I. Resnick (2007). "Limit laws for random vectors with an extreme component". In: *Annals of Applied Probability* 17.2, pp. 537–571 (cit. on p. 22).

Heffernan, Janet E. and Jonathan A. Tawn (2004). "A conditional approach to modelling multivariate extreme values". In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 66.3, pp. 497–546 (cit. on pp. i, iii, 2, 3, 5, 6, 21, 23, 24, 33–38, 41, 43, 75, 93).

Hüsler, Jürg and Rolf Dieter Reiss (1989). "Maxima of bivariate random vectors: Between independence and complete dependence". In: *Statistics and Probability Letters* 7, pp. 283–286 (cit. on p. 27).

Jenkinson, A. F. (1955). "The frequency distribution of the annual maximum (or minimum) values of meteorological elements". In: *Journal of the Royal Meteorological Society* 81.348, pp. 158–171 (cit. on p. 14).

Joe, Harry, R.J. Smith, and Ishay Weissman (1992). "Bivariate Threshold Methods for Extremes". In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 54.1, pp. 171–183 (cit. on pp. 27, 4).

Johannessen, Kenneth, Trond Stokka Meling, and Sverre Haver (2002). "Joint Distribution for Wind and Waves in the Northern North Sea". In: *Journal of Offshore and Polar Engineering* 12.1, pp. 1–8 (cit. on p. 1).

Jonathan, Philip, Kevin C. Ewans, and George Z. Forristall (2008). "Statistical estimation of extreme ocean environments: The requirement for modelling

directionality and other covariate effects". In: *Ocean Engineering* 35.11-12, pp. 1211–1225 (cit. on pp. 2, 12).

Jonathan, Philip, Kevin C. Ewans, and David Randell (2014). "Non-stationary conditional extremes of northern North Sea storm characteristics". In: *Environmetrics* 25.3, pp. 172–188 (cit. on pp. iii, 3, 5, 6, 65, 88, 89, 92, 93, 101, 103, 104).

Keef, Caroline, Ioannis Papastathopoulos, and Jonathan A. Tawn (2013). "Estimation of the conditional distribution of a multivariate variable given that one of its components is large: Additional constraints for the Heffernan and Tawn model". In: *Journal of Multivariate Analysis* 115, pp. 396–404 (cit. on pp. iii, 3, 6, 23, 34, 35, 39–41, 51, 52, 54, 62–64, 76, 84, 88, 102, 105, 15, 53).

Kourbatov, Alexei (2014). "The distribution of maximal prime gaps in Cramer's probabilistic model of primes". In: *arXiv.org* 3.2, pp. 18–29 (cit. on p. 18).

Lang, Stefan and Andreas Brezger (2004). *Bayesian P-Splines*. Vol. 13. 1. Institute of Mathematical Statistics, pp. 183–212 (cit. on pp. 5, 90).

Liu, Y. and Jonathan A. Tawn (2014). "Self-consistent estimation of conditional multivariate extreme value distributions". In: *Journal of Multivariate Analysis* 127, pp. 19–35 (cit. on p. 43).

Lugrin, T., Anthony C. Davison, and Jonathan A. Tawn (2016). "Bayesian Uncertainty Management in Temporal Dependence of Extremes". In: *Extremes* 19.3, pp. 491–515 (cit. on pp. 35, 36, 39, 41, 75, 102).

Mejzler, D. G. (1956). "On the problem of the limit distribution for the maximal term of a variational series". In: *Lvov Politechn. Inst. Naucn. Zap. Ser. Fiz.-Mat* 38.1, pp. 90–109 (cit. on p. 17).

Metropolis, Nicholas, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller (1953). "Equation of state calculations by fast computing machines". In: *Journal Chemical Physics* 21.6, pp. 1087–1092 (cit. on p. 67).

Mitra, Abhimanyu and Sidney I. Resnick (2013). "Modeling multiple risks: Hidden domain of attraction". In: *Extremes* 16.4, pp. 507–538 (cit. on p. 24).

Pickands, James (1975). "Statistical Inference using Extreme Order Statistics". In: *The Annals of Statistics* 3.1, pp. 119–131 (cit. on p. 18).

Raghupathi, Laks, David Randell, and Kevin C. Ewans (2016). "Consistent Design Criteria for South China Sea with a Large-Scale Extreme Value Model". In: *Offshore Technology Conference*. Kuala Lumpur (cit. on pp. 2, 12).

Randell, David, Kathryn Turnbull, Kevin C. Ewans, and Philip Jonathan (2016). "Bayesian inference for non-stationary marginal extremes". In: *Environmetrics* 27.1, pp. 439–450 (cit. on p. 1).

Rao, C. Radhakrishna (1945). "Information and Accuracy Attainable in the Estimation of Statistical Parameters". In: *Bulletin of Calcutta Mathematical Society* 37, pp. 81–91 (cit. on p. 71).

Reeds, James A., Jeffrey C. Lagarias, Margaret H. Wright, and Paul E. Wright (1998). "Convergence Properties of the Nelder–Mead Simplex Method in Low Dimensions". In: *SIAM Journal on Optimization* 9.1, pp. 112–147 (cit. on p. 15).

Reiss, Rolf Dieter and Michael Thomas (2007). *Statistical Analysis of Extreme Values*. Birkhäuser (cit. on p. 10).

Rényi, Alfréd (1953). "On the Theory of Order Statistics". In: *Acta Mathematica Hungarica* 4.3-4, pp. 191–231 (cit. on p. 19).

Resnick, Sidney I. (1987). *Extreme Values, Regular Variation and Point Processes*. Springer (cit. on pp. 10, 26).

Rothenberg, Thomas J . (1971). "Identification in Parametric Models". In: *Econometrica* 39.3, pp. 577–591 (cit. on p. 55).

Self, Steven G and Kung-yee Liang (1987). "Asymptotic Properlies of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions". In: *Journal of the Amer* 82.398, pp. 605–610 (cit. on p. 54).

Stephenson, Alec G. (2003). "Simulating Multivariate Extreme Value Distributions of Logistic Type". In: *Extremes* 6.1, pp. 49–59 (cit. on p. 47).

Tawn, Jonathan A. (1988). "Bivariate extreme value theory: Models and estimation". In: *Biometrika* 75.3, pp. 397–415 (cit. on p. 26).

Tawn, Jonathan A. and A. W. Ledford (1996). "Statistics for Near Independence in Multivariate Extreme Values". In: *Biometrika* 83.1, pp. 169–187 (cit. on pp. 30, 31).

Tawn, Jonathan A., Ser-Huang Poon, and Michael Rockinger (2003). "Modelling Extreme-Value Dependence in International Stock Markets". In: *Statistica Sinica* 13.4, pp. 929–953 (cit. on pp. 1, 21).

Von Mises, Richard (1936). "La distribution de la plus grande n valeurs". In: *Review of the Mathematical Union Interbalcanique* 1.1, pp. 141–160 (cit. on pp. 14, 16).

Wadsworth, Jennifer L. and Jonathan A. Tawn (2013). "A new representation for multivariate tail probabilities". In: *Bernoulli* 19.5B, pp. 2689–2714 (cit. on p. 32).

Zheng, Feifei, Seth Westra, Michael Leonard, and Scott A. Sisson (2014). "Modeling dependence between extreme rainfall and storm surge to estimate coastal flooding risk". In: *Water Resources Research* 50.1, pp. 2050–2071 (cit. on p. 22).