



UNIL | Université de Lausanne

Unicentre

CH-1015 Lausanne

<http://serval.unil.ch>

Year : 2020

Statistics of extremes, matrix distributions and applications in non-life insurance modeling

Bladt Martin

Bladt Martin, 2020, Statistics of extremes, matrix distributions and applications in non-life insurance modeling

Originally published at : Thesis, University of Lausanne

Posted at the University of Lausanne Open Archive <http://serval.unil.ch>

Document URN : urn:nbn:ch:serval-BIB_F8A1A37248765

Droits d'auteur

L'Université de Lausanne attire expressément l'attention des utilisateurs sur le fait que tous les documents publiés dans l'Archive SERVAL sont protégés par le droit d'auteur, conformément à la loi fédérale sur le droit d'auteur et les droits voisins (LDA). A ce titre, il est indispensable d'obtenir le consentement préalable de l'auteur et/ou de l'éditeur avant toute utilisation d'une oeuvre ou d'une partie d'une oeuvre ne relevant pas d'une utilisation à des fins personnelles au sens de la LDA (art. 19, al. 1 lettre a). A défaut, tout contrevenant s'expose aux sanctions prévues par cette loi. Nous déclinons toute responsabilité en la matière.

Copyright

The University of Lausanne expressly draws the attention of users to the fact that all documents published in the SERVAL Archive are protected by copyright in accordance with federal law on copyright and similar rights (LDA). Accordingly it is indispensable to obtain prior consent from the author and/or publisher before any use of a work or part of a work for purposes other than personal use within the meaning of LDA (art. 19, para. 1 letter a). Failure to do so will expose offenders to the sanctions laid down by this law. We accept no liability in this respect.



UNIL | Université de Lausanne

FACULTÉ DES HAUTES ÉTUDES COMMERCIALES
DÉPARTEMENT DE SCIENCES ACTUARIELLES

**STATISTICS OF EXTREMES,
MATRIX DISTRIBUTIONS AND APPLICATIONS IN
NON-LIFE INSURANCE MODELING**

THÈSE DE DOCTORAT

présentée à la

Faculté des Hautes Études Commerciales
de l'Université de Lausanne

pour l'obtention du grade de
Docteur ès Sciences actuarielles

par

Martin BLADT

Directeur de thèse
Prof. Hansjörg Albrecher

Jury

Prof. Felicitas Morhart, présidente
Prof. Valérie Chavez-Demoulin, experte interne
Prof. Jan Beirlant, expert externe
Prof. Alexander McNeil, expert externe

LAUSANNE
2020



UNIL | Université de Lausanne

FACULTÉ DES HAUTES ÉTUDES COMMERCIALES

DÉPARTEMENT DE SCIENCES ACTUARIELLES

**STATISTICS OF EXTREMES,
MATRIX DISTRIBUTIONS AND APPLICATIONS IN
NON-LIFE INSURANCE MODELING**

THÈSE DE DOCTORAT

présentée à la

Faculté des Hautes Études Commerciales
de l'Université de Lausanne

pour l'obtention du grade de
Docteur ès Sciences actuarielles

par

Martin BLADT

Directeur de thèse
Prof. Hansjörg Albrecher

Jury

Prof. Felicitas Morhart, présidente
Prof. Valérie Chavez-Demoulin, experte interne
Prof. Jan Beirlant, expert externe
Prof. Alexander McNeil, expert externe

LAUSANNE
2020

IMPRIMATUR

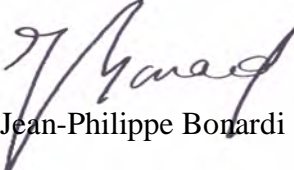
Sans se prononcer sur les opinions de l'auteur, la Faculté des Hautes Etudes Commerciales de l'Université de Lausanne autorise l'impression de la thèse de Monsieur Martin BLADT, titulaire d'un bachelor en Mathématiques de l'Université Nationale Autonome du Mexique et d'un master en Statistique de l'Université de Copenhague, en vue de l'obtention du grade de docteur ès Sciences actuarielles.

La thèse est intitulée :

STATISTICS OF EXTREMES, MATRIX DISTRIBUTIONS AND APPLICATIONS IN NON-LIFE INSURANCE MODELING

Lausanne, le 04 mai 2020

Le doyen


Jean-Philippe Bonardi



Members of the Jury

Prof. **Felicitas Morhart**

President of the jury, University of Lausanne, Department of Marketing.

Prof. **Hansjoerg Albrecher**

Thesis director, University of Lausanne, Department of Actuarial Science.

Prof. **Valérie Chavez-Demoulin**

Internal Expert, University of Lausanne, Department of Operations.

Prof. **Jan Beirlant**

External Expert, Catholic University of Leuven, Department of Mathematics.

Prof. **Alexander J. McNeil**

External Expert, University of York, York Management School.

University of Lausanne
Faculty of Business and Economics

PhD in Actuarial Science

I hereby certify that I have examined the doctoral thesis of

Martin BLADT

and have found it to meet the requirements for a doctoral thesis.

All revisions that I or committee members
made during the doctoral colloquium
have been addressed to my entire satisfaction.

Signature:

A handwritten signature in blue ink, appearing to read 'Hansjörg Albrecher', written over a light blue rectangular background.

Date: April 30, 2020

Prof. Hansjörg ALBRECHER
Thesis supervisor

University of Lausanne
Faculty of Business and Economics

PhD in Actuarial Science

I hereby certify that I have examined the doctoral thesis of

Martin BLADT

and have found it to meet the requirements for a doctoral thesis.

All revisions that I or committee members
made during the doctoral colloquium
have been addressed to my entire satisfaction.

Signature: _____



Date: ____29.04.2020_____

Prof. Valérie CHAVEZ
Internal member of the doctoral committee

University of Lausanne
Faculty of Business and Economics

PhD in Actuarial Science

I hereby certify that I have examined the doctoral thesis of

Martin BLADT

and have found it to meet the requirements for a doctoral thesis.

All revisions that I or committee members
made during the doctoral colloquium
have been addressed to my entire satisfaction.

Signature:



Date:

29/04/2020

Prof. Jan BEIRLANT
External member of the doctoral committee

University of Lausanne
Faculty of Business and Economics

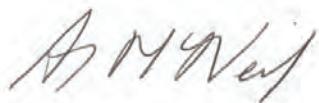
PhD in Actuarial Science

I hereby certify that I have examined the doctoral thesis of

Martin BLADT

and have found it to meet the requirements for a doctoral thesis.

All revisions that I or committee members
made during the doctoral colloquium
have been addressed to my entire satisfaction.



Signature:

Date: 30th April 2020

Prof. Alexander MCNEIL
External member of the doctoral committee

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my doctoral supervisor Hansjörg Albrecher, for always being there during the entire process. He has taught me a great deal during these past couple of years, both on a professional and personal level.

I would also like to thank the co-authors of the research papers which constitute the present work as well as of research items developed during the course of my PhD studies but omitted in the thesis. In alphabetical order, they are: Nicole Bäuerle, Jan Beirlant, Mogens Bladt, Dalit Daily-Amir, Dominik Kortschak, Alexander J. McNeil, Franz Prettenthaler, Tina Swierczynski and Eleni Vatamidou. Their support and insight has been very valuable.

Special thanks to all the members of the jury for taking the time to read earlier versions of the manuscript, and for their interesting comments during the private defense.

On a more personal note, I would like to thank Pernille and Sia, for being my constant source of support and motivation, and to both of our families for their kindness and generosity. Finally, I am grateful for the support that Guille, Mogens and Thomas have selflessly and constantly provided throughout my entire life.

Contents

1	General Introduction	21
1.1	Events underlying non-life insurance	21
1.2	Collective risk theory	22
1.2.1	Exact solutions	22
1.2.2	Simulation	24
1.3	Severity modeling	24
1.3.1	Extreme Value Theory	25
1.3.2	Matrix distributions	28
2	Flood occurrence change-point analysis	31
	Abstract	31
2.1	Introduction	32
2.2	Study Area	34
2.3	Data and Methods	35
2.3.1	Flood data from natural geochronology	35
2.3.2	Statistical analysis	36
2.3.3	Change-points and dependence	38
2.3.4	Fitting a renewal model	41
2.4	Results and Discussion	42
2.4.1	Testing for a Poisson process	42
2.4.2	Change-points and dependence	43
2.4.3	Fitting a renewal model	46
2.4.4	Comparison to an analysis with lower resolution data	48
2.4.5	The waiting time until the next flood	50
2.5	Conclusion and Outlook	52
3	Dividends: From Refracting to Ratcheting	55
	Abstract	55
3.1	Introduction	55
3.2	The Spectrally-Negative Lévy Risk Model	57
3.2.1	Refracting Strategy	58
3.2.2	Ratchet Strategy	59
3.3	Brownian approximation	62
3.3.1	Refracting Strategy	62
3.3.2	Ratchet Strategy	64
3.4	The Cramér-Lundberg model with hyper-exponential claims	70

3.5	The expected time to ruin	73
3.6	Conclusion and Future Research	77
4	Efficient simulation of ruin probabilities	79
	Abstract	79
4.1	Introduction	79
4.2	Model description and preliminaries	81
4.3	Control variate techniques	83
	4.3.1 Max of heavy-tails	84
	4.3.2 Conditional Monte Carlo	88
	4.3.3 Comparison with the Pollaczek-Khinchine expansion	89
4.4	Numerical experiments	91
	4.4.1 Mixture of exponential and Pareto claim sizes	91
	4.4.2 Parameters	91
	4.4.3 Results	92
4.5	Conclusion	94
5	Combined Tail Estimation	97
	Abstract	97
5.1	Introduction	98
5.2	Derivation and properties	100
5.3	Penalization seen as a Bayesian prior	104
5.4	Extreme Value Theory	105
5.5	Simulation Study and MTPL Insurance	111
	5.5.1 Simulation Study	112
	5.5.2 Insurance Data	113
5.6	Conclusion	121
	Appendix	122
6	Threshold selection and trimming in extremes	125
	Abstract	125
6.1	Introduction	125
6.2	A lower-trimmed Hill statistic	128
	6.2.1 Derivation	128
	6.2.2 A lower-trimmed Hill plot	130
6.3	Regularly varying tails	131
	6.3.1 Distribution of the average	133
	6.3.2 Optimal k in the Hall class	137
	6.3.3 Interpretation of \bar{T}_k as a weighted Hill estimator	138
6.4	A ratio statistic	139
6.5	Simulations	140
6.6	Insurance data	143
6.7	Conclusion	150
6.8	Proofs	153

7	Trimmed EV estimators for censored data	159
	Abstract	159
7.1	Introduction	159
7.2	Trimmed estimators for ξ	162
	7.2.1 Trimming tail estimators	162
	7.2.2 Averaging and kernels	163
7.3	Asymptotic representations	164
7.4	Optimal choice of k when estimating ξ	166
7.5	Simulations	167
7.6	Insurance Application: censored claims data vs ultimates	167
7.7	Conclusion	168
8	Novelty detection	175
	Abstract	175
8.1	Introduction	175
8.2	Problem Formulation	178
8.3	Extreme Value Theory	178
	8.3.1 Domains of attraction and GPD	179
	8.3.2 Point Processes	181
8.4	Novelty detection for randomly censored data	182
	8.4.1 Point Process of Exceedances	183
	8.4.2 Law of the likelihood	186
	8.4.3 The Bonferroni correction	191
	8.4.4 A special class of regularly varying distributions	193
8.5	Performance	193
8.6	Conclusion	195
9	Matrix Mittag–Leffler distributions	197
	Abstract	197
9.1	Introduction	197
9.2	Some relevant background	200
	9.2.1 Mittag–Leffler functions	200
	9.2.2 Mittag–Leffler distributions	201
	9.2.3 Phase–type distributions	202
9.3	Matrix Mittag–Leffler distributions	203
9.4	Sample path representations	210
	9.4.1 Random time-inhomogeneous phase–type distributions	210
	9.4.2 Semi–Markov framework	212
9.5	Statistical modeling using MML distributions	216
9.6	Conclusion	220
10	Multivariate Matrix ML distributions	221
	Abstract	221
10.1	Introduction	221
10.2	Phase–type distributions	223
	10.2.1 Notation	223
	10.2.2 Univariate phase–type distributions	223

10.2.3	Multivariate phase-type distributions	224
10.2.4	Matrix Mittag-Leffler distributions	228
10.3	Generalized matrix Mittag-Leffler distributions	229
10.4	The multivariate matrix Mittag-Leffler distribution	231
10.5	Special structures and examples	235
10.6	Conclusion	242
11	Multivariate fractional PH distributions	245
	Abstract	245
11.1	Introduction	245
11.2	Background	248
11.2.1	Phase-type distributions (PH)	248
11.2.2	Multivariate phase-type distributions (MPH*)	249
11.2.3	Univariate fractional phase-type distributions (PH _α)	250
11.3	Multivariate fractional phase-type distributions	251
11.3.1	The construction	251
11.3.2	Denseness properties of the MPH* _α class and an extension	255
11.3.3	A product representation	256
11.3.4	Distribution of projections	256
11.4	Two specific examples	258
11.4.1	The feed-forward case	258
11.4.2	A two-dimensional explicit example with tail dependence	259
11.5	Conclusion	261
	References	263

Chapter 1

General Introduction

Insurance is the business dedicated to indemnifying random financial losses in exchange of charging premiums. One can also regard it as an attempt to tame the uncertainty which underlies many aspects of human existence. These risks can be of very different nature, but the classical division is made between life insurance and non-life insurance. The present work is dedicated to various aspects of the latter, mainly through the specification of new probabilistic models. These models not only give insight into the nature of the possible drivers behind the uncertain events in insurance, but often lead to a statistical approach which can be applied in practice. The main themes throughout are the analysis of flood data, the mathematical modeling of the surplus and dividends of an insurance company, and the statistical modeling of claim severities. While the challenges addressed in this thesis are motivated by those insurance applications, the developed methods may also be interesting in their own right, and for other application areas. In the following sections of this chapter, we will expand further on the topics covered in the remainder of the thesis.

1.1 Events underlying non-life insurance

Nowadays a wide variety of insurance products exist, and what all the lines of business have in common is that they cover losses incurred by random phenomena. It is clear that a deeper understanding of the driving mechanisms would be ideal. Apart from the fundamental modern physics theories, which assume that there is an intrinsic randomness, which is inherent at the sub-atomic level, there exists the idea that if we just had enough information, many things that appear to be random would turn out to actually be structured. From that perspective, insurance companies design products that make us feel in control of our own ignorance. But not even the insurers can know all the processes behind random events, so two techniques are adopted in an attempt to understand the nature of the insured risks. The first one is proposing a model based on a physical or logical belief, and the second is recollecting data from the past, and extrapolating by making the assumption that the future will behave in a similar way. Good models often combine the two approaches in a sound way. One thing is at least intuitively clear: a better understanding of the insured risk will result in a better way of managing

the incurred risks by incorporating this information into a statistical model (see also Prettenthaler et al. (2017)). Let us now proceed to a very important line of business, related to natural disasters and to the subject of Chapter 2 of this thesis.

One of the most common and widespread natural disasters in the world is flooding. Thousands of people are killed every year worldwide due to floods, often placing it as the most deadly natural events, closely followed by storms and landslides (cf. Wallemacq (2018)). The number of people reached by floods is also the highest among natural disasters, and the effects are being exacerbated as climate change intensifies (cf. Field and Van Aalst (2014); Jongman et al. (2014)). It is clear that these changes have led insurers in this line of business to put more effort into learning from past flood records, but a key drawback is that catastrophic events do not happen often, so the available data is typically scarce (cf. Jones et al. (2012); Blöschl et al. (2017); Merz et al. (2018); Schmocker-Fackel and Naef (2010); Swierczynski et al. (2017) for flood series records), and the loss estimates can fluctuate a lot (cf. Prettenthaler et al. (2015)). New technologies in paleoflood hydrology has permitted the analysis of longer and more detailed time series of floods (cf. Wilhelm et al. (2018b); Arnaud et al. (2012); Swierczynski et al. (2012); Czymzik et al. (2013); Swierczynski et al. (2013); Wirth et al. (2013); Sabatier et al. (2017)).

The statistical tools that have been employed on flood records focus on trends (cf. Merz and Blöschl (2003); Petrow and Merz (2009); Mudelsee et al. (2003)) and clustering (Merz et al. (2016)) with the aim of detecting abrupt changes in flood occurrence, but have not been applied to the newer, longer records. For surveys of statistical methods for time series of climate data, we refer to Mudelsee (2014); Wilhelm et al. (2018a). In Chapter 2 we derive a new approach for modeling trends and change-points and proceed to analyze the flood frequency record from Lake Mondsee sediments dating back 7100 years (cf. Swierczynski et al. (2013)), which is of high resolution and more reliable than previous archives.

1.2 Collective risk theory

Collective risk theory deals with a mathematical representation of an the surplus process of an insurance portfolio over time, and aims to solve certain optimality problems based on model parameters. The typically used models oversimplify the behaviour of insurers, often making strong assumptions in order to obtain explicit solutions. However, these models do not aim to model real-life insurance companies' behaviour accurately, but instead serve the purpose of hopefully gaining insight into problems which are too complicated to solve mathematically, by solving related but simpler problems.

1.2.1 Exact solutions

Since the introduction of the compound Poisson process by Filip Lundberg over a century ago, a large amount of models in risk theory have been proposed. Modern risk theory now often uses a class of stochastic processes $X = \{X_t\}_{t \geq 0}$, called Lévy risk processes (see Kyprianou (2014)), typically with the property that jumps may

only be negative. The initial capital is denoted by $X_0 = x \geq 0$, and the drift of the process is assumed to be positive.

For a fixed $b \geq 0$ the first passage times

$$\tau_0^- := \inf\{t \geq 0 : X_t < 0\}, \quad \tau_b^+ := \inf\{t \geq 0 : X_t > b\},$$

have the property that (cf. (Kyprianou, 2014, Ch.8))

$$\mathbb{E}_x \left(e^{-\delta \tau_b^+}; \tau_b^+ < \tau_0^- \right) = \frac{W(x)}{W(b)}$$

and

$$\mathbb{E}_x \left(e^{-\delta \tau_0^-}; \tau_0^- < \tau_b^+ \right) = Z(x) - Z(b) \frac{W(x)}{W(b)},$$

for any $0 \leq x \leq b$, where W is a scale function, defined as the inverse Laplace transform of the reciprocal of the (shifted) Laplace exponent

$$\psi(\theta) := \log \mathbb{E} e^{\theta X_1},$$

and Z is the second scale function, which is a scaled version of the integral of W . The above expressions are very useful when dealing with fluctuations of Levy processes, and as a special case one can obtain the ruin probability, $\mathbb{P}_x(\tau_0^- < \infty)$, in terms of the scale functions.

Since de Finetti's work (cf. de Finetti (1957)), it has also been of interest to consider dividend payout strategies in collective risk theory, and more precisely, the specification of the optimal strategy to pay dividends over the life-time of the insurance portfolio (see Gerber (1969)). In Kyprianou and Loeffen (2010) the expected sum of discounted dividends until ruin under a refraction strategy is derived explicitly, that is

$$\mathbb{E}_x \left[\int_0^\tau e^{-\delta s} c 1_{\{U_s > b\}} ds \right]$$

is given in terms of the scale functions, where $\tau = \inf\{t > 0 : U_t < 0\}$, and

$$U_t = X_t - c \int_0^t 1_{\{U_s > b\}} ds, \quad t \geq 0.$$

The interpretation is that no dividends are paid while X is below the threshold b , and dividends are paid at rate c above the threshold. The importance of this result is historical, since it arises as the optimal dividend payout strategy for diffusions (Jeanblanc-Picqué and Shiryaev (1995); Asmussen and Taksar (1997)) and some compound Poisson processes (Gerber and Shiu (2006a)) when the dividend rate is bounded. In the unbounded case, it is known that band strategies are optimal, and in some cases they collapse to a barrier strategy (cf. Gerber (1969); Shreve et al. (1984); Schmidli (2008); Azcue and Muler (2014), see also Avanzi (2009); Albrecher and Thonhauser (2009) for surveys on optimality for dividend strategies).

In Chapter 3 we consider a strategy where a barrier is set, but once it is crossed and the dividend rate is increased, it cannot be decreased again. The expected discounted dividend payments until ruin are considered, an optimality criterion for the barrier of this model is obtained, and the solution is shown to be close to the classical optimal solution for some specific scale functions. This addresses the concern of some shareholders who prefer to never have a decrease in the dividend rate, even if such strategy might not be optimal.

1.2.2 Simulation

When an exact solution is not available, an approach which has gained strength with increased computer capabilities is simulation. The numerical evaluation of ruin probabilities and their accuracy using pseudo-random numbers is a branch of so-called *stochastic simulation* methods (cf. Asmussen and Glynn (2007) for a comprehensive overview of the field). A particular advantage of simulation for stochastic models is that one can often allow for more complex models in the setup of the problem. One such model will be presented in Chapter 4, where the exact solution of the processes considered is not available, but a correlated process with explicit ruin probability can be efficiently used to decrease the variance of the simulation procedure. In a general setting, Asmussen and Kroese (2006); Hartinger and Kortschak (2009); Chan and Kroese (2011); Asmussen and Kortschak (2012); Ghamami and Ross (2012); Asmussen and Kortschak (2015) give results for error rates and efficient rare event simulation. We will use some of the methodology (in particular, the Asmussen-Kroese estimator) and apply it to a case where claims are mixtures of heavy- and light-tailed distributions (a setting that was previously investigated in Vatamidou et al. (2013)).

1.3 Severity modeling

From an applied point of view, premium calculation often relies on the law of large numbers or quantile estimation, in the sense that the expected loss (plus some safety loading) or some quantile (Value-at-Risk) is often used as an ingredient in the associated risk analysis and management. Models which take into account the frequency and severity of claims at the same time do exist, but in the vast majority of occasions the two are modeled separately. While the net premium is often determined simply as the product of the average claim size and the average rate of occurrence, for a more sophisticated analysis an accurate estimation of the entire distribution of claim sizes is crucial. For instance, if we deal with heavy-tailed risks, the mean might not even exist, or a slight change in the tail parameter can have large effects on the VaR. In this section we will elaborate more on some useful probabilistic and statistical tools which can and will be applied to severity modeling.

1.3.1 Extreme Value Theory

Often, the largest observation in a claim severity dataset is significantly larger than any other data point. Loosely speaking, this is what we refer to as having a heavy-tail, or in other words, the probability of observing very large claims is not negligible. Extreme value theory (cf. Embrechts et al. (1997); Beirlant et al. (2004); de Haan and Ferreira (2007) for classical texts on the subject and Gomes and Guillou (2015) for an overview of the univariate case) is concerned with the analysis of rare but large events. When modeling severity, it is of crucial importance to analyze the tail of the distribution precisely. The mean value and quantiles are highly sensitive to this characteristic, and hence the premium calculation, as well as any other properties that one derives based on the inferred model.

Consider an independent and identically distributed (i.i.d.) sequence of random variables X_1, X_2, \dots , such that the the following relation holds

$$c_n^{-1}(M_n - d_n) \stackrel{d}{=} X_1$$

for some norming constants c_n, d_n and where $M_n = \max\{X_1, \dots, X_n\}$. If this is satisfied, we say that they follow a max-stable distribution. The concept of max-stability is the backbone of extreme-value theory, since it can be related with the limiting distributions of normalized maxima. By the convergence to types theorem (cf. e.g. Embrechts et al. (1997)), it can be shown that a distribution is the non-degenerate limit of normalized maxima, as the sample size grows, if and only if it is max-stable. This means that if we desire to investigate how the maximum of a large sample behaves, we can use the asymptotic theory to do inference. In this regard, the following result (cf. Fisher and Tippett (1928)) is very useful, since it guarantees a fully explicit way of writing the associated limiting distribution:

Theorem 1.3.1. *For X_i an i.i.d. sequence of random variables, if there exist non-degenerate H and norming constants such that $c_n^{-1}(M_n - d_n) \xrightarrow{d} H$, then necessarily H is one of the following:*

Frechet: $\Phi_\alpha(x) = e^{-x^{-\alpha}} \mathbf{1}_{(0, \infty)}(x),$

Weibull: $\Psi_\alpha(x) = e^{-(-x)^\alpha} \mathbf{1}_{(-\infty, 0]}(x) + \mathbf{1}_{(0, \infty)}(x)$

Gumbel: $\Lambda(x) = e^{-e^{-x}}.$

Of course, any given dataset will not fit exactly one of these three limit distributions, but we can say that a distribution is for instance in the Frechet maximum domain of attraction ($\text{MDA}(\Phi_\alpha)$) if their maximum asymptotically and correctly normalized converges to the Frechet distribution. The concept extends in a similar way to the other two distributions. There is a way of characterizing whether a distribution belongs to a certain domain of attraction. For us to understand the result, we need to define two useful concepts. Say that a distribution function is regularly varying with index α ($F \in \mathcal{R}_{-\alpha}$) if

$$\overline{F}(x) = x^{-\alpha} \ell(x), \tag{1.1}$$

where $\lim_{x \rightarrow \infty} \ell(tx)/\ell(x) = 1$, for all $t > 0$. The mean excess function is defined as

$$s_X(u) = \mathbb{E}(X - u | X > u) = \frac{\int_u^\infty \bar{F}(y) dy}{\bar{F}(u)}.$$

Theorem 1.3.2. • $F \in \text{MDA}(\Phi_\alpha) \Leftrightarrow \bar{F} \in \mathcal{R}_{-\alpha}$. The norming constants can be taken as $d_n = 0$ and $c_n = F^{\leftarrow}(1 - n^{-1})$.

• $F \in \text{MDA}(\Psi_\alpha) \Leftrightarrow x_F < \infty$ and $\bar{F}(x_F - x^{-1}) \in \mathcal{R}_{-\alpha}$. The norming constants can be taken as $d_n = x_F$ and $c_n = x_F - F^{\leftarrow}(1 - n^{-1})$.

• $F \in \text{MDA}(\Lambda) \Leftrightarrow \bar{F}(x) = c(x) \exp\left(-\int_z^x \frac{g(t)}{a(t)} dt\right)$, with $g \rightarrow 1$, $c \rightarrow c_0 > 0$, $a' \rightarrow 0$. The norming constants can be taken as $d_n = F^{\leftarrow}(1 - n^{-1})$ and $c_n = a(d_n)$. One can choose a to be the mean excess function $s_X(u)$.

We refer to Embrechts et al. (1997) for further details. For our purposes, we are mostly interested in the $F \in \text{MDA}(\Phi_\alpha)$ case, and consequently in the estimation of the tail parameter α of the class $F \in \mathcal{R}_{-\alpha}$.

In the exact Pareto case, the function ℓ in (1.1) is identically equal to one. In other cases, however, although slowly varying, the function ℓ might incur a large bias on estimators constructed for the Pareto distribution. Consequently, a central question in univariate extreme value theory is the determination of the threshold above which one can sensibly regard the effect of the function ℓ as negligible, with respect to a pure power law tail. In Chapter 6 of this thesis we tackle this classical problem by modifying the Hill estimator (the mean of the log-observations above a certain threshold) through the deletion of low importance observations and compensating deterministically.

When confining ourselves to the Hall subclass (cf. Hall (1982); Hall et al. (1985)), or in other words, making second order assumptions on the tail of the distribution, many explicit calculations are available in terms of the given parametrization. Hence, it is a popular assumption in the literature when deriving asymptotic results, and most of the frequently used heavy-tailed distributions in the Frechet domain of attraction belong to this class. Formally it is defined by its tail being of the form

$$\mathbb{P}(X > u) = Cu^{-\alpha} (1 + Du^{-\nu_1}(1 + o(1))), \text{ for } u \rightarrow \infty,$$

for some constants C, D . Estimation of the constants C, D and the index ν_1 is a difficult task, which is necessary but not the primary focus of the analysis, which is rather concerned with the index α . The reason for this distinction is that the truly large claims will have a behaviour dictated primarily by α . In Chapter 8 we will consider a related but slightly different problem. Given a dataset, we are interested in the estimation of α solely for the purpose of *testing* whether it can be regarded to be the same or different from the one of a reference heavy-tailed distribution. The latter problem is usually denoted novelty detection (see Pimentel et al. (2014); Markou and Singh (2003a,b) for reviews) and the use of extreme value theory in this area is recent (cf. Clifton et al. (2011, 2013, 2014); Luca et al. (2014, 2016, 2018)). Novelty detection lies in the intersection of extreme value theory and signal processing (in engineering), and we will further consider censored data as well, which we now proceed to introduce.

Censoring

Censoring is a statistical inconvenience that has for long been observed in real-life applications. One of the most common situations where one encounters censoring is when investigating whether a new drug is effective or not for terminally ill patients. A study has to be made involving two groups: one using the drug and one not using it, but one cannot wait until all patients die in order to take their average death age by group. After all, the drug might be effective and one might end up waiting for recovered patients to live the rest of their healthy lives. Instead, during the past century, methods for accounting for the fact that some observations are censored were included within the statistical analysis. The most common censoring mechanism in research (although not necessarily in real-life) is right-censoring completely at random. This is written mathematically as

$$Z = \min\{X, Y\}, \quad e = 1\{Z = X\},$$

where X is the variable of interest, Y is a censoring mechanism, independent of X , and Z and e is the actual observed data. That is, only the censored variable Z and the censoring indicator e can be used for estimating X . One of the most cited papers of all history, Kaplan and Meier (1958), estimates the tail of X from the pair (Z, e) .

For insurance, censoring takes a different interpretation. While before X was time (often a lifetime, hence the name *survival analysis* for these methods; see Bogaerts et al. (2018) for censoring in survival analysis), it can also be considered to be a monetary unit. If we consider a third-party liability insurer, often claims which are very large take many years (e.g. due to open legal cases) to come to a close. However, actuaries need to estimate the distribution of the claim sizes in any given year. Thus, observation of (Z, e) is commonplace.

Methods such as the classical chain-ladder (cf. Mack et al. (1994) for the stochastic model behind this method, and Wüthrich and Merz (2008) for a general overview) try to account for the rest of the development of an open claim, and make an estimate of the total claim size at closure. On the other hand, the statistical methodology from survival analysis can be applied directly, giving different results. The estimates of the tail index of a heavy-tailed claim can be wildly different depending on which method one uses. In Chapter 5, a compromise between the two approaches is made, giving rise to a hybrid estimator, which can be used in practice.

Estimation in the context of extreme value theory under random right-censoring has been considered in Beirlant et al. (2007) using a modified version of the Hill estimator (see also Einmahl et al. (2008) for a classical overview of censoring in extreme value theory), in Worms and Worms (2014) using a different approach, based on the Kaplan-Meier estimator, and in Ameraoui et al. (2016) from a Bayesian perspective. In Beirlant et al. (2018), the bias reduction version of such estimators within the Hall class is considered. In Chapter 7 of this thesis we adapt the trimming methodology from the non-censored case, which leads to family of kernel estimators for the extreme value index of randomly right-censored observations.

1.3.2 Matrix distributions

As remarked earlier, the estimation of the tail of the distribution is a central problem in insurance, and is also a main reason why extreme value theory methods are so useful in this connection (cf. Embrechts et al. (1997)). However, many of the extreme value distributions fail to capture the true behaviour of real life data, especially at lower quantiles. One way to circumvent this problem has been to select a threshold (Wan and Davis (2019); Bladt et al. (2019)) and then splice the distribution above such a value (cf. Albrecher et al. (2017); Pigeon and Denuit (2011)). Suitable models for the body of the distribution are then application-dependent. On the other hand, within the established area of matrix analytic methods in applied probability, it has recently been proposed to consider a global model with no threshold selection by transforming a dense class of light-tailed distributions into the heavy-tailed domain (cf. Albrecher and Bladt (2019), see also Bladt and Rojas-Nandayapa (2018) for another approach). The underlying light-tailed distributions are the building blocks of the construction, and are referred to as phase-type distributions. In the later chapters of this thesis, we will consider a modified version of such building blocks, such that they are heavy-tailed to begin with, and a transformation is possibly only needed in order to fine-tune the model.

In the univariate case (univariate risk modelling), a phase-type distribution (PH) is defined as the time to absorption of a fine-state Markov jump process, with one absorbing state, and all other states being transient (cf. Bladt and Nielsen (2017) for a recent comprehensive treatment on PH and matrix analytic methods). In Chapter 9, we take an analytic approach to matrix distributions in that we modify the Laplace transform of a PH distribution in a way which results in the law of a stable-mixed version of a (power of a) PH random variable. The resulting distribution is peculiar in that it can additionally be seen as an absorption time of a renewal process (generally non-Markovian). It is also heavy-tailed (a property which it inherits from the stable distribution), and thus for an arbitrary regularly varying index, it suffices to consider power transforms. Another property which is inherited by the power transforms from PH distributions is denseness. In a nutshell, the resulting family (the Matrix Mittag-Leffler (MML) family, based on the Mittag-Leffler function, cf. Mittag-Leffler (1904)) will be tractable, heavy-tailed, and dense.

Concerning the multivariate counterpart of PH distributions, the most commonly considered family has been the MPH* class, which consists of a reward-collection scheme in which a random vector has entries which collect state-dependent rewards at different linear rates until absorption. If all rates are equal, the resulting vector has identical entries. For a Markov jump-process generated as the augmentation of two other Markov jump-processes, if rewards of a bivariate random vector collect rewards according to each of the two sub-processes of the enlarged Markov jump-process, then the entries are independent. In fact, any possible dependence structure can be approximated as accurately as required, since the MPH* is dense on the set of distributions of random vectors with positive coordinates. In Chapter 10, we again take an analytic approach to generalizing the MPH* class into a tractable (only for some special cases, with feed-forward structure, analogously to the MPH* case), heavy-tailed and dense class. The idea, however, is not equivalent to using

the path representation of MML distributions and applying a reward system directly (this different line is explored in Chapter 11, where we introduce fractional phase-type distributions). It is, instead, equivalent to plugging a MPH* vector instead of a PH variable into the mixing decomposition of a MML variable. Both in the univariate and multivariate cases, the theory of matrix Mittag-Leffler distributions and fractional phase-type distributions draws considerable insight from fractional calculus methods (see for instance Kozubowski (2001); Garrappa and Popolizio (2018) for details as to how the Mittag-Leffler function arises in solutions to fractional differential equations).

All subsequent chapters of this thesis have either been published or submitted for publication in peer-reviewed journals. Respective details will be given at the beginning of each chapter.

Chapter 2

Flood occurrence change-point analysis in the paleoflood record from lake Mondsee (NE Alps)

This chapter is based on the following article:

Albrecher, H., Bladt, M., Kortschak, D., Prettenthaler, F., & Swierczynski, T. (2019). Flood occurrence change-point analysis in the paleoflood record from Lake Mondsee (NE Alps). *Global and Planetary Change*, 178, 65-76.

Abstract

Knowledge about changes of flood occurrence patterns is important for risk estimation of the future. Robust and well-calibrated paleoflood records, derived e.g. from lake sediments, are excellent natural archives to investigate flood variability of the past and to use the data for further modelling. In this paper, we analyse a 7100 year summer flood record recovered from Lake Mondsee (NE Alps), using a statistical approach. We identify a point process of renewal type, with a significant change-point of the occurrence pattern around 350 AD, switching from the overlay of two mechanisms of event recurrences of 5 and 50 years before to 2 and 17 years after this change-point. This change-point approach enables a comparison to other flood records, and possibly to relate event frequencies to climatic conditions. We also highlight how lower temporal resolution of flood records can hamper the analysis of relations to climatic signals. Hence high-resolution records with robust chronologies and flood information (e.g. seasonality and event characteristics) are essential to improve the understanding of the interplay between climatic signals and flood occurrences, which is an important ingredient for proper risk estimation and risk management.

2.1 Introduction

Floods rank among the most wide-reaching and commonly occurring natural hazards worldwide. According to Wallemacq (2018), flood events, with 3 331 of killed people in 2017, accounted for the largest number of deaths due to natural risks globally in this year, followed by storms (2 510) and landslides (2 312). Of all disaster types, with 55 million people affected, floods were also the leading cause of suffering in 2017, if measured in these terms, followed by storms (25 million) and droughts (10 million). The longer term picture for the preceding decade 2007-2016 was similar: 85 million people affected by floods, 73 million by droughts and 33 million by storms. In addition, also climate change is expected to intensify the impacts of flooding (Field and Van Aalst (2014)). Hence, whilst representing a major issue already today, managing the risk of flooding is expected to become an even more important topic in the future.

The economic burden of all this poses a significant challenge to societies all over the globe, calling for a more efficient flood risk management. For Europe alone, e.g. Jongman et al. (2014) fear that observed extreme flood losses could more than double in frequency by 2050 under future climate change and socio-economic development. This requires a combination of risk reduction, risk retention and risk transfer. As shown in Prettenthaler et al. (2017) for all these tasks, a good local and regional quantification of flood risk is a key requisite for reducing the risk, also e.g. by international cooperation in insuring the risks. A good quantification of local flood risks is hindered by rather short damage records usually, though. Prettenthaler et al. (2015) showed that the range in loss estimates can be large, depending on modelling scale. This is one reason why in general there is some interest from insurance and actuarial science in putting more effort into learning from data sources that mirror flood processes over time periods longer than the archives of insurance companies. The second reason was mentioned already, it is the threat posed by climate change. Since the climate is changing, one of the most common assumptions in local flood risk analysis, namely that flood processes are stationary, might just not be appropriate at larger time scales. But only very long records potentially contain enough information to discern whether a number of events can be explained by one stationary process with statistical significance or whether the underlying regime is indeed changing. Correspondingly, a look into the far past can potentially help to better understand the future.

In Europe, instrumental flood series usually cover 30-50 years (Blöschl et al. (2017)). Records from larger rivers (e.g. Danube, Rhine) sometimes present longer series with 70-150 years (Merz et al. (2018)). Pre-instrumental data from historical archives (Schmocker-Fackel and Naef (2010)) or natural geoarchives, e.g. lake sediments (Swierczynski et al. (2017)) and fluvial sediments (Jones et al. (2012)) are of great interest to study floods of the past. In contrast to sub-daily resolution of instrumental flood data obtained from gauging stations, flood records from natural geoarchives provide long data series for pre-instrumental periods reaching back 10'000 years. However, these data have a lower temporal resolution reaching from

seasonal, annual to decadal timescale and present distinguished flood information (e.g. flood type, magnitude, frequency). During the last decade, large advances in paleoflood hydrology (Wilhelm et al. (2018b)) resulted in a higher number of paleorecords with datasets of improved temporal resolution, which enables great potential for risk research to explore the natural variability of paleofloods, sometimes with seasonal resolution. Numerous studies have been undertaken in the Alpine region (Arnaud et al. (2012); Swierczynski et al. (2012); Czymzik et al. (2013); Swierczynski et al. (2013); Wirth et al. (2013); and Sabatier et al. (2017)). When compared to climate information, many datasets suggest that recorded floods exhibit a pronounced sensitivity to changes of climatic conditions. Most records exhibit higher flood frequencies during cooling episodes, e.g. during the Little Ice Age (ca. 1300-1900 AD). However, the linkage of climate change and flood occurrence is not straightforward, as changes in flood frequencies show temporal as well as regional differences, and uncertainties in chronologies and the nature of recorded floods in sediments have to be considered (Swierczynski et al. (2017)).

Statistical properties of flood records help to better understand flood patterns and to further investigate abrupt changes, long-term trends, natural variability, flood episodes etc. For instrumental flood series, flood clustering (Merz et al. (2016)) as well as trend analysis of floods have been applied (Merz and Blöschl (2003); Petrow and Merz (2009)). In some cases, historical flood series have been used to detect trends as well (Mudelsee et al. (2003)). Mudelsee (2014) provides a nice survey for classical statistical techniques to deal with time series analysis of climate-related data, see also the recent survey by Wilhelm et al. (2018a). Statistical properties, such as inhomogeneous intensities of data series have been investigated by Merz et al. (2016). That study analysed flood records over the past 80 years across different locations in Germany and identified significant deviations from homogeneous flood occurrence patterns. However, long flood series have not been analyzed so far.

Lake Mondsee sediments exhibit a summer flood record (April-August) for the last 7100 years (Swierczynski et al. (2013)). The flood record from Lake Mondsee sediments is based on a robust chronology of annual resolution and include intercalated summer flood layers which are triggered by extreme precipitation events (Kämpf et al. (2014)). The sediment record from Lake Mondsee presents an excellent study site to investigate statistical properties of the long paleoflood record in detail. The focus of this study is on frequency, since the reconstruction of the magnitudes of flood events is limited from this study site (Kämpf et al. (2014)). Previous studies (e.g. Frances et al. (1994); Payrastre et al. (2011)) used historical flood information for flood frequency analysis. These authors argued that the usefulness of integrating historical flood information archives for flood frequency analysis strongly depends on characteristics of the time series, such as length of time series, return periods, magnitude of historical flood and threshold level of perception (Frances et al. (1994)). While historical floods reflect extreme floods of highest, often unknown magnitudes and low recurrence rates (>100 years), the paleoflood record from Lake Mondsee sediments reflects summer floods with higher flood re-

currence (Swierczynski et al. (2013)). Based on sediment calibration with 30 years of discharge data, the paleoflood record reflects summer floods above $Q = 80 \text{ m}^3/\text{s}$ (Kämpf et al. (2014)). Furthermore, Lake Mondsee presents a long high resolution summer flood series of the 7100 years with tight uncertainty bands (Swierczynski et al. (2013)) favoring robust statistical analysis.

Having such a long reliable and detailed track record of flood occurrences available, our goal in this paper is hence to investigate the marginal time series of floods by statistical techniques, and to study to what extent a simple self-contained parsimonious model can already capture the observed flood dynamics of Lake Mondsee over the large time span of the series in a reasonable way. We would like to emphasize that this study is done on a marginal level, i.e. the findings are based by solely looking at the sediment flood record stand-alone, and not relating the flood activity to other climate indicators, which will be the subject of a future study. In any case, the simplicity of the obtained models in the present study looks striking in its own right, and is meant to serve as a contribution to further more detailed discussions of the topic.

The paper is organized as follows. Section 2.2 describes the study area and Section 2.3.1 the available data. Sections 2.3.2–2.3.4 then introduce and discuss the statistical techniques used for the analysis in the sequel. In particular, we present a novel regression approach based on generalized linear models that models trends and change-points at the same time and then enables to give statements about the significance of either of these in the presence of the other. Section 2.4 then performs a statistical analysis for the occurrence of flood events for the Lake Mondsee sediment record.

Section 2.5 concludes. This study is a cooperation of an interdisciplinary team bringing disciplines and competences together from Geosciences, Mathematics, Statistics and Socioeconomics to explore the potential of long flood series as recorded in lake sediments.

2.2 Study Area

Lake Mondsee is a pre-alpine lake in Upper Austria/NE Alps ($47^\circ 49' \text{N}$, $13^\circ 24' \text{E}$). The lake is located at 481 m asl and has a surface of 14.2 km^2 (Fig. 2.1). The lake catchment of ca. 241 km^2 is characterised by siliciclastic Flysch sediments in the northern part with maximum elevation up to 1100 m asl (ca. 75%) and Triassic Main Dolomite and Mesozoic limestones of the Northern Calcareous Alps (ca. 25%) in the southern part with maximum elevation of 1783 m asl (van Husen, 1989). Three main rivers mainly drain the northern part leading to siliciclastic sediment input during flood events. Smaller rivers drain the steep parts of the southern catchment (see further information in Swierczynski et al. (2013) and Kämpf et al. (2014), Lauterbach et al. (2011)). Lake Mondsee is exposed to moist airmasses from the Atlantic and the Mediterranean Sea. Flood events preferably occur in summer after heavy precipitation events.

2.3 Data and Methods

2.3.1 Flood data from natural geoarchives

At Lake Mondsee, a long sediment core of 15 m has been retrieved in 2005 using a piston corer (Lauterbach et al. (2011)). These sediments exhibit biochemical calcite varves with seasonal resolutions and thin intercalated clastic layers (Fig. 2.1). A combination of microfacies analysis and geochemical element analysis enables to identify and allocate thin flood layers composed of siliciclastic sediments which have been transported from the catchment into the lake during flood events (Fig. 2.1). Detecting the event layers in the sediments, a flood chronology of the last 7100 years has been established (Swierczynski et al. (2013)). The chronology of the flood events is based on robust varve chronology with an error ca 1,25% for the last 4000 years. A calibration study for the sediment deposition of flood layers in the distal coring location as deposited between 1976—2013 (Kämpf et al. (2014); Swierczynski et al., in prep.) document that these summer flood events are caused by heavy precipitation events. In this study we use the flood event chronology from Lake Mondsee which has been previously published in the database PANGAEA (<https://doi.pangaea.de/10.1594/PANGAEA.818922>).

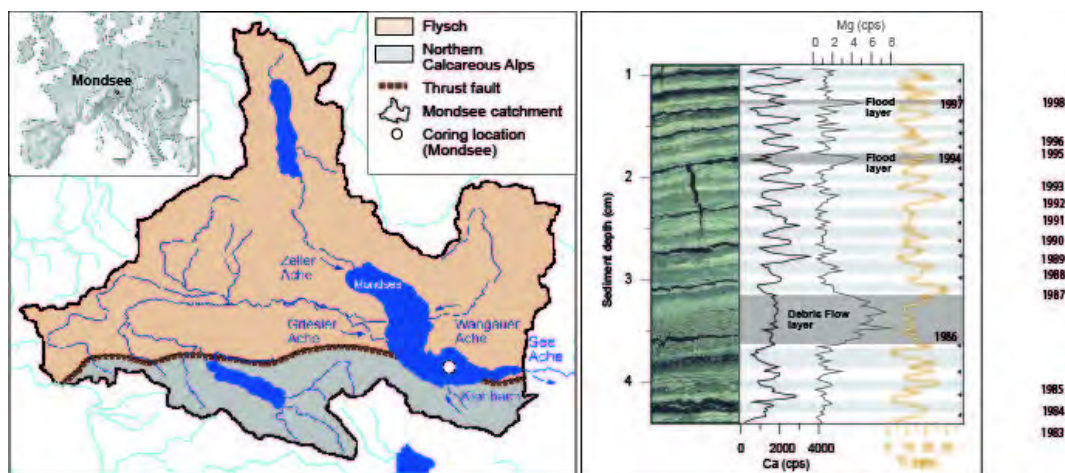


Figure 2.1: Left: Location of Mondsee, geological characteristics and lake catchment. Three main tributaries Griesler Ache, Wangauer Ache and Zeller Ache (coring location as a white circle); Right: Thin section from Lake Mondsee sediments (modified from Swierczynski et al., 2012) with spring/summer sublayers enriched in endogenic Calcium (Ca) and autumn/winter layers enriched in titanium (Ti), microscopic flood layers are enriched in magnesium (Mg) and titanium (Ti) transported from the lake catchment while thicker debris flow layers are enriched in magnesium (Mg) indicating local provenance of sediment from Northern Calcareous Alps (Dolomite and Limestone).

2.3.2 Statistical analysis

The Mondsee record contains information on the occurrence of a flood in any given year for the last 7100 years. Over-all there are 271 recorded flood events. Figure 2.2 depicts the 270 inter-flood times (i.e., the years between two consecutive floods) in chronological order. It is now interesting to investigate the properties of this time series and to test model assumptions for the nature of the stochastic process underlying the occurrence of the flood events. In particular, the goal is to investigate properties of stationarity and dependence in time from a statistical perspective and in a second step interpret the observed structures. The analysis of this paper can be adapted and extended whenever additional reliable covariate information is available. For most of the implementations below, we use routines with the software package R (described in more detail at the respective places).

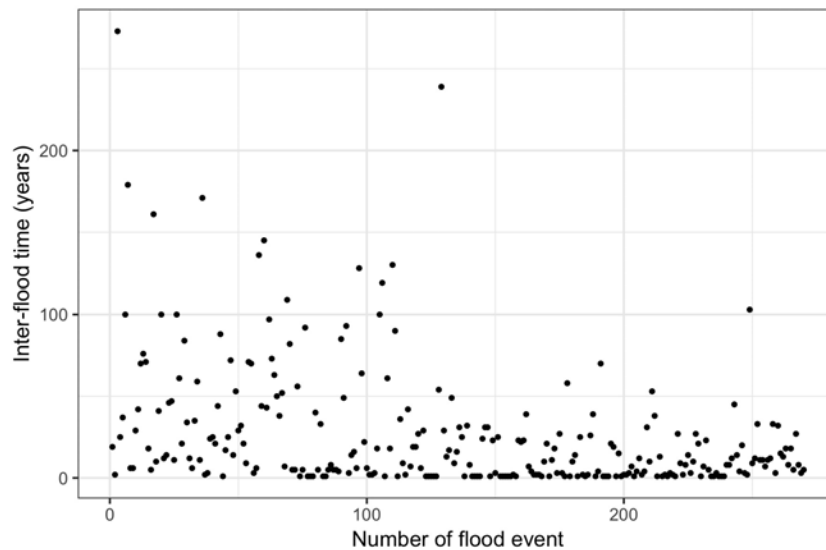


Figure 2.2: Inter-flood times (in years) of the Mondsee record arranged in chronological order.

Testing for a Poisson process

While the resolution of the Mondsee record is in years, we decide to still consider continuous-time models for the counting process underlying the flood occurrences. In fact, when modelling counting processes, observations are in general restricted to a discrete observation grid, and in the present situation the length of the time series (and the average time between observations) is such that this simplification should not be problematic. In any case, we later also test for the model sensitivity due to this binning of data points.

For a general reference to counting process models, see e.g. Daley and Vere-Jones (2007). The simplest stochastic process for the occurrence of events is a Poisson process, which is built on the assumption of lack of memory. More formally, let $N(t), t \geq 0$, be the number of flood events up to time t (where $t = 0$ refers to

the start of the observation period). Then $N(t)$ is called a homogeneous Poisson process with intensity λ , if it has stationary and independent increments, and each increment $N(t) - N(s)$ ($0 \leq s \leq t$) is Poisson-distributed with parameter $\lambda(t - s)$, i.e. $\mathbb{P}(N(t) - N(s) = n) = (\lambda(t - s))^n e^{-\lambda(t - s)} / n!$. A particular consequence of this definition is that the times between events follow an identical exponential distribution (with rate λ) throughout the entire time span, their values are independent, and the latter properties can be tested statistically. For the test of exponentiality of inter-arrival times we will use a Kolmogorov-Smirnov test.

In case of rejection, it is common as a next step in such a situation to then suspect an inhomogeneous Poisson process to be a viable alternative model. This process takes into account changes (like trends) in the intensity over time, but otherwise keeps the lack-of-memory property of the counting process in place (see e.g. Merz et al. (2016) for a respective assumption for German flood records). Let $\lambda(t)$ be the time-inhomogeneous intensity function which needs to be estimated from the data series. A non-parametric approach is a kernel estimation of the intensity function following

$$\hat{\lambda}(t) = m^{-1} \sum_{i=1}^n K\left(\frac{t - T_i}{m}\right), \quad (2.1)$$

where K is a kernel function, n is the number of observed flood events (in our case $n = 271$), T_i are the times of floods, and m is a bandwidth that determines the time horizon in the observed event series that is used for the estimation of the intensity at any time t , see also Merz et al. (2016). Typical choices for the kernel function are Gaussian and uniform shapes (see e.g. Diggle and Marron (1988)). For the confidence interval around the estimate $\hat{\lambda}(t)$, we follow the procedure of Merz et al. (2016). That is, we randomly select with replacement years with a flood from the set of all years with floods to get a new series of event times with the same number of events. Then the estimation procedure is repeated on all these samples and a 95% confidence bound is built from these. If the resulting confidence interval around $\hat{\lambda}(t)$ at some point in time does not contain the estimated homogeneous intensity, this is an indication of the non-homogeneity of the process. However, while previous studies in the literature stop at this point (see e.g. Wilhelm et al. (2018a)), one still needs to test whether the flood time series can reasonably be modelled by an inhomogeneous Poisson process. In this paper, this is done by time-transformation. With a suitable time transformation based on the inhomogeneous intensity, every inhomogeneous Poisson process can be expressed as a homogeneous Poisson process with resulting exponential inter-arrival times in 'operational' time (see e.g. (Albrecher et al., 2017, Ch.V) for details). For this purpose we prefer to work with an estimate based on a uniform kernel

$$\hat{\lambda}(t) = \frac{\#\text{years with floods between years } t - m/2 \text{ and } t + m/2}{m}, \quad (2.2)$$

i.e., we simply count the number of floods in year-windows of size m (for small m , it can happen that for some t -values this empirical estimate is zero, in which case we replace it by 10^{-2} ; at the left and right end of the interval, the average is taken with

the respectively fewer available 0/1 entries). Since the result is potentially quite sensitive to the choice of the bandwidth m , we vary m across a large number of feasible choices and choose the one that turns out to be 'closest' to an inhomogeneous Poisson behavior and then study its significance. That latter significance test is then simply one for a homogeneous Poisson behavior, but now for the new time scale.

2.3.3 Change-points and dependence

It will often occur that neither the homogeneous nor the non-homogeneous Poisson assumptions hold. When looking for alternative models, there are several directions possible. In the context of this paper, we strive for a simple model that can already explain the observed data reasonably well. One such possibility is to consider a renewal model, where inter-flood times are independent and identically distributed, but possibly follow another than exponential distribution (see e.g. Albrecher et al. (2017)).

Such models are homogeneous in time, so we implement a modified version that allows for changes throughout time. Concretely, we allow for change-points, before and after which the dynamics can possibly be described by a simple model, but with respective different characteristics. Such change-points could e.g. be interpreted as a sudden external event or a change of conditions that forces an abrupt change of the dynamics underlying the occurrence pattern of flood events.

In order to identify possible change-points, various statistical approaches are possible, see e.g. Brodsky and Darkhovsky (1993); Chen and Gupta (2011) for a general reference on parametric and non-parametric methods.

Classical mean-variance test

A classical likelihood ratio test based on changes in both the mean and variance of the inter-flood times is performed using the function `cpt.meanvar` in R. The variant of the test used here is tailored towards exponential and independent data. The exponential assumption can be challenged, but in a number of cases the final model later on will result in inter-arrival times of similar shape to the exponential, so that in those cases the test can be seen reasonably appropriate. The test also allows for the detection of more than one change-point (using the option `BinSeg`, which performs a clustering algorithm, cf. Chen and Gupta (2011)).

CUSUM Test

A second classical test that we perform is the non-parametric sequential Cumulative Sums (CUSUM) test, which tests for detecting a change in the mean based on cumulative sums and on a normality assumption, see e.g. Csörgö and Horváth (1997).

A regression approach to change-point detection

Whenever one suspects a sudden change-point, an alternative is a smoother trend, so we would like to have a test at hand that allows to directly compare whether the inclusion of a drift or a change-point is significant in the presence of the other. For that purpose, we need a nested model. There exist change-point detection algorithms for linear regression models, which to some extent provide that property (see e.g. Chen and Gupta (2011); Csörgö and Horváth (1997)). We will extend that approach here by allowing for a (more general) Generalized Linear Model (GLM) setting, since our variables are not normally distributed and the trend is not necessarily linear over time. The algorithm is simple and transparent. We present it in a general form, as it may be useful in other contexts as well.

Assume we have n observations y_1, \dots, y_n , (say, inter-flood times) and time events $x_1 \leq \dots \leq x_n$ (say, flood number or flood year). The algorithm is as follows:

1. For $K = 1, \dots, n - 1$:
 Choose a regression model which is most appropriate with the distributional characteristics of the data. Generalized Linear Models (GLM's) are a flexible and suggested family (see e.g. McCullagh and Nelder (1989); in the present case, the Gamma GLM is appropriate given the results of the statistical analysis on the particular data), and their implementation in R is through the `glm` function. The covariates should include the binary variable $z_k = 1_{x_k \leq x_K}$ ($z_k = 1$ when x_k is smaller or equal than x_K , and otherwise $x_k = 0$), $k = 1, \dots, n$, which allows for the inclusion of a change-point. The additional inclusion of x_k corresponds to the introduction of a trend.
2. Find the $k := k_0$ that leads to the largest t-value (equivalently, the smallest p-value) for the hypothesis of not having a change-point, i.e. find the most significant coefficient (from $k = 1, \dots, n$) of the variable z_k with respect to the t-value (Wald test) that is outputted automatically when applying the R function `summary` to a `glm` object.
3. Test for the significance of the selected change-point k_0 by performing Analysis of Variance (ANOVA) against the model without that change-point. That is, fit the model k_0 with the deletion of the covariate z_k and compare the two models with the `anova` function in R. A related useful function is `drop1`, which does the two steps automatically (for every variable in the original model).
4. On the basis of the above test, decide for a model with a trend, with a change-point, or both.

Remark 2.3.1. The above procedure tests for difference in means. If a difference in variance is desired, it is best to first fit a model to the original data and then perform the change-point analysis on the residuals of the fitted model.

Remark 2.3.2. The covariates included in Step (1) above do not necessarily enter linearly into the modelling of the mean. The way they affect the variate is through

the link function, and this is another parameter to be chosen in Step (1). The covariate choices $\beta_1 z_i$ or $\beta_1 z_i + \beta_2 x_i$ are good choices.

Remark 2.3.3. Finally, note one further advantage of this regression approach above: both the classical mean-variance test and the CUSUM test rely on independent inter-flood times, so that their application for lakes with a higher degree of correlation becomes doubtful. However, with the regression above, some of the dependence might be explained through the covariates so that the residuals may in fact be close to independent. In other words, the regression approach is applicable in a larger set of scenarios than the previous methods to detect change-points.

For the lake sediment record, we will apply the above procedure in Section 2.4 with the Gamma family with canonical link function (negative inverse function), see e.g. Frees (2009). Let $x_k = k$ be the observation number and y_k the inter-flood times. For $K = 1, \dots, n$ we then model the mean as

$$\mathbb{E}(y_k) = -\frac{1}{\beta_0 + \beta_1 x_k + \beta_2 \mathbf{1}_{x_k \leq K}}, \quad \beta_i \in \mathbb{R}, \quad (2.3)$$

and we identify the K with the largest absolute resulting t -value. The resulting coefficients for β_1 and β_2 then decide whether a (in our case decreasing) trend and/or a change-point for the distribution of inter-flood times are present in the time series, cf. Section 2.4.

Dependence between inter-flood times

In order to assess the (in)dependence between inter-flood times we look at the corresponding auto-correlation function (ACF) and partial auto-correlation function (PACF).

Furthermore, once a change-point is detected, the next task is then to test whether before and after that change-point there is sufficient evidence for independence among inter-flood times. We also do that initially with an ACF and PACF analysis.

Since we are not only interested in correlations but more generally in the independence assumption between inter-flood times, we further lag the series and use an empirical copula test of independence between the resulting lagged vectors. This is done by calling the `serialIndepTest` in R, which is an implementation of the statistics developed in Genest and Rémillard (2004). The basic idea behind this test is to compare the empirical copula of the data against the copula of independent data, through a suitable measure, which can then be arranged in a plot according to the lagged vectors that were considered. The standard way of plotting the test is through the `dependogram` (`dependogram` function in R), which is merely a plot of the values of the serial independence statistic for different selections of lagged vectors. Critical values are automatically given (in the same way that bands are given in ACF and PACF), which are such that the simultaneous acceptance region has probability 0.95 under the null hypothesis of independence.

2.3.4 Fitting a renewal model

Once a change-point is detected, and independence is plausible in each resulting period, the next task is to identify an inter-flood time distribution for each of the two periods that provides a reasonable fit to the sediment record. That is, within each period, the inter-flood times W_i , $i = 1, \dots, n$ are considered to be independent and identically distributed variables. For the class of distributions in which we look for the best fit, we choose the general class of *phase-type* distributions. This class extends the idea of an exponential distribution (with lack of memory) to a concatenation of memory-less components by introducing κ states and a homogeneous Markov chain that transits between these states according to fixed transition intensities. The realization of the random variable is then the total amount of time a trajectory spends in the chain from the initial state until it reaches an additional final (absorption) state (see e.g. Asmussen and Albrecher (2010) for details). The previous exponential distribution is then the special case of having only one state before moving to absorption. The Markov chain starts according to the probability vector π containing the κ probabilities to start in state j ($j = 1, \dots, \kappa$) and the dynamics of the chain are given by a sub-intensity matrix T . It is any matrix that satisfies to be non-positive in the diagonal, non-negative outside of the diagonal, and row-wise sums to zero. The element T_{ij} for $i \neq j$ of the matrix has the interpretation of being the transition rate from state i to j . The resulting survival function is

$$\mathbb{P}(W_i > x) = \pi \exp(Tx)1, \quad x \geq 0,$$

where 1 is the $(\kappa \times 1)$ -vector of 1's. Phase-type distributions are not only an intuitive way to extend the idea of memorylessness, but also represent a very broad family of distributions on the positive real line. In particular, any other distribution can be arbitrarily closely approximated by a phase-type distribution (if only the number κ of states and the intensity matrix T is chosen appropriately). However, the phase-type class is quite rich on its own and often it is the case that an exact distribution from this class produces a nice fit. For instance, one of the most tractable phase-type distributions is the hyper-exponential distribution, which corresponds to T being zero in the off-diagonal elements and is interpreted as a mixture of exponential distributions. The data will determine the needed size κ , and hence the complexity of the model.

We fit such a model within each period using maximum likelihood via an EM algorithm (utilizing the C program `EMpht`, see e.g. Asmussen et al. (1996)).

One advantage of such an explicit model is that one can make quantitative statements about quantities like the distribution of the residual waiting time until the next flood, given the time since the last flood (under the assumption that the model represents the actual dynamics well). For any phase-type distribution, one can calculate the probability that the next flood does not appear within the next $x - y$ years given that the previous flood occurred y years ago (see e.g. Asmussen and Albrecher (2010)) as

$$S(x|y) = \frac{\pi \exp(T(x))1}{\pi \exp(Ty)1}, \quad x \geq y,$$

and the respective expected value of number of years until the next flood is then given by

$$-\frac{\pi \exp(Ty)}{\pi \exp(Ty)1} T^{-1} 1.$$

Further details on the implemented methods are given in Section 2.4.

2.4 Results and Discussion

2.4.1 Testing for a Poisson process

We first focus on the identical distribution behavior and will deal with dependence later. One can see even with the naked eye from Figure 2.2 that there does not seem to be a homogeneous behavior throughout time. Indeed, a respective statistical test on exponentiality of the interflood times rejects this hypothesis with a p -value below 10^{-15} , so the assumption of a homogeneous Poisson process for the occurrence of floods for the Mondsee record can immediately be rejected, and flood events do show some degree of clustering.

Figure 2.3 depicts the resulting estimate $\hat{\lambda}(t)$ for a Gaussian kernel function (cf. (2.1)) with bandwidth $m = 100$ years (i.e., the standard deviation of the kernel). The rugs represent the actual flood events. The process was reflected at the borders to compensate for missing values outside the considered time horizon.

The plot also contains the estimate of a constant intensity under a homogeneous Poisson assumption together with bootstrap confidence intervals from resampling the (supposedly) exponential interflood times.

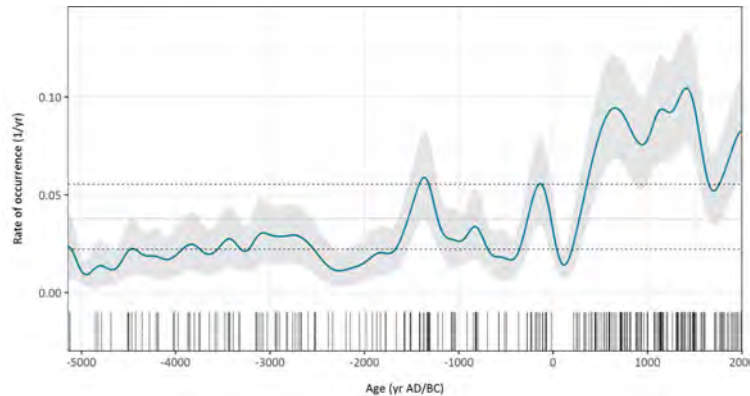


Figure 2.3: Estimated inhomogeneous intensity together with confidence intervals. The rug depicts the actual occurrence of floods in the Mondsee record

One can read the resulting plot in various ways. First, it reconfirms that the hypothesis of a homogeneous Poisson process can be rejected, with the estimated intensity leaving the confidence interval around the homogeneous estimate clearly and for a significant proportion of the time. Vice versa, the estimate of the homogeneous intensity is outside the confidence interval around the inhomogeneous

estimate for most of the time. More importantly, one sees a strongly time-varying pattern of the intensity function with a long-term tendency of increase.

As described in Section 2.3.2, we now time-transform the process according to the intensity estimate (2.2). Figure 2.4 depicts the inter-flood times after time transformation for a small, medium and large value of window size m ($m = 6, 50, 400$). If the inhomogeneous Poisson assumption were appropriate, these points should now follow a unit exponential distribution. The smooth lines depict a natural cubic spline with four knots fitted to each of the three (now homogenous) datasets, indicating (particularly for larger m) a reasonably constant unit-mean behavior. The choice of m that minimizes the Kolmogorov-Smirnov criterion for an exponential fit is $m = 50$. However, even in this case, the resulting p-value is only 0.007, rejecting the assumption of an inhomogeneous Poisson process for the Mondsee lake record. Hence, while clustering of flood events is observed in the data, the inhomogeneous Poisson process is not the appropriate model to account for this clustering.

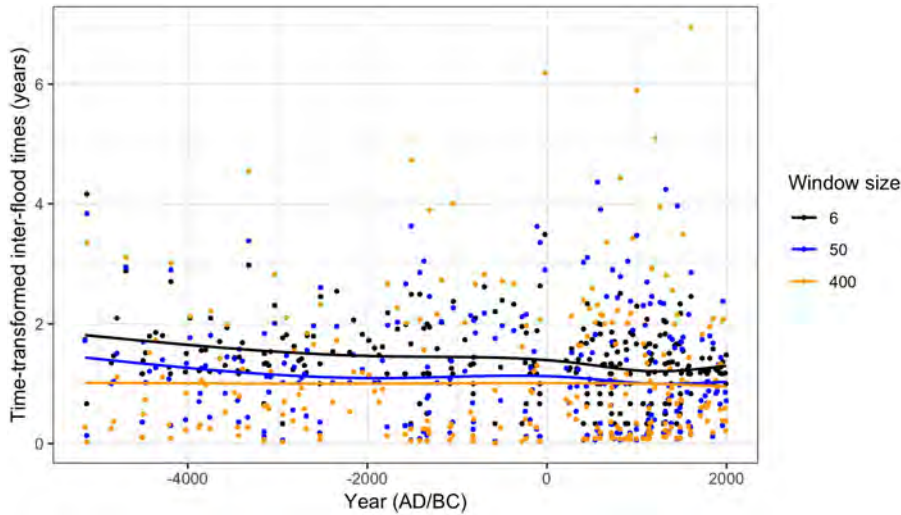


Figure 2.4: Inter-flood times after being time-transformed according to an empirical moving average of window size 6 (black), 50 (blue), and 400 (orange), respectively. By color, the lines are four-spline regression curves added to the points for visual tracking of their mean throughout time.

2.4.2 Change-points and dependence

We saw above that the assumption of a (homogeneous or inhomogeneous) Poisson process has to be rejected on the basis of the resulting marginal distribution of inter-flood times not being exponentially distributed. For this conclusion, the dependence pattern between the inter-flood times had yet to be tested. In order to find a suitable alternative model, the ACF and the PACF for the series of inter-flood times was examined and it was readily concluded that across the entire time range the inter-flood times are not sufficiently uncorrelated to directly pursue a renewal model with independent (and non-exponential) inter-flood times.

We hence look for possible change-points. As outlined in Section 2.3.3, we start with a classical test based on both the mean and variance of the inter-flood times, which allows for the detection of also more than one change-point. The result is a single significant change-point at the observation number 133, corresponding to year 350 AD (cf. Figure 2.5). Indeed, the identified change-point in Figure 2.5 could roughly be detected even by visual inspection of the time series. In order to

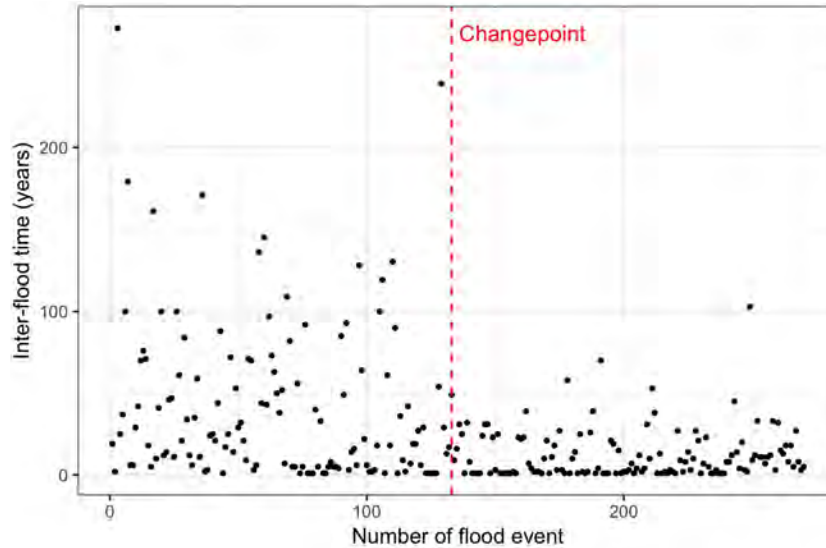


Figure 2.5: Inter-flood times (in years) with the suggested change-point at observation number 133.

test the presence of a change-point against an alternative description by a smooth trend over time, we now apply the regression approach to change-point detection as outlined in Section 2.3.3, using the Gamma family with canonical link function (inverse function). This choice is justified by the fact that the actual distribution of inter-flood times is close to that of a Gamma distribution, and the latter is within the exponential dispersion family. Here we have $x_k = k$, the y_k are the inter-flood times and $K = 1, \dots, 269$. We model the mean according to (2.3), and we look for the largest absolute t -value of the β_2 parameter (the absolute value of the ratio between the coefficient β_2 and its standard error). The latter is plotted in Figure 2.6 as a function of K , and one sees that the maximum is achieved for observation 133. Note that this is in remarkable correspondence with the change-point obtained by the classical mean-variance procedure above. The resulting model includes both a change-point and a trend, see the dashed line in Figure 2.7. The plot also contains the best model fits without a change-point (solid line) and without trend terms, respectively (dashed-dotted line). Due to the nested model construction, one can now test for significance of keeping or rejecting terms. An ANOVA likelihood ratio test tells that dropping the change-point is highly rejected (p-value of 0.001953) while the drift term is close to being insignificant, its dropping not being rejected at a 5% significance level (p-value of 0.08001). Note that the best fit can lead to very different decrease rates before and after the change-point (since the value of β_2 is not restricted), and indeed the downward trend is virtually inexistent after observation 133 (correspondingly, the absolute values of the best fit with and without

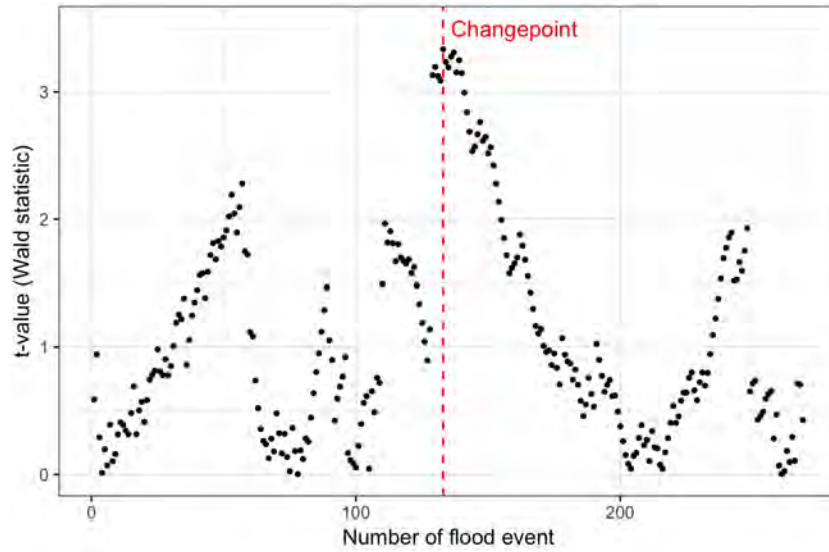


Figure 2.6: t -values as a function of observation, for the change-point coefficient in the regression modelling of inter-flood times of Lake Mondsee. The change-point is detected as before at observation number 133.

trend virtually coincide after the change-point). That is, only the drift term during the first period is responsible for the near-significance of keeping the trend in the model when tested against the simpler model with no drift.

We also compared the change-point detection with the CUSUM method, which identifies a change-point at 131, which is again close to the value of 133.

Altogether, we hence settle for the single change-point at observation 133 (i.e., year 350 AD). In addition, in view of the almost insignificant trend term in the above model, and in favour of model simplicity, we decide to remove the trend from the final model, which allows us to view the inter-flood times as identically distributed random variables within their period. This assumption is in any case not restrictive for the second period, where the drift term is not present at all.

The next task is then to check whether the two periods (before and after the change-point) exhibit dependence or independence among their inter-flood times. The result of the ACF suggests that the change-point divides the data into two time series that are each virtually un-autocorrelated. Hence, the autocorrelation that was present in the full series was rather due to the data being comprised of two homogeneous periods of different behaviour (namely of different mean), rather than from the independence assumption being violated at each of the two periods. The PACF suggested the same behaviour. It is also worth mentioning that the ACF and PACF of the square of the inter-flood times and of the successive differences of the inter-flood times are also within the confidence bounds. Analogous results were obtained using dependograms, so we consider it reasonable to assume independence between inter-flood times.

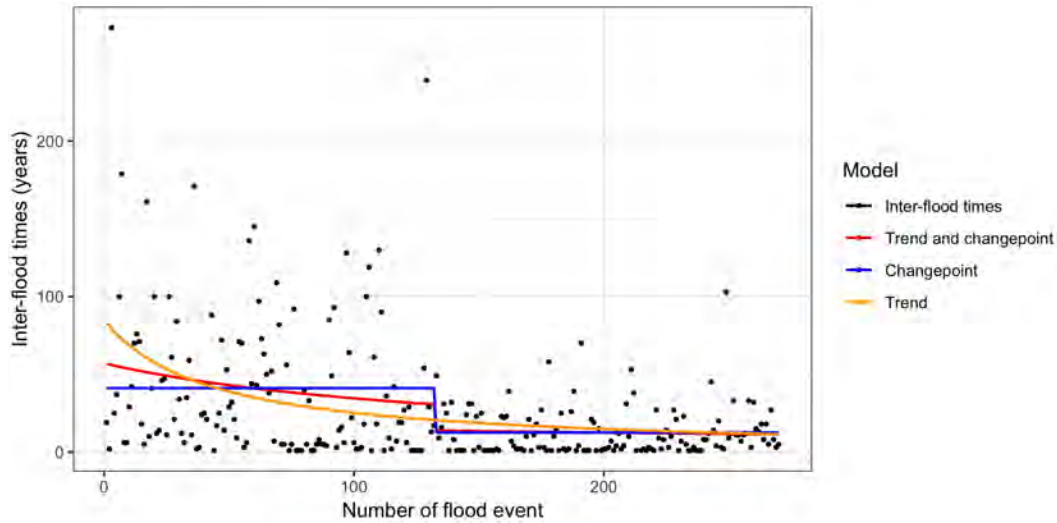


Figure 2.7: Regression models for change-point and drift modelling.

2.4.3 Fitting a renewal model

In view of the above findings, the next task is then to identify an inter-flood time distribution for each of the two periods that provides a reasonable fit to the sediment record. That is, within each period, the inter-flood times are considered to be independent and identically distributed variables. A simple exponential distribution (with a different parameter for before and after the change-point) for W_i is amply rejected (with p-values below $1e-15$). As described in Section 2.3.4, we hence consider phase-type maximum likelihood fitting via an EM algorithm for the subclass of hyper-exponential distributions (which are mixtures of exponential distributions) gives, for each of the two periods, an excellent fit for two states already: for the first period the log-likelihood (LL) is -618.042 for both two and three states, and -617.924 for a general phase-type distribution with three states. For the second period the corresponding values are -469.371, -469.371 and -468.266. This may be considered rather surprising, as it shows that two states (and a mixture of two exponential distributions) already suffice to provide a reasonable fit to the data, and one does not improve by allowing a third component in the mixture, and improves very little by allowing an arbitrary three-state phase-type distribution. Any slightly parsimonious selection method – let alone AIC or BIC – will unequivocally choose this two-state hyper-exponential model. The resulting parameter estimates are:

$$\begin{aligned} \pi_1 &= (0.211, 0.789), & \text{Diag}(T_1) &= (-0.251, -0.020), \\ \pi_2 &= (0.305, 0.695), & \text{Diag}(T_2) &= (-0.521, -0.060), \end{aligned}$$

Correspondingly, the resulting model can be interpreted in the following way: in the first period, after each flood occurrence, with a probability of roughly 0.79 there follows an exponential waiting time with mean $1/0.02=50$ years and with probability $1-0.79=0.21$ there is an exponential waiting time with mean $1/0.2=5$ years. In the second period, after each flood occurrence, with a probability of roughly 0.69

there follows an exponential waiting time with mean $1/0.06=16.6$ years and, alternatively, with probability $1-0.69=0.31$ there follows an exponential waiting time with mean $1/0.5=2$ years. In other words, one may interpret this as two mechanisms present, one with a long and one with a much shorter return period, and after the change-point around 350 AD the presence of these two mechanisms stays present with very similar likelihood of appearance, but the return period is significantly smaller for both of them.

For completeness, let us also consider a phase-type fit for a renewal model over the entire 7100 years of the Mondsee record (ignoring the autocorrelation present on the entire series) In that case, a three-state hyper-exponential model with parameters

$$\begin{aligned}\pi_F &= (0.375, 0.258, 0.367), \\ \text{Diag}(T_F) &= (-0.053, -0.398, -0.020)\end{aligned}\tag{2.4}$$

fits the data with a log-likelihood value of -1114.991, while a four-state fit does not improve this LL, and a general phase-type distribution with three states also yields the same LL. A general four state fit yields a LL of -1114.340. This can be seen as a quite convincing indication that (2.4) would be a reasonable model, i.e. three co-existing mechanisms with exponential waiting time of mean $100/1.9 \approx 52.6$, $100/5.3 \approx 18.8$ and $100/39 \approx 2.6$ years, respectively. Hence, the full series contains two mechanisms (states) that describe big inter-flood times, roughly corresponding to the respective state of of the first and the second period. The third state is the one responsible for the small return periods (mixing the two states with small return periods of the two separate series) and the mean lies between the respective values of the small return periods of the two separate periods (the concrete mixture probabilities are determined by the fact that the first period with about 5400 years contains more instances than the second with about 1700 years, so that the value is closer to the former; similarly for the initial probabilities). One may also argue in this context that for the full time period there was not enough statistical evidence for the increased complexity of keeping both states with small return periods separate in terms of a maximum likelihood tradeoff, while for the longer return periods this is the case. In turn, the detection of the change-point allowed to disentangle the effect of the mechanism with small return period into two separate mechanisms that are different before and after the change-point. In addition, this increased granularity of the model suggests that the mechanism responsible for the larger return period changed in time towards one with significantly smaller return period, which contributes to the increase of flood events. This nicely illustrates how refined models can lead to a better understanding of involved flood risk.

It is quite surprising that a rather simple renewal structure with one change-point after about 5400 years suffices to develop a very reasonable model. In particular, the pattern observed over the last 1700 years can reasonably be described by a model without a trend, but rather the co-existence of two regimes, one with a longer return period (2 years) and one with a shorter return period (17 years). One may

interpret this result as the (again possibly surprising) finding that from the available long-time record one can not conclude in a statistically significant way that the flood activity has increased lately, as only the one change-point 1700 years ago is identified, and the dynamics since then can still be accommodated within the same simple mechanism. However, we would also like to emphasize that this statement is about statistical significance with respect to the long time horizon of 7100 years. As is seen in Figure 2.3, on the intensity level one does observe a substantial increase of recent flood activity, following a lower activity some centuries ago. However, it is interesting that the parsimonious model identified in this paper demonstrates a mechanism that allows to accommodate these changes within a stationary view on longer time scales. Note that this is only possible after a change-point in history about 1700 years ago, and for the 5400 years before that, again a similarly simple model can capture the empirically observed flood occurrence pattern.

As the set of considered models was quite large, it is indeed remarkable that the resulting model turns out to be so parsimonious. On the other hand, the non-identifiability of phase-type distributions makes it impossible to make inference on the parameters directly. A Monte Carlo confidence band for the empirical distribution function is the usual alternative and is shown in Figure 2.8. The black step function corresponds to the empirical cumulative hazard and the 95% confidence bands were obtained by simulating empirical hazards from the fitted distribution.

Let us finally look at the sensitivity of the resulting model due to the binning of flood events into the annual grid. For that purpose, one can assume that inter-flood times are interval-censored at plus and minus one year of the current estimate. The resulting fit is easily handled by the EM algorithm and yields an almost identical distribution function (and, equivalently, cumulative hazard function). Figure 2.8 plots both fitted hazard rates for the dataset (binned observations in red, observations assuming censoring in blue) together with the empirical cumulative hazard of the data in black. The confidence band is for the fitted phase-type curve and was generated with 10^6 Monte Carlo simulations. One observes that the resulting fit is rather satisfactory and that the effect of the annual binning of the flood event date seems not to be problematic.

2.4.4 Comparison to an analysis with lower resolution data

Previous studies using sediment data often had a lower resolution available, see for instance Wirth et al. (2013). In order to illustrate the advantage of the high resolution of the Lake Mondsee sediment record, let us consider in the following a situation with mildly lower resolution.

Concretely, we bin the inter-flood times into blocks of three observations (with the final block containing four inter-flood times). Hence, instead of 270 data points we now deal with 90 data points, each of which comprises the sum of three inter-flood times (i.e., instead of knowing each inter-flood time, we only know the time when three inter-flood times have passed). Additionally, each such sum is allowed

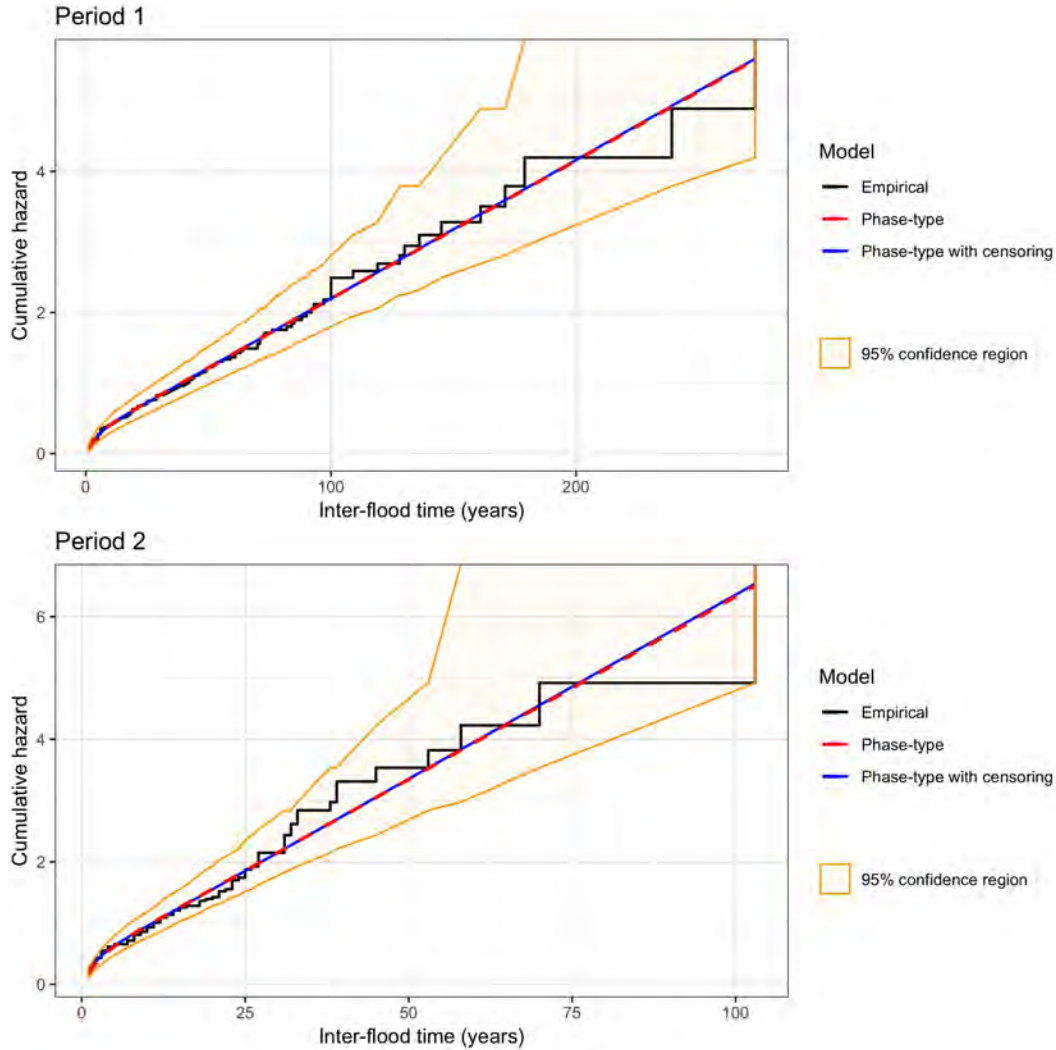


Figure 2.8: Fitted cumulative hazard of the (phase-type) hyper-exponential model with two states without (red) and with (blue) censoring.

an inaccuracy of plus minus 3 years (which is dealt with by censoring techniques).

Interestingly, the `cpt.meanvar` test identifies again one (and only one) statistically significant change-point at the year 350, which is identical to the one identified for the full resolution data. Figure 2.9 depicts the resulting change-point.

After subdividing the data into the two periods, independence is again not rejected from the correlation and dependogram plots. A statistical fit of the aggregated inter-flood times identifies a generalised Erlang distribution of order 3 as an excellent fit for both periods, which is a sum of 3 independent, but non-identical, exponential random variables. The corresponding LL is -176.343 for the first period and -124.536 for the second period. For comparison, a general phase-type distribution with nine states yields a LL of -175.153 and -123.078 , respectively. This is a minor improvement, while the resulting model would contain $3^2 - 3 + 2 = 8(!)$ additional parameters, so that any criterion like AIC, BIC chooses the generalised Erlang model at a glimpse. The resulting exponential parameters of the three states

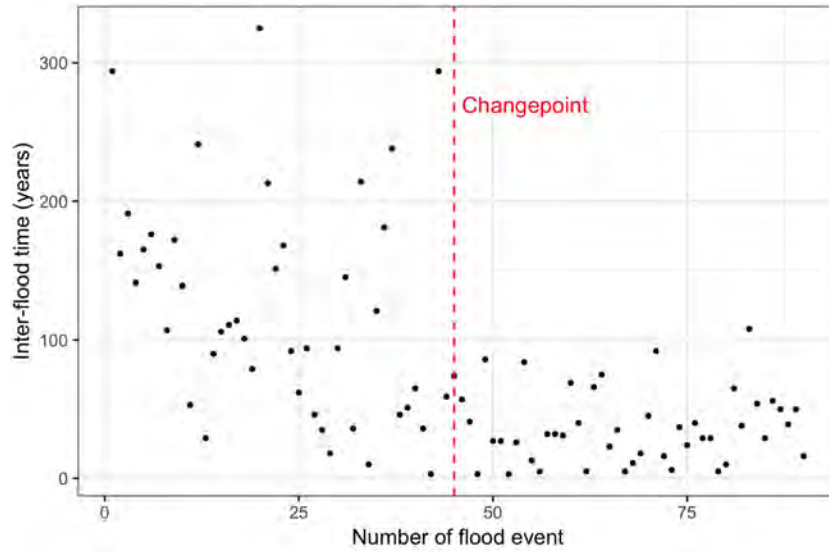


Figure 2.9: Binned inter-flood data with a suggested change-point.

are

$$\text{Diag}(T_1) = (-0.0164, -43.596, -0.016)$$

$$\text{Diag}(T_2) = (-0.054, -92.395, -0.054).$$

Note that for each period there are two states with virtually identical intensity values, which are very similar to the intensity value of the state with the longer return period in the full-resolution model. However, the difference is that in the binned model two such components with longer expected return periods appear with certainty, whereas in the model calibrated from full resolution they only have a probability 0.78 and 0.69, respectively, to occur. Both before and after the change-point, the third state in the binned model (which also occurs deterministically) has an extremely small (in view of the annual time grid underlying the model, one may even say degenerate) expected time until the next flood. Consequently the latter would have to be considered an artefact and, in essence, the coexistence of two parallel mechanisms can not be recovered in a meaningful way in the lower resolution model. As the expected time of the short (degenerate) inter-flood time may be considered negligible, the mixture of three exponentials in the full resolution model is merely replaced by the sum of two exponentials of the regime with longer return periods in the lower resolution model, and the respective conclusions one would draw from the modelling are significantly different.

This illustration emphasizes the advantage of the availability of the higher resolution, leading to a model with much more structure, yet still low complexity.

2.4.5 The waiting time until the next flood

As already outlined above, a phase-type renewal model enables us to compute expected values of quantities of interest, like the number of years until the next flood. For instance, in the second period, if already 20 years have passed since the

last flood, the expected waiting time until the next flood can be calculated explicitly to be roughly 17 years. Figure 2.10 depicts this expected value as function of years since the last flood (purple line). It may seem counter-intuitive at first that this number increases, the longer the time since the last flood. However, the nature of the exponential distribution with its lack of memory is responsible for this effect, as with time passing it becomes more likely that the mechanism with longer return period is present, and the latter has lack of memory with mean around 17 years. In case our phase-type fit had identified some sum (rather than the mixture) of exponentials as the feasible model, then the situation would be different and the expected time until the next flood would decrease as a function of time since the last flood. To illustrate this, consider the binned model (with lower resolution) studied in Section 2.4.4, where a generalized Erlang(3) distribution was the obtained model for the aggregate time of three-inter-flood times. For the expected value of the next flood under that (lower) model resolution one has to divide the aggregate mean by 3 (since the aggregated variables are exchangeable). Figure 2.10 depicts the result, again for the period after the change-point (blue line). Note that the line converges to $100/(3 \times 5.4446) = 6.1222$, which is one-ninth of the mean of the largest individual exponential mean in the sum. Both curves start close to the sample mean of the inter-flood times (black line), and the model implications are dramatically different for larger number of years since the last flood. The additional resolution hence allowed to decipher the fine structure of the model and its consequences in a way that was not to be expected from the aggregated data situation. This is a further illustration for the sensitivity of these results with respect to the model and the available resolution of flood record data.

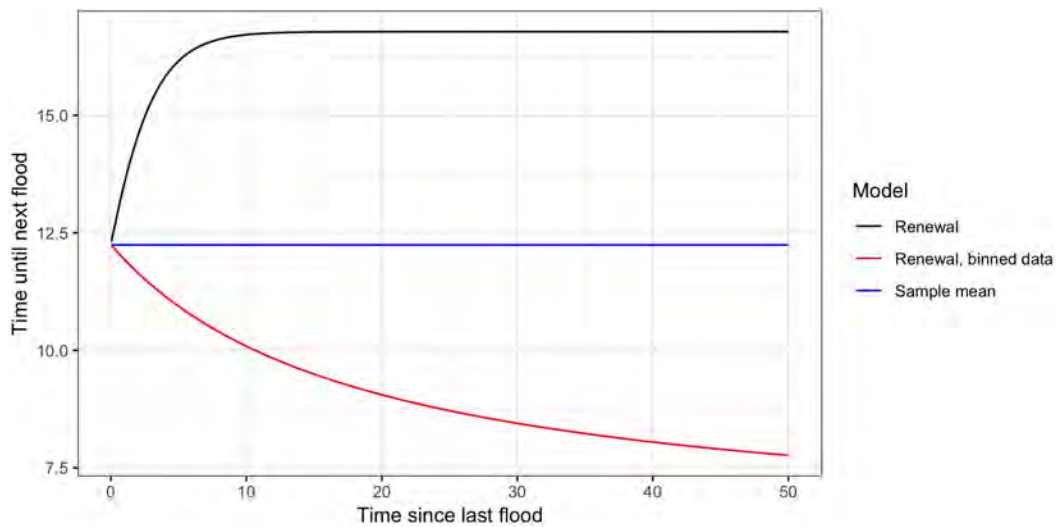


Figure 2.10: Expected time until next flood as a function of years since the last flood for the data from the second period.

2.5 Conclusion and Outlook

This paper investigated statistical properties of the well-dated flood record of Lake Mondsee covering the last 7100 years. It was shown that neither a homogeneous nor an inhomogeneous Poisson process are appropriate models to characterise the paleoflood records. Instead, we identified a significant change-point around 350 AD, together with a strikingly simple renewal structure before and after that change-point, leading to an increase of flood intensity for the second period. We developed a more versatile nested regression approach to distinguish between trends and change-points. We indeed obtain that change-points are a statistically more satisfactory description of the dynamics than a trend for the Mondsee paleoflood record under study. We emphasize that the change-point in flood dynamics is observed on a purely statistical basis. The question of whether there is a causal relation to other physical changes linked to the climate around this period is beyond the scope of this study and will be subject of future research. But if such investigations, as to whether causal relationships in the complex dynamics of the atmosphere / geosphere and hydrosphere nexus can be detected, are carried out and further empirical inquiry into other sorts of paleodata could eventually explain the change-point that we detected here, this will come as a valuable contribution to the broader questions mentioned in the introduction: With a changing climate, which change-points do we have to expect in our flood regimes and how can we thus reduce the uncertainty in flood risk modelling facing climate change? One such important issue to be solved for all the local flood regimes that paleoflood data are such an impressive witness of, is identifying the respective meteorological regimes that lead to, or were in the past likely to have led to high precipitation events. This is beyond the scope of this study, but we believe that the results presented can encourage further interdisciplinary studies.

When sticking to the target of learning more about potential future flood damages from the study of paleoflood records certainly another challenge will be the following: Understand well, how modern records of flood processes such as gauge data or damage data can be reasonably linked to the sediment data on the actual end of the time series. One fascinating aspect about sediment data for a statistician certainly is the fact that lakes kept the memories of past floods for such a long time. While the approach here was a purely statistical one, it will be interesting to link the more recent flood records to human interventions such as river regulation etc.

For now we conclude, that for investigating climatic and hydrological changes in well-dated paleoflood series, high-resolution data are crucial. The statistical analysis of Lake Mondsee sediments performed in this paper demonstrates that the resolution of a paleoflood record is important for the detection of change-points. Similar to rare historical floods, the analysis of such time series is affected by the recurrence times of floods. We showed that the flood dynamics inferred from the data can be quite different under the higher resolution as compared to models that would have been suggested by lower-resolution data. Consequently, the completeness as well as the uncertainty within a record is an important information for adapting an appropriate method.

One of the main findings of the present study is that for the Lake Mondsee record under study, it turns out that the use of change-points leads to a surprisingly parsimonious satisfactory statistical description of the flood occurrence pattern over time. We hence encourage the use of change-points for robust flood chronologies to improve risk assessments. Change-point analysis can also be used for other paleoflood records and hydro-climatic records. The PAGES Initiative (Floods Working Group) promotes the collection of paleoflood datasets with related metadata (<http://www.pages-igbp.org>). Information about dating chronology and information about completeness of floods as well as further flood information beyond the paleoflood record (calibration studies) presents an excellent opportunity to use the PAGES database of past floods for further statistical treatment. Further investigations of the past can then lead to an increased insight in mechanisms and hence also the understanding of the future to come.

Chapter 3

Dividends: From Refracting to Ratcheting

This chapter is based on the following article:

Albrecher, H., Bäuerle, N., & Bladt, M. (2018). Dividends: From refracting to ratcheting. *Insurance: Mathematics and Economics*, 83, 47-58.

Abstract

In this paper we consider an alternative dividend payment strategy in risk theory, where the dividend rate can never decrease. This addresses a concern that has often been raised in connection with the practical relevance of optimal classical dividend payment strategies of barrier and threshold type. We study the case where once during the lifetime of the risk process the dividend rate can be increased and derive corresponding formulae for the resulting expected discounted dividend payments until ruin. We first consider a general spectrally-negative Lévy risk model, and then refine the analysis for a diffusion approximation and a compound Poisson risk model. It is shown that for the diffusion approximation the optimal barrier for the ratcheting strategy is characterized by an unexpected relation to the case of refracted dividend payments. Finally, numerical illustrations for the diffusion case indicate that with such a simple ratcheting dividend strategy the expected value of discounted dividends can already get quite close to the respective value of the refracted dividend strategy, the latter being known to be optimal among all admissible dividend strategies.

3.1 Introduction

Starting with de Finetti's work de Finetti (1957), the study of optimal dividend payout strategies in collective risk theory has been a very active field of research over the last 60 years. It is nowadays well-known that in order to maximize the expected aggregate discounted dividends until ruin, it is optimal to pay dividends according to a band strategy, which in a number of cases collapses to a barrier

strategy, see e.g. Gerber (1969), Shreve et al. (1984), Schmidli (2008) and Azcue and Muler (2014). When this optimal control problem is considered with an upper bound on the dividend rate, then in many cases a *refracting* dividend strategy (or, synonymously, a *threshold* strategy) is optimal (i.e., no dividend payments up to a certain barrier level, and dividend payments at maximal allowed rate above that barrier), see for instance Jeanblanc-Picqué and Shiryaev (1995), Asmussen and Taksar (1997) for diffusions and Gerber and Shiu (2006a) and Lin and Pavlova (2006) for the compound Poisson process. Since then numerous extensions and variations of the dividend problem have been considered, many of which leading to intricate and interesting mathematical problems, see e.g. Albrecher and Thonhauser (2009) and Avanzi (2009) for surveys on the topic.

Among these variations, some also address the issue that the theoretically optimal refraction strategy may not be realistic when it comes to implementation in practice. One constraint is that dividends can only be paid out in a discrete rather than a continuous-time fashion, see Albrecher et al. (2011) for a contribution in this direction for random discrete payment times. Another problem with the threshold strategy is the strong variability of payment patterns across time. In Avanzi and Wong (2012) it was proposed for a diffusion process to consider dividend payments that are proportional to the current surplus level, which leads to much smoother dividend streams. Recently, the respective analysis of performance measures was extended to the compound Poisson model in Albrecher and Cani (2017). In Bäuerle and Jaśkiewicz (2015) and Bäuerle and Jaśkiewicz (2017) the aim was to maximize risk-sensitive dividend payments in discrete time which, in case of exponential utility, results in optimizing a weighted criterion of expectation and variance of the dividends.

In this paper we would like to consider another constraint that is often mentioned in discussions with practitioners. Concretely, it is at times considered desirable to have a dividend payment stream that does not decrease over time (which we will refer to as a *ratcheting* strategy, see e.g. Dybvig (1995)). One reason is that a decrease may have a negative psychological impact on shareholders and the considered value of the company in general. Sticking to the analytically much more tractable situation of a continuous-time model, the over-all mathematical question could then be to find the optimal dividend payment pattern with a non-decreasing dividend rate. This general optimal control problem is a considerable challenge, as one immediately is put into a two-dimensional setting with one variable keeping track of the current dividend rate, so that the usually advantageous Markovian structure of the surplus process is lost even for simple processes. We hence in this paper will consider only a first step towards improving the understanding of this general problem, namely to allow for a constant dividend rate through-out the lifetime of the risk process, which once during the lifetime can be increased to a higher level. The task then is to find the optimal rates before and after that change, and the surplus level at which this change should take place, so that the expected over-all discounted dividend payments are maximized. The only constraint in this context will be that the dividend rate must never exceed the drift rate of the original process. This setting leads to quite explicit results and will give some insight into

the nature of the problem. A particular focus will be given on the comparison of the resulting optimal ratcheting strategy with the best refracting strategy, which is known to be optimal among all admissible dividend strategies in the absence of such a ratcheting constraint.

We will first deal with the general case of spectrally-negative Lévy processes in Section 3.2. After collecting some necessary preliminaries, we adapt an existing formula for refraction strategies that also allows for dividend payments below the barrier. We then derive a general formula for the expected discounted dividend payments until ruin according to a ratcheting strategy in terms of the scale function of the underlying Lévy process, and derive a criterion for the optimal ratcheting barrier. Subsequently, in Section 7.3 we refine the analysis for the case of a pure Brownian motion with drift (diffusion approximation). While many of the respective formulas can in fact be obtained from Section 3.2 by substituting the simple form of the respective scale function, we derive a number of formulas in that section in a self-contained, sometimes more direct, way. This adds another perspective to the problems under study, and also allows to read parts of that section independently of the general Lévy fluctuation theory that underlies the approach in Section 3.2. We then give some numerical illustrations on the performance of both the refracting and ratcheting strategy, which somewhat surprisingly indicate that the best ratcheting procedure is not far behind the best refracting (and hence over-all best) dividend strategy. We also derive a somewhat surprising criterion for the optimal ratcheting barrier in terms of a matching with a refracting strategy. In Section 7.4 we then proceed to work out the formulas of Section 3.2 to the particular case of a Cramér-Lundberg risk model with hyper-exponential claim sizes. We also show that the optimality criterion for the diffusion case mentioned above no longer holds in the compound Poisson setting, the reason being the non-differentiability of the refraction value function at the barrier level.

One natural concern for the implementation of a ratcheting strategy of the above kind is that after the switching, the higher dividend rate will always stay, even when the surplus level gets low. Correspondingly one may expect that ruin is more likely or happens earlier, particularly when the optimal barrier level is chosen according to the profitability criterion of maximal expected dividend payments. Also, the optimal ratcheting barrier is in general higher than the optimal refraction barrier. In Section 3.5 we discuss this issue further and quantify it in terms of expected ruin time, given that ruin occurs. The numerical results in fact indicate that, conditional on ruin to occur in finite time, the expected time to ruin is larger for the ratcheting strategy. Finally, Section 3.6 contains some conclusions.

3.2 The Spectrally-Negative Lévy Risk Model

Let us consider a spectrally-negative Lévy process $Y = \{Y_t\}_{t \geq 0}$, i.e. a Lévy process with only negative jumps, and which is not a.s. a non-increasing process. We assume that $Y_0 = x \geq 0$ and the drift of this process is positive, and call such a Y a *Lévy risk process*. Our focus in this paper will be on risk processes constantly paying out

dividends at rate $c_1 \geq 0$ (so possibly $c_1 = 0$), and in certain periods (to be specified) at an increased rate $c_1 + c_2$, with $c_2 > 0$. Correspondingly, the Lévy processes of interest are

$$X_t = Y_t - c_1 t, \quad \tilde{X}_t = Y_t - (c_1 + c_2)t.$$

for all $t \geq 0$. Define the Laplace exponent

$$\psi(\theta) := \log \mathbb{E} e^{\theta X_1},$$

which is finite for at least all $\theta \geq 0$, and denote by $\Phi(\delta)$ the largest root of the equation $\psi(\theta) = \delta$. The δ -scale functions $W(x)$ and $Z(x)$ of the process X are defined for any $\delta \geq 0$ as

$$\int_0^\infty e^{-ux} W(x) dx = \frac{1}{\psi(u) - \delta}, \quad u > \Phi(\delta),$$

$$Z(x) = 1 + \delta \int_0^x W(y) dy.$$

The respective δ -scale functions for the process \tilde{X} will be denoted by $\mathbb{W}(x)$ and $\mathbb{Z}(x)$, respectively, and $\phi(\delta)$ shall be the corresponding largest root of $\psi(\theta) - c_2 \theta = \delta$. Define now for a fixed $b \geq 0$ the first passage times

$$\tau_0^- := \inf\{t \geq 0 : X_t < 0\}, \quad \tau_b^+ := \inf\{t \geq 0 : X_t > b\}.$$

We will be working on a canonical probability space for all the processes involved, consisting of the space of all right-continuous functions with left-sided limits, with a probability law denoted by \mathbb{P}_x , and associated conditional expectation \mathbb{E}_x , given that the process starts at $x \geq 0$. It is well-known from Lévy fluctuation theory that one has

$$\mathbb{E}_x \left(e^{-\delta \tau_b^+}; \tau_b^+ < \tau_0^- \right) = \frac{W(x)}{W(b)} \quad (3.1)$$

and

$$\mathbb{E}_x \left(e^{-\delta \tau_0^-}; \tau_0^- < \tau_b^+ \right) = Z(x) - Z(b) \frac{W(x)}{W(b)}, \quad (3.2)$$

for any $0 \leq x \leq b$, see for instance (Kyprianou, 2014, Ch.8).

3.2.1 Refracting Strategy

For reasons of comparison, let us now first recollect a formula from Kyprianou and Loeffen (2010) for the expected sum of discounted dividends until ruin under a refraction strategy. For any threshold level $b \geq 0$, the respective modified Lévy risk process is given by

$$U_t = X_t - c_2 \int_0^t 1_{\{U_s > b\}} ds, \quad t \geq 0. \quad (3.3)$$

The interpretation is that dividends are paid at rate c_1 while the original process Y is below the threshold b , and at rate $c_1 + c_2$ above the threshold. Note that the

existence of the process (3.3) is in fact not as straightforward as one may expect, and one has to require

$$c_2 \in \left(0, \gamma + \int_{(0,1)} x \Pi(dx) \right),$$

whenever X has paths of bounded variation, where γ is the canonical drift coefficient in the Lévy-Khintchine representation of X , and Π is the corresponding Lévy measure, see Theorem 1 of Kyprianou and Loeffen (2010) for details.

Define

$$\tau = \inf\{t > 0 : U_t < 0\}$$

as the time of ruin of the process U defined in (3.3). A slight adaptation of Eq. (10.25) of Kyprianou and Loeffen (2010) (which corresponds to the case $c_1 = 0$) then leads to a formula for the expected sum of discounted dividends until ruin under the threshold strategy (3.3), for general $x, b \geq 0$:

$$\begin{aligned} V(x, c_1, c_2, b) &:= \mathbb{E}_x \left[\int_0^\tau e^{-\delta s} c_2 1_{\{U_s > b\}} ds \right] + c_1 \mathbb{E}_x \left[\int_0^\tau e^{-\delta s} ds \right] \\ &= \frac{c_2}{\delta} (1 - \mathbb{Z}(x - b)) + \frac{W(x) + c_2 \int_b^x \mathbb{W}(x - y) W'(y) dy}{\phi(\delta) \int_0^\infty e^{-\phi(\delta)y} W'(y + b) dy} \\ &\quad + \frac{c_1}{\delta} \left[1 - Z(x) - \delta c_2 \int_b^x \mathbb{W}(x - y) W(y) dy \right] \\ &\quad + \frac{c_1}{\delta} \left[\frac{W(x) + c_2 \int_b^x \mathbb{W}(x - y) W'(y) dy}{e^{-\phi(\delta)b} \int_0^\infty e^{-\phi(\delta)y} W'(y + b) dy} \right] \delta \int_b^\infty e^{-\phi(\delta)y} W(y) dy. \end{aligned} \quad (3.4)$$

This is a somewhat involved, but completely explicit expression for $V(x, c_1, c_2, b)$, which can be evaluated whenever the scale function of the underlying Lévy process is available.

3.2.2 Ratchet Strategy

We now turn to the study of the following ratcheting strategy: Dividends are paid at a fixed constant rate $c_1 \geq 0$ until the first time the surplus process hits a barrier b , at which point the dividend rate is increased (ratcheted) to $c_1 + c_2$ for a fixed constant $c_2 > 0$, and stays at this higher level until the time of ruin. The modified Lévy risk process under this ratcheting strategy is then given by

$$U_t^R = Y_t - \int_0^t (c_1 + c_2 1_{\{M_s \geq b\}}) ds = X_t - c_2 \int_0^t 1_{\{M_s \geq b\}} ds, \quad (3.5)$$

where $M_t = \sup_{0 \leq s \leq t} Y_t$. In contrast to the refracting case, the existence of U_t^R is straightforward for any $c_2 > 0$. Such a ratcheting strategy takes into account the fact that shareholders prefer to not experience a decrease in the rate of their dividend stream. This strategy is no longer Markovian, but depends on the history

of the process.

Define by

$$\tau^R = \inf\{t > 0 : U_t^R < 0\}$$

the time of ruin and by

$$V^R(x, c_1, c_2, b) = \mathbb{E}_x \left[\int_0^{\tau^R} e^{-\delta s} (c_1 + c_2 1_{\{M_s \geq b\}}) ds \right] \quad (3.6)$$

the expected value of the aggregate discounted dividend payments under such a ratcheting strategy.

Theorem 3.2.1. *The expected value of the aggregate discounted dividend payments until ruin under a ratcheting strategy for a Lévy risk model is given by*

$$V^R(x, c_1, c_2, b) = \begin{cases} \frac{c_1 + c_2}{\delta} [1 - \mathbb{Z}(x) + \frac{\delta}{\phi(\delta)} \mathbb{W}(x)], & 0 \leq b \leq x, \\ \frac{c_1 + c_2}{\delta} [1 - \mathbb{Z}(b) + \frac{\delta}{\phi(\delta)} \mathbb{W}(b)] \frac{W(x)}{W(b)} \\ + \frac{c_1}{\delta} [1 - Z(x) + (Z(b) - 1) \frac{W(x)}{W(b)}], & 0 \leq x < b. \end{cases} \quad (3.7)$$

Proof. Consider first the case $x \geq b$. Then the higher dividend rate $c_1 + c_2$ is paid out on from the beginning until ruin, i.e.

$$V^R(x, c_1, c_2, b) = \mathbb{E}_x \left[\int_0^{\tau^R} e^{-\delta s} (c_1 + c_2) ds \right] = \frac{c_1 + c_2}{\delta} [1 - \mathbb{E}_x(e^{-\delta \tilde{\tau}_0^-})],$$

where

$$\tilde{\tau}_0^- = \inf\{t \geq 0 : \tilde{X}_t < 0\}$$

is the ruin time of the risk process when the original drift is reduced by $c_1 + c_2$. But then the result follows from (3.2) and the fact that $\lim_{b \rightarrow \infty} \mathbb{Z}(b)/\mathbb{W}(b) = \delta \cdot \lim_{b \rightarrow \infty} \mathbb{W}(b)/\mathbb{W}'(b) = \delta/\phi(\delta)$.

For $x < b$, we have to distinguish whether the process will reach b before ruin or not, and in the former case we apply the strong Markov property at that point in time, on from which the process dynamics change to the drift being reduced by $c_1 + c_2$. We thus have

$$V^R(x, c_1, c_2, b) = c_1 \mathbb{E}_x \left[\int_0^{\tau_b^+ \wedge \tau^R} e^{-\delta s} ds \right] + \mathbb{E}_x(e^{-\delta \tau_b^+}; \tau_b^+ < \tau_0^-) \cdot V^R(b, c_1, c_2, b),$$

and $V^R(b, c_1, c_2, b)$ is given above. The result then follows from

$$\delta \mathbb{E}_x \left[\int_0^{\tau_b^+ \wedge \tau^R} e^{-\delta s} ds \right] = 1 - \mathbb{E}_x \left[e^{-\delta(\tau_b^+ \wedge \tau^R)} \right] = 1 - \mathbb{E}_x \left[e^{-\delta \tau_b^+}; \tau_b^+ < \tau_0^- \right] - \mathbb{E}_x \left[e^{-\delta \tau_0^-}; \tau_0^- < \tau_b^+ \right],$$

and again using (3.1) and (3.2). □

It is interesting to try to identify the barrier level b , for which $V^R(x, c_1, c_2, b)$ is maximized. The natural criterion for that purpose is to look for a solution of

$$\frac{\partial V^R(x, c_1, c_2, b)}{\partial b} = 0 \quad (3.8)$$

One should keep in mind, however, that the necessity and sufficiency of this criterion depends on the analytical properties of V^R , which are inherited from the scale function W . Correspondingly, it is not possible to characterize such a barrier level in full generality, but for most cases of practical interest the scale function structure is such that the above derivative condition for the optimal barrier is the relevant one (see also Loeffen (2008)). Also, if the optimal barrier is positive, it represents a necessary condition. For simplicity we will hence in the following refer to a barrier that fulfills (3.8) as optimal. Theorem 3.2.1 can now be used to derive a criterion for a barrier level b to be optimal for the ratcheting strategy:

Proposition 3.2.2. (*Optimal barrier*) *For fixed c_1, c_2 , the barrier b that maximizes (3.7) does not depend on x and is characterized by the equation*

$$\frac{d}{db} \left(\frac{W(b)}{\mathcal{W}(b)} \right) = 0, \quad (3.9)$$

where

$$\mathcal{W}(x) = \frac{c_1 Z(x) + c_2}{c_1 + c_2} - \mathbb{Z}(x) + \frac{\delta}{\phi(\delta)} \mathbb{W}(x).$$

Proof. By the nature of the ratcheting strategy, for every initial capital x , a barrier level $b < x$ is equivalent to a barrier level $b = x$, so that we can w.l.o.g. consider the case $0 \leq x \leq b$ only. Differentiation of expression (3.7) with respect to b and equating it to zero shows that all terms depending on x disappear and that

$$\frac{W(b)}{W'(b)} = \frac{\mathcal{W}(b)}{\mathcal{W}'(b)}$$

or equivalently Equation (3.9) must hold. \square

One can easily see from (3.7) that the function $V^R(x, c_1, c_2, b)$ is continuous at $x = b$. On the other hand, the derivative of V^R with respect to x does not have to be continuous in that point. The following result shows that this derivative is, however, continuous in the optimal barrier level. There is hence an alternative way to identify the optimal barrier:

Theorem 3.2.3. (*Smooth pasting*) *In the ratcheting dividend problem, the optimal barrier b^R is exactly the one which makes the value function continuously differentiable.*

Proof. Taking the derivative of V^R of (3.6) with respect to x on both sides of the barrier b , evaluating at b and equating both expressions yields after some algebra precisely the criterion (3.9). \square

In the next sections we will now refine the analysis for the case of Brownian motion and for a compound Poisson process with hyper-exponential jumps.

3.3 Brownian approximation

Consider now a risk process

$$Y_t = x + \mu t + \sigma B_t, \quad t \geq 0, \quad (3.10)$$

where $\mu > 0$ is a constant drift, $\sigma > 0$ and $(B_t)_{t \geq 0}$ denotes standard Brownian motion. Clearly, this is a special case of the Lévy risk model considered in Section 3.2, and by substituting the corresponding scale function (which in this case is a linear combination of two exponential terms), one can retrieve the respective formulas for the refracting and ratcheting strategies in this more particular setting. We prefer, however, to give here a self-contained, more direct derivation for the diffusion case, and then use the resulting formulas for a more detailed analysis of the performance of the ratcheting dividend strategy. In particular, we will also establish an unexpected connection between the refracted and the ratcheting case which does not hold for general Lévy risk processes.

3.3.1 Refracting Strategy

Consider a diffusion risk reserve process with continuous dividend payout

$$dU_t = (\mu - c_t)dt + \sigma dB_t,$$

where dividends are paid at rate

$$c_t = c_1 1_{\{U_t \leq b\}} + (c_1 + c_2) 1_{\{U_t > b\}} \quad (3.11)$$

with $b, c_1, c_2 \geq 0$. I.e., whenever the risk reserve is above level b we pay at rate $c_1 + c_2 > 0$, otherwise only at rate c_1 . Thus the payment rate changes at b which could be physically understood as a refraction. We first derive the corresponding value of this strategy, measured in terms of the expected aggregate discounted dividend payments

$$V(x, c_1, c_2, b) := \mathbb{E}_x \left[\int_0^\tau e^{-\delta s} c_s ds \right],$$

where as before

$$\tau := \inf\{t \geq 0 : X_t = 0\},$$

$\delta \geq 0$ is a discount rate and \mathbb{E}_x is the conditional expectation given that $X_0 = x$. Here $X_t = Y_t - c_1 t$. For $c_1 = 0$, a formula for V is well-known, see e.g. Gerber and Shiu (2006b). In the following we establish the corresponding extension for $c_1 > 0$.

Denote by $\theta_1 > 0 > \theta_2$ the roots of

$$\frac{1}{2}\sigma^2 z^2 + (\mu - c_1)z - \delta = 0, \quad (3.12)$$

and by $\tilde{\theta}_1 > 0 > \tilde{\theta}_2$ the roots of

$$\frac{1}{2}\sigma^2 z^2 + (\mu - c_1 - c_2)z - \delta = 0. \quad (3.13)$$

Moreover, let $\kappa := ((\mu - c_1)^2 + 2\sigma^2\delta)^{-\frac{1}{2}}$ and

$$W(x) := \kappa(e^{\theta_1 x} - e^{\theta_2 x}), \quad x \geq 0.$$

Note that $W(x)$ is the scale function of the process $(X_t)_{t \geq 0}$.

Theorem 3.3.1. *The value function under the fixed threshold strategy (3.11) in the diffusion case is given by*

$$V(x, c_1, c_2, b) = \begin{cases} B \cdot W(x) + \frac{c_1}{\delta}(1 - e^{\theta_2 x}), & 0 \leq x \leq b \\ \frac{c_1 + c_2}{\delta} + D e^{\tilde{\theta}_2 x}, & x \geq b \end{cases}$$

where

$$B := \frac{1}{\delta} \cdot \frac{c_1 e^{\theta_2 b}(\theta_2 - \tilde{\theta}_2) - c_2 \tilde{\theta}_2}{W'(b) - \tilde{\theta}_2 W(b)} \quad (3.14)$$

$$D := B e^{-\tilde{\theta}_2 b} W(b) - \frac{c_1}{\delta} e^{(\theta_2 - \tilde{\theta}_2)b} - \frac{c_2}{\delta} e^{-\tilde{\theta}_2 b}. \quad (3.15)$$

Proof. In what follows we write $V(x)$ instead of $V(x, c_1, c_2, b)$ since c_1, c_2, b are fixed. First note that we can decompose the value function into

$$V(x) = c_1 \mathbb{E}_x \left[\int_0^\tau e^{-\delta s} ds \right] + V(x, \mu - c_1),$$

where $V(x, \mu - c_1)$ is the value function of the expected discounted dividends of a process X_t with drift $\mu - c_1$ where nothing is paid below the barrier b and above b the rate c_2 is paid. This is the usual refracting case (see e.g. Gerber and Shiu (2006b)). From this observation it follows that $V \in C^1$, i.e. V is continuously differentiable.

Now, since the process

$$\int_0^{t \wedge \tau} e^{-\delta s} c_s ds + e^{-\delta(t \wedge \tau)} V(X_{t \wedge \tau})$$

is a martingale, the drift of the process has to vanish and the value function has to satisfy the following differential equations below and above the barrier:

$$\frac{1}{2}\sigma^2 V''(x) + (\mu - c_1)V'(x) - \delta V(x) + c_1 = 0, \quad 0 \leq x \leq b, \quad (3.16)$$

$$\frac{1}{2}\sigma^2 V''(x) + (\mu - c_1 - c_2)V'(x) - \delta V(x) + c_1 + c_2 = 0, \quad x \geq b, \quad (3.17)$$

with boundary conditions $V(0) = 0$ and $\lim_{x \rightarrow \infty} V(x) = \frac{c_1 + c_2}{\delta}$. The general solution of (3.16) is

$$V(x) = \frac{c_1}{\delta} + B_1 e^{\theta_1 x} + B_2 e^{\theta_2 x}$$

with constants $B_1, B_2 \in \mathbb{R}$. Using $V(0) = 0$ we obtain $B_2 = -\frac{c_1}{\delta} - B_1$. The general solution of (3.17) is

$$V(x) = \frac{c_1 + c_2}{\delta} + D_1 e^{\tilde{\theta}_1 x} + D_2 e^{\tilde{\theta}_2 x}$$

with constants $D_1, D_2 \in \mathbb{R}$. Using the second boundary condition we obtain $D_1 = 0$. The smooth fit condition that V and V' are continuous at b yields

$$B_1 e^{\theta_1 b} - \left(\frac{c_1}{\delta} + B_1\right) e^{\theta_2 b} = \frac{c_2}{\delta} + D_2 e^{\tilde{\theta}_2 b} \quad (3.18)$$

$$B_1 \theta_1 e^{\theta_1 b} - \theta_2 \left(\frac{c_1}{\delta} + B_1\right) e^{\theta_2 b} = D_2 \tilde{\theta}_2 e^{\tilde{\theta}_2 b}. \quad (3.19)$$

The first equation implies

$$D_2 e^{\tilde{\theta}_2 b} = B_1 (e^{\theta_1 b} - e^{\theta_2 b}) - \frac{c_1}{\delta} e^{\theta_2 b} - \frac{c_2}{\delta} \quad (3.20)$$

which can be inserted into the second equation to obtain

$$\frac{B_1}{\kappa} W'(b) = \frac{B_1 \tilde{\theta}_2}{\kappa} W(b) + \frac{c_1}{\delta} e^{\theta_2 b} (\theta_2 - \tilde{\theta}_2) - \frac{c_2}{\delta} \tilde{\theta}_2.$$

This implies that $B_1 = \kappa B$ with B as in (3.14). Inserting the expression for B_1 into (3.20) yields $D := D_2$ in (3.15). \square

Remark 3.3.2. The special case $c_1 = 0, c_2 = c > 0$ has been studied intensively before. In this case we obtain the formula

$$V(x, b) := V(x, 0, c, b) = \begin{cases} \frac{c}{\delta} W(x) \frac{\tilde{\theta}_2}{\theta_2 W(b) - W'(b)}, & 0 \leq x \leq b \\ \frac{c}{\delta} \left(1 + \frac{W'(b)}{\theta_2 W(b) - W'(b)} e^{\tilde{\theta}_2(x-b)}\right), & x \geq b. \end{cases}$$

This result can e.g. be found in Gerber and Shiu (2006b) as equation (2.22), (2.23). Taking the derivative w.r.t. b and equating to zero establishes the optimal barrier b^* that maximizes $V(x, b)$ as

$$b^* = \frac{1}{\theta_1 - \theta_2} \log \left(\frac{\theta_2(\theta_2 - \tilde{\theta}_2)}{\theta_1(\theta_1 - \tilde{\theta}_2)} \right)$$

if $\mu + \frac{\sigma^2}{2} \tilde{\theta}_2 > 0$ and $b^* = 0$ otherwise (see e.g. Asmussen and Taksar (1997)). In fact, in Gerber and Shiu (2006b), various other characterizations of b^* have been shown: First b^* can be characterized as the unique b s.t. $V(x, b)$ is twice continuously differentiable in x . Moreover it is the unique b s.t. the value function $V(x, b)$ coincides with the value function of the dividends according to a horizontal barrier strategy with barrier b , i.e. the case $c = \mu$.

3.3.2 Ratchet Strategy

Let us now look into the value function for the diffusion case under the ratcheting strategy:

Theorem 3.3.3. *The value function under the ratchet dividend strategy with barrier b for the diffusion case is given by*

$$V^R(x, c_1, c_2, b) = \begin{cases} \frac{c_1}{\delta} (1 - e^{\theta_2 x}) + \frac{1}{\delta} \frac{e^{\theta_1 x} - e^{\theta_2 x}}{e^{\theta_1 b} - e^{\theta_2 b}} (c_2 + c_1 e^{\theta_2 b} - (c_1 + c_2) e^{\tilde{\theta}_2 b}), & 0 \leq x \leq b, \\ \frac{c_1 + c_2}{\delta} (1 - e^{\tilde{\theta}_2 x}), & x \geq b. \end{cases} \quad (3.21)$$

Proof. In the diffusion case the scale functions are given by

$$W(x) = \kappa(e^{\theta_1 x} - e^{\theta_2 x}), \quad \mathbb{W}(x) = \tilde{\kappa}(e^{\tilde{\theta}_1 x} - e^{\tilde{\theta}_2 x}),$$

with $\kappa = ((\mu - c_1)^2 + 2\sigma^2\delta)^{-\frac{1}{2}}$ and $\tilde{\kappa} := ((\mu - c_1 - c_2)^2 + 2\sigma^2\delta)^{-\frac{1}{2}}$. Substituting these expressions into (3.7) in Theorem 3.2.1, together with some algebraic manipulations, gives the result. Note that here $\phi(\delta) = \tilde{\theta}_1$ and $Z(x) = \frac{\delta}{\tilde{\theta}_1}W(x) + e^{\theta_2 x}$.

For $x \geq b$ there is also a direct way: Since in this case we start immediately to pay out at rate $c_1 + c_2$, the value function is

$$V^R(x, c_1, c_2, b) = \frac{c_1 + c_2}{\delta}(1 - \mathbb{E}_x(e^{-\delta\tilde{\tau}_0^-})).$$

But by the same arguments as in Theorem 3.3.1, the quantity $m(x) := \mathbb{E}_x(e^{-\delta\tilde{\tau}_0^-})$ satisfies, for any $x > 0$, the differential equation

$$\frac{1}{2}\sigma^2 m''(x) + (\mu - c_1 - c_2)m'(x) - \delta m(x) = 0,$$

with boundary condition $m(0) = 1$ and $\lim_{x \rightarrow \infty} m(x) = 0$. Hence $m(x) = Ae^{\tilde{\theta}_1 x} + Be^{\tilde{\theta}_2 x}$ with $A = 0$ and $B = 1$, giving $V^R(x, c_1, c_2, b) = \frac{c_1 + c_2}{\delta}(1 - e^{\tilde{\theta}_2 x})$ for $x \geq b$. \square

Note that V^R is continuous in x .

Remark 3.3.4. For $c_1 = 0, c_2 = c > 0$, (3.21) simplifies to the formula

$$V^R(x, 0, c, b) = \begin{cases} \frac{c}{\delta} \frac{e^{\theta_1 x} - e^{\theta_2 x}}{e^{\theta_1 b} - e^{\theta_2 b}} (1 - e^{\tilde{\theta}_2 b}), & 0 \leq x \leq b, \\ \frac{c}{\delta} (1 - e^{\tilde{\theta}_2 x}), & x \geq b. \end{cases}$$

From Theorem 3.2.3 we already know that the barrier b^R which maximizes the payout in the ratcheting case, i.e. $V^R(x, c_1, c_2, b^R) := \sup_b V^R(x, c_1, c_2, b)$ is the one for which the value function is continuously differentiable. It turns out, that for the diffusion case another somewhat surprising characterization of the optimal barrier b^R can be found:

Theorem 3.3.5. *In the ratcheting dividend problem, the optimal barrier b^R is exactly the one for which the value function coincides with the value function in the refracting case, i.e.*

$$V^R(x, c_1, c_2, b^R) = V(x, c_1, c_2, b^R), \quad x \geq 0.$$

Proof. Inspecting the value function of the ratcheting problem we see that we can write it as

$$V^R(x, c_1, c_2, b) = \begin{cases} \gamma W(x) + \frac{c_1}{\delta}(1 - e^{\theta_2 x}), & 0 \leq x \leq b \\ \frac{c_1 + c_2}{\delta}(1 - e^{\tilde{\theta}_2 x}), & x \geq b, \end{cases}$$

with a suitable constant γ . The value function in the refracting problem is given by

$$V(x, c_1, c_2, b) = \begin{cases} B \cdot W(x) + \frac{c_1}{\delta}(1 - e^{\theta_2 x}) & , 0 \leq x \leq b \\ \frac{c_1 + c_2}{\delta} + D e^{\tilde{\theta}_2 x} & , x \geq b \end{cases}$$

with suitable constants B and D . Since $V^R \in C^1$, if we plug in $b = b^R$ and since $V \in C^1$ for all $b \geq 0$ we deduce that the following equations hold:

$$\gamma W(b^R) - \frac{c_1}{\delta} e^{\theta_2 b^R} = \frac{c_2}{\delta} - \frac{c_1 + c_2}{\delta} e^{\tilde{\theta}_2 b^R} \quad (3.22)$$

$$\gamma W'(b^R) - \frac{c_1}{\delta} \theta_2 e^{\theta_2 b^R} = -\frac{c_1 + c_2}{\delta} \tilde{\theta}_2 e^{\tilde{\theta}_2 b^R} \quad (3.23)$$

$$B \cdot W(b^R) - \frac{c_1}{\delta} e^{\theta_2 b^R} = \frac{c_2}{\delta} + D e^{\tilde{\theta}_2 b^R} \quad (3.24)$$

$$B \cdot W'(b^R) - \frac{c_1}{\delta} \theta_2 e^{\theta_2 b^R} = D \tilde{\theta}_2 e^{\tilde{\theta}_2 b^R}. \quad (3.25)$$

Subtracting (3.24) from (3.22) we obtain

$$W(b^R)(\gamma - B) = -e^{\tilde{\theta}_2 b^R} \left(\frac{c_1 + c_2}{\delta} + D \right) \quad (3.26)$$

and subtracting (3.25) from (3.23) we obtain

$$W'(b^R)(\gamma - B) = -\tilde{\theta}_2 e^{\tilde{\theta}_2 b^R} \left(\frac{c_1 + c_2}{\delta} + D \right). \quad (3.27)$$

These last two equations yield that

$$(\gamma - B)(W'(b^R) - \tilde{\theta}_2 W(b^R)) = 0.$$

Since $W(x) = \kappa(e^{\theta_1 x} - e^{\theta_2 x}) > 0$ and $W'(x) = \kappa(\theta_1 e^{\theta_1 x} - \theta_2 e^{\theta_2 x}) > 0$ for $x \geq 0$ (note that $\theta_2 < 0$ and $\tilde{\theta}_2 < 0$), we obtain that $B = \gamma$, but in view of (3.24) and (3.22) the latter implies that $D = -\frac{c_1 + c_2}{\delta}$. Hence at this barrier level b^R both value functions coincide. \square

Example 3.3.1. Consider the case where $c_1 = 0, c_2 = 5$ and the parameters of the risk reserve process are given by $\mu = 10, \sigma = 4$ and the discount rate is $\delta = 0.999$. In Figure 3.1 we see on the left-hand side the value function $V(x, 0, 5, b)$ as a function of the initial state and the barrier level. Note that $x \mapsto V(x, 0, 5, b)$ is here differentiable for all b . On the right-hand side we see the value function $V^R(x, 0, 5, b)$ as a function of the initial state and the barrier level. The mapping $x \mapsto V(x, 0, 5, b)$ has a kink for all but one b (note that the unusually high value of δ was chosen here to visually amplify this phenomenon).

In Figure 3.2 we see both functions in one picture (left). There is exactly one b for which these functions coincide. In the picture on the right-hand side one can see the difference $V^R(x, 0, 5, b) - V(x, 0, 5, b)$. The optimal barrier is $b^R = 2.41$ in this case.

In Figure 3.3 the two value functions can be seen as a function of b for fixed $x = 0.5$. One nicely observes that the value b^R which maximizes V^R indeed coincides with the one where both values coincide.

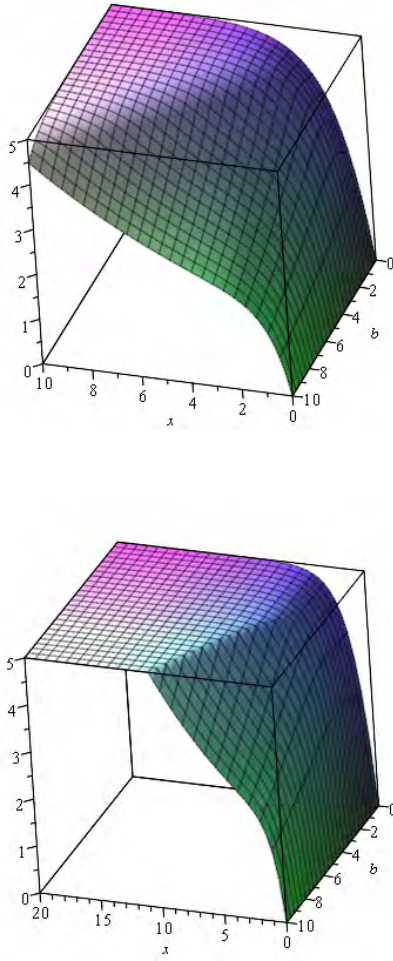


Figure 3.1: Value functions $V(x, 0, 5, b)$ (left) and $V^R(x, 0, 5, b)$ (right) with $\mu = 10$, $\sigma = 4$, $\delta = 0.999$ as functions of the initial state and the barrier level.

In the example above, when comparing the ratcheting strategy with the refraction strategy we fixed $c_1 = 0$, i.e. no dividend payments before the first hitting time of the barrier. While $c_1 = 0$ is optimal for the refraction strategy (since the resulting strategy is known to maximize the expected discounted dividend payments until ruin among all admissible dividend strategies), it may not be a fair way to compare the performance of the two types of dividend strategies, since for the ratcheting case it may very well be possible that a positive c_1 is preferable. Let us therefore compare the best refraction strategy with the best ratcheting strategy. For that purpose, one can determine the optimal threshold value $b^*(c_1, c_2)$ for the refraction strategy and the optimal barrier value $b^R(c_1, c_2)$ for the ratcheting case for each fixed c_1 and c_2 . In a second step, an optimization of

$$V^R(x, c_1, c_2, b^R(c_1, c_2)) \quad \text{and} \quad V(x, c_1, c_2, b^*(c_1, c_2)) \quad (3.28)$$

with respect to c_1 and c_2 in some constrained region of choice will yield the (trivariate) optimum of V^R and V for fixed initial surplus value x . Unfortunately, even

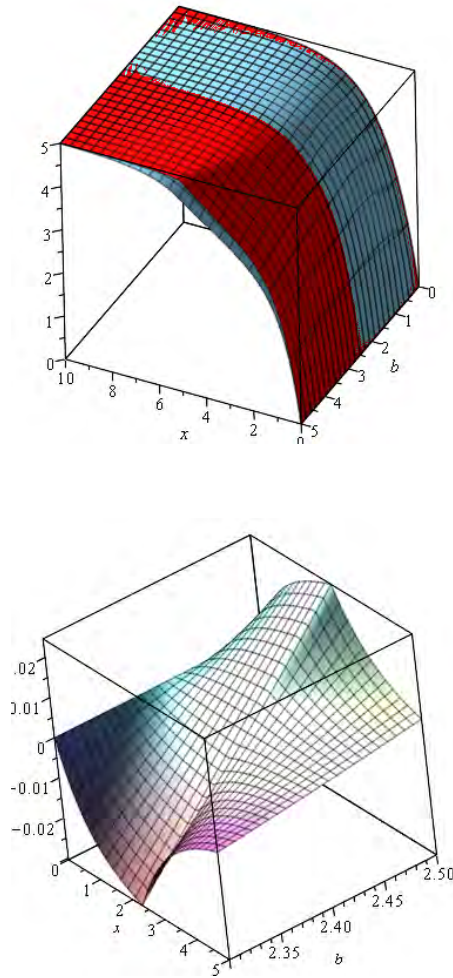


Figure 3.2: Value functions $V(x, 0, 5, b)$ and $V^R(x, 0, 5, b)$ with $\mu = 10, \sigma = 4, \delta = 0.999$ both in one picture (left) and the difference of both functions (right).

for the diffusion case, identifying the optimal trivariate choice of c_1, c_2, b for a given x in an analytic way seems out of reach since c_1, c_2 enter into the equations in an intricate way through the roots of the Laplace equation, and the characterization of the optimal ratcheting barrier level also does not lead to an explicit formula. We will, however, illustrate the behaviour of (3.28) numerically in the following example.

Example 3.3.2. Let $\mu = 10, \sigma = 6, \delta = 0.1$ and $x = 1$. Figure 3.4 depicts the two functions in (3.28) as well as their difference, for all dividend rates in the region

$$(c_1, c_2) \in [0, 5]^2$$

(note that $c_1 + c_2 = 10$ corresponds to the drift μ of the original risk process, in which case the refracting strategy turns into a horizontal dividend barrier, which in the absence of any constraint on c_1, c_2 is the optimal dividend strategy). The two

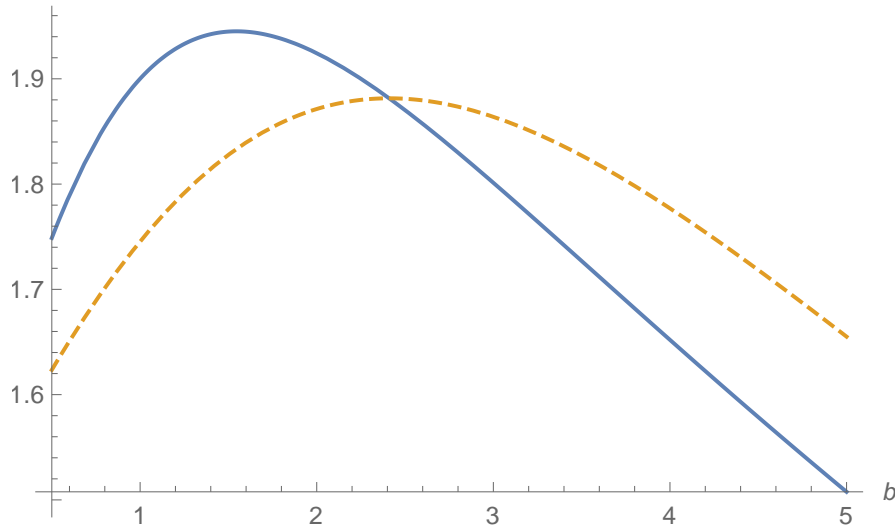


Figure 3.3: Value functions $V(0.5, 0, 5, b)$ (solid curve) and $V^R(0.5, 0, 5, b)$ (dashed curve) with $\mu = 10, \sigma = 4, \delta = 0.999, c = 5$ as functions of b .

functions agree for $c_2 = 0$, and are strictly increasing in c_2 for fixed c_1 . However, for fixed c_2 they are both not monotone in c_1 (note that increasing c_1 also increases the larger dividend rate $c_1 + c_2$). The maximum absolute difference between the two functions is achieved for the largest values of c_1 and c_2 .

In order to compare the performance of the ratcheting strategy with optimal ratcheting barrier b^R and optimal choices of c_1^R and c_2^R to the threshold strategy with optimal threshold b^* (with $c_1 = 0$ and c_2 being as large as feasible, which is known to be optimal in that case), we plot those two functions for each given upper bound $k = c_1 + c_2$ on the maximal dividend rate in Figure 3.5 (again for $x = 1, \mu = 10, \sigma = 6, \delta = 0.1$). Note that each curve depicts the respective overall best possible performance among ratcheting and refraction strategies for a given upper bound $k, k \in (0, 10)$. It is quite remarkable that the performance of the ratcheting strategy is so close to the refracting strategy (which is the overall optimal strategy), albeit the type of ratcheting is very simple (only based on one switch). Note that here the optimal choice of c_1^R also turns out to be zero, and the optimal choice of c_2^R is k , just as for the refracting case. This may lead to the conjecture that $c_1^R = 0$ and $c_2^R = k$ is always optimal for ratcheting. In view of the intricate structure, a proof of this conjecture seems, however, difficult.

If the barrier is not optimal, one also observes non-monotonicity in c_2 , cf. Figure 3.6, where V^R and V are plotted as a function of c_1, c_2 for a fixed and non-optimal $b = 1$.

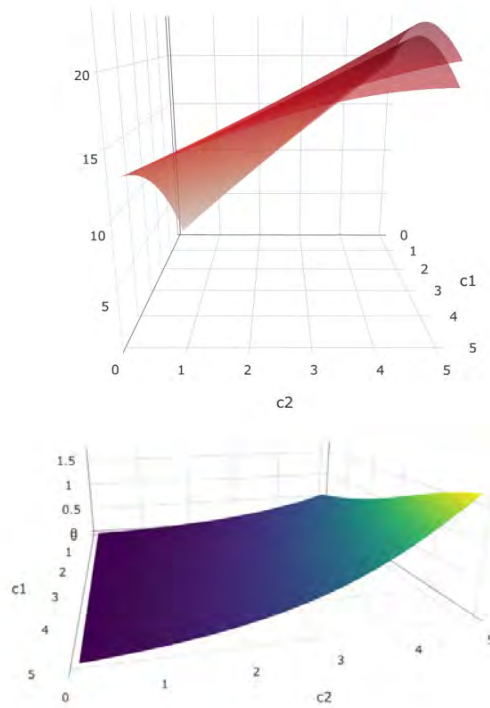


Figure 3.4: Plot of the functions (3.28) (left) and their difference (right) for $x = 1$, $\mu = 10$, $\sigma = 6$, $\delta = 0.1$.

3.4 The Cramér-Lundberg model with hyper-exponential claims

In the compound Poisson case, we have the usual Crámer-Lundberg setting, where

$$Y_t = x + ct - \sum_{i=1}^{N_t} Z_i$$

where $N = \{N_t\}_{t \geq 0}$ is a Poisson process with intensity $\lambda > 0$ and Z_i are i.i.d. non-negative random variables which are independent of N . Assume also that $c_1 + c_2 \leq c$. We focus on the case where Z_1 is hyper-exponential, i.e. a mixture of exponentials with

$$\psi(\theta) = c\theta - \lambda + \lambda \sum_{k=1}^n \frac{A_k}{1 + \theta/\alpha_k}, \quad A_k \geq 0, \quad \sum_{k=1}^n A_k = 1, \quad \alpha_k > 0, \quad n \in \mathbb{N},$$

for $\theta > -\min_k \{\alpha_k\}$. The scale functions of the process $X_t = Y_t - c_1 t$ are then given by

$$W(x) = \sum_{k=0}^n D_k e^{\theta_k x}, \quad Z(x) = 1 + \delta \sum_{k=0}^n \frac{D_k}{\theta_k} (e^{\theta_k x} - 1),$$

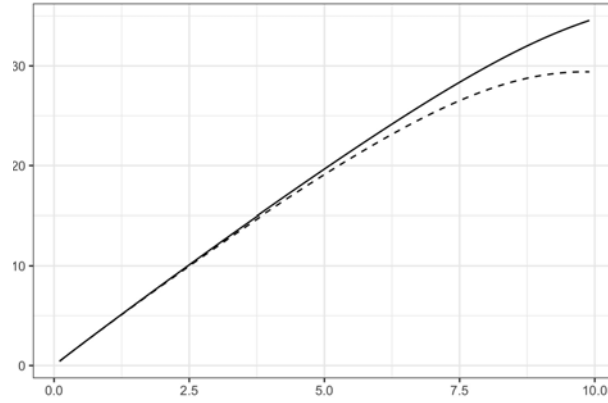


Figure 3.5: The optimal value of V (solid) and V^R (dashed) as a function of the upper limit $k = c_1 + c_2$ for $x = 1$, $\mu = 10$, $\sigma = 6$, $\delta = 0.1$.

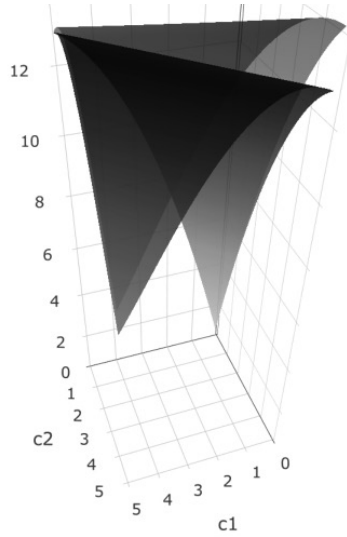


Figure 3.6: Plot of V and V^R for a fixed, non-optimal $b = 1$ and $x = 1$, $\mu = 10$, $\sigma = 6$, $\delta = 0.1$.

where θ_k , $k = 1, \dots, n + 1$ are the $n + 1$ roots, in decreasing order, of the function

$$f(\theta) = (c - c_1)\theta + \lambda + \lambda \sum_{k=1}^n \frac{A_k}{1 + \theta/\alpha_k} - \delta,$$

and

$$D_k^{-1} = \frac{df}{d\theta}(\theta_k).$$

Similarly we obtain \mathbb{W} , \mathbb{Z} in terms of $\tilde{\theta}_k$ and \tilde{D}_k by replacing c_1 by $c := c_1 + c_2$ in the above formulae. Substituting these expressions into the formulas derived in Section 3.2, it follows that

$$V^R(x, c_1, c_2, b) = \begin{cases} \frac{c}{\delta} + c \sum_{k=1}^{n+1} G_k e^{\tilde{\theta}_k x}, & 0 \leq b \leq x \\ \frac{c_1}{\delta} - c_1 \sum_{k=1}^{n+1} \frac{D_k}{\tilde{\theta}_k} e^{\tilde{\theta}_k x} + HW(x), & 0 \leq x < b, \end{cases} \quad (3.29)$$

where

$$G_k = \tilde{D}_k \left(\frac{1}{\tilde{\theta}_1} - \frac{1}{\tilde{\theta}_k} \right),$$

$$H = \frac{1}{W(b)} \left[\frac{c_2}{\delta} + c \sum_{k=1}^{n+1} \tilde{D}_k \left(\frac{1}{\tilde{\theta}_1} - \frac{1}{\tilde{\theta}_k} \right) e^{\tilde{\theta}_k b} + c_1 \sum_{k=1}^{n+1} \frac{D_k}{\theta_k} e^{\theta_k b} \right]$$

Similarly, for the refraction strategy, it is given by

$$V(x, c_1, c_2, b) = \begin{cases} \frac{c}{\delta} + c_2 \sum_{j=1}^{n+1} \zeta_j e^{\tilde{\theta}_j x} + c_1 \sum_{j=1}^{n+1} \chi_j^1 e^{\theta_j x} \\ + c_1 \sum_{j,i} (\chi_{ij}^2 + \chi_{ij}^3) (e^{\theta_i(x-b)} - e^{\tilde{\theta}_j(x-b)}), & 0 \leq b \leq x \\ \frac{c_1}{\delta} + (\eta + c_1 \xi) W(x) + c_1 \sum_{j=1}^{n+1} \chi_j^1 e^{\theta_j x} \\ + c_1 \sum_{j,i} \chi_{ij}^3 (e^{\theta_i(x-b)} - e^{\tilde{\theta}_j(x-b)}), & 0 \leq x < b, \end{cases} \quad (3.30)$$

where

$$\eta = \left(\tilde{\theta}_1 \sum_{k=1}^{n+1} \frac{D_k \theta_k}{\tilde{\theta}_1 - \theta_k} e^{\theta_k b} \right)^{-1}, \quad \zeta_j = \left\{ \eta \sum_{k=1}^{n+1} \frac{\tilde{D}_j D_k \theta_k}{\tilde{\theta}_j - \theta_k} e^{\theta_k b} - \frac{\tilde{D}_j}{\tilde{\theta}_j} \right\} e^{-\tilde{\theta}_j b},$$

$$\xi = e^{\tilde{\theta}_1 b} \tilde{\theta}_1 \eta \sum_{k=1}^{n+1} \frac{D_k}{\tilde{\theta}_1 - \theta_i} e^{\theta_k b}, \quad \chi_j^1 = -D_j / \theta_j, \quad \chi_{ij}^2 = c_2 \frac{D_i \tilde{D}_j}{\tilde{\theta}_j - \theta_i} e^{\theta_i b}, \quad \chi_{ij}^3 = c_2 \frac{D_i \tilde{D}_j \theta_i}{\tilde{\theta}_j - \theta_i} e^{\theta_i b} \xi.$$

Remark 3.4.1. In the compound Poisson case it generally does not hold anymore that

$$V^R(x, c_1, c_2, b^R) = V(x, c_1, c_2, b^R), \quad x \geq 0,$$

where b^R is the optimal barrier under the ratcheting strategy. To see this, consider $c_1 = 0$ and $n = 1$. In that case, (3.29) and (3.30) simplify to

$$V^R(x, c_1, c_2, b) = \begin{cases} \frac{c_2}{\delta} + c_2 G_2 e^{\tilde{\theta}_2 x}, & 0 \leq b \leq x \\ H \cdot W(x), & 0 \leq x < b, \end{cases} \quad (3.31)$$

and

$$V(x, c_1, c_2, b) = \begin{cases} \frac{c_2}{\delta} + c_2 \zeta_2 e^{\tilde{\theta}_2 x}, & 0 \leq b \leq x \\ \eta W(x), & 0 \leq x < b, \end{cases} \quad (3.32)$$

Arguing as in the diffusion case, the two analogous equations to (3.26) and (3.27) would then be

$$(\eta - H)W(b^R) = c_2(\zeta_2 - G_2)e^{\tilde{\theta}_2 b^R} \quad (3.33)$$

$$(\eta - H)W'(b^R) = c_2(\zeta_2 - G_2)\tilde{\theta}_2 e^{\tilde{\theta}_2 b^R} \quad (3.34)$$

and one may be inclined to think that the identity then follows in the same way. However, in contrast to the diffusion case, the derivative of V is not continuous in b unless b is the optimal barrier b^* of the threshold case (cf. Gerber and Shiu

(2006a)), which differs from b^R . Hence equation (3.34) is in fact *not* valid and the identity does no longer hold.

It is instructive to study the weak limit of the compound Poisson process with exponential claims towards a Brownian motion (which can be achieved by driving up the intensity and reducing the mean claim size at the appropriate speed). Concretely, fix the mean and variance

$$\mu = c - \frac{\lambda}{\alpha}, \quad \sigma^2 = \frac{2\lambda}{\alpha^2}. \quad (3.35)$$

Thus, for every choice of α we set $\lambda(\alpha) = \alpha^2\sigma^2/2$ and $c(\alpha) = \mu + \alpha\sigma^2/2$. Then, as $\alpha \rightarrow \infty$ one reaches the diffusion case in the limit and we have

$$D_1 \rightarrow -\kappa, \quad D_2 \rightarrow \kappa, \quad G_1 = 0, \quad G_2 \rightarrow \delta^{-1},$$

and the corresponding roots have the form

$$\theta_{1,2} = \frac{\frac{\delta}{\alpha} - \mu + c_1 \pm \sqrt{(\frac{\delta}{\alpha} - \mu + c_1)^2 + 2\sigma^2\delta + \frac{4(\mu - c_1)\delta}{\alpha}}}{\sigma^2 + \frac{2(\mu - c_1)}{\alpha}},$$

which coincide with the ones of the Brownian case defined in Section 7.3. This implies that V^R and V indeed converge to the ones of the Brownian case as $\alpha \rightarrow \infty$. We can hence observe the optimality condition of Theorem 3.3.5 to gradually come into place and being valid in the limit. Indeed, in the limit V is continuously differentiable at any barrier level b . This transition is illustrated in Figure 3.7, where $V^R(1, 0, 8, b)$ (green) and $V(1, 0, 8, b)$ (purple) are plotted as a function of b . The solid line corresponds to the diffusion case with $\mu = 10$, $\sigma = 6$ and $\delta = 0.1$, the dashed-and-dotted and the dashed lines correspond to the compound Poisson case with $\alpha = 5, 10$, respectively, where λ, c are chosen according to (3.35) in each case. The crossing of the solid lines at the maximum of the green curve exemplifies Theorem 3.3.5. The other crossings do not share this property.

3.5 The expected time to ruin

The fact that an optimal ratcheting strategy can perform nearly as well in some cases as the optimal refracting strategy is remarkable, especially since shareholders are guaranteed payments from a certain point onwards (until the time of ruin). The drawback is that the optimal ratcheting barrier is in general higher than the optimal refracting barrier (see for instance Figure 3.7), so that in the ratcheting case shareholders will have to wait longer until receiving the increased payments (or any payments at all if $c_1 = 0$). Furthermore, the distribution of the time until ruin and hence the length of the overall period of dividend payments will differ, but due to discounting this difference becomes less relevant the larger the time of ruin is. In this section we intend to quantify this tradeoff.

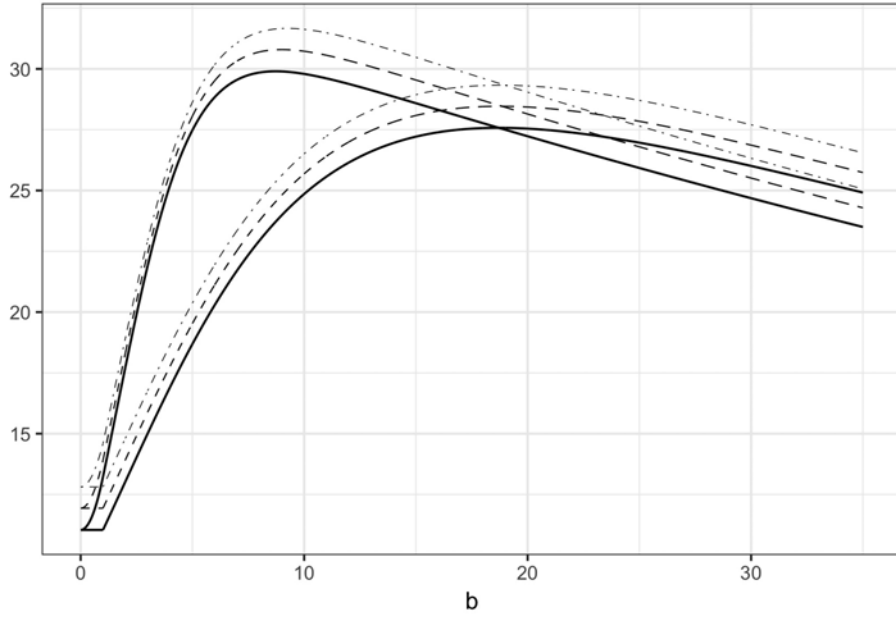


Figure 3.7: Plot of $V^R(1, 0, 8, b)$ and $V(1, 0, 8, b)$ as a function of the threshold b , for the diffusion case with $\mu = 10$, $\sigma = 6$ and $\delta = 0.1$ (solid lines) and for the compound Poisson case with $\alpha = 5$ (dashed-dotted lines) and $\alpha = 10$ (dashed lines), respectively, with λ, c chosen according to (3.35).

Whenever $c_1 + c_2$ is smaller than the drift of the original risk process, there is a positive probability to not have ruin at all, so we will confine our analysis here to the time of ruin given that ruin occurs (a more refined analysis could look into occupation time distributions of certain surplus ranges). Consider an initial surplus x below barrier b . The expected time of ruin, given it occurs in finite time, under a ratcheting strategy is then given by

$$\mathbb{E}_x[\tau^R; \tau^R < \infty] = \mathbb{E}_x[\tau_0^-; \tau_0^- < \tau_b^+] + \mathbb{E}_x[\tau_b^+ + \widehat{\tau}_0^-; \tau_b^+ < \tau_0^-, \tau^R < \infty],$$

where $\widehat{\tau}_0^-$ is the time to ruin, starting from b under the increased dividend rate $c_1 + c_2$. This leads to

$$\begin{aligned} \mathbb{E}_x[\tau^R; \tau^R < \infty] &= \mathbb{E}_x[\tau_0^-; \tau_0^- < \tau_b^+] + \mathbb{P}_b(\widetilde{\tau}_0^- < \infty) \mathbb{E}_x[\tau_b^+; \tau_b^+ < \tau_0^-] \\ &\quad + \mathbb{P}_x(\tau_b^+ < \tau_0^-) \mathbb{E}_b[\widetilde{\tau}_0^-; \widetilde{\tau}_0^- < \infty]. \end{aligned}$$

All these quantities can be recovered from the identities (3.1) and (3.2).

For the refracted strategy, for $x \leq b$ one has directly from Theorem 5(ii) of Kyrianiou and Loeffen (2010) that

$$\mathbb{E}_x(e^{-\delta\tau_0^-}; \tau_0^- < \infty) = Z(x) - \left[\frac{W(x)}{e^{-\phi(\delta)b} \int_0^\infty e^{-\phi(\delta)y} W'(y+b) dy} \right] \delta \int_b^\infty e^{-\phi(\delta)y} W(y) dy,$$

from which the expected ruin time can be obtained by differentiation w.r.t. δ and evaluating the result at $\delta = 0$.

Example 3.5.1. In the compound Poisson case with exponential claims and the safety loading condition $\mathbf{c} - c_1 - c_2 > \lambda/\alpha$, we have for the first term

$$\begin{aligned} \mathbb{E}_x[\tau_0^-; \tau_0^- < \tau_b^+] = & \\ & \frac{(\mathbf{c} - c_1)\lambda e^{\frac{b\lambda}{(\mathbf{c}-c_1)}} \left(\lambda(\alpha x + 1)e^{\frac{b\lambda}{(\mathbf{c}-c_1)}} + \alpha e^{\alpha b}(\alpha b(\mathbf{c} - c_1) - \alpha(\mathbf{c} - c_1)x + b\lambda + (\mathbf{c} - c_1)) \right)}{(\mathbf{c} - c_1)(\alpha(\mathbf{c} - c_1) - \lambda) \left(\alpha(\mathbf{c} - c_1)e^{\alpha b} - \lambda e^{\frac{b\lambda}{(\mathbf{c}-c_1)}} \right)^2} \\ & + \frac{\lambda e^{\alpha b - \alpha x + \frac{\lambda x}{(\mathbf{c}-c_1)}} \left(\lambda e^{\frac{b\lambda}{(\mathbf{c}-c_1)}}(\alpha b(\mathbf{c} - c_1) + \lambda(b - x) + (\mathbf{c} - c_1)) + \alpha(\mathbf{c} - c_1)e^{\alpha b}((\mathbf{c} - c_1) + \lambda x) \right)}{(\mathbf{c} - c_1)(\alpha(\mathbf{c} - c_1) - \lambda) \left(\alpha(\mathbf{c} - c_1)e^{\alpha b} - \lambda e^{\frac{b\lambda}{(\mathbf{c}-c_1)}} \right)^2}, \end{aligned}$$

for the second

$$\begin{aligned} \mathbb{P}_b(\tilde{\tau}_0^- < \infty) &= \frac{\lambda}{\alpha(\mathbf{c} - c_1 - c_2)} e^{\left[\frac{\lambda}{\mathbf{c}-c_1-c_2} - \alpha \right] b}, \\ \mathbb{E}_x[\tau_b^+; \tau_b^+ < \tau_0^-] &= \\ & \frac{e^{\alpha(b-x)} \left(e^{\frac{\lambda(b+x)}{\mathbf{c}-c_1}} (b-x)\lambda^3 + \alpha(\mathbf{c}-c_1) \left(e^{\alpha b + \frac{\lambda x}{\mathbf{c}-c_1}} \lambda((\alpha b + 2)(\mathbf{c}-c_1) + \lambda x) + e^{\alpha x} \left(\alpha^2(\mathbf{c}-c_1)^2 e^{\alpha b}(x-b) - e^{\frac{b\lambda}{\mathbf{c}-c_1}} \lambda(b\lambda + (\mathbf{c}-c_1)(\alpha x + 2)) \right) \right) \right)}{(\mathbf{c}-c_1)(\alpha(\mathbf{c}-c_1) - \lambda) \left(\alpha(\mathbf{c}-c_1)e^{\alpha b} - e^{\frac{b\lambda}{\mathbf{c}-c_1}} \lambda \right)^2}, \end{aligned}$$

and for the third

$$\begin{aligned} \mathbb{P}_x(\tau_b^+ < \tau_0^-) &= \frac{\alpha(\mathbf{c} - c_1) - \lambda e^{\frac{\lambda x}{\mathbf{c}-c_1} - \alpha x}}{\alpha(\mathbf{c} - c_1) - \lambda e^{\frac{\lambda b}{\mathbf{c}-c_1} - \alpha b}}, \\ \mathbb{E}_b[\tilde{\tau}_0^-; \tilde{\tau}_0^- < \infty] &= \frac{\mathbf{c} - c_1 - c_2 + \lambda b}{(\mathbf{c} - c_1 - c_2)((\mathbf{c} - c_1 - c_2)\alpha - \lambda)} \cdot \frac{\lambda}{\alpha(\mathbf{c} - c_1 - c_2)} e^{\left[\frac{\lambda}{\mathbf{c}-c_1-c_2} - \alpha \right] b}. \end{aligned}$$

Similar formulae can be derived in a simpler manner for the ruin probabilities both in the ratcheting and refracted case. Taking the ratio then yields

$$\mathbb{E}_x[\tau^R | \tau^R < \infty] = \frac{\mathbb{E}_x[\tau^R; \tau^R < \infty]}{\mathbb{P}_x(\tau^R < \infty)} \quad \text{and} \quad \mathbb{E}_x[\tau | \tau < \infty] = \frac{\mathbb{E}_x[\tau; \tau < \infty]}{\mathbb{P}_x(\tau < \infty)}, \quad (3.36)$$

cf. also (Gerber and Shiu, 1998, p.59). Figure 3.8 depicts the behaviour of these two quantities as a function of b for initial capital $x = 1$ and parameters $\mathbf{c} = 6$, $c_1 = c_2 = 2$, $\lambda = \alpha = 1$. One observes that the expected ruin time (given ruin occurs in finite time) is, for the same barrier, typically larger for the ratcheting case, which at first sight may look counter-intuitive, since the refraction strategy increases the drift again when the process is below b . However, this indicates that in the refraction case those sample paths that do not lead to ruin quickly, will more likely escape ruin also later, so the conditioning on the event of ruin is essential here.

Finally, we compare these properties of sample paths for the refracting and ratcheting strategies when the respective optimal barrier is chosen. Figure 3.9 shows the probability to reach the optimal barrier before ruin as well as the expected time to reach the optimal barrier, given that it is reached before ruin for the two

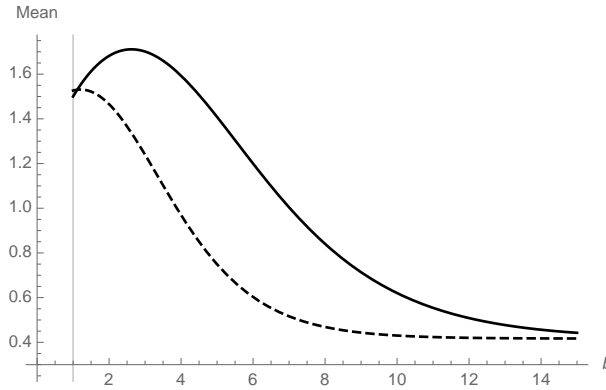


Figure 3.8: The expected time to ruin, given that it occurs, for the ratcheting (solid line) and the refracted (dashed line) strategies as a function of barrier level b ($x = 1$, $c = 6$, $c_1 = c_2 = 2$, $\lambda = \alpha = 1$).

strategies, as a function of initial capital x . Note that for the used parameters, the respective optimal barrier levels are $b^R = 4.604602$ and $b^* = 2.723496$. One observes that despite the higher value of b^R , the probability to reach that level (and hence the probability to increase the dividend rate) is not much less than for the respective refraction strategy, whereas the expected time to get there roughly doubles.

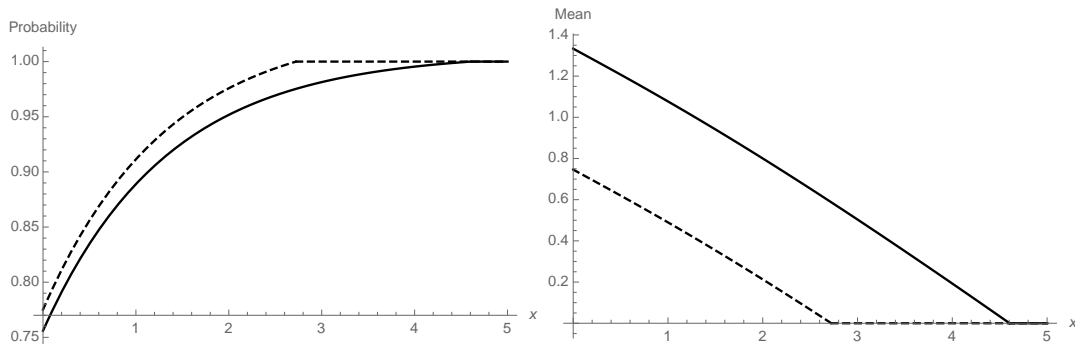


Figure 3.9: The probability of hitting the optimal barrier before ruin (left) and the expected hitting time of the optimal barrier, given that it happens before ruin (right) as a function of initial capital x , for ratcheting (solid line) and refracting (dashed line), $c = 6$, $c_1 = c_2 = 2$, $\lambda = \alpha = 1$.

Remark 3.5.1. Observe that, as a corollary of the first formula in Example 3.5.1, by taking the limit to the diffusion process (3.10) with drift μ and variance σ^2 (using the parametrization (3.35)) as well as taking $b \rightarrow \infty$, one retrieves the simple expression

$$\mathbb{E}_x(\tau_0^- | \tau_0^- < \infty) = \frac{x}{\mu}.$$

Since this formula is interesting in its own right, and seems not to have been considered in actuarial circles before, we derive it here also directly using an alternative approach. Consider $S_t = x - Y_t = -\mu t - \sigma B_t$, with drift $-\mu$. By exponential tilting

by θ we have that in the new measure $\tilde{\mathbb{P}}$ the process S has drift $-\mu + \theta\sigma^2$ and the likelihood ratio

$$\exp(\theta S_t - t\psi(\theta)) = \exp(\theta S_t - t(-\mu + \theta\sigma^2)\theta - t\sigma^2\theta^2/2)$$

is a martingale wrt $\tilde{\mathbb{P}}_x$, the law of Y_t starting at zero, where ψ is the Laplace exponent given as in the beginning of Section 3.2 (here with respect to $\tilde{\mathbb{P}}_x$). Inserting now $\theta = \mu/\sigma^2$ simplifies to zero drift and

$$\exp(\theta S_t - t\psi(\theta)) = \exp((\mu/\sigma^2)S_t - t\mu^2/(2\sigma^2)).$$

Optional stopping holds for this martingale if and only if the stopping times are finite (see e.g. (Asmussen and Albrecher, 2010, Ch.IV.4)), and the time of ruin τ_0^- is such a time, since under $\tilde{\mathbb{P}}_x$ the drift is zero. Hence we have

$$\tilde{\mathbb{E}}_x(\exp((\mu/\sigma^2)S_{\tau_0^-} - \tau_0^- \mu^2/(2\sigma^2))) = 1.$$

But $S_{\tau_0^-} = x$, and

$$\exp((\mu/\sigma^2)x - \tau_0^- \mu^2/(2\sigma^2))$$

is bounded around any finite neighbourhood of μ , hence uniformly integrable for any sequence $\mu_n \rightarrow \mu$ so we may take the derivative with respect to μ and get

$$\tilde{\mathbb{E}}_x \left(\left\{ \frac{x}{\sigma^2} - \frac{\tau_0^- \mu}{\sigma^2} \right\} \exp((\mu/\sigma^2)x - \tau_0^- \mu^2/(2\sigma^2)) \right) = 0.$$

This translates in the original measure to

$$\mathbb{E}_x \left(\left\{ \frac{x}{\sigma^2} - \frac{\tau_0^- \mu}{\sigma^2} \right\}; \tau_0^- < \infty \right) = 0,$$

and a rearrangement yields indeed

$$\frac{x}{\mu} = \frac{\mathbb{E}_x(\tau_0^-; \tau_0^- < \infty)}{\mathbb{P}_x(\tau_0^- < \infty)} = \mathbb{E}_x(\tau_0^- | \tau_0^- < \infty).$$

3.6 Conclusion and Future Research

In this paper we considered a ratcheting dividend strategy in an insurance risk theory context, where the dividend rate can be raised once during the lifetime of the surplus process. We derived analytical formulas for the expected discounted dividend payments until ruin for a general Lévy risk model, and refined the results for a diffusion approximation and a compound Poisson model with hyper-exponential claims. The numerical illustrations indicate that the performance of such a ratcheting strategy is in fact not far behind the optimal refraction strategy, and also in terms of expected ruin time the resulting performance seems rather competitive.

There are many possible directions for extensions and generalizations from here. In a future paper we will consider the case of multiple barriers, where the ratcheting strategy will mean a gradual increase of the dividend rate. Another question

of interest is to analytically show that the performance of the ratcheting strategy is monotone in the choice of the dividend rate increase c_2 at the switching time. Finally, to solve the general stochastic control problem of identifying the optimal ratcheting strategy (which possibly leads to a continuous function $c(x)$ as a function of first hitting of the surplus level x) will be an interesting challenge for future research.

Chapter 4

Efficient simulation of ruin probabilities when claims are mixtures of heavy and light tails

This chapter is based on the following article, currently submitted for publication:

Albrecher, H., Bladt, M., & Vatamidou, E. (2020). Efficient simulation of ruin probabilities when claims are mixtures of heavy and light tails. Preprint, University of Lausanne.

Abstract

We consider the classical Cramér-Lundberg risk model with claim sizes that are mixtures of phase-type and subexponential variables. Exploiting a specific geometric compound representation, we propose control variate techniques to efficiently simulate the ruin probability in this situation. The resulting estimators perform well for both small and large initial capital. We quantify the variance reduction as well as the efficiency gain of our method over another fast standard technique based on the classical Pollaczek-Khinchine formula. We provide a numerical example to illustrate the performance, and show that for more time-/consuming conditional Monte Carlo techniques, the new series representation also does not compare unfavorably to the one based on the Pollaczek-Khinchine formula.

4.1 Introduction

The study of ruin probabilities for insurance risk models is a classical topic in applied probability, see e.g. Rolski et al. (1999). Explicit formulas for ruin probabilities are available only in specific situations. One such instance is the classical Cramér-Lundberg risk model when claim sizes are of phase-type, see e.g. Asmussen and Albrecher (2010) for more details. However, the tail of such phase-type distributions is exponentially bounded Neuts (1994), whereas insurance data often suggest heavy-

tails Albrecher et al. (2017). In the presence of heavy-tails one then typically has to resort to approximations or simulations, and to achieve accuracy for either of the two can be challenging. While highly efficient simulation techniques for ruin probabilities for exponentially bounded claims are available for a long time already (e.g. using Lundberg conjugation (Asmussen and Albrecher, 2010, Ch.XV)), the field of efficient simulation for heavy-tails has only advanced significantly in more recent years and is an active field of research (cf. Asmussen and Kortschak (2015); Ghamami and Ross (2012); Juneja (2007); Nguyen and Robert (2014) and Asmussen and Glynn (2007) for an overview).

Among the many possible modelling approaches for insurance claim sizes, in this paper we will be interested in mixture models, where with a certain probability ϵ a new claim is of a heavy-tailed type and with probability $1 - \epsilon$ it is of a certain light-tailed type. Such a co-existence of heavy and light tails is very intuitive in practice, see e.g. Lee et al. (2012); Tzougas et al. (2014). For small ϵ , Vatamidou et al. (2013) used a perturbation approach to devise a numerical approximation scheme for the determination of ruin probabilities in the presence of heavy-tails in the spirit of corrected phase-type approximations. Their approach relied on an alternative representation of the Pollaczek-Khinchine (PK) formula that converges more quickly as $\epsilon \rightarrow 0$, see also Geiger and Adekpedjou (2019). Inspired by this approach, in this paper we want to study the potential of such an alternative representation for general mixture models and not necessarily small ϵ . The focus here will be to see whether large claim approximations can be used more efficiently as control variates in a simulation procedure than for algorithms based on the classical PK formula. We will show both theoretically and in a numerical implementation that this is indeed the case. The results in principle apply to any situation where claim sizes are a mixture between a tractable light-tailed and a heavy-tailed distribution for which the convolution of the two can be calculated explicitly. Moreover, even if the latter convolution can not be evaluated explicitly, the series representation can be advantageous.

We will also study the performance of the alternative series representation for a conditional Monte Carlo method developed by Asmussen & Kroese Asmussen and Kroese (2006). The latter can be applied to the PK formula and leads to a significant reduction of variance for the ruin probability estimator, but at a considerable additional computational cost. It will turn out that for this case, our series representation has no significant advantage over the classical PK approach, but the performance is not worse either.

The rest of the paper is organised as follows. Section 4.2 describes the risk model based on the mixture of light- and heavy-tailed claims and provides some preliminaries. In Section 4.3, we then construct a new control variate estimator for the ruin probability based on subexponential properties, which can exploit the advantage of exact ruin probability formulas for the light-tailed component in the mixture. We provide error bounds, investigate the tail behaviour, and quantify the resulting variance reduction when using the control variates, as well as the advantage of our approach to the analogous one based on the PK formula. We also consider the introduction of this alternative series representation for a conditional Monte Carlo framework in the spirit of Asmussen and Kroese (2006). In Section

4.4, we then perform numerical experiments and analyse the results. Finally, we conclude in Section 4.5.

4.2 Model description and preliminaries

Consider the classical Cramér-Lundberg risk model for the surplus process of an insurance portfolio. The premium inflow is assumed at a constant rate (w.l.o.g. 1 per unit time) and claims arrive according to a homogeneous Poisson process $\{N(t)\}_{t \geq 0}$ with rate λ . The claim sizes $U_k \stackrel{\mathcal{D}}{=} U$ are i.i.d. with common distribution function G , and are independent of $\{N(t)\}$. If u is the initial capital, the surplus at time t is then given by

$$R(t) = u + t - \sum_{k=1}^{N(t)} U_k.$$

We also define the claim surplus process $S(t) = u - R(t)$ and its maximum $M = \sup_{0 \leq t < \infty} S(t)$. The probability $\psi(u)$ of ultimate ruin is then

$$\psi(u) = \mathbb{P}(M > u). \quad (4.1)$$

In addition, we assume that the safety loading condition $\rho = \lambda \mathbb{E}U < 1$ holds and thus the well-known Pollaczek-Khinchine (PK) formula

$$1 - \psi(u) = (1 - \rho) \sum_{k=0}^{\infty} \rho^k (G^e)^{*k}(u) \quad (4.2)$$

can be used for the evaluation of the ruin probability. Here $G^e(u) = \int_0^u (1 - G(x)) / \mathbb{E}U$ is the distribution function of the stationary excess claim size U^e , see e.g. Asmussen and Albrecher (2010).

In this paper, we assume that claim sizes are of a mixture type. Concretely, U is phase-type with probability $1 - \epsilon$ and heavy-tailed with probability ϵ , where $\epsilon \in (0, 1)$. The phase-type claim sizes $B_k \stackrel{\mathcal{D}}{=} B$ and the heavy-tailed claim sizes $C_k \stackrel{\mathcal{D}}{=} C$ are both assumed to have finite means μ_B and μ_C , respectively. Denote by $\tilde{G}^e(s)$, $\tilde{F}_p^e(s)$, and $\tilde{F}_h^e(s)$ the Laplace transforms of the stationary excess claim sizes $U_k^e \stackrel{\mathcal{D}}{=} U^e$, $B_k^e \stackrel{\mathcal{D}}{=} B^e$, and $C_k^e \stackrel{\mathcal{D}}{=} C^e$, respectively. Moreover, we set $\delta := \lambda \mu_B$ and $\theta := \lambda \mu_C$, which means that the phase-type and heavy-tailed claims are responsible for expected aggregate claim size $(1 - \epsilon)\delta$ and $\epsilon\theta$ per unit time, respectively. The expected overall aggregate claim size is then given by $\rho = (1 - \epsilon)\delta + \epsilon\theta$. In terms of Laplace transforms, the Pollaczek-Khinchine formula can be written as

$$\mathbb{E}e^{-sM} = (1 - \rho) \sum_{k=0}^{\infty} \rho^k (\tilde{G}^e(s))^k = \frac{1 - \rho}{1 - \rho \tilde{G}^e(s)} = \frac{1 - (1 - \epsilon)\delta - \epsilon\theta}{1 - (1 - \epsilon)\delta \tilde{F}_p^e(s) - \epsilon\theta \tilde{F}_h^e(s)}. \quad (4.3)$$

Using representation (4.3), it was shown in Vatamidou et al. (2013) that $\psi(u)$ can be expressed as a series expansion involving the ruin probability of a risk process

with purely phase-type claim sizes (base model). One easy way to establish a phase-type base model is by simply considering that $G(x) = (1 - \epsilon)F_p(x) + \epsilon$, $x \geq 0$, i.e. discard all heavy-tailed claim sizes. This base model, for which the claim size distribution has an atom at zero, is equivalent to the compound Poisson risk model in which claims arrive at rate $(1 - \epsilon)\lambda$ and follow the distribution of B . We denote by M^\bullet the supremum of its corresponding claim surplus process and we set $\rho^\bullet = (1 - \epsilon)\delta$. The PK formula for this base model takes the form

$$\mathbb{E}e^{-sM^\bullet} = \frac{1 - \rho^\bullet}{1 - \rho^\bullet \tilde{F}_p^e(s)}. \quad (4.4)$$

We denote by $\psi^\bullet(u)$ the phase-type approximation of $\psi(u)$ that is obtained when we apply Laplace inversion to (4.4). The following series expansion of $\psi(u)$ for the general risk process was shown in (Vatamidou et al., 2013, Th.1). In order to keep this paper self-contained, we repeat the short proof here in the present notation.

Theorem 4.2.1 (Vatamidou et al. (2013)). *We have*

$$\psi(u) = \frac{1 - \rho}{1 - \rho^\bullet} \psi^\bullet(u) + \frac{1 - \rho}{1 - \rho^\bullet} \sum_{k=1}^{\infty} \left(\frac{\epsilon\theta}{1 - \rho^\bullet} \right)^k \mathcal{A}_k(u), \quad (4.5)$$

where $\mathcal{A}_k(u) = \mathbb{P}(M_0^\bullet + M_1^\bullet + \dots + M_k^\bullet + C_1^e + \dots + C_k^e > u)$ and $M_k^\bullet \stackrel{\mathcal{D}}{=} M^\bullet$. This expansion converges for all values of u .

Proof. It can easily be derived that $U^e = IB_k^e + (1 - I)C_k^e$, where $I \sim \text{Bernoulli}(\rho^\bullet / (\rho^\bullet + \epsilon\theta))$. Therefore $\tilde{G}^e(s) = \frac{\rho^\bullet}{\rho^\bullet + \epsilon\theta} \tilde{F}_p^e(s) + \frac{\epsilon\theta}{\rho^\bullet + \epsilon\theta} \tilde{F}_h^e(s)$, and we find by virtue of the binomial identity

$$(\tilde{G}^e(s))^\ell = \frac{1}{(\rho^\bullet + \epsilon\theta)^\ell} \sum_{k=0}^{\ell} \binom{\ell}{k} (\rho^\bullet)^{\ell-k} (\tilde{F}_p^e(s))^{\ell-k} (\epsilon\theta)^k (\tilde{F}_h^e(s))^k.$$

Combining (4.3), (4.4), we get

$$\begin{aligned} \mathbb{E}e^{-sM} &= (1 - \rho^\bullet - \epsilon\theta) \sum_{\ell=0}^{\infty} \sum_{k=0}^{\ell} \binom{\ell}{k} (\rho^\bullet)^{\ell-k} (\tilde{F}_p^e(s))^{\ell-k} (\epsilon\theta)^k (\tilde{F}_h^e(s))^k \\ &= (1 - \rho^\bullet - \epsilon\theta) \sum_{k=0}^{\infty} (\epsilon\theta)^k (\tilde{F}_h^e(s))^k \sum_{\ell=k}^{\infty} \binom{\ell}{k} (\rho^\bullet)^{\ell-k} (\tilde{F}_p^e(s))^{\ell-k} \\ &= (1 - \rho^\bullet - \epsilon\theta) \sum_{k=0}^{\infty} (\epsilon\theta)^k (\tilde{F}_h^e(s))^k \frac{1}{(1 - \rho^\bullet \tilde{F}_p^e(s))^{k+1}} \\ &= (1 - \rho^\bullet - \epsilon\theta) \sum_{k=0}^{\infty} (\epsilon\theta)^k (\tilde{F}_h^e(s))^k \frac{1}{(1 - \rho^\bullet)^{k+1}} \left(\mathbb{E}e^{-sM^\bullet} \right)^{k+1} \\ &= \frac{1 - \rho}{1 - \rho^\bullet} \sum_{k=0}^{\infty} \left(\frac{\epsilon\theta}{1 - \rho^\bullet} \right)^k (\tilde{F}_h^e(s))^k \left(\mathbb{E}e^{-sM^\bullet} \right)^{k+1}. \end{aligned}$$

We obtain the provided series expansion for $\psi(u)$ via Laplace inversion and using $\psi^\bullet(u) = \mathbb{P}(M_0^\bullet > u)$. The convergence is granted by $|\mathbb{E}e^{-sM^\bullet}| \leq 1$ and $|\tilde{F}_h^e(s)| \leq 1$, while $\epsilon\theta < 1 - \rho^\bullet$ due to the stability condition $\rho < 1$. \square

Theorem 4.2.1 provides an alternative interpretation for M , i.e. $M \stackrel{\mathcal{D}}{=} \sum_{k=0}^N (M_k^\bullet + C_k^e)$, where $C_0^e := 0$ and N is a geometric random variable $N \sim \text{Geom}\left(\frac{1-\rho}{1-\rho^\bullet}\right)$. Note that in addition to $k = 0$, for various subexponential random variables C_k^e also the term for $k = 1$ in (4.5) can be calculated explicitly, so that

$$\begin{aligned} \psi(u) &= \underbrace{\frac{1-\rho}{1-\rho^\bullet} \psi^\bullet(u) + \frac{1-\rho}{1-\rho^\bullet} \frac{\epsilon\theta}{1-\rho^\bullet} \mathbb{P}(M_0^\bullet + M_1^\bullet + C_1^e > u)}_{\text{explicit}} \\ &\quad + \frac{1-\rho}{1-\rho^\bullet} \sum_{k=2}^{\infty} \left(\frac{\epsilon\theta}{1-\rho^\bullet}\right)^k \mathcal{A}_k(u). \end{aligned}$$

Thus, to approximate $\psi(u)$, we only need to have an estimate for

$$\begin{aligned} \varphi(u) &:= \frac{1-\rho}{1-\rho^\bullet} \sum_{k=2}^{\infty} \left(\frac{\epsilon\theta}{1-\rho^\bullet}\right)^k \mathcal{A}_k(u) = \left(\frac{\epsilon\theta}{1-\rho^\bullet}\right)^2 \mathbb{E}\mathcal{A}_{N+2}(u) \\ &= \left(\frac{\epsilon\theta}{1-\rho^\bullet}\right)^2 \mathbb{P}(M_0^\bullet + M_1^\bullet + \dots + M_{N+2}^\bullet + C_1^e + \dots + C_{N+2}^e > u), \end{aligned} \quad (4.6)$$

which we want to approximate by simulating the tail of

$$V \stackrel{\mathcal{D}}{=} M_0^\bullet + M_1^\bullet + C_1^e + \sum_{k=2}^{N+2} (M_k^\bullet + C_k^e), \quad (4.7)$$

with $N \sim \text{Geom}\left(\frac{1-\rho}{1-\rho^\bullet}\right)$.

Using the above representation, we propose in Section 4.3 efficient variance reduction techniques for this simulation based on suitably chosen control variates.

4.3 Control variate techniques

Let $Z(u)$ be the random variable we must simulate in order to calculate its expectation $\varphi(u) = \mathbb{E}Z(u)$. The idea of a control variate is to use another random variable $W(u)$, which has a known expectation $\mathbb{E}W(u)$ and is strongly correlated with $Z(u)$. Thus, the deviation of the simulated from the exact value of $W(u)$ may be used for improving the simulation accuracy for $Z(u)$. If $(Z_i(u), W_i(u))$, $i = 1, 2, \dots, \kappa$, are independent copies of $(Z(u), W(u))$, then an efficient control variate estimator is defined as

$$\hat{\varphi}_\kappa(u) := \hat{z}_\kappa(u) + \hat{\alpha}_\kappa(\hat{w}_\kappa(u) - \mathbb{E}W(u)), \quad (4.8)$$

where

$$\hat{z}_\kappa(u) = \frac{\sum_{i=1}^{\kappa} Z_i(u)}{\kappa}, \hat{w}_\kappa(u) = \frac{\sum_{i=1}^{\kappa} W_i(u)}{\kappa}, \quad (4.9)$$

$$\hat{\alpha}_\kappa = -\frac{\sum_{i=1}^{\kappa} (Z_i(u) - \hat{z}_\kappa(u))(W_i(u) - \hat{w}_\kappa(u))}{\sum_{i=1}^{\kappa} (W_i(u) - \hat{w}_\kappa(u))^2}. \quad (4.10)$$

Note that this choice of $\hat{\alpha}_\kappa$ based on the empirical correlation of $Z(u)$ and $W(u)$ optimizes the variance gain, see e.g. Albrecher et al. (2017); Asmussen and Glynn (2007). We assume now that the distribution of C^e belongs to the class of subexponential distributions \mathcal{S} , i.e. for any $n \in \mathbb{N}$,

$$\overline{F^{*n}}(u) \sim n\overline{F}(u), \quad \text{as } u \rightarrow \infty, \quad (4.11)$$

where $\overline{F}(u) = 1 - F(u)$, see e.g. Teugels (1975). The construction of the concrete $W(u)$ below is inspired by the following well-known asymptotic property of subexponential distributions (see e.g. (Foss et al., 2013, Cor.3.18) or (Asmussen and Albrecher, 2010, Cor.X.1.11)):

Property 4.3.1. *Let $F \in \mathcal{S}$ and let A be any distribution with a lighter tail, i.e. $\overline{A}(u) = o(\overline{F}(u))$. Then for the convolution $A * F$ of A and F we have $A * F \in \mathcal{S}$ and $\overline{(A * F)}(u) \sim \overline{F}(u)$.*

In other words, for sufficiently large u , only the maximum of the subexponential claims will substantially contribute to the probability (4.6).

4.3.1 Max of heavy-tails

It is immediately obvious from (4.6) that

$$Z(u) = \left(\frac{\epsilon\theta}{1 - \rho^\bullet} \right)^2 \mathbf{1}_{\{V > u\}}, \quad (4.12)$$

while for a fixed $n \in \mathbb{N}$, we consider the random variable

$$V_n := \max\{C_1^e, \dots, C_{N+2}^e\} \mathbf{1}_{\{N+2 \leq n\}}. \quad (4.13)$$

Definition 4.3.2. *For a fixed $n \in \mathbb{N}$, define the control variate*

$$W(u) = \left(\frac{\epsilon\theta}{1 - \rho^\bullet} \right)^2 \mathbf{1}_{\{V_n > u\}}. \quad (4.14)$$

The n th order approximation $\varphi_n(u) = \mathbb{E}W(u)$ of $\varphi(u)$ is then

$$\varphi_n(u) = \left(\frac{1 - \rho}{1 - \rho^\bullet} \right) \sum_{k=2}^n \left(\frac{\epsilon\theta}{1 - \rho^\bullet} \right)^k \mathbb{P}(\max\{C_1^e, \dots, C_k^e\} > u). \quad (4.15)$$

By construction, $\varphi_n(u)$ underestimates $\varphi(u)$. Next we collect some properties of this approximation.

Properties of the approximation

The following lower and upper bounds for the approximation error can be obtained.

Proposition 4.3.3 (Error bounds). *The error of the approximation $\varphi_n(u)$, $n \in \mathbb{N}$, is bounded from above and below as follows:*

$$\begin{aligned} \left(\frac{\epsilon\theta}{1-\rho^\bullet}\right)^{n+1} \mathcal{A}_1(u) \leq \varphi(u) - \varphi_n(u) \leq & \left(\frac{\epsilon\theta}{1-\rho^\bullet}\right)^{n+1} \\ & + \left(1 - \frac{\epsilon\theta}{1-\rho^\bullet}\right) \left(\frac{\epsilon\theta}{1-\rho^\bullet} F_h^e(u)\right)^2 \frac{1 - \left(\frac{\epsilon\theta}{1-\rho^\bullet} F_h^e(u)\right)^{n-1}}{1 - \frac{\epsilon\theta}{1-\rho^\bullet} F_h^e(u)}. \end{aligned}$$

Proof. For simplicity of notation, we set $p := \frac{\epsilon\theta}{1-\rho^\bullet}$. The error of the approximation is equal to

$$\begin{aligned} \varphi(u) - \varphi_n(u) &= (1-p) \sum_{k=2}^{\infty} p^n \mathcal{A}_k(u) - (1-p) \sum_{k=2}^n p^k \mathbb{P}(\max\{C_1^e, \dots, C_k^e\} > u) \\ &= (1-p) \sum_{k=2}^n p^k \left(\mathcal{A}_k(u) - \mathbb{P}(\max\{C_1^e, \dots, C_k^e\} > u) \right) + (1-p) \sum_{k=n+1}^{\infty} p^k \mathcal{A}_k(u). \end{aligned}$$

For the upper bound, we use $\mathbb{P}(\max\{C_1^e, \dots, C_k^e\} > u) = 1 - (F_h^e(u))^k$ and $\mathcal{A}_k(u) \leq 1$ to obtain

$$\begin{aligned} \varphi(u) - \varphi_n(u) &\leq (1-p) \sum_{k=2}^n p^k \mathbb{P}(\max\{C_1^e, \dots, C_k^e\} \leq u) + (1-p) \sum_{k=n+1}^{\infty} p^k \\ &= p^{n+1} + (1-p) \sum_{k=2}^n (p F_h^e(u))^k. \end{aligned}$$

For the lower bound, we take $\mathcal{A}_k(u) \geq \mathbb{P}(\max\{C_1^e, \dots, C_k^e\} > u)$ when $k \leq n$ and $\mathcal{A}_k(u) \geq \mathcal{A}_1(u)$ otherwise, to calculate

$$\varphi(u) - \varphi_n(u) \geq (1-p) \sum_{k=n+1}^{\infty} p^k \mathcal{A}_1(u) = p^{n+1} \mathcal{A}_1(u),$$

and the proof is complete. \square

Proposition 4.3.4 (Tail behaviour). *For $C^e \in \mathcal{S}$, the n th approximation*

$$\psi_n(u) := \frac{1-\rho}{1-\rho^\bullet} \psi^\bullet(u) + \frac{1-\rho}{1-\rho^\bullet} \frac{\epsilon\theta}{1-\rho^\bullet} \mathbb{P}(M_0^\bullet + M_1^\bullet + C_1^e > u) + \varphi_n(u)$$

of the target ruin probability $\psi(u)$ has the following tail behaviour:

$$\psi_n(u) \sim \frac{\epsilon\theta}{1-\rho} \left(1 - (n+1) \left(\frac{\epsilon\theta}{1-\rho^\bullet}\right)^n + n \left(\frac{\epsilon\theta}{1-\rho^\bullet}\right)^{n+1} \right) \overline{F}_h^e(u), \quad u \rightarrow \infty.$$

Proof. The approximation $\psi^\bullet(u)$ has a phase-type representation; therefore, it is of order $o(\overline{F}_h^e(u))$. The same holds for the tail of the distribution of $M_0^\bullet + M_1^\bullet$. Moreover, since $C^e \in \mathcal{S}$, from Property 4.3.1 we obtain $\mathbb{P}(M_0^\bullet + M_1^\bullet + C_1^e > u) \sim \overline{F}_h^e(u)$. Finally, from $\mathbb{P}(\max\{C_1^e, \dots, C_n^e\} > u) \leq \mathbb{P}(C_1^e + \dots + C_n^e > u)$ and (4.11), we deduce that $\mathbb{P}(\max\{C_1^e, \dots, C_n^e\} > u) \sim n\overline{F}_h^e(u)$, which leads to the following result by inserting these asymptotic estimates into Definition 4.3.2:

$$\begin{aligned} \psi_n(u) &\sim \left(1 - \frac{\epsilon\theta}{1 - \rho^\bullet}\right) \sum_{k=1}^n k \left(\frac{\epsilon\theta}{1 - \rho^\bullet}\right)^k \overline{F}_h^e(u) \\ &= \frac{\frac{\epsilon\theta}{1 - \rho^\bullet} \left(1 - (n+1) \left(\frac{\epsilon\theta}{1 - \rho^\bullet}\right)^n + n \left(\frac{\epsilon\theta}{1 - \rho^\bullet}\right)^{n+1}\right)}{1 - \frac{\epsilon\theta}{1 - \rho^\bullet}} \overline{F}_h^e(u) \\ &= \frac{\epsilon\theta}{1 - \rho} \left(1 - (n+1) \left(\frac{\epsilon\theta}{1 - \rho^\bullet}\right)^n + n \left(\frac{\epsilon\theta}{1 - \rho^\bullet}\right)^{n+1}\right) \overline{F}_h^e(u). \end{aligned}$$

□

Proposition 4.3.4 (in comparison with Theorem 5 in Vatamidou et al. (2013)) shows that $\psi_n(u)$ nearly captures the asymptotic behaviour of the exact ruin probability

$$\psi(u) \sim \frac{\epsilon\theta}{1 - \rho} \overline{F}_h^e(u), \quad (4.16)$$

being off by a factor $\left(1 - (n+1) \left(\frac{\epsilon\theta}{1 - \rho^\bullet}\right)^n + n \left(\frac{\epsilon\theta}{1 - \rho^\bullet}\right)^{n+1}\right) \in (0, 1)$. As expected, the tail of $\psi_n(u)$ underestimates the tail of $\psi(u)$.

Variance reduction

We consider now the bivariate simulation of i.i.d. copies of the random variables V and V_n :

$$(V^{(i)}, V_n^{(i)}), \quad i = 1, 2, \dots, \kappa. \quad (4.17)$$

For each fixed $n \in \mathbb{N}$, the estimator (4.8) takes the form

$$\hat{\varphi}_\kappa^n(u) := \hat{z}_\kappa(u) + \hat{\alpha}_\kappa(\hat{w}_\kappa(u) - \varphi_n(u)). \quad (4.18)$$

We can now establish our main result.

Theorem 4.3.5 (Variance reduction). *For each fixed $n \in \mathbb{N}$, the variance of the estimator (4.18) behaves asymptotically as*

$$\text{Var}(\hat{\varphi}_\kappa^n(u)) \sim \left(\frac{\epsilon\theta}{1 - \rho^\bullet}\right)^{n+3} \frac{1 + n \left(\frac{1-\rho}{1-\rho^\bullet}\right)}{\frac{1-\rho}{1-\rho^\bullet}} \cdot \frac{\overline{F}_h^e(u)}{\kappa}, \quad \text{as } u \rightarrow \infty \quad (4.19)$$

and satisfies

$$\frac{\text{Var}(\hat{\varphi}_\kappa^n(u))}{\text{Var}(\hat{z}_\kappa(u))} \rightarrow \left(\frac{\epsilon\theta}{1 - \rho^\bullet}\right)^{n-1} \frac{1 + n \left(\frac{1-\rho}{1-\rho^\bullet}\right)}{1 + \frac{1-\rho}{1-\rho^\bullet}}, \quad \text{as } u \rightarrow \infty. \quad (4.20)$$

Proof. Since $\mathbb{E}\hat{z}_\kappa(u) = \varphi(u)$, we know from Asmussen and Glynn (2007) that the proposed estimator has variance

$$\frac{1 - (r(u))^2}{\kappa} \text{Var}Z(u), \quad (4.21)$$

with correlation coefficient $r(u) = \text{Corr}(Z(u), W(u))$. By the definition of V_n , $\{V_n > u\} \subseteq \{V > u\}$ and consequently $\mathbb{1}_{\{V > u\}} \cdot \mathbb{1}_{\{V_n > u\}} = \mathbb{1}_{\{V_n > u\}}$. We calculate,

$$\begin{aligned} \text{Cov}(Z(u), W(u)) &= \left(\frac{\epsilon\theta}{1 - \rho^\bullet}\right)^4 \text{Cov}\left(\mathbb{1}_{\{V > u\}} \cdot \mathbb{1}_{\{V_n > u\}}\right) \\ &= \left(\frac{\epsilon\theta}{1 - \rho^\bullet}\right)^4 \left(\mathbb{E}\left(\mathbb{1}_{\{V > u\}} \mathbb{1}_{\{V_n > u\}}\right) - \mathbb{E}\mathbb{1}_{\{V > u\}} \mathbb{E}\mathbb{1}_{\{V_n > u\}}\right) \\ &= \left(\frac{\epsilon\theta}{1 - \rho^\bullet}\right)^4 \left(\mathbb{P}(V_n > u) - \mathbb{P}(V > u) \mathbb{P}(V_n > u)\right) \\ &= \left(\frac{\epsilon\theta}{1 - \rho^\bullet}\right)^4 \mathbb{P}(V_n > u) \mathbb{P}(V \leq u). \end{aligned}$$

Similarly, we find

$$\begin{aligned} \text{Var}(Z(u)) &= \left(\frac{\epsilon\theta}{1 - \rho^\bullet}\right)^4 \mathbb{P}(V > u) \mathbb{P}(V \leq u), \text{ and} \\ \text{Var}(W(u)) &= \left(\frac{\epsilon\theta}{1 - \rho^\bullet}\right)^4 \mathbb{P}(V_n > u) \mathbb{P}(V_n \leq u). \end{aligned}$$

Hence, it is immediate that

$$1 - (r(u))^2 = \frac{1 - \mathbb{P}(V_n > u) / \mathbb{P}(V > u)}{1 - \mathbb{P}(V_n > u)}. \quad (4.22)$$

Following Proposition 4.3.4, we calculate

$$\begin{aligned} \mathbb{P}(V_n > u) &\sim \left(1 - \frac{\epsilon\theta}{1 - \rho^\bullet}\right) \sum_{k=2}^n k \left(\frac{\epsilon\theta}{1 - \rho^\bullet}\right)^{k-2} \overline{F}_h^\epsilon(u) \\ &= \frac{2 - \frac{\epsilon\theta}{1 - \rho^\bullet} - (n+1) \left(\frac{\epsilon\theta}{1 - \rho^\bullet}\right)^{n-1} + n \left(\frac{\epsilon\theta}{1 - \rho^\bullet}\right)^n}{1 - \frac{\epsilon\theta}{1 - \rho^\bullet}} \overline{F}_h^\epsilon(u) \end{aligned}$$

and

$$\mathbb{P}(V > u) \sim \frac{2 - \frac{\epsilon\theta}{1 - \rho^\bullet}}{1 - \frac{\epsilon\theta}{1 - \rho^\bullet}} \overline{F}_h^\epsilon(u),$$

as $u \rightarrow \infty$. We finally obtain

$$\frac{\mathbb{P}(V_n > u)}{\mathbb{P}(V > u)} \rightarrow 1 - \frac{(n+1) \left(\frac{\epsilon\theta}{1 - \rho^\bullet}\right)^{n-1} - n \left(\frac{\epsilon\theta}{1 - \rho^\bullet}\right)^n}{2 - \frac{\epsilon\theta}{1 - \rho^\bullet}},$$

so that

$$1 - (r(u))^2 \rightarrow \frac{(n+1) \left(\frac{\epsilon\theta}{1-\rho^\bullet}\right)^{n-1} - n \left(\frac{\epsilon\theta}{1-\rho^\bullet}\right)^n}{2 - \frac{\epsilon\theta}{1-\rho^\bullet}},$$

and the statement of the theorem follows. \square

The above theorem quantifies the asymptotic variance reduction for fixed n as u increases, this reduction being arbitrarily large when n is increased sufficiently.

4.3.2 Conditional Monte Carlo

While the approach of Section 4.3.1 is the focus of this paper, for purposes of comparison and completeness we are also interested in the performance of the alternative series representation for the conditional Monte Carlo estimate and its variance reduction proposed in Asmussen and Kroese (2006). To that end, let us recap here its idea and present its application to our series representation. Define $X_0^\star = M_0^\bullet$ and $X_k = M_k^\bullet + C_k^e$, $k = 1, 2, \dots$, so that $V \stackrel{\mathcal{D}}{=} X_0^\star + \sum_{k=1}^{N+2} X_k$, where $N \sim \text{Geom}\left(\frac{1-\rho}{1-\rho^\bullet}\right)$ as before. (4.6) can then be written as

$$\varphi(u) = \left(\frac{\epsilon\theta}{1-\rho^\bullet}\right)^2 \mathbb{P}(X_0^\star + X_1 + \dots + X_{N+2} > u).$$

Note that for fixed $k \geq 1$ and $m_k := \max\{X_1, \dots, X_k\}$, we have

$$\begin{aligned} \mathbb{P}(X_0^\star + X_1 + \dots + X_k > u) &= k \mathbb{P}(S_k > u - X_0^\star, X_k = m_k) \\ &= k \mathbb{P}(X_k > m_{k-1}, X_k > u - X_0^\star - S_{k-1}) = k \mathbb{E}\bar{F}_X(m_{k-1} \vee (u - X_0^\star - S_{k-1})), \end{aligned}$$

where \bar{F}_X is the common c.c.d.f. of the X_k 's and $S_\ell = \sum_{k=1}^\ell X_k$, $S_0 = 0$. Consequently, the random variable $Z(u)$ becomes

$$Z^\star(u) = \left(\frac{\epsilon\theta}{1-\rho^\bullet}\right)^2 (N+2)\bar{F}_X(m_{N+1} \vee (u - X_0^\star - S_{N+1})).$$

We can further introduce $N\bar{F}_X(u)$ as a control variate for the number of summands (see e.g. Ghamami and Ross (2012)).

Definition 4.3.6. *We use the control variate*

$$W^\star(u) = \left(\frac{\epsilon\theta}{1-\rho^\bullet}\right)^2 (N+2)\bar{F}_X(u).$$

The resulting approximation $\varphi^\star(u) = \mathbb{E}W^\star(u)$ of $\varphi(u)$ then is

$$\varphi^\star(u) := \left(\frac{\epsilon\theta}{1-\rho^\bullet}\right)^2 \left(\frac{\epsilon\theta}{1-\rho} + 2\right)\bar{F}_X(u).$$

This control variate leads to the following Asmussen-Kroese (AK)-type estimator:

$$\hat{\psi}_\kappa^*(u) := \hat{z}_\kappa^*(u) + \hat{\alpha}_\kappa^*(\hat{w}_\kappa^*(u) - \varphi^*(u)), \quad (4.23)$$

where $\hat{z}_\kappa^*(u)$, $\hat{w}_\kappa^*(u)$, and $\hat{\alpha}_\kappa^*$ are calculated via (4.9) using $Z^*(u)$ and $W^*(u)$.

Remark 4.3.7. *An alternative approach is to set $X_1^* = M_0^\bullet + M_1^\bullet + C_1^e$ and $X_k = M_k^\bullet + C_k^e$, $k = 2, 3, \dots$ and write (4.6) as*

$$\varphi(u) = \left(\frac{\epsilon\theta}{1 - \rho^\bullet} \right)^2 \mathbb{P}(X_1^* + X_2 + \dots + X_{N+2} > u).$$

Observe that all the random variables on the right hand side of this equation are heavy-tailed and independent, but not identically distributed. Thus, using the AK estimator for non i.i.d. random variables established in Chan and Kroese (2011), we could instead construct a control variate based on the conditional Monte Carlo estimator

$$Z^*(u) = \left(\frac{\epsilon\theta}{1 - \rho^\bullet} \right)^2 \left(\overline{F}_{X_1^*}(m_{-1}^* \vee (u - S_{N+1} + X_1)) + (N+1)\overline{F}_X(m_{-(N+2)}^* \vee (u - S_N - X_0^*)) \right),$$

where $m_{-1}^* = \max\{X_2, \dots, X_{N+2}\}$ and $m_{-k}^* = \max\{X_1^*, X_2, \dots, X_{k-1}, X_{k+1}, X_{N+2}\}$.

4.3.3 Comparison with the Pollaczek-Khinchine expansion

For reference and the purpose of comparison, we also consider the estimators analogous to the ones in Sections 4.3.1 and 4.3.2 using the usual PK series expansion of the ruin probability in (4.2), which we rewrite as

$$\psi(u) = \underbrace{(1 - \rho)\rho\overline{G^e}(u)}_{\text{explicit}} + \underbrace{(1 - \rho) \sum_{k=2}^{\infty} \rho^k (1 - (G^e)^{*k}(u))}_{:=\varphi^\circ(u)}.$$

Define the random variables $N^\circ \sim \text{Geom}(1 - \rho)$,

$$V^\circ = \sum_{k=1}^{N^\circ+2} U_k^e,$$

$$V_n^\circ = \max\{U_1^e, \dots, U_{N^\circ+2}^e\} \mathbb{1}_{\{N^\circ \leq n-2\}},$$

and let $S_n^\circ = \sum_{k=1}^n U_k^e$ as well as $m_k^\circ = \max\{U_1^e, \dots, U_k^e\}$. With this notation, the following equations define the analogous control variate estimators of $\varphi^\circ(u)$:

$$\begin{aligned} Z^\circ(u) &= \rho^2 \mathbb{1}_{\{V^\circ > u\}} & Z^{\circ,*}(u) &= \rho^2 (N^\circ + 2) \overline{G^e}(m_{N^\circ+1}^\circ \vee (u - S_{N^\circ+1}^\circ)) \\ W^\circ(u) &= \rho^2 \mathbb{1}_{\{V_n^\circ > u\}} & W^{\circ,*}(u) &= \rho^2 (N^\circ + 2) \overline{G^e}(u), \end{aligned}$$

and the associated empirical estimator

$$\hat{\varphi}_\kappa^{\circ,n}(u) := \hat{z}_\kappa^\circ(u) + \hat{\alpha}_\kappa^\circ(\hat{w}_\kappa^\circ(u) - \varphi_n^\circ(u)). \quad (4.24)$$

Observe now that the distributional behaviour of the variable U^e is slightly different than that of C_k^e . Recall that $U^e = IB_k^e + (1-I)C_k^e$, where $I \sim \text{Bernoulli}(\rho^\bullet / (\rho^\bullet + \epsilon\theta))$. Hence,

$$\mathbb{P}(U^e > u) = \frac{\rho^\bullet}{\rho^\bullet + \epsilon\theta} \mathbb{P}(B^e > u) + \frac{\epsilon\theta}{\rho^\bullet + \epsilon\theta} \mathbb{P}(C^e > u) \sim \frac{\epsilon\theta}{\rho^\bullet + \epsilon\theta} \mathbb{P}(C^e > u),$$

as $u \rightarrow \infty$. Moreover, since C^e is subexponential, the above relation implies that U^e is subexponential as well. Consequently,

$$\mathbb{P}(m_k^\circ > u) \sim k \frac{\epsilon\theta}{\rho^\bullet + \epsilon\theta} \mathbb{P}(C^e > u) = k \frac{\epsilon\theta}{\rho^\bullet + \epsilon\theta} \overline{F}_h^e(u).$$

Using the above asymptotic and following the proof of Theorem 4.3.5, we obtain the next result.

Theorem 4.3.8. *For each fixed $n \in \mathbb{N}$, the variance of the estimator (4.24) behaves asymptotically as*

$$\text{Var}(\hat{\varphi}_\kappa^{\circ,n}(u)) \sim \rho^{n+3} \frac{1+n(1-\rho)}{1-\rho} \cdot \frac{\epsilon\theta}{\rho} \cdot \frac{\overline{F}_h^e(u)}{\kappa}, \quad \text{as } u \rightarrow \infty,$$

and satisfies

$$\frac{\text{Var}(\hat{\varphi}_\kappa^{\circ,n}(u))}{\text{Var}(\hat{z}_\kappa^\circ(u))} \rightarrow \rho^{n-1} \frac{1+n(1-\rho)}{1+(1-\rho)}, \quad \text{as } u \rightarrow \infty. \quad (4.25)$$

It follows that we can compare the asymptotic effect on the variance between the two different series expansions for the ruin probability, as well as the effect on the proportion of variance reduction due to the use of control variates:

Corollary 4.3.9. *For each fixed $n \in \mathbb{N}$, the following relations hold:*

$$\frac{\text{Var}(\hat{\varphi}_\kappa^n(u))}{\text{Var}(\hat{\varphi}_\kappa^{\circ,n}(u))} \sim \left[\frac{\epsilon\theta}{1-\rho^\bullet} / \rho \right]^{n+2} \frac{1+n\left(\frac{1-\rho}{1-\rho^\bullet}\right)}{1+n(1-\rho)}, \quad \text{as } u \rightarrow \infty, \quad (4.26)$$

and

$$\left[\frac{\text{Var}(\hat{\varphi}_\kappa^n(u))}{\text{Var}(\hat{z}_\kappa(u))} \right] \cdot \left[\frac{\text{Var}(\hat{\varphi}_\kappa^{\circ,n}(u))}{\text{Var}(\hat{z}_\kappa^\circ(u))} \right]^{-1} \rightarrow \left[\frac{\epsilon\theta}{1-\rho^\bullet} / \rho \right]^{n-1} \frac{1+n\left(\frac{1-\rho}{1-\rho^\bullet}\right)}{1+n(1-\rho)} \cdot \frac{1+(1-\rho)}{1+\left(\frac{1-\rho}{1-\rho^\bullet}\right)},$$

as $u \rightarrow \infty$.

Notice that the inequality $\frac{\epsilon\theta}{1-\rho^\bullet} < \rho$ is actually equivalent to the net profit condition $\rho < 1$. As a consequence, the terms involving powers of $\frac{\epsilon\theta}{1-\rho^\bullet} / \rho < 1$ in the above result guarantee (for large n) a better performance of our new series representation over the classical Pollaczek-Khinchine expansion.

Remark 4.3.10. *Note that the term ρ^\bullet also depends on ϵ and $\frac{\epsilon\theta}{1-\rho^\bullet}$ is not necessarily monotone in ϵ . Correspondingly, it will depend not only on the fraction ϵ of heavy-tailed claims in the mixture but also on the value of all other involved parameters how large an improvement our series representation provides.*

4.4 Numerical experiments

In this section, we test and numerically illustrate the efficiency of our proposed technique, and compare it to the analogous classical simulation techniques based on the PK representation (4.2) (see also (Asmussen and Albrecher, 2010, Ch.XV.2)).

To perform our numerical experiments, we need to specify a mixture claim size distribution for which the distributions of $M_0^\bullet + M_1^\bullet + C_1^e$ and $M_1^\bullet + C_1^e$ can be evaluated explicitly; note that the second convolution is only required for the AK estimator.

4.4.1 Mixture of exponential and Pareto claim sizes

For the phase-type claim sizes we choose an exponential distribution with rate μ , i.e. $\overline{F}_p(u) = \overline{F}_p^e(u) = e^{-\mu u}$, and $\mu_B = 1/\mu$. For the heavy-tailed claim sizes we consider a shifted Pareto distribution with shape parameter $a > 1$ and scale $b > 0$, i.e. $\overline{F}_h(u) = (1 + u/b)^{-a}$ and $\overline{F}_h^e(u) = (1 + u/b)^{-(a-1)}$, $u \geq 0$, with $\mu_C = b/(a-1)$.

The two tail probabilities of the aforementioned sums of variables are explicitly available. For instance, for $\mu = 3$, $a = 2$, $b = 1$, $\epsilon = 0.1$ and $\rho = 0.99$ they are given by

$$\begin{aligned} \mathbb{P}(M_0^\bullet + M_1^\bullet + C_1^e > u) &= \frac{1}{25.600.000.000(1+u)} \times \left(-41200(223427 + 264627u) \right. \\ &\quad \left. + 297(1+u) \left(400e^{-309u/400}(292973 + 91773u) \right. \right. \\ &\quad \left. \left. + 31827(1691 + 891u)e^{-309(1+u)/400} \left(\text{Ei}\left(\frac{309(1+u)}{400}\right) - \text{Ei}\left(\frac{309}{400}\right) \right) \right) \right) \\ \mathbb{P}(X_1 > u) &= \frac{103}{400(1+u)} + \frac{297}{320000} \times \left(800e^{-309u/400} \right. \\ &\quad \left. + 618e^{-309(1+u)/400} \left(\text{Ei}\left(\frac{309(1+u)}{400}\right) - \text{Ei}\left(\frac{309}{400}\right) \right) \right), \end{aligned} \quad (4.27)$$

where $\text{Ei}(z) = -\int_{-z}^{\infty} \frac{e^{-t}}{t} dt$ is the exponential integral. For all other parameters that we consider, analogous formulas are used. Finally, we calculate $\mathbb{P}(\max\{C_1^e, \dots, C_k^e\} > u) = 1 - (1 - (1 + u/b)^{-(a-1)})^k$.

4.4.2 Parameters

In all our experiments, we fixed $\mu = 3$ and $b = 1$, while we considered various combinations for the remaining parameters. Motivated by Vatamidou et al. (2013), we focused mainly on the cases $\rho \in \{0.9, 0.99, 0.999\}$, where simulations involving heavy-tails can be considerably problematic (known as the heavy-traffic regime in the related queueing context, cf. Asmussen (2003)) and where the first two terms of (4.15) are known to be unable to close the gap between the approximation and

the exact ruin probability even for values of $\epsilon = 0.1$. For the remaining parameters we tested $\epsilon \in \{0.1, 0.7\}$ and $a \in \{2, 3, 4\}$.

4.4.3 Results

In all the presented examples, the order of $\psi_n(u)$ is equal to $n = 100$ and the number of simulations is $\kappa = 10,000$.

We plot in Figure 4.1 the simulated ruin probability that is obtained using the Monte Carlo estimator (4.12) together with the heavy-tail approximation (4.16). The dashed black lines depict the error bounds in Proposition 4.3.3. We observe in both graphs that the lower bound converges to the heavy-tail approximation (4.16) as $u \rightarrow \infty$. This behaviour is observed for any n and is in accordance with theory. A similar statement holds for any u as $n \rightarrow \infty$. Further empirical tests show that this convergence in n is remarkably fast. However, one cannot draw a safe conclusion for which choice of parameters the lower bound is below or above the heavy-tail approximation. Finally, we observe in the left graph that the upper bound is not very tight, as expected by Proposition 4.3.3, since the chosen parameters give $\epsilon\theta/(1 - \rho^\bullet) = 0.875$. The bound becomes tighter in the right graph, where $\epsilon\theta/(1 - \rho^\bullet) = 0.25$.

From this point on, let us fix the parameters to $a = 2$, $\epsilon = 0.1$, and $\rho = 0.99$ to allow for comparability between Figures 4.2 and 4.3. Moreover, we use a log-log scale. In Figure 4.2, we plot the MC estimate (4.12) (blue solid line) together with the control variate extension (4.18) (black dashed line) against the heavy-tail approximation (4.16). We observe that the control variate technique outperforms the crude estimate (4.12) across the entire range of u (see the variance plot on the right). Figure 4.2 also compares the simulation results with the ones based on the classical PK formula described in Section 4.3.3. For the crude version, the latter are competitive for large u , but perform worse for small u . However, for the control variate, our new approach is always significantly and convincingly better. This nicely illustrates the theoretical asymptotic results of Section 4.3: note that for the present choice of parameters the control variate asymptotically reduces the variance by a factor 0.09 (the constant on the right-hand side of (4.20)) for our series representation, to be compared with 0.73 for the analogous constant on the right-hand side of (4.25) for the PK representation. Related to that, the constant on the right-hand side of (4.26) in Corollary 4.3.9 is 0.12, which means that our series representation reduces the asymptotic variance by almost 90%, when control variates are used in both cases.

In Figure 4.3, we plot the simulated ruin probability with the AK estimator (blue solid line) and its control variate extension in Section 4.3.2 against the heavy-tail approximation (4.16), both for the PK and our new series expansion. One recognises that the asymptotic behaviour according to (4.16) (red dotted line) is recovered for all four estimators for sufficiently large u . The right graph illustrates that the introduction of the control variate is a significant improvement in terms of variance reduction for both the PK and our series, and that the two perform similarly. The overall variance is much lower than for the method underlying Figure

4.2. However, one should keep in mind that in terms of computation time the AK estimator in Figure 4.3 is much more time-consuming (about 20–50 times in our implementations), as the integrals (4.27) have to be evaluated κ times, whereas for the method in Figure 4.2 only once for the explicit term in front.

For large ρ , the number of summands tends to be large, and the results of the presented simulations suggest that the approximation

$$\mathbb{P}(M_0^\bullet + M_1^\bullet + \dots + M_k^\bullet + C_1^e + \dots + C_k^e > u) \approx \mathbb{P}(\max\{C_1^e, \dots, C_k^e\} > u) \quad (4.28)$$

is better than the one employed using the usual PK series expansion

$$\mathbb{P}(U_1^e + \dots + U_k^e > u) \approx \mathbb{P}(\max\{U_1^e, \dots, U_k^e\} > u). \quad (4.29)$$

Intuitively, the latter is comprised of mixtures of heavy-tailed and light-tailed variables, and hence the number of heavy-tailed variables is thinned down, which is a drawback that our new method does not have. This is further supported by the plot in the left panel of Figure 4.4, where the empirical correlations between the control variates are given. Concretely, when simulating from (4.29), only $100 \cdot \frac{\epsilon\theta}{\rho}\%$ of our U_k^e 's will actually be heavy-tailed and thus one loses too much information from the original presence of heavy-tailed C_k^e 's, in contrast to (4.28) where only the light tails are omitted and all heavy-tails are kept. Consequently, the new control variate is much more efficient, cf. the factors $\epsilon\theta/\rho$ in Corollary 4.3.9. In contrast, for the AK estimator the control variate does not significantly differ for the two series representations, and therefore – while the control variate itself is a huge improvement over the crude estimate (cf. Figure 4.3 (right)) – there is no improvement from using the alternative representation.

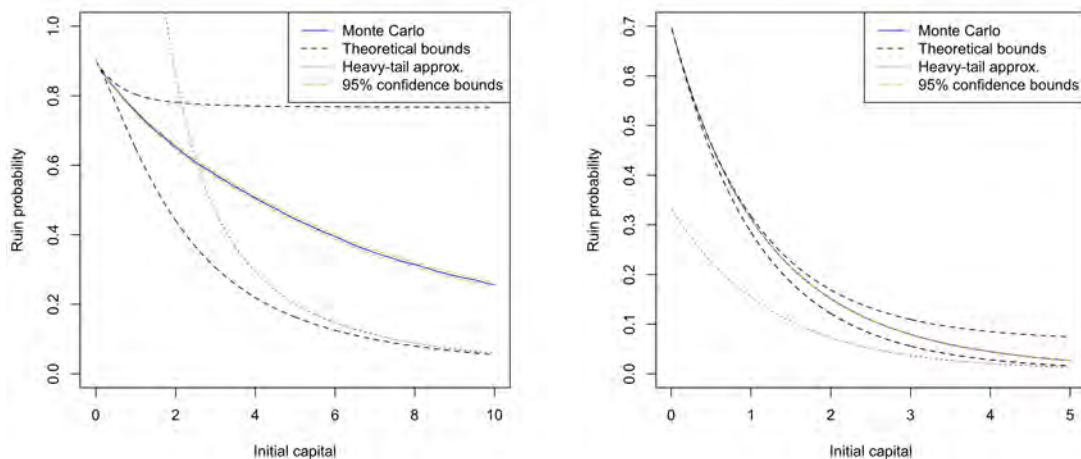


Figure 4.1: The simulated ruin probability with MC estimator (4.12) (blue solid line together with its 95% confidence interval in orange) and the heavy-tail approximation (red dotted line), as a function of the initial capital u . The black dashed lines represent the error bounds in Theorem 4.3.3. Model parameters: $a = 3$ (both) and $\{\epsilon, \rho\} = \{0.7, 0.9\}$ (left) or $\{\epsilon, \rho\} = \{0.1, 0.7\}$ (right).

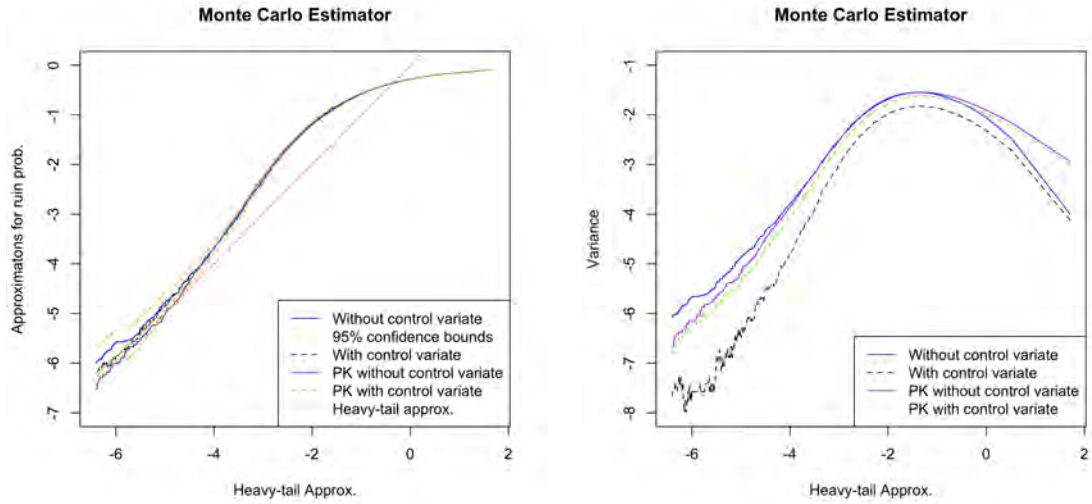


Figure 4.2: The simulated ruin probability with MC estimator (4.12) (blue solid line) together with its 95% (blue solid line) and the control variate extension (4.18) (black dashed line) against the heavy-tail approximation. The corresponding estimates based on the PK formula are depicted in pink and dashed green. Model parameters: $a = 2$, $\epsilon = 0.1$, and $\rho = 0.99$. The respective empirical variances are presented on the right.

4.5 Conclusion

In this paper, we introduced an alternative series expansion for the PK formula in the Cramér-Lundberg model for the case when claims are mixtures of distributions with heavy and light tails. We showed that this can give rise to a significant improvement of simulation algorithms based on this series, both for large and small values of initial capital.

When using the AK conditional Monte Carlo technique, the new series representation performs similarly as the original one based on the PK formula. Both these AK procedures (and particularly their control variate extensions w.r.t. N) have a significantly lower variance for a fixed simulation size when compared to the method of Section 4.3.1. However, the AK estimator is quite slow because it has to evaluate an improper integral in every iteration for the chosen mixture model. Hence, whenever time matters, the first simulation method based on (4.18) is to be preferred, and there our new series is a significant improvement. The latter is particularly the case also in the heavy-traffic regime where simulation is known to be difficult. In addition, the performance is quite convincing also for moderate and low initial capital.

In addition, it is hard or even impossible to use the AK estimator when the distribution of $M_1^\bullet + C_1^e$ is not known explicitly. On the other hand, our estimator can be used even if the probability $\mathbb{P}(M_0^\bullet + M_1^\bullet + C_1^e > u)$ cannot be calculated in a closed form. In such cases, one can simply simulate that latter probability as well and adapt the theoretical results in Sections 4.3.1 and 4.3.1 accordingly.

In addition, although we concretely considered a mixture of a phase-type and a

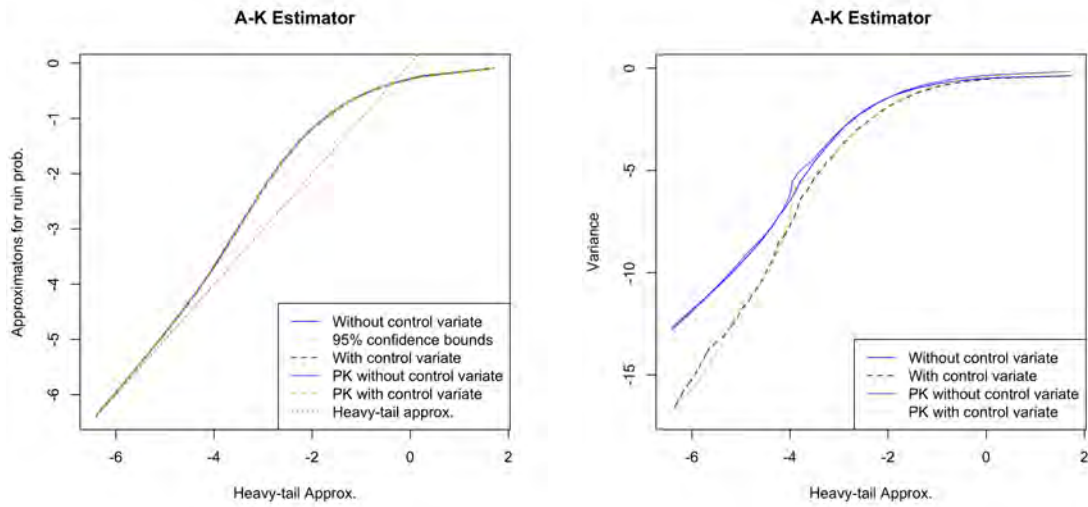


Figure 4.3: The simulated ruin probability with the AK estimator (blue solid line) and its control variate extension (4.23) (black dashed line) against the heavy-tail approximation. The corresponding estimates based on the PK formula are depicted in pink and dashed green. Model parameters: $a = 2$, $\epsilon = 0.1$, and $\rho = 0.99$. The respective empirical variances are presented on the right.

subexponential distribution in this paper, the results still hold if we replace $F_p(x)$ by any distribution for which $\psi^\bullet(u) = \mathbb{P}(M_0^\bullet > u)$ has a closed form, e.g. matrix-exponential distributions (cf. Bladt and Nielsen (2017)). In addition, one can further modify our approach in order to evaluate $\psi^\bullet(u)$ via simulation for any other light-tailed distribution, which is known to produce effortlessly reliable simulation outputs.

Finally, we would like to point out that the ruin probability of the more general Sparre Andersen model also has a Pollaczek-Khinchine-type formula with respect to the ladder height distribution ((Asmussen and Albrecher, 2010, Ch.VI)). Our estimator is also valid for this model as long as the ladder height distribution can be found explicitly, which is for instance the case when the inter-occurrence times belong to the class of distributions with rational Laplace transform.

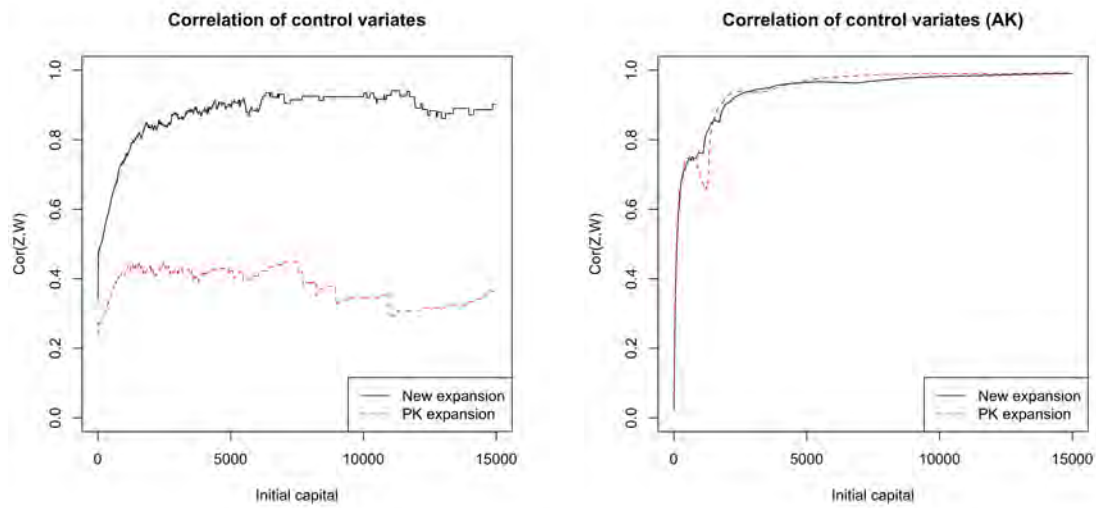


Figure 4.4: The correlation of the control variate using the new series expansion (black solid line) and the classical one (red dashed line) for the traditional Monte Carlo approach (4.3.1, left panel) and for the AK approach (4.3.2, right panel). Model parameters: $a = 2$, $\epsilon = 0.1$, and $\rho = 0.99$.

Chapter 5

Combined Tail Estimation Using Censored Data and Expert Information

This chapter is based on the following article:

Bladt, M., Albrecher, H., & Beirlant, J. (2019). Combined tail estimation using censored data and expert information. *Scandinavian Actuarial Journal*, to appear.

Abstract

We study tail estimation in Pareto-like settings for datasets with a high percentage of randomly right-censored data, and where some expert information on the tail index is available for the censored observations. This setting arises for instance naturally for liability insurance claims, where actuarial experts build reserves based on the specificity of each open claim, which can be used to improve the estimation based on the already available data points from closed claims. Through an entropy-perturbed likelihood we derive an explicit estimator and establish a close analogy with Bayesian methods. Embedded in an extreme value approach, asymptotic normality of the estimator is shown, and when the expert is clair-voyant, a simple combination formula can be deduced, bridging the classical statistical approach with the expert information. Following the aforementioned combination formula, a combination of quantile estimators can be naturally defined. In a simulation study, the estimator is shown to often outperform the Hill estimator for censored observations and recent Bayesian solutions, some of which require more information than usually available. Finally we perform a case study on a motor third-party liability insurance claim dataset, where Hill-type and quantile plots incorporate ultimate values into the estimation procedure in an intuitive manner.

5.1 Introduction

In applied statistics, one is often faced with the need to combine different types of information to produce a single decision. For instance, in credibility theory, the weights that link a relevant but small dataset with a big but not-so-relevant dataset are looked for, or similarly in bioinformatics one may use a dataset containing different cell types in the estimation of a cell, and scale down their importance in various ways.

The present paper is motivated by a specific problem in liability insurance. In that line of business, claim size data usually have a high percentage of censored observations, as policies take years, or even decades, to be finally settled. Due to the limited number of claims, one still would like to take into account available information about the open claims in the estimation of claim size distributions (see e.g. Albrecher et al. (2017)). On the one hand, experts typically project the final amount of open claims, i.e. they propose *incurred values*, or also *ultimates* based on covariate information or other (objective or subjective) considerations which are not in the payment dataset that arrives at a statistician's table. On the other hand, statisticians have standard ways of dealing with censored observations, for instance the Peaks over Threshold method when one is interested in extremes, as well as the Hill estimator for heavy and Pareto-like tails. This research has started in Beirlant et al. (2007) and Einmahl et al. (2008) and has received more attention recently, see e.g. Worms and Worms (2014), Amaraoui et al. (2016), Beirlant et al. (2018). However, in that line of extreme value methods expert information has not been incorporated. In Albrecher et al. (2017), incurred values were used to derive upper bounds for the open claims and survival analysis methods for interval censored data were implemented. See also Bogaerts et al. (2018) for frequentist and Bayesian analysis of interval censored data.

One often faces the question of whether to conduct the analysis from the right-censored observations point of view, or whether to use the imputed ultimate (expert) values into the dataset and treat it as a fully-observed dataset. The latter is typically an easy (and cheap) solution. Figure 5.1 illustrates a possible situation of available data for motor third-party liability (MTPL) insurance claims of a direct insurance company operating in the EU, cf. Section 5.5 for more details, where this data set will be studied. In what follows we are interested in developing a procedure that combines both approaches, without making any assumptions on the quality or method used to obtain the expert information.

To that end, we assume that for each censored observation (open claim), we have a tail parameter β_i which reflects the belief of the expert on the heaviness of the tail of this particular (unsettled) observation. The typical situation may be that all the β_i are equal or that there is an upper and a lower bound for all of them. However, we develop the theory for the general case, and we embed these indices into a statistical framework, where a single tail parameter is estimated for the entire dataset. At a philosophical level, the proposal of different β_i is not an ill-posed problem, it rather shows a prior variability of a belief on the tail index.

The difference with the Bayesian paradigm is the fact that we only make the assumption for censored observations, such that increasing sample sizes but constant

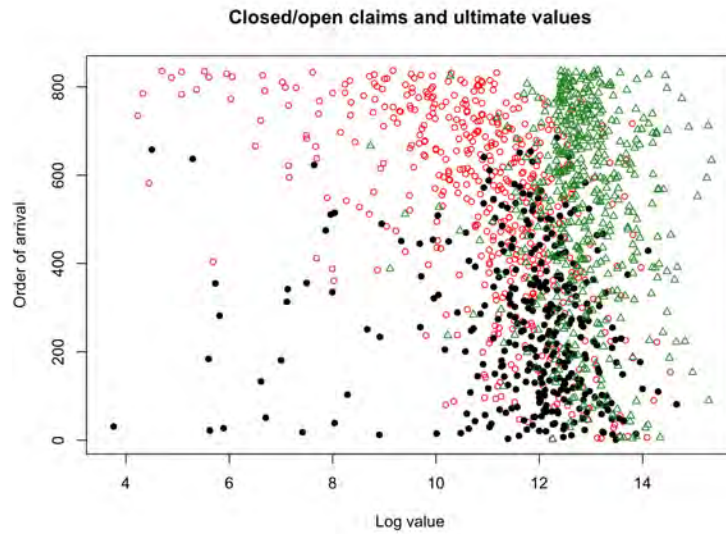


Figure 5.1: Motor third-party liability insurance: log-claims in vertical order of arrival, showing the paid amount for both open (red, circle) and closed (black, dot) claims, as well as ultimate values for the open claims (green, triangle).

censored proportions will keep the importance of the expert guesses constant, with respect to the rest of the data. In mathematical terms we will see that our approach will have a Bayesian interpretation when the prior distribution of the parameter depends on the sample through the censoring indicators.

We propose a perturbation of the likelihood via an exponential factor and use the relative entropy between two densities as a dissimilarity measure. The resulting maximum perturbed-likelihood estimate has an explicit formula which resembles the Hill estimator adapted for censoring (cf. Hill (1975); Beirlant et al. (2007)), degenerates into it as the perturbation becomes small, and converges to the mean of the expert tail indices as the perturbation becomes large. Thus, in a similar way as in the prior specification for Bayesian estimation, if experts have additional information on the quality of their belief, the perturbation parameter can be tuned accordingly. However, we propose a method which does not assume such additional prior knowledge, apart from the original expert knowledge.

Penalization is a prevalent idea which has gained popularity in the age of cheap computational power. The idea behind it is to impose beliefs on the statistical estimation which can yield a better estimation or an estimator which has more acceptable properties for the application at hand. It can help to impute a control or perturbation parameter which in turn helps to tailor estimators towards a certain degree of convenience. For instance, in a different statistical setting, the Lasso or Ridge regression imposes the belief or need of a data scientist to reduce the number of covariates included in a covariate-response analysis. In some cases this procedure helps to remove nuisance covariates, but in others it might be too aggressive and exclude truly informative variables. This bottleneck is specific to each application – and even to each dataset – which suggests that a fully automated procedure is not recommended. In the same vein, the effectiveness of the proposed method depends

on the quality of the provided expert tail information, and this is something which is not always available or quantifiable. In any case, it is recommended that the tail inference is done side-by-side with the experts.

We derive the asymptotic properties of the perturbed-likelihood estimator, and although the asymptotic mean square error (AMSE) is available, the parameter which minimizes it depends on subjective considerations. One such consideration is the strength of the belief in the expert information provided: if the belief is certain, the penalization parameter goes to infinity, which means that the data should be ignored and the expert information should be used instead, based on the AMSE criterion. To avoid assumptions which are not realistically available in practice, we instead suggest selecting the penalizing parameter in a convenient way, as the one which reduces the perturbed estimator to a simple sum of identifiable components. When the expert information is precise (degenerate at the true parameter) it holds that the penalization weight equal to 1 always leads to a lower AMSE, the latter also having a formula with a pleasant interpretation. When substituting this penalization parameter into the original formula, a very simple interpretation of the (inverse of the) estimator is available (Theorem 5.4.4): it is the combination of the Hill estimator and the expert information, the weights being the proportion of non-censored and censored data-points, respectively. Such a simple combination estimator is shown for a variety of common heavy-tailed distributions and for a variety of parameters to perform very well alongside competing methods, which need information to tune their own parameters.

The remainder of the paper is structured as follows. In Section 5.2, for the exact Pareto distribution, we introduce the notation and perturbation that we will deal with, as well as deriving simple expressions for the maximum perturbed-likelihood estimators and showing how they bridge theory and practice in a smooth manner. In Section 5.3 we establish a close link with Bayesian statistics. In Section 8.3 we extend the methodology to the case where the data only exhibit Pareto-type behaviour in the tail, and we derive the asymptotic distributional properties of the perturbed-likelihood estimator and unveil a simple combination formula. In this more general heavy-tailed case, we naturally deal with estimators that use only a fraction of upper order statistics, and introduce as benchmarks some recent Bayesian estimators that have been proposed for censored datasets. In Section 5.5 we perform a simulation study and a real-life motor third-party liability insurance application. The latter data has been studied in the literature from both the expert information and the censored dataset viewpoints, but not yet in a joint manner, as we do here. We conclude in Section 5.6.

5.2 Derivation and properties

Consider the estimation from a censored sample following an exact Pareto distribution. That is, we observe the randomly censored data-points and the binary indicator censoring variables:

$$(Z_1, e_1), (Z_2, e_2), \dots, (Z_n, e_n).$$

Contrary to classical survival analysis, the Z_i here correspond to payment sizes rather than times. The density and tail of the non-censored underlying data (which is not observed) are given by

$$f_\alpha(x) = \frac{\alpha x_0^\alpha}{x^{\alpha+1}}, \quad \bar{F}_\alpha(x) = \frac{x_0^\alpha}{x^\alpha}, \quad x \geq x_0 > 0, \quad (5.1)$$

with unknown tail parameter α and known scale parameter x_0 . The latter assumption poses no restriction, since we are interested only in the estimation of the tail index, and as we will see, the Hill-type estimators based on upper order statistics that will be considered depend only on the log spacings of the data, which are independent of the scale parameter.

Additionally, and in contrast to classical survival analysis, we assume that we are given for each (right-)censored data-point an expert information of the possible tail parameter, i.e. we have knowledge of $\beta_i > 0$ for $i = 1, \dots, n$. This can arise, for instance, when the data are collected from different sources, or the realization of a data-point showing some pattern due to a particular settlement history, or some covariate information that can not be included in a more direct way. However, it is believed that all data points eventually come from one underlying distribution (or at least one aims for such a modelling description). We are also primarily interested in the case where all the β_i are the same, since more often than not, expert information can come in this format.

When ignoring the information from the data, natural estimates of α are given by the weighted arithmetic and harmonic mean

$$\hat{\alpha}_{\text{am}} = \frac{\sum_{i=1}^n (1 - e_i) \beta_i}{\sum_{i=1}^n (1 - e_i)}, \quad \hat{\alpha}_{\text{hm}} = \frac{\sum_{i=1}^n (1 - e_i)}{\sum_{i=1}^n (1 - e_i) / \beta_i}, \quad (5.2)$$

respectively, where $e_i = 0$ if Z_i is right-censored, and $e_i = 1$ otherwise. On the other hand, in the context of survival analysis it is a standard approach to maximize the following likelihood based purely on the data

$$\mathcal{L}(\alpha; z) = \prod_{i=1}^n f_\alpha(z_i)^{e_i} \bar{F}_\alpha(z_i)^{1-e_i} = \prod_{i=1}^n \left(\frac{\alpha x_0^\alpha}{z_i^{\alpha+1}} \right)^{e_i} \left(\frac{x_0^\alpha}{z_i^\alpha} \right)^{1-e_i}.$$

The maximum likelihood estimator is then given by

$$\hat{\alpha}^{MLE} = \frac{\sum_{i=1}^n e_i}{\sum_{i=1}^n \log \left(\frac{Z_i}{x_0} \right)}, \quad (5.3)$$

which is an adaptation (see Beirlant et al. (2007)) to the censoring case of the famous Hill estimator (cf. Hill (1975)) from extreme value theory obtained by Peaks-over-Threshold modelling; see Embrechts et al. (1997) or Beirlant et al. (2004) for a broader treatment on Pareto-type tail estimation, see also Section 4.

The two aforementioned approaches to the estimation of the tail parameter are in practice separated. That is, an expert will take only one of the two approaches, based on factors such as reliability of the expert information or data availability. This is an especially difficult decision when there is a high percentage of censored

and large observations, which present a key problem in the estimation in statistics in general, see for instance Leung et al. (1997).

In the present paper, we introduce an estimator which bridges the previous estimators. We will proceed by the perturbation of the likelihood function, and consider a penalized likelihood:

$$\begin{aligned}\mathcal{L}^P(\alpha; z) &= \prod_{i=1}^n f_\alpha(z_i)^{e_i} \bar{F}_\alpha(z_i)^{1-e_i} e^{-(1-e_i)\lambda D(\alpha, \beta_i)} \\ &= \prod_{i=1}^n \left(\frac{\alpha x_0^\alpha}{z_i^{\alpha+1}} \right)^{e_i} \left(\frac{x_0^\alpha}{z_i^\alpha} e^{-\lambda D(\alpha, \beta_i)} \right)^{1-e_i},\end{aligned}$$

where the factor $e^{-\lambda D(\alpha, \beta_i)}$ penalizes the contribution of the censored observations according to some measure of dissimilarity between f_α and the Pareto distribution with parameter β_i , denoted by $D(\alpha, \beta_i)$, and the $\lambda \geq 0$ models the strength of the penalization imposed by $D(\alpha, \beta_i)$. We propose to use the relative entropy as a dissimilarity measure:

$$D(\beta_i, \alpha) = \int_{x_0}^{\infty} \log \left(\frac{g_i(s)}{f_\alpha(s)} \right) g_i(s) ds = \frac{\alpha}{\beta_i} - 1 - \log \left(\frac{\alpha}{\beta_i} \right) \geq 0, \quad (5.4)$$

where g_i is a Pareto density with tail index β_i and scale parameter x_0 . The associated log-likelihood is then given by

$$\log(\mathcal{L}^P(\alpha, z)) = \sum_{i=1}^n e_i \log \left(\frac{\alpha x_0^\alpha}{z_i^{\alpha+1}} \right) + \sum_{i=1}^n (1 - e_i) \log \left(\frac{x_0^\alpha}{z_i^\alpha} \right) - \sum_{i=1}^n \lambda (1 - e_i) D(f_\alpha, g_i). \quad (5.5)$$

Equation (5.5) turns out to have an explicit minimizer when using D from (5.4) (we omit the details), given by

$$\hat{\alpha}^P(\lambda) = \frac{\sum_{i=1}^n (e_i + \lambda(1 - e_i))}{\sum_{i=1}^n (\log(Z_i/x_0) + \lambda(1 - e_i)/\beta_i)}.$$

Notice that if we flip the densities in the entropy penalization and consider instead

$$D(\alpha, \beta_i) = \int_{x_0}^{\infty} \log \left(\frac{f_\alpha(s)}{g_i(s)} \right) f_\alpha(s) ds = \frac{\beta_i}{\alpha} - 1 - \log \left(\frac{\beta_i}{\alpha} \right) \geq 0,$$

the associated penalized likelihood has the explicit solution

$$\begin{aligned}\hat{\alpha}^I(\lambda) &= \frac{\sum_1^n (e_i - \lambda(1 - e_i)) + \sqrt{[\sum_1^n (e_i - \lambda(1 - e_i))]^2 + 4 \sum_1^n \log \left(\frac{Z_i}{x_0} \right) \cdot \sum_1^n \beta_i \lambda (1 - e_i)}}{2 \sum_1^n \log \left(\frac{Z_i}{x_0} \right)},\end{aligned} \quad (5.6)$$

which is less appealing, with more complicated asymptotic properties.

Remark 5.2.1. The particular choice of entropy penalization is mathematical in nature, since the resulting explicit and simple form of the maximum likelihood estimator permits a deeper analysis than other choices. For instance, the significantly more complicated explicit estimators (5.6) or (5.13) lead to a much more involved analysis.

Remark 5.2.2. In general, with lack of any other type of information, giving equal weight to each censored observation is the most natural way to deal with them. If the expert has an idea of the importance of each data point and their corresponding tail indices β_i , the selection can be done on a single parameter λ through

$$\lambda_i = \lambda \omega_i.$$

Note that then

$$\lim_{\lambda \rightarrow \infty} \hat{\alpha}^P(\lambda) = \frac{\sum_{i=1}^n (1 - e_i) \omega_i}{\sum_{i=1}^n (1 - e_i) \omega_i / \beta_i}, \quad (5.7)$$

and

$$\lim_{\lambda \rightarrow \infty} \hat{\alpha}^I(\lambda) = \frac{\sum_{i=1}^n (1 - e_i) \omega_i \beta_i}{\sum_{i=1}^n (1 - e_i) \omega_i}, \quad (5.8)$$

i.e. the information brought by the data becomes irrelevant and we take a weighted average of the expert guesses. Taking uniform weights, i.e., giving equal importance to each censored observation, will result in (5.2).

If no weights are naturally suggested one can always tackle the multi-dimensional selection problem on all λ_i . In this more general case we have that

$$\lim_{\lambda_i \rightarrow 0; i=1, \dots, n} \hat{\alpha}^P(\lambda_1, \dots, \lambda_n) = \frac{\sum_{i=1}^n e_i}{\sum_{i=1}^n \log \left(\frac{Z_i}{x_0} \right)}. \quad (5.9)$$

which can readily be deduced directly from (5.5), since it is the classical non-penalized estimator. Similarly

$$\lim_{\lambda_i \rightarrow 0; i=1, \dots, n} \hat{\alpha}^I(\lambda_1, \dots, \lambda_n) = \frac{\sum_{i=1}^n e_i}{\sum_{i=1}^n \log \left(\frac{Z_i}{x_0} \right)}. \quad (5.10)$$

As a consequence of the limits (5.7), (5.8), (5.9) and (5.10), we readily get for $\lambda_1 = \dots = \lambda_n =: \lambda \geq 0$ that

$$\lim_{\lambda \rightarrow \infty} \hat{\alpha}^P(\lambda) = \hat{\alpha}_1, \quad \lim_{\lambda \rightarrow \infty} \hat{\alpha}^I(\lambda) = \hat{\alpha}_2, \quad \lim_{\lambda \rightarrow 0} \hat{\alpha}^P = \lim_{\lambda \rightarrow 0} \hat{\alpha}^I(\lambda) = \hat{\alpha}^{MLE},$$

which confirms that the estimator bridges the estimation of α and the proposal of the β_i , and that the parameter λ reflects in some sense the strength of the belief on the expert information. The next section will touch upon this interpretation in a more precise manner.

5.3 Penalization seen as a Bayesian prior

We will use a single λ value in practice, but here we assume the most general setting where the λ_i could be different, at no complexity cost. The penalized likelihood that gives rise to $\hat{\alpha}^P$ is given by

$$\begin{aligned} \mathcal{L}^P(\alpha; z) &= \prod_{i=1}^n \left(\frac{\alpha x_0^\alpha}{z_i^{\alpha+1}} \right)^{e_i} \left(\frac{x_0^\alpha}{z_i^\alpha} \right)^{1-e_i} e^{-\lambda_i(1-e_i)(\alpha/\beta_i - 1 - \log(\alpha/\beta_i))} \\ &= \left[\prod_{i=1}^n \left(\frac{\alpha x_0^\alpha}{z_i^{\alpha+1}} \right)^{e_i} \left(\frac{x_0^\alpha}{z_i^\alpha} \right)^{1-e_i} \right] \cdot \left[\alpha^{\sum_{i=1}^n \lambda_i(1-e_i)} e^{-\alpha \sum_{i=1}^n \lambda_i(1-e_i)/\beta_i} \right] \\ &\quad \times \left[\prod_{i=1}^n \beta_i^{-\lambda_i(1-e_i)} e^{\lambda_i(1-e_i)} \right] \\ &= \left[\alpha^{\sum_{i=1}^n (e_i + \lambda_i(1-e_i))} e^{-\alpha \sum_{i=1}^n (\lambda_i(1-e_i)/\beta_i + \log(z_i/x_0))} \right] \\ &\quad \times \left[\prod_{i=1}^n \beta_i^{-\lambda_i(1-e_i)} e^{\lambda_i(1-e_i)} z_i^{-e_i} \right]. \end{aligned}$$

Note that the second factor after the last equality sign does not depend on α , and the first one is proportional to a gamma density. We thus recognize that the penalized maximum likelihood estimator can be seen as the posterior mode arising from a Pareto likelihood and the conjugate gamma prior with hyper-parameters

$$\alpha_0 = \sum_{i=1}^n \lambda_i(1 - e_i) + 1, \quad \beta_0 = \sum_{i=1}^n \lambda_i(1 - e_i)/\beta_i,$$

and corresponding posterior parameters

$$\alpha^* = \sum_{i=1}^n (e_i + \lambda_i(1 - e_i)) + 1, \quad \beta^* = \sum_{i=1}^n (\lambda_i(1 - e_i)/\beta_i + \log(z_i/x_0)).$$

The hyper-parameters of the prior, however, do depend on the sample, namely on the censoring indicators e_i , so we are not in the classical Bayesian setting. Nonetheless, we will continue to call it a prior, for simplicity.

In this context we also have the following interpretation of the effects of the selection of the λ_i . The mode of the prior distribution is given by

$$\frac{\sum_{j=1}^n \lambda_j(1 - e_j)}{\sum_{i=1}^n \lambda_i(1 - e_i)/\beta_i} = \left(\sum_{i=1}^n \frac{\lambda_i(1 - e_i)}{\sum_{j=1}^n \lambda_j(1 - e_j)} \beta_i^{-1} \right)^{-1}, \quad (5.11)$$

and one sees that the proportions of the λ_i give the weights which will determine this mode. However, we can multiplicatively scale these λ_i and the mode will remain unchanged. The magnitude of the λ_i , in contrast, does play a role for the variance of the prior:

$$\frac{\sum_{i=1}^n \lambda_i(1 - e_i) + 1}{\left(\sum_{i=1}^n \lambda_i(1 - e_i)/\beta_i \right)^2}, \quad (5.12)$$

since the larger the λ_i , the smaller the prior variance. Thus, a single estimate as an expert information will trump the ability to effectively determine the magnitude of the penalization parameter. This is a problem which is often encountered in Bayesian statistics, and a prior is often selected nonetheless, making frequentists doubtful of this philosophical leap of faith.

Note that the gamma distribution has two parameters, and any two descriptive statistics (presently we used the mode and variance) which bijectively map to the mode and variance can be used to give alternate full explanations as to how the proportions $\lambda_i(1 - e_i)/\sum_{j=1}^n \lambda_j(1 - e_j)$ and the sizes of the λ_i play a role in the modification of the prior distribution, and hence on the expert belief.

Remark 5.3.1. If instead of using the penalization given by (5.4) we simply use squared penalization given by

$$D(\beta_i, \alpha) = \frac{(\alpha - \beta_i)^2}{2} \geq 0,$$

then the maximum perturbed-likelihood estimate will again be explicit and given by

$$\hat{\alpha}^{Sq} = \frac{\sum_{i=1}^n (\lambda_i(1 - e_i)\beta_i - \log(Z_i/x_0))}{\sum_{i=1}^n \lambda_i(1 - e_i)} + \frac{\sqrt{[\sum_{i=1}^n (\lambda_i(1 - e_i)\beta_i - \log(Z_i/x_0))]^2 + 4 \sum_{i=1}^n \lambda_i(1 - e_i) \cdot \sum_{i=1}^n e_i}}{\sum_{i=1}^n \lambda_i(1 - e_i)}, \quad (5.13)$$

which naturally leads to a Gaussian prior interpretation when the λ_i are equal. This estimator also converges to the Hill estimator as $\lambda_i \rightarrow 0$, $i = 1, \dots, n$, but it can have numerical instabilities when the denominator becomes very small.

5.4 Extreme Value Theory

We now move on to a more general heavy-tail approach and consider the case of regularly varying distributions with tail of the form

$$x^{-\alpha} \ell(x), \quad \alpha > 0,$$

where ℓ is a slowly varying function, i.e. $\frac{\ell(vx)}{\ell(x)} \rightarrow 1$, as $x \rightarrow \infty$ for every $v > 1$. We also assume now that censoring is done at random and the data is generated as the minimum of two independent random variables

$$Z_i = \min\{X_i, L_i\},$$

with regularly varying tails:

$$\begin{aligned} \mathbb{P}(X_i > u) &= u^{-\alpha} \ell(u), \\ \mathbb{P}(L_i > u) &= u^{-\alpha_2} \ell_2(u). \end{aligned}$$

It follows that

$$\mathbb{P}(Z_i > u) = u^{-\alpha_c} \ell_c(u), \quad \alpha_c = \alpha + \alpha_2, \quad (5.14)$$

and slowly varying function $\ell_c = \ell \ell_2$. Here we confine ourselves to the so-called *Hall class* (cf. Hall (1982)). This popular second-order assumption in extreme value theory often makes asymptotic identities tractable:

$$\begin{aligned} \mathbb{P}(X_i > u) &= C_1 u^{-\alpha} (1 + D_1 u^{-\nu_1} (1 + o(1))) \text{ for } u \rightarrow \infty, \\ \mathbb{P}(L_i > u) &= C_2 u^{-\alpha_2} (1 + D_2 u^{-\nu_2} (1 + o(1))) \text{ for } u \rightarrow \infty, \end{aligned} \quad (5.15)$$

where ν_1, ν_2, C_1, C_2 are positive constants and D_1, D_2 real constants. Then, with

$$C = C_1 C_2, \quad \nu_* = \min(\nu_1, \nu_2)$$

and

$$D_* = \begin{cases} D_1, & \nu_1 < \nu_2 \\ D_2, & \nu_2 < \nu_1 \\ D_1 + D_2, & \nu_1 = \nu_2, \end{cases}$$

we have that

$$\mathbb{P}(Z_i > u) = C u^{-\alpha_c} (1 + D_* u^{-\nu_*} (1 + o(1))),$$

that is, the censored dataset is again in the Hall class.

Denote the quantile function of Z by Q and consider the tail quantile function $U(x) = Q(1 - x^{-1})$, $x > 1$. Then we have that

$$U(x) = C^{1/\alpha_c} \left(1 + \frac{D_*}{\alpha_c} C^{-\nu_*/\alpha_c} x^{-\nu_*/\alpha_c} (1 + o(1)) \right).$$

The order statistics of the data will be denoted by

$$Z^{(1)} \geq \dots \geq Z^{(n)},$$

with associated censoring indicators $e^{(i)}$ and expert information $\beta^{(i)}$. Given a high threshold $u > x_0$, the Hill estimator adapted for censoring is

$$\hat{\alpha}_u^H = \frac{\sum_{i=1}^n e_i 1\{Z_i > u\}}{\sum_{i=1}^n \log\left(\frac{Z_i}{u}\right) 1\{Z_i > u\}}. \quad (5.16)$$

Taking $Z^{(k)}$ for some $1 \leq k \leq n$, as a (random) threshold u , we obtain the alternative order statistics version

$$\hat{\alpha}_k^{MLE} = \frac{\sum_{i=1}^k e^{(i)}}{\sum_{i=1}^k \log\left(\frac{Z^{(i)}}{Z^{(k+1)}}\right)} = \frac{\hat{p}_k}{H_k}, \quad (5.17)$$

where

$$\hat{p}_k = \frac{1}{k} \sum_{i=1}^k e^{(i)}$$

is the proportion of non-censored observations in the largest k observations of Z , and

$$H_k = \frac{1}{k} \sum_{i=1}^k \log \left(\frac{Z^{(i)}}{Z^{(k+1)}} \right)$$

is the classical Hill estimator based on the largest k observations. For details on these censored versions of the Hill estimator, we refer to (Einmahl et al., 2008, Sec.2).

The asymptotic distribution of H_k has been studied intensively in the literature under the above second-order assumptions (see for instance (Beirlant et al., 2004, Ch.4)): assuming

$$\sqrt{k}(k/n)^{\nu_*/\alpha_c} \rightarrow \delta \geq 0, \quad (5.18)$$

as $k, n \rightarrow \infty$ with $k/n \rightarrow 0$, we have that

$$\sqrt{k} \left(H_k - \frac{1}{\alpha_c} \right) \xrightarrow{d} Y_0 \sim \mathcal{N} \left(-C^{-\nu_*/\alpha_c} D_* \frac{\nu_* \delta}{\alpha_c (\alpha_c + \nu_*)}, \alpha_c^{-2} \right). \quad (5.19)$$

As discussed in Einmahl et al. (2008), the asymptotic bias of \hat{p}_k follows from the leading term in $\frac{1}{k} \sum_{i=1}^k p(U(n/i)) - p$, where

$$p(z) = \mathbb{P}(e = 1 | Z = z),$$

and p denotes the asymptotic probability of non-censoring

$$p = \lim_{z \rightarrow \infty} p(z) = \frac{1/\alpha_2}{1/\alpha + 1/\alpha_2} = \frac{\alpha}{\alpha + \alpha_2}.$$

Under the Hall class (5.15), we have with the definition

$$(D/\alpha)_* = \begin{cases} D_1/\alpha, & \nu_1 < \nu_2 \\ -D_2/\alpha_2, & \nu_2 < \nu_1 \\ D_1/\alpha - D_2/\alpha_2, & \nu_1 = \nu_2, \end{cases}$$

that as $x \rightarrow \infty$

$$p(U(x)) - p = p(1-p)(D/\alpha)_* \nu_* C^{-\nu_*/\alpha_c} x^{-\nu_*/\alpha_c} (1 + o(1)). \quad (5.20)$$

From this, assuming that $\sqrt{k}(k/n)^{\nu_*/\alpha_c} \rightarrow \delta$ as $k, n \rightarrow \infty$ with $k/n \rightarrow 0$, one gets

$$\sqrt{k}(\hat{p}_k - p) \xrightarrow{d} \mathcal{N} \left(p(1-p) C^{-\nu_*/\alpha_c} (D/\alpha)_* \frac{\alpha_c \nu_* \delta}{\alpha_c + \nu_*}, p(1-p) \right).$$

In Einmahl et al. (2008) it was also derived that asymptotically H_k and \hat{p}_k are independent, so that under the condition (5.18) as $k, n \rightarrow \infty$ with $k/n \rightarrow 0$,

$$\sqrt{k} \left(\frac{1}{\hat{\alpha}_k^{MLE}} - \frac{1}{\alpha} \right) \xrightarrow{d} \mathcal{N} \left(-\frac{\delta \nu_*}{\alpha_c + \nu_*} C^{-\nu_*/\alpha_c} [D_* (\alpha_c^{-1} + \alpha^{-1}) + \frac{\alpha_2}{\alpha} (D/\alpha)_*], \frac{1}{p\alpha^2} \right). \quad (5.21)$$

In the same manner we can define a version of $\hat{\alpha}^P$ which perturbs at censored data-points and which considers only large claims. We consider as before that $\lambda_i = \lambda$, and, in analogy to the exact Pareto setting, define the two estimators

$$\hat{\alpha}_u^P = \frac{\sum_{i=1}^n (e_i + \lambda(1 - e_i)) 1\{Z_i > u\}}{\sum_{i=1}^n (\log(Z_i/u) + \lambda(1 - e_i)/\beta_i) 1\{Z_i > u\}},$$

and the order statistics version

$$\hat{\alpha}_k^P = \frac{\sum_{i=1}^k (e^{(i)} + \lambda(1 - e^{(i)}))}{\sum_{i=1}^k \left(\log\left(\frac{Z^{(i)}}{Z^{(k)}}\right) + \lambda(1 - e^{(i)})/\beta^{(i)} \right)}.$$

Theorem 5.4.1. *Assume (5.15). Set $\lambda_i = \lambda \geq 0$, $\beta_i = \beta > 0$. As $\sqrt{k}(k/n)^{\nu_*/\alpha_c} \rightarrow \delta$, as $k, n \rightarrow \infty$ with $k/n \rightarrow 0$,*

$$\sqrt{k} \left(\frac{1}{\hat{\alpha}_k^P} - \frac{\lambda\alpha_2/\beta + 1}{\lambda\alpha_2 + \alpha} \right)$$

is asymptotically normal with asymptotic mean

$$\mathcal{M} = -\frac{\delta\nu_*C^{-\nu_*/\alpha_c}}{1 - r_1} \left(\frac{D_*/\alpha_c + \lambda p(1 - p)(D/\alpha)_*\alpha_c/\beta}{\nu_* + \alpha_c} + \frac{\lambda r_2 + \alpha_c^{-1}}{1 - r_1} p(1 - p)(D/\alpha)_* \left(\frac{\alpha_c}{\alpha_c + \nu_*} \right) \right)$$

and variance

$$\mathcal{V} = \frac{1}{\alpha_c^2(1 - r_1)^2} + \frac{1}{(1 - r_1)^4} \left(\frac{\lambda}{\beta(1 - \lambda)} + \frac{1}{\alpha_c} \right)^2 (1 - \lambda)^2 p(1 - p), \quad (5.22)$$

where $r_1 = (1 - p)(1 - \lambda)$ and $r_2 = (1 - p)/\beta$. The asymptotic bias of $1/\hat{\alpha}_k^P$ equals

$$\mathcal{B} = \frac{\lambda\alpha_2/\beta + 1}{\lambda\alpha_2 + \alpha} - \frac{1}{\alpha} + O\left((k/n)^{\nu_*/\alpha_c}\right) \quad (5.23)$$

as $k, n \rightarrow \infty$ and $k/n \rightarrow 0$.

Proof. See Appendix A. □

Remark 5.4.2. Notice that estimates of α_2 or α_c are available using basic survival analysis techniques, cf. (5.14). Consequently, we can use the plug-in method for the estimation of any of the above formulas that involve these quantities.

Remark 5.4.3. As a sanity check, observe that in Theorem 5.4.1, whenever $\beta = \alpha$ and $\delta = 0$, the bias vanishes.

In the same spirit, even more can be said:

Corollary 5.4.4. (Combination) Assume the conditions of Theorem 5.4.1, and further $\delta = 0$, with $\beta = \alpha$. Then the estimator $\hat{\alpha}_k^P$ with $\lambda = 1$ is unbiased and can be written as

$$\hat{\alpha}_k^P = \left(\frac{\sum_{i=1}^k e^{(i)}}{k} \cdot \frac{\sum_{i=1}^k \log(Z^{(i)}/Z^{(k+1)})}{\sum_{i=1}^k e^{(i)}} + \frac{\sum_{i=1}^k (1 - e^{(i)})}{k} \cdot \beta^{-1} \right)^{-1}.$$

In words, $1/\hat{\alpha}_k^P$ is the weighted average of the MLE estimator and the expert information, the weights being the proportion of non-censored (and censored, respectively) observations above the threshold $T^{(k)}$. Moreover, its inverse has asymptotic variance (and hence mean square error) given by

$$\text{Var}(1/\hat{\alpha}_k^P) = \frac{1}{kp(\alpha + \alpha_2)^2},$$

which, when compared to (5.21), is seen to enhance estimation.

The proof of Corollary 5.4.4 is immediate.

Remark 5.4.5. Observe that in a Bayesian setting, whenever we are aware that a parameter lies within an interval, a natural estimator is constructed as follows. We set a uniform prior on $[b_1, b_2]$ and together with the Pareto likelihood we use the posterior mean as an estimate. Such a mean is given by

$$\begin{aligned} & \frac{\int_{b_1}^{b_2} \alpha^{1+\sum_{i=1}^n e_i} e^{-\alpha \sum_{i=1}^n \log\left(\frac{z_1}{x_0}\right)} d\alpha}{\int_{b_1}^{b_2} \alpha^{\sum_{i=1}^n e_i} e^{-\alpha \sum_{i=1}^n \log\left(\frac{z_1}{x_0}\right)} d\alpha} \\ &= \frac{\sum_{i=1}^n e_i + 1}{\sum_{i=1}^n \log\left(\frac{z_1}{x_0}\right)} \\ & \times \left[\frac{\gamma\left(\sum_{i=1}^n e_i + 2, b_2 \sum_{i=1}^n \log\left(\frac{z_1}{x_0}\right)\right) - \gamma\left(\sum_{i=1}^n e_i + 2, b_1 \sum_{i=1}^n \log\left(\frac{z_1}{x_0}\right)\right)}{\gamma\left(\sum_{i=1}^n e_i + 1, b_2 \sum_{i=1}^n \log\left(\frac{z_1}{x_0}\right)\right) - \gamma\left(\sum_{i=1}^n e_i + 1, b_1 \sum_{i=1}^n \log\left(\frac{z_1}{x_0}\right)\right)} \right], \end{aligned}$$

where

$$\gamma(u, v) = \frac{\int_0^v t^{u-1} e^{-t} dt}{\Gamma(u)}$$

is the (normalized) lower incomplete gamma function. One can go one step further and define the order statistics version of the above estimator. However, despite being theoretically neat, the latter estimator is numerically unstable for both large ($k > 100$) and small ($k < 5$) number of upper order statistics, and hence we will not pursue it in the simulation section.

Remark 5.4.6. In Ameraoui et al. (2016), several Bayesian approaches for heavy-tail estimation were considered (see also Beirlant et al. (2018)) under the random

censoring assumption. We will use two of them as a benchmark. The first one arises from the posterior mean of a Pareto likelihood and the conjugate Gamma(a, b) prior:

$$\hat{\alpha}^{BG} = \frac{a + \sum_{i=1}^k e^{(i)}}{b + \sum_{i=1}^k \log(Z^{(i)}/Z^{(k+1)})}. \quad (5.24)$$

In the presence of a single expert estimate β of the tail index, the prior parameters can be tuned by moment matching, where the variance will need to be imposed subjectively. That is, solve $\beta = a/b$ and $\sigma^2 = a/b^2$ for an expert opinion on σ^2 .

The second one arises from the maximal data information prior, and leads to the estimator

$$\hat{\alpha}^{BM} = \frac{1 + \sum_{i=1}^k e^{(i)} + \sqrt{(1 + \sum_{i=1}^k e^{(i)})^2 + 4 \sum_{i=1}^k \log(Z^{(i)}/Z^{(k+1)})}}{2 \sum_{i=1}^k \log(Z^{(i)}/Z^{(k+1)})}. \quad (5.25)$$

Notice that the latter does not admit tuning the prior to additional data.

Quantile estimation

With the last result at hand it is natural to propose a quantile estimator based on the approach taken in Weissman (1978). Recall that we denote the quantile function of a regularly varying tail by $Q(p)$. Exploiting the fact that

$$\frac{Q(1-p)}{Q(1-k/n)} \sim \frac{p^{-1/\alpha}}{(k/n)^{-1/\alpha}} = \left(\frac{k}{np}\right)^{1/\alpha}, \quad p \downarrow 0, k/n \rightarrow 0, np = o(k), \quad (5.26)$$

the Weissman estimator based on k order statistics (and without expert information) arises naturally as

$$\hat{Q}_k^{MLE}(1-p) = \hat{Q}^{KM}(1-k/n) \cdot \left(\frac{k}{np}\right)^{1/\hat{\alpha}_k^{MLE}}, \quad (5.27)$$

where \hat{Q}^{KM} is the quantile function derived from the Kaplan-Meier estimator

$$\hat{S}(z) = \prod_{i: Z_i \leq z} \left(1 - \frac{d_i}{n_i}\right),$$

for the survival curve of the censored dataset in question, (Z_i, e_i) , $i = 1, \dots, n$, where the Z_i are payments (which would correspond to times in classical survival analysis terminology). Here, d_i is the number of closed claims of a given size z , and n_i is the number of payments, which irrespectively of censoring, are above z . In the case of no censoring this reduces to the empirical quantiles of the dataset, since the Kaplan-Meier curve is then just the empirical distribution function. Similarly, in the case of pure expert information an estimator can be proposed as

$$\hat{Q}_k^{EX}(1-p) = \hat{Q}^{EX}(1-k/n) \cdot \left(\frac{k}{np}\right)^{1/\beta}, \quad (5.28)$$

where \hat{Q}^{EX} is either an expert-given cumulative distribution function, or in absence of it, simply the Kaplan-Meier quantiles. To combine these two results, Theorem 5.4.4 leads the way. For the choice $\lambda = 1$, we see that the Pareto part of the tail splits for the perturbed estimator according to

$$\left(\frac{k}{np}\right)^{1/\hat{\alpha}^P} = \left(\frac{k}{np}\right)^{\hat{p}_k/\hat{\alpha}_k^{MLE}} \cdot \left(\frac{k}{np}\right)^{(1-\hat{p}_k)/\beta},$$

where

$$\hat{p}_k = \frac{1}{k} \sum_{i=1}^k e^{(i)},$$

and hence the following estimator is proposed for the overall tail

$$\hat{Q}_k^P(1-p) = \left[\hat{Q}^{KM}(1-k/n) \cdot \left(\frac{k}{np}\right)^{1/\hat{\alpha}_k^{MLE}} \right]^{\hat{p}_k} \cdot \left[\hat{Q}^{EX}(1-k/n) \cdot \left(\frac{k}{np}\right)^{1/\beta} \right]^{1-\hat{p}_k} \quad (5.29)$$

$$= \hat{Q}^P(1-k/n) \cdot \left(\frac{k}{np}\right)^{1/\hat{\alpha}_k^P}, \quad (5.30)$$

where

$$\hat{Q}^P(1-k/n) = (\hat{Q}^{KM}(1-k/n))^{\hat{p}_k} (\hat{Q}^{EX}(1-k/n))^{1-\hat{p}_k}.$$

Observe that in the absence of expert information for the quantile function, we merely have

$$\hat{Q}_k^P(1-p) = \hat{Q}^{KM}(1-k/n) \cdot \left(\frac{k}{np}\right)^{1/\hat{\alpha}_k^P}.$$

5.5 Simulation Study and MTPL Insurance

We perform in this section a simulation study and apply our method to a motor third party liability insurance dataset (cf. (Albrecher et al., 2017, Sec.1.3.1)). In order to make our results comparable with existing studies and existing analysis of the aforementioned dataset, we will consider estimation of

$$\xi = \frac{1}{\alpha},$$

and thus we will make use of the estimators

$$\hat{\xi}_k^{MLE} = \frac{1}{\hat{\alpha}_k^{MLE}}, \quad \hat{\xi}_k^P = \frac{1}{\hat{\alpha}_k^P}, \quad \hat{\xi}_k^{BG} = \frac{1}{\hat{\alpha}_k^{BG}}, \quad \hat{\xi}_k^{BM} = \frac{1}{\hat{\alpha}_k^{BM}}. \quad (5.31)$$

5.5.1 Simulation Study

We consider three heavy-tails belonging to the Hall class (5.15), and compare $\hat{\xi}_k$ and the quantile estimator

$$\hat{Q}_k^P(1-p) = \hat{Q}^{KM}(1-k/n) \cdot \left(\frac{k}{np}\right)^{\hat{\xi}_k},$$

where $\hat{\xi}_k$ is one of the four estimators in (5.31), and for $p = 0.005$. For any tail estimator $\hat{\xi}_k$, we generically refer to \hat{Q}_k^P as the corresponding Weismann estimator, since it was derived using the general principle of equation (5.26).

Concretely, we simulate two independent i.i.d. samples of size $n = 200$, corresponding to the variables X_i and L_i , $i = 1, \dots, n$, in (5.15). We repeat the procedure $N_{sim} = 1000$ times. The following three distributions are employed, with two sub-cases for each distribution, for varying parameters:

- The exact Pareto distribution, defined in (5.1), for $\xi = 1, 1/2$.
- The Burr distribution, with tail given by

$$\bar{F}(x) = \left(\frac{\eta}{\eta + x^\tau}\right)^\lambda, \quad x > 0, \quad \eta, \tau, \lambda > 0,$$

We consider $\eta = 1, \lambda = 2, \tau = 1/2$; $\eta = 2, \lambda = 1, \tau = 2$; and $\eta = 2$. Notice that $\xi = 1/(\lambda\tau)$

- The Fréchet distribution with tail

$$\bar{F}(x) = 1 - \exp(-x^{-\alpha}), \quad \alpha > 0.$$

We consider $\xi = 1, 1/2$.

For the expert information we draw a single random number from a Gaussian distribution centered at the true ξ and with standard deviation 0.2, and define that value as $1/\beta$. Then, by moment matching, using a variance of 0.04, we obtain the parameters a, b needed for $\hat{\xi}_k^{BG}$. Notice that we input the true value of the variance, and hence we are giving additional information to the Bayesian setting, opposed to $\hat{\xi}_k^P$, where we make no such assumptions and we use the combination with $\lambda = 1$. Additional studies (which we omit here) show that if the Bayesian variance is not correctly specified (for instance, set at 1 or 0.5), the Bayesian solution behaves almost identically to the censored Hill estimator $\hat{\xi}_k^{MLE}$.

Also notice the misspecification of the Gamma prior in the derivation of $\hat{\xi}_k^{BG}$ with respect to the Gaussian distribution from which the expert information is actually simulated. Using a Gaussian prior would not only make explicit posterior formulas not available (and hence the need to resort to MCMC sampling methods as Gibbs

sampling), but would add more information than what we have assumed is available throughout the paper (we have not even assumed knowledge of the variance).

We then plot the empirical bias and MSE of each resulting estimator as a function of k (comparing the estimates with respect to the true value). We write expressions such as $\text{Burr}(\xi = 1)$ to indicate that the parameters of the distribution are not the focus, but rather the resulting tail index from the Hall class (which is a function of the parameters). The results are given in Figures 5.2, 5.3 and 5.4. We observe how the fact that the perturbation will affect estimation based on the proportion of censored observations as opposed to the total amount of data-points performs well for $k > 10$. As a result, a substantial amount of bias and MSE is removed. This is especially the case for the heavy-tail case $\xi = 1$. For the lighter tail $\xi = 0.5$, the perturbed estimator has either the best or second best performance bias-wise, and its only major drawback is the MSE for the exact Pareto case for $k > 50$, where the Bayesian gamma solution performs even worse. When considering quantiles, the perturbed estimator behaves better than the Hill estimator and on par with the other two benchmarks for the heavy-tail exact Pareto case. In the lighter tail case it performs the worst for large order statistics, recovering and behaving as in the previous case for $k < 60$. In all other non-exact Pareto tail cases, the perturbation was superior to all methods.

Notice that one assumption made in this study was that the expert guess was centered and with relatively good quality (mean ξ and standard deviation 0.2). If the latter conditions are changed, it is easy to construct a simulation study where both the perturbed and the Bayesian gamma solution perform much worse. Consequently, the findings of this simulation study suggest that insurers that are very confident in their expert opinions might benefit from using the combination estimator $\hat{\xi}_k^P$ with $\lambda = 1$.

Remark 5.5.1. For the adaptive selection of λ , the procedure of cross-validation may naturally come to mind. However, the latter is based on averages rather than extremes. For instance, in a 10-fold cross-validation, the 9 parts of the data that do not contain the maximum will tend to prefer lighter tail indices, and only the one part with the maximum will suggest a heavy-tail index, so that the overall index will be underestimated. Correspondingly, cross-validation is not a method of choice in this context.

5.5.2 Insurance Data

We consider a dataset from Motor Third Party Liability Insurance (MTPL) from a direct insurance company operating in the EU (cf. (Albrecher et al., 2017, Sec.1.3.1)), consisting of yearly paid amounts to policyholders during the period 1995-2010. At 2010 we have roughly 60% right-censored (open) observations out of the total 837 claims. The data are reported as soon as the incurred value exceeds the reporting threshold given in Figure 1.2 in Albrecher et al. (2017), and the histogram of the IBNR delays is given in Figure 1.3 in Albrecher et al. (2017). We also have an *ultimate* estimate which is the company's expert estimation of the eventual size of the claim. In Figure 5.5 we have several descriptive statistics of the

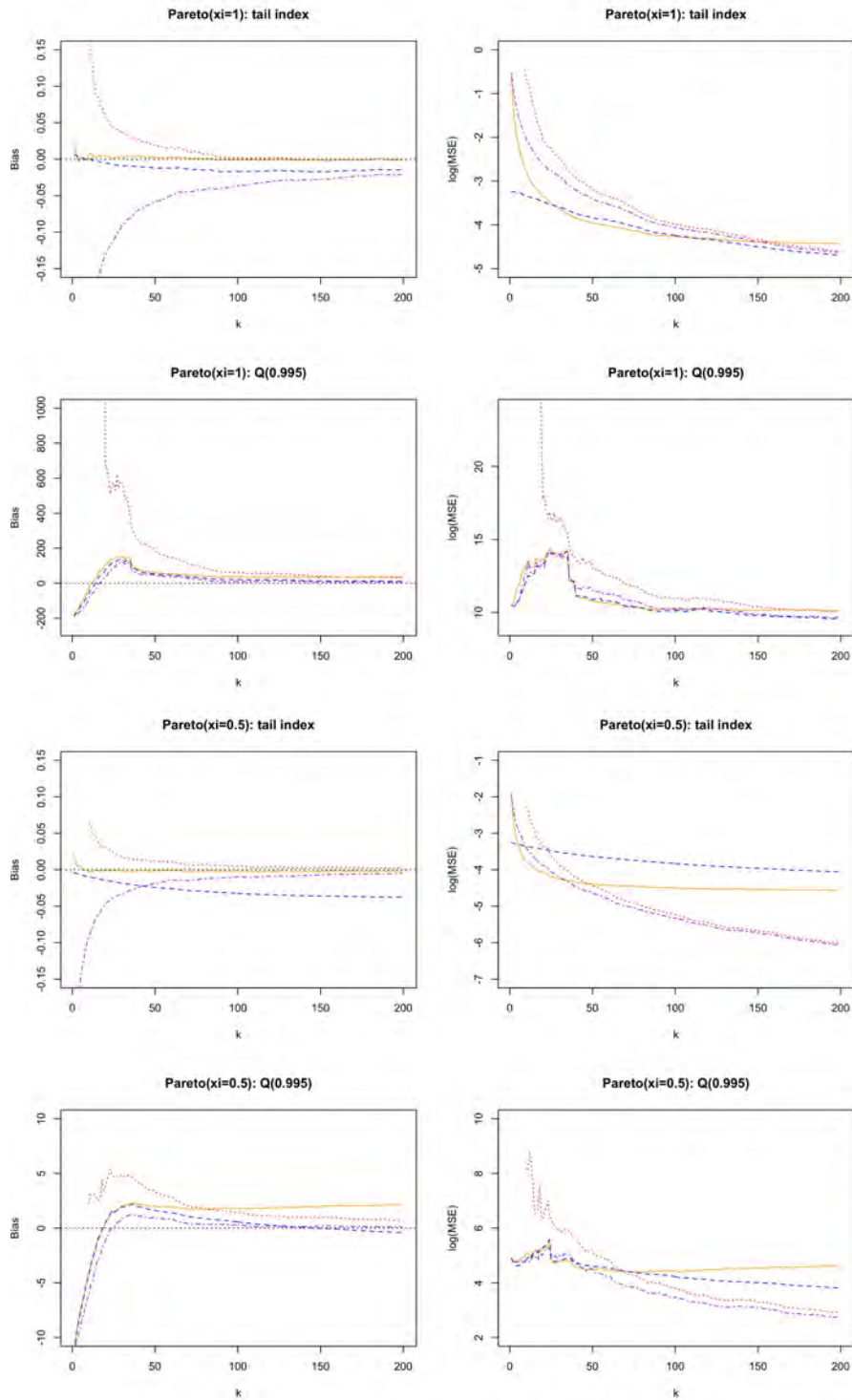


Figure 5.2: Bias and (log) Mean Square Error for the exact Pareto distribution, for varying parameters. We compare $\hat{\xi}_k^P$ (orange, solid), $\hat{\xi}_k^{MLE}$ (red, dotted), $\hat{\xi}_k^{BG}$ (blue, dashed) and $\hat{\xi}_k^{BM}$ (purple, dashed and dotted), as well as the associated Weissman quantile estimator.

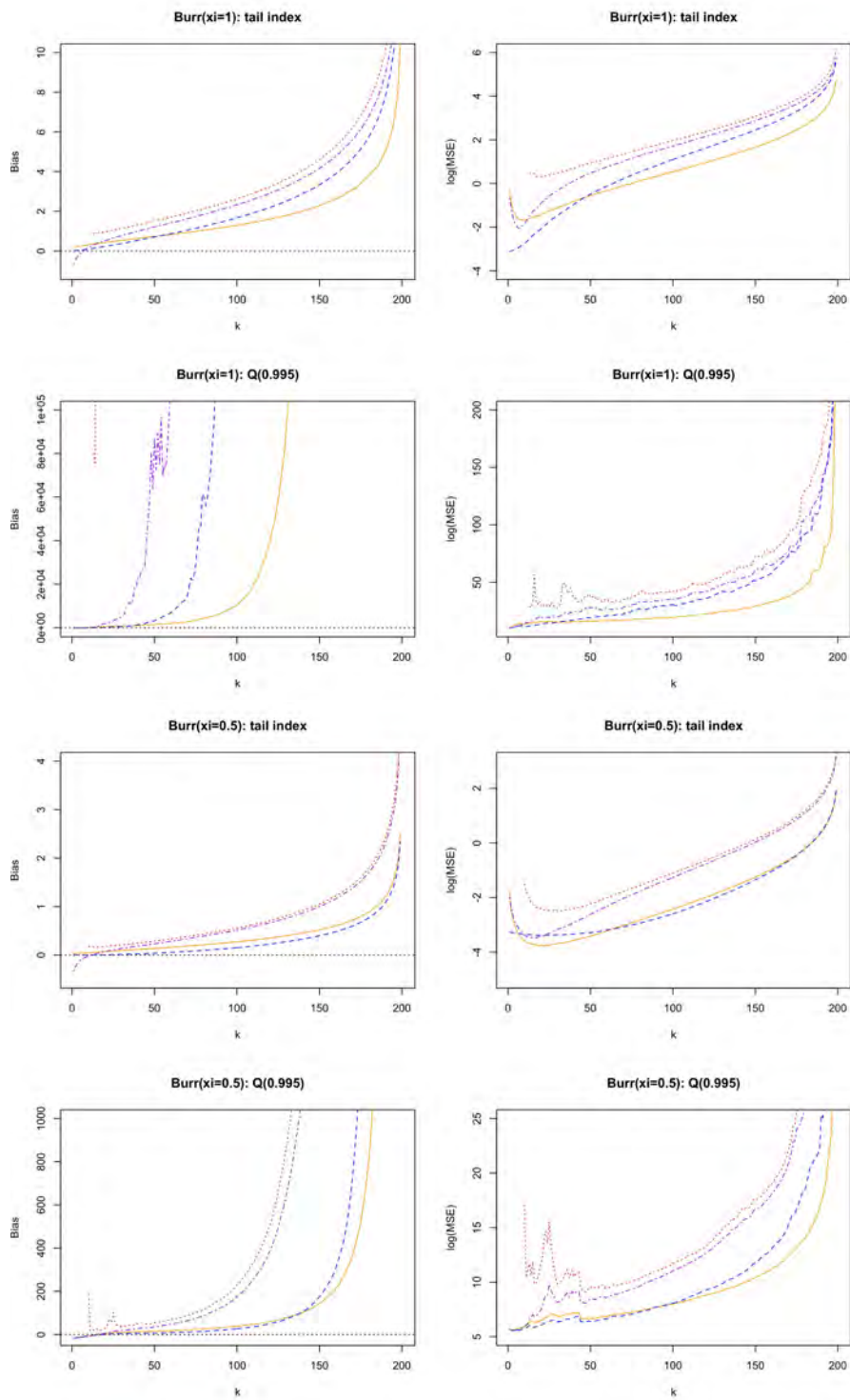


Figure 5.3: Bias and (log) Mean Square Error for the Burr distribution, for varying parameters. We compare $\hat{\xi}_k^P$ (orange, solid), $\hat{\xi}_k^{MLE}$ (red, dotted), $\hat{\xi}_k^{BG}$ (blue, dashed) and $\hat{\xi}_k^{BM}$ (purple, dashed and dotted), as well as the associated Weissman quantile estimator.

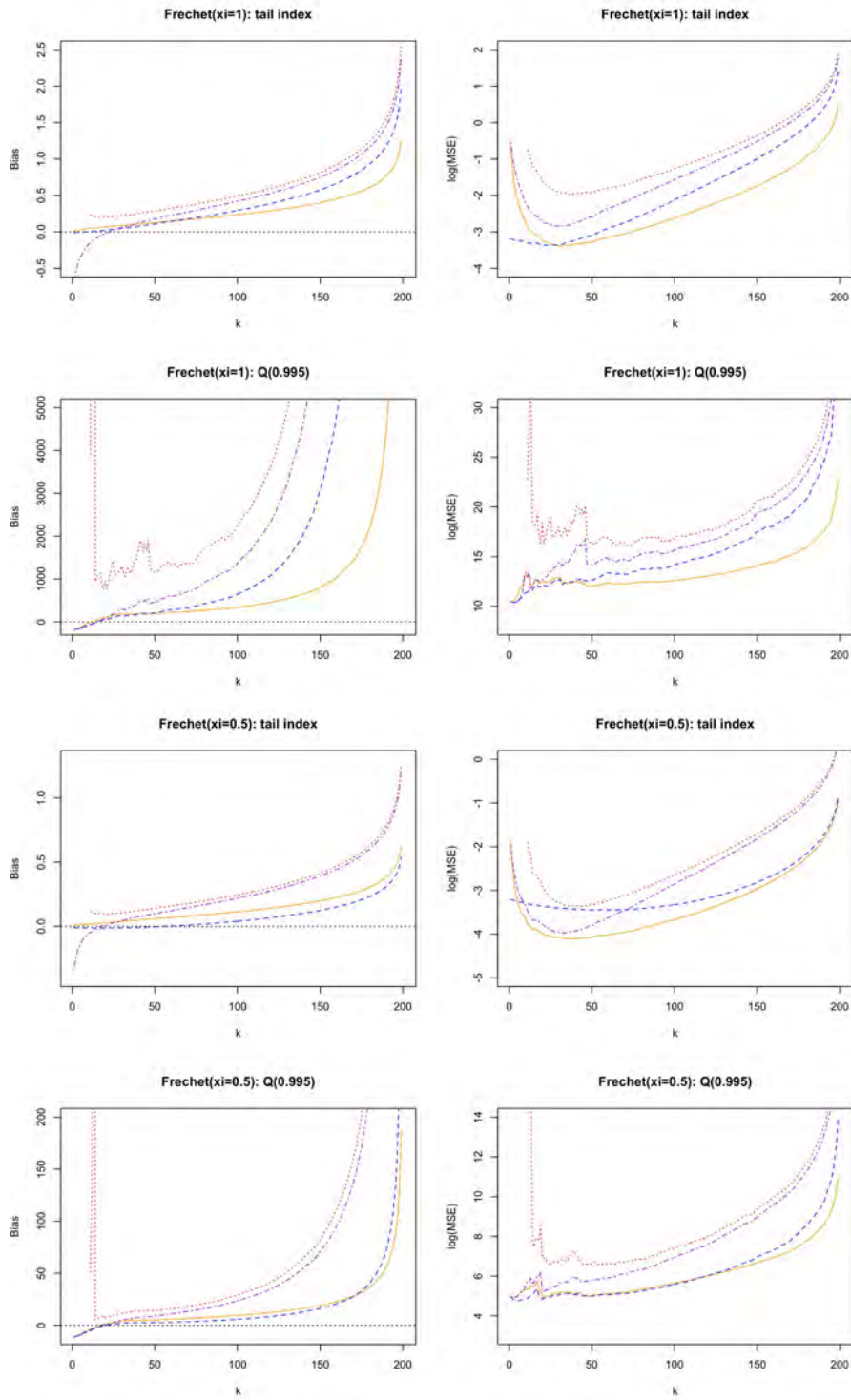


Figure 5.4: Bias and (log) Mean Square Error for the Frechet distribution, for varying parameters. We compare $\hat{\xi}_k^P$ (orange, solid), $\hat{\xi}_k^{MLE}$ (red, dotted), $\hat{\xi}_k^{BG}$ (blue, dashed) and $\hat{\xi}_k^{BM}$ (purple, dashed and dotted), as well as the associated Weissman quantile estimator.

data: the log-claim sizes, the Kaplan-Meier estimator of the data (cf. Kaplan and Meier (1958)), the proportion of non-censoring (closed claims) as a function of the order statistics of the claims, and a QQ-plot for the log-claims against theoretical exponential quantiles. We observe that censoring at random is not a far-fetched assumption to make, since a horizontal behaviour of the proportion of closed claims as a function of the number of upper order statistics does not reject the possibility of the sizes of claims and the probabilities of censoring of those claims being independent. The Pareto tail behaviour of large (above 0.45 million, possibly censored) claim sizes seems to hold, based on the QQ-plot of their logarithm against theoretical exponential quantiles. Standard tests also do not reject the exponential hypothesis of the logarithm of these large claims (Kolmogorov-Smirnov p-value of 0.50).

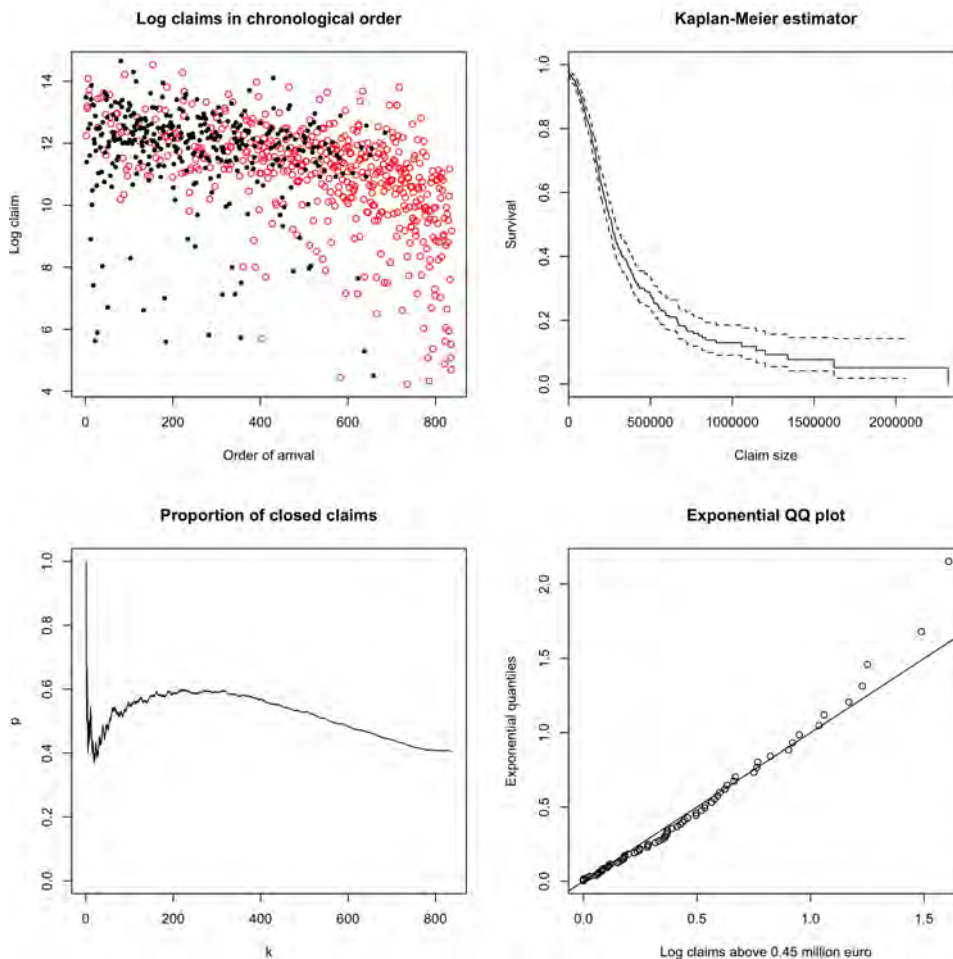


Figure 5.5: Descriptive statistics of the insurance data. Top left: log-claims in order of arrival, showing both open (red, circle) and closed (black, dot) claims. Top right: Kaplan-Meier survival probability estimator for the claims. Bottom left: proportion of closed claims as a function of the top k order statistics of the claim sizes. Bottom right: QQ plot of the logarithm of the claims larger than 0.54 million euro, against the theoretical exponential quantiles with the same mean.

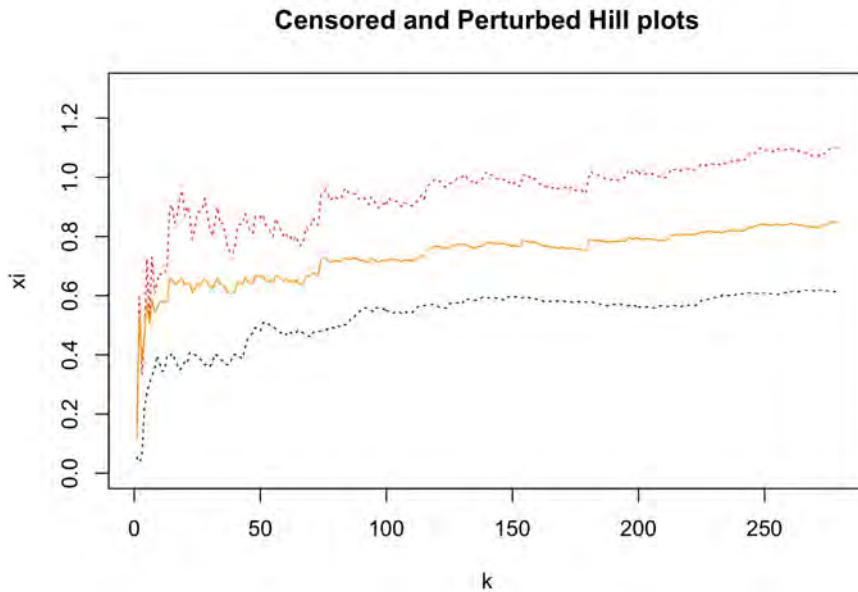


Figure 5.6: Hill plot of the ultimates (black, dashed), censored Hill estimator $\hat{\xi}_k^{MLE}$ for the claims (red, dashed), and the combined estimator $\hat{\xi}_k^P$ (orange, solid) with $\lambda = 1$ and $\beta = 1/0.48$.

Now we would like to know how the ultimates can help to estimate the tail parameter. In Bladt et al. (2019), the ultimates of this dataset were explored, where it was observed that they are Pareto in the tail. Furthermore, using developments in threshold selection, using trimming techniques, it was shown that $\xi = 0.48$ is a good estimate for the heaviness of the tail of the ultimates. In Figure 5.6 we show the Hill plot for the ultimates, together with the chosen expert ξ value, and the censored Hill $\hat{\xi}_k^{MLE}$ and the perturbed version $\hat{\xi}_k^P$ with $\lambda = 1$ and $\beta = 1/0.48$. Notice that in this case, we know how β is obtained, and this additional knowledge could be useful. However, our method does not assume any specific structure, which means that any other method can be used to obtain β , and we merely give the current one for the sake of example. We observe a particularly stable region when k is between 20 and 70, which suggests a heavier tail (roughly 0.65) than the ultimates alone predict. We will see how this affects the quantiles alike.

As a way of validating our estimation procedure we perform the following check. We consider all claims arriving in the shorter period of time 1995-2000 and we follow exclusively these 310 older claims until 2010. The proportion of censoring at 2010 drops to roughly 29.5%. We examine the censored Hill estimator and perturbed estimator (using the same $\lambda = 1$ and $1/\beta = 0.48$ as before) for this reduced data, and plot it in Figure 5.7, together with the corresponding estimators using the full data which we had previously obtained. We observe that the censored Hill estimator for the reduced data dropped its value in the most stable region by about 0.2, almost reaching the perturbed estimator for both the complete and reduced datasets, showing that as the proportion of censoring decreases, the estimators come closer together. Notice that the perturbed estimator remained surprisingly stable,

even when the penalization parameter stayed at the same value but the sample size decreased, due to the fact that the proportion of censored claims controlled the strength of the penalization in a natural way.

Finally, we add the corresponding analysis of the 99.5% quantile (which is relevant for Value-at-Risk considerations) for the case where the expert quantile information is given by the empirical distribution function of the ultimates, and combine the Hill estimator and the expert information by means of Equation (5.29). That is,

$$\begin{aligned}\hat{Q}_k^P(1-p) &= \left[\hat{Q}^{KM}(1-k/n) \cdot \left(\frac{k}{np} \right)^{\hat{\xi}_k^{MLE}} \right]^{\hat{p}_k} \cdot \left[\hat{Q}^{ULT}(1-k/n) \cdot \left(\frac{k}{np} \right)^{1/\beta} \right]^{1-\hat{p}_k} \\ &= \left((\hat{Q}^{KM}(1-k/n))^{\hat{p}_k} (\hat{Q}^{EX}(1-k/n))^{1-\hat{p}_k} \right)^{\hat{\xi}_k^P},\end{aligned}$$

where $\hat{p}_k = \frac{1}{k} \sum_{i=1}^k e^{(i)}$, \hat{Q}^{KM} is the quantile function associated with the Kaplan-Meier curve of the claims, and \hat{Q}^{ULT} is the quantile function associated with the empirical distribution function of the ultimates.

The quantile coming from the ultimates alone is given by

$$\hat{Q}_k^{ULT}(1-p) = \hat{Q}^{ULT}(1-k/n) \cdot \left(\frac{k}{np} \right)^{H_k^U},$$

where H_k^U is the Hill estimator of the ultimates. Finally, without any expert information (ignoring the ultimates), the quantile is given by

$$\hat{Q}_k^{KM}(1-p) = \hat{Q}^{KM}(1-k/n) \cdot \left(\frac{k}{np} \right)^{\hat{\xi}_k^{MLE}}.$$

Note that, due to missing IBNR data at the later accident years, some care is needed concerning the interpretation of these quantile estimates as the outcome levels which are exceeded in $100 \times 0.5\%$ of the reported cases. However, as these IBNR data concern smaller losses, the influence of these omissions is limited as can be verified by restricting the proposed approach to the claims from earlier accident years and comparing with the present results. The result is gathered as a function of the number k of upper order statistics in Figure 5.8. The combined estimation of the high quantile results is a stable compromise between the under-estimated quantiles from the expert opinions and the pure Weissman approach, which has higher variability. Such under-estimation of the size of the claims at closure by the ultimates was also observed empirically while exploring the data (details are omitted). Observe also Figure 5.9, where the reduction of the data which was used above to validate the procedure was applied to the quantiles, and an analogous interpretation applies. This suggests that the current reserving could benefit from a re-evaluation. However, this analysis is made without knowing the actual process behind the calculation of the ultimates, and a deeper understanding of this process could in the future elucidate whether there is something being overlooked by the experts or by the statisticians.

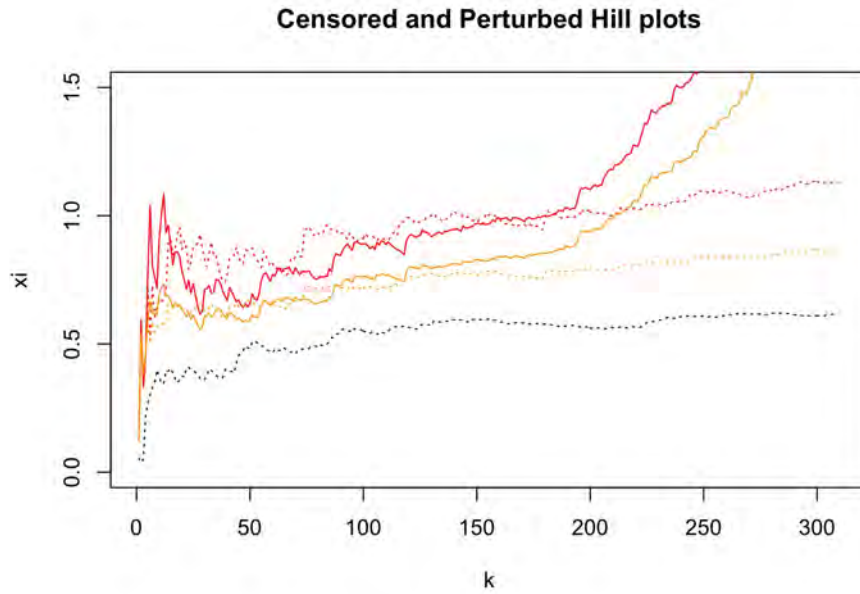


Figure 5.7: Hill plot of the ultimates (black), for the reduced (solid) and complete (dashed) datasets: censored Hill estimator $\hat{\xi}_k^{MLE}$ for the claims (red), and the combined estimator $\hat{\xi}_k^P$ (orange) with $\lambda = 1$ and $\beta = 1/0.48$.

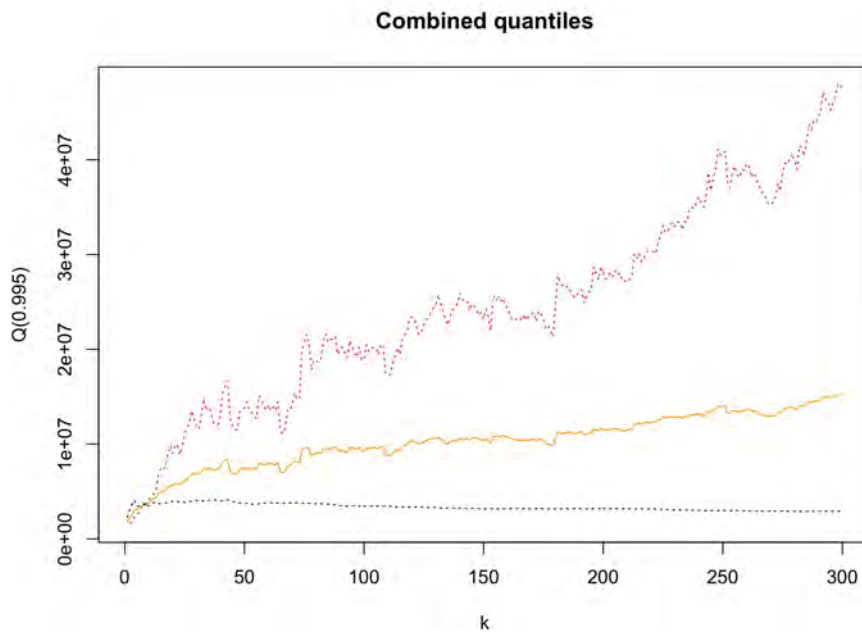


Figure 5.8: 99.5% quantile estimator using the censored approach ($\hat{Q}_k^{KM}(0.005)$, red) for the claims, expert information ($\hat{Q}_k^{ULT}(0.005)$, black) and their combination via $\hat{\xi}_u^P$, with the selection $\lambda = 1$ ($\hat{Q}_k^P(0.005)$, orange).

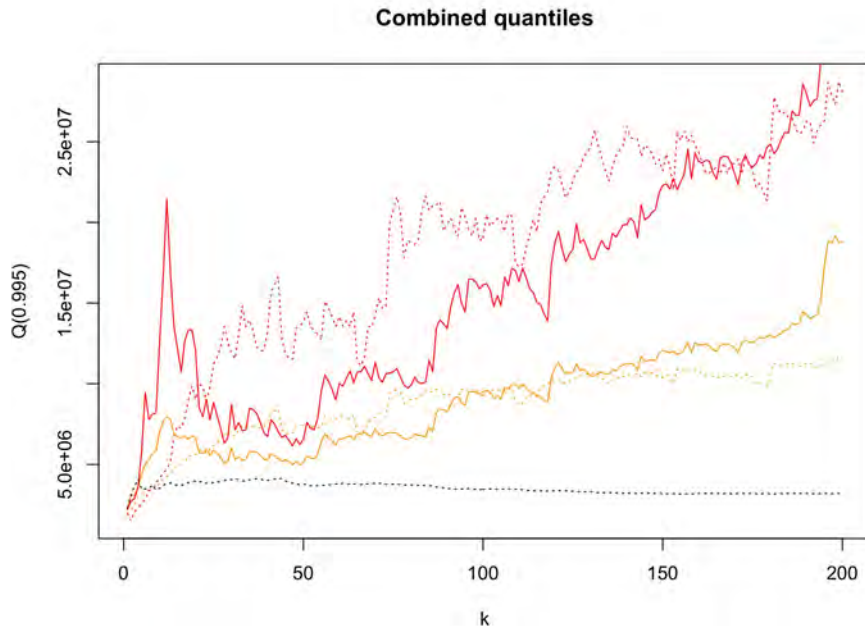


Figure 5.9: Plot of the 99.5% quantile for the ultimates \hat{Q}_k^{ULT} (black), and for the reduced (solid) and complete (dashed) datasets: censored Hill estimation \hat{Q}_k^{KM} for the claims (red), and the combined estimator \hat{Q}_k^P (orange) with $\lambda = 1$ and $\beta = 1/0.48$.

5.6 Conclusion

We have derived a flexible estimator that bridges statistical theory and practice when it comes to tail estimation. The results also apply to adaptation of quantile estimation techniques both when more expert information is available (for instance when an expert cumulative distribution function is available) and also when it is lacking. Like in Bayesian statistics, the strength of the belief of the expert is often subjective and in many cases unquantifiable, especially when provided with a single point estimate. As discussed in the paper, our method is in fact closely related to Bayesian techniques, but it is driven by the proportion of censoring, rather than by the number of total observations. The developed estimator represents a statistically sound method for making a compromise between expert information and likelihood methods, without the need of any additional prior assumptions, and its performance depends on the quality of the expert guess. In particular, we suggested a convenient approach to avoid selection of a tuning parameter for the linking of expert information and Hill estimation. The methods developed can readily be adapted for the selection of the tuning parameter using more complex methods (such as moment matching) whenever there is more expert information available than presently assumed.

For heavy-tails, the estimator is shown to be asymptotically normal, and has further desirable properties when the tuning parameter is chosen to be 1. Indeed, Theorem 5.4.4 can serve as a simple rule of thumb in practice for combining the

two sources of information, and suggests that using good quality expert information can reduce the variance while keeping the bias at bay. This rule appears to be rather natural, and the approach in this paper enables to embed this intuitive combination within the theory of perturbed likelihood estimation. A more detailed analysis would depend on the specific application at hand, and on the quantifiability of the strength of beliefs, which in the present liability insurance dataset, and more generally in any analysis made by statisticians without the experts present, is commonly lacking.

A simulation study showed that when the guess is close to being correct, the estimator fares very favorably, compared to the Hill estimator and two recently proposed Bayesian solutions. Moreover, the estimator seems to be quite stable with respect to the chosen threshold, which is of particular interest since the choice of an appropriate threshold is a classical problem in extreme value analysis. Concerning quantiles, and for the simulated examples, the estimator was favorable to all the benchmarks for virtually all sample fractions for non-exact Pareto tails. Trimming techniques have recently been proposed to address threshold selection, and a future line of research will be to consider lower-trimmed versions of the proposed perturbed estimator to aid in the visual and automatic sample fraction selection.

Finally, the application of the method to actual motor third-party liability liability insurance data illustrates that decision makers with a strong belief in a point estimate of the tail parameter could be less reluctant to use the tail parameter and quantiles suggested by the inclusion of data-points proposed by our method than the one from the pure censored Hill estimation of the data.

Appendix

Proof of Theorem 5.4.1

Proof. Define

$$V_k = \frac{1}{k} \sum_{i=1}^k (1 - e^{(i)})(1 - \lambda), \quad W_k = \frac{1}{k} \sum_{i=1}^k (1 - e^{(i)})/\beta. \quad (5.32)$$

First note that

$$V_k \xrightarrow{d} r_1, \quad W_k \xrightarrow{d_p} r_2.$$

Concerning the asymptotic distribution of V_k and W_k we make use of the method developed in Einmahl et al. (2008) introducing i.i.d. uniform (0,1) random variables U_i , $i \geq 1$, independent of the Z_i sequence, and corresponding indicators being equal to 1 if $U \leq p(Z)$, and 0 otherwise. When denoting the U variable induced by $Z^{(i)}$ by $U^{(i)}$ we have

$$V_k \stackrel{d}{=} \frac{1}{k} \sum_{j=1}^k 1_{\{U^{(j)} > p(Z^{(j)})\}} (1 - \lambda) \quad \text{and} \quad W_k \stackrel{d}{=} \frac{1}{\beta k} \sum_{j=1}^k 1_{\{U^{(j)} > p(Z^{(j)})\}},$$

which, under (5.20), can be replaced asymptotically by

$$\hat{V}_k = \frac{1}{k} \sum_{j=1}^k 1_{\{U^{(j)} > p(U(n/i))\}}(1 - \lambda) \text{ and } \hat{W}_k = \frac{1}{\beta k} \sum_{j=1}^k 1_{\{U^{(j)} > p(U(n/i))\}},$$

for which

$$\sqrt{k}(\hat{V}_k - r_1) \xrightarrow{d} Y_1, \quad \sqrt{k}(\hat{W}_k - r_2) \xrightarrow{d} Y_2,$$

where

$$Y_1 \sim \mathcal{N}\left(-\frac{\delta\nu_*\alpha_c}{\alpha_c + \nu_*}p(1-p)C^{-\nu_*/\alpha_c}(D/\alpha)_*(1-\lambda), p(1-p)(1-\lambda)^2\right),$$

$$Y_2 \sim \mathcal{N}\left(-\frac{\delta\nu_*\alpha_c}{\alpha_c + \nu_*}p(1-p)C^{-\nu_*/\alpha_c}(D/\alpha)_*/\beta, p(1-p)/\beta^2\right),$$

which are independent of Y_0 , defined in (5.19). Then we obtain

$$\begin{aligned} \sqrt{k} \left(\frac{H_k + \lambda \hat{W}_k}{1 - \hat{V}_k} - \frac{1 + \lambda \alpha_2 / \beta}{\alpha + \lambda \alpha_2} \right) &= \frac{1}{1 - \hat{V}_k} \sqrt{k} \left(H_k - \frac{1}{\alpha_c} \right) + \frac{\lambda}{1 - \hat{V}_k} \sqrt{k} (\hat{W}_k - r_2) \\ &\quad + \left(\lambda r_2 + \frac{1}{\alpha_c} \right) \sqrt{k} \left(\frac{1}{1 - \hat{V}_k} - \frac{1}{1 - r_1} \right) \\ &= \frac{1}{1 - \hat{V}_k} \sqrt{k} \left(H_k - \frac{1}{\alpha_c} \right) + \frac{\lambda}{1 - \hat{V}_k} \sqrt{k} (\hat{W}_k - r_2) \\ &\quad + \frac{\lambda r_2 + \frac{1}{\alpha_c}}{(1 - r_1)(1 - \hat{V}_k)} \sqrt{k} (\hat{V}_k - r_1) \\ &\xrightarrow{d} \frac{1}{1 - r_1} Y_0 + \frac{\lambda}{1 - r_1} Y_2 + \frac{\lambda r_2 + \frac{1}{\alpha_c}}{(1 - r_1)^2} Y_1 \\ &= \frac{1}{1 - r_1} Y_0 + \left[\frac{\lambda}{1 - r_1} \frac{1}{\beta(1 - \lambda)} + \frac{\lambda r_2 + \frac{1}{\alpha_c}}{(1 - r_1)^2} \right] Y_1 \end{aligned}$$

The mean and variance are then computed from the last expression, since Y_0 and Y_1 are independent. \square

Chapter 6

Threshold selection and trimming in extremes

This chapter is based on the following article, currently submitted for publication:

Bladt, M., Albrecher, H., & Beirlant, J. (2019). Trimming and threshold selection in extremes. arXiv preprint arXiv:1903.07942.

Abstract

We consider removing lower order statistics from the classical Hill estimator in extreme value statistics, and compensating for it by rescaling the remaining terms. Trajectories of these trimmed statistics as a function of the extent of trimming turn out to be quite flat near the optimal threshold value. For the regularly varying case, the classical threshold selection problem in tail estimation is then revisited, both visually via trimmed Hill plots and, for the Hall class, also mathematically via minimizing the expected empirical variance. This leads to a simple threshold selection procedure for the classical Hill estimator which circumvents the estimation of some of the tail characteristics, a problem which is usually the bottleneck in threshold selection. As a by-product, we derive an alternative estimator of the tail index, which assigns more weight to large observations, and works particularly well for relatively lighter tails. A simple ratio statistic routine is suggested to evaluate the goodness of the implied selection of the threshold. We illustrate the favourable performance and the potential of the proposed method with simulation studies and real insurance data.

6.1 Introduction

The use of Pareto-type tails has been shown to be important in different areas of risk management, such as for instance in computer science, insurance and finance. In social sciences and linguistics the model is referred to as Zipf's law. This model corresponds to the max-domain of attraction of a generalized extreme value distri-

bution with a positive extreme value index (EVI) ξ :

$$1 - F(x) = x^{-1/\xi} \ell(x), \quad \xi > 0, \quad (6.1)$$

where ℓ denotes a slowly varying function at infinity:

$$\lim_{x \rightarrow \infty} \frac{\ell(ux)}{\ell(x)} = 1, \quad \text{for every } u > 0. \quad (6.2)$$

Since the appearance of the paper of Hill (1975) in which the EVI estimator

$$H_{k,n} = \frac{1}{k} \sum_{i=1}^k \log X_{n-j+1,n} - \log X_{n-k,n} \quad (6.3)$$

was proposed with

$$X_{n,n} \geq X_{n-1,n} \geq \cdots \geq X_{n-i+1,n} \geq \cdots \geq X_{1,n}$$

denoting the ordered statistics of a random sample from F , the literature on estimation of $\xi > 0$ and other tail quantities such as extreme quantiles and tail probabilities has increased exponentially. We refer to Embrechts et al. (1997), Beirlant et al. (2004), de Haan and Ferreira (2007) and Gomes and Guillou (2015) for detailed discussions and reviews of these estimation problems. Next to the proposal of numerous estimators, focus has gradually shifted to selection methods of k and to the construction of bias-reduced estimators which exhibit plots of estimates which, as a function of k , are as stable as possible. Indeed, plots of estimators of ξ as a function of k that are consistent under the large semi-parametric model (6.1) are hard to interpret. In case of the Hill estimator some authors refer to Hill horror plots. While it has been frequently suggested to choose a 'stable' area (see for instance Drees et al. (2000) and De Sousa and Michailidis (2004)), such a stable part is often absent or hard to find. Sometimes more than one stable section is present, like in some insurance applications as we will discuss later.

The typical available guidelines for the choice of k to be used in the implementation of the EVI estimators depend strongly on the properties of the tail itself, and k needs to be estimated adaptively from the data. This problem can be compared with choosing a bandwidth parameter in density estimation. It is typically suggested that the optimal value of k should be the one that minimizes the mean-squared error (MSE). However, this optimum depends on the sample size, the unknown value of ξ as well as on the nature of ℓ , as was first described in Hall et al. (1985). Bootstrap methods were proposed in Hall (1990), Draisma et al. (1999), Danielsson et al. (2001), and Gomes and Oliveira (2001). Beirlant et al. (1996, 2002) derived regression diagnostic methods on a Pareto quantile plot. Other selection procedures can be found in Drees and Kaufmann (1998) and Guillou and Hall (2001). Possible heuristic choices are provided in Gomes and Pestana (2007), Gomes et al. (2008) and Beirlant et al. (2011). Almost all authors consider the adaptive choice of k for the Hill estimator, while methods can be adapted to other estimators as well.

In this paper we consider trimming of the Hill estimator, omitting some of the lower order statistics in $X_{n-k+1,n}, \dots, X_{n,n}$, which leads to statistics of the type

$$T_{b,k} = \sum_{i=1}^b c_i(b, k) \log \left(\frac{X_{n-i+1,n}}{X_{n-k,n}} \right), \quad (6.4)$$

for some $1 \leq b \leq k$ and suitable constants $c_i(b, k)$. This kind of kernel-type statistics have been previously proposed (cf. Csörgő et al. (1985)) as estimators of ξ . However, the implementation of the optimal kernel is not an easy task nor our focus in this paper. Instead, we propose a special form of the kernel that leads to an identity which aids in the threshold estimation problem. In Section 2 we derive the coefficients $c_i(b, k)$ which make $T_{b,k}$ unbiased when ℓ is constant and when we force the coefficients $c_i(b, k) = c(b, k)$ not to depend on i . We present a novel lower-trimmed Hill plot which provides significant graphical support for the estimation problem of ξ , as we illustrate with both simulations and real world data. We also provide mathematical evidence that, as a function of b , the variability of the $T_{b,k}$ statistics is lower than the one in the Hill plot. In Section 3, we examine the asymptotic characteristics of $T_{b,k}$ in (6.4) under the general model (6.1). The asymptotic expected empirical variance of $T_{b,k}$ is shown to be less sensitive on the tail parameter ξ than the asymptotic mean-squared error (AMSE) of the usual Hill estimator (6.3). We identify a link between the corresponding two optimal k -choices which allows to bypass the specification of ξ and other characteristics of the tail behavior for the identification of the optimal threshold in the classical Hill estimate, and the resulting procedure turns out to be simple to implement in practice. Subsequently, we study the estimator \bar{T}_k obtained by averaging the trimmed Hill estimators over $b = 1, \dots, k$. This latter estimator naturally assigns more weight to the larger observations, the weights being only moderately changed when increasing k . Furthermore, the specification of these weights is independent of the distribution F . Note that, in contrast, earlier criteria for reweighting terms in the Hill estimator (such as e.g. Csörgő et al. (1985) in terms of kernel estimates, see also (Beirlant et al., 2002, Sec.3)) had to heavily rely on the tail parameter ξ . In Section 4 we then present a simple ratio statistic as a tool to evaluate the goodness of selection of k . Section 5 confirms the good performance of the proposed methods using simulations, where \bar{T}_k turns out to outperform the classical Hill estimator in almost all cases. Note that our approach eventually suggests a fully automated procedure for the threshold selection, also in the absence of knowledge about, or assumptions on, the tail characteristics. Section 6 favorably illustrates this on a set of real-life motor third party liability insurance data. We would like to emphasize that the approach proposed in this paper suggests a general procedure that can in principle also be applied to other estimators in extreme value analysis.

6.2 A lower-trimmed Hill statistic

6.2.1 Derivation

Assume first, for simplicity, that we have independent and identically distributed (i.i.d.) exact Pareto random variables, X_1, X_2, \dots, X_n , with tail given by

$$\bar{F}(x) = (x/\sigma)^{-1/\xi}, \quad x \geq \sigma, \quad \xi, \sigma > 0, \quad (6.5)$$

and we are interested in robust estimation of the tail index ξ .

A main tool used throughout the paper is the well-known *Rényi representation*, which states (in the second distribution equality below), that for the order statistics of a random sample X_1, \dots, X_n from the distribution (6.5), one has, for $k \leq n$,

$$\left(\log \left(\frac{X_{n,n}}{X_{n-k,n}} \right), \dots, \log \left(\frac{X_{n-k+1,n}}{X_{n-k,n}} \right) \right) \stackrel{d}{=} (E_{k,k}, \dots, E_{1,k}) \stackrel{d}{=} \left(\sum_{j=1}^k \frac{E_j^*}{k-j+1}, \dots, \frac{E_1^*}{k} \right). \quad (6.6)$$

Here, $E_{k,k} \geq \dots \geq E_{1,k}$ are the order statistics of an independent i.i.d. exponential sample E_1, \dots, E_k with mean ξ , and E_1^*, \dots, E_k^* is another independent i.i.d. exponential sample with mean ξ .

Bhattacharya et al. (2017) recently proposed linear estimators of the form

$$\hat{\xi}_{k_0,k} = \sum_{i=k_0+1}^k c_{k_0,k}(i) \log(X_{n-i+1,n}/X_{n-k,n}), \quad 0 \leq k_0 < k < n,$$

in order to trim the upper order statistics in outlier-contaminated samples, where the constants $c_{k_0,k}(i)$ are chosen in a way to ensure that the resulting estimator for ξ is unbiased. For fixed k_0, k , the problem can then be recast into that of finding suitable weights δ_i such that one can write

$$\hat{\xi}_{k_0,k} = \sum_{i=k_0+1}^k c_{k_0,k}(i) E_{k-i+1,k} = \sum_{i=1}^{k-k_0} \delta_i E_{i,k}.$$

Using the Rényi representation (6.6) and solving some elementary linear equations, they derived $\delta_i = \frac{1}{r}$, $i < r$, and $\delta_r = (k-r+1)/r$. This led them to the so-called *trimmed Hill estimator*

$$\hat{\xi}_{k_0,k} = \frac{k_0+1}{k-k_0} \log(X_{n-k_0,n}/X_{n-k,n}) + \frac{1}{k-k_0} \sum_{i=k_0+2}^k \log(X_{n-i+1,k}/X_{n-k,n}),$$

which is shown to be quite useful in outlier detection under (6.1).

In a similar way, but for a different purpose, in this paper we investigate trimming from the left. Concretely, we consider estimators of the form

$$T_{b,k} = \sum_{i=1}^b c_i(b,k) \log(X_{n-i+1,n}/X_{n-k,n}), \quad 0 < b \leq k,$$

where $c_i(b, k)$ are constants to be determined. As above, we would like to find suitable weights γ_i such that

$$T_{b,k} = \sum_{i=1}^b c_i(b, k) E_{k-i+1, k} = \sum_{i=k-b+1}^k \gamma_i E_{i, k} \quad (6.7)$$

Setting $q = k - b + 1$, the Rényi representation (6.6) yields

$$\begin{aligned} T_{b,k} &= \sum_{i=q}^k \gamma_i E_{i, k} = \sum_{i=q}^k \gamma_i \sum_{j=1}^i \frac{E_j^*}{k-j+1} \\ &= \sum_{j=1}^k E_j^* \sum_{i=j \vee q}^k \frac{\gamma_i}{k-j+1} = \sum_{j=1}^k \bar{\gamma}_j E_j^* \end{aligned}$$

with $\bar{\gamma}_j := \sum_{i=j \vee q}^k \frac{\gamma_i}{k-j+1}$. Here we use the notation $j \vee q = \max\{j, q\}$. Unfortunately, the set of equations

$$\bar{\gamma}_j = \frac{1}{k}, \quad j = 1, \dots, k,$$

has no solution (for $j \leq q$ the left-hand-side cannot remain constant in j). Instead, we choose to set

$$\gamma_q = \gamma_{q+1} = \dots = \gamma_k =: \frac{1}{\bar{\omega}(q, k)} \quad (6.8)$$

and

$$\mathbb{E}(T_{b,k}) = \xi \quad (6.9)$$

as the defining equations. The solution of (6.8) and (6.9) is given by

$$\bar{\omega}(q, k) = \sum_{j=1}^k \frac{k-j \vee q + 1}{k-j+1}. \quad (6.10)$$

Plugging (6.10) into (6.7), we then arrive at the following definition of a *lower-trimmed Hill statistic* $T_{b,k}$:

$$\begin{aligned} T_{b,k} &= \sum_{i=q}^r \frac{\log(X_{n-k+i, n}/X_{n-k, n})}{\bar{\omega}(q, k)} = \frac{\sum_{i=1}^b \log(X_{n-i+1, n}/X_{n-k, n})}{\bar{\omega}(k-b+1, k)} \\ &= \frac{\sum_{i=1}^b \log(X_{n-i+1, n}/X_{n-k, n})}{b(1 + \sum_{j=b+1}^k j^{-1})}, \quad b = 1, \dots, k, \quad k < n, \end{aligned} \quad (6.11)$$

where we use the convention $\sum_{j=k+1}^k j^{-1} := 0$.

6.2.2 A lower-trimmed Hill plot

$T_{b,k}$ defined above is unbiased for any b, k , $b \leq k$, by construction. Analogous to the Hill plot, in which $T_{k,k}$ is plotted as a function of k , we now exploit the second degree of freedom and plot, for selected values of k , $T_{b,k}$ as a function of b . That is, the plot is constructed by overlaying the trajectories

$$(b, T_{b,k}), \quad b = 1, \dots, k,$$

for a selection of k values. The lower variance of these trajectories comes from the fact that the normalizing order statistic is fixed, and hence a non-constant behaviour is easier to identify visually than in the classical Hill plot. As a particular consequence, the selection of k that makes the tail resemble a pure Pareto tail is easier to determine, by examining when the trajectories start to be constant.

The following Proposition provides mathematical evidence for the above observations.

Proposition 6.2.1. *As a function of the number b of order statistics being used, in the exact Pareto case (6.5) the estimator $T_{b,k}$ has lower variance than the classical Hill estimator $T_{b,b}$. More precisely,*

$$\mathbb{V}(T_{b,b}) = \frac{\xi^2}{b} \quad \text{and} \quad \mathbb{V}(T_{b,k}) \leq \frac{\xi^2}{\sum_{j=1}^{k-b+1} \left(\frac{b}{k-j+1}\right)^2 + b}.$$

As an illustration, we now compare the performance of these lower-trimmed Hill (LTH) plots for Pareto, near-Pareto and spliced Pareto distributions. The latter is defined through its cumulative distribution function (c.d.f.)

$$F(x; \xi_0, r, c) = \frac{(1 - x^{-1/\xi_0 - r} \mathbf{1}_{\{x \geq c\}}) - \mathbf{1}_{\{x \geq c\}}(c^{-1/\xi_0} - c^{-1/\xi_0 - r})}{1 - c^{-1/\xi_0} + c^{-1/\xi_0 - r}}, \quad x \geq 1 \quad (6.12)$$

for $c \geq 1$ and $r > -1/\xi_0$, which is the c.d.f. of a Pareto random variable with tail index ξ_0 up to some splicing point c , continuously pasted with the c.d.f. of a Pareto random variable with another tail index $\xi = (1/\xi_0 + r)^{-1}$ thereafter. Splicing models (also sometimes referred to as composite models) are for instance popular in reinsurance modelling, cf. (Albrecher et al., 2017, Ch.4).

Concretely we simulated a sample of size $n = 1000$ from a:

- pure Pareto $\xi = \sigma = 1$ sample, defined in (6.5).
- spliced Pareto sample, defined in (6.12), with parameters $\xi = 1$, $\xi_0 = 4$ and splicing point $c = 1.3$.
- spliced Pareto sample, defined in (6.12), with parameters $\xi = 1$, $\xi_0 = 1/4$ and splicing point $c = 1.3$.

- Burr sample with tail $\bar{F}(x) = \frac{1}{1+x}$, $x > 0$ (which amounts to a shifted Pareto).
- Loggamma with logshape parameter $3/2$ and lograte parameter 1 .

The LTH plots together with usual Hill plots are shown in the top panels of Figures 6.1–6.5. The LTH plots are made for a selection of k , from 1 to 1000 by spacings of 50 (1,51,101,...), as a function of the lower trimming b . Recall that $b \leq k$, so the lines have different domains on the x-axis. Observe that the right end-point of each of the overlaid lines corresponds to the respective point in the Hill plot.

For the spliced distributions in Figures 6.2 and 6.3 observe how the LTH estimator becomes horizontal as a function of b when k is close to the (rank of the) splicing point. For smaller k , the plot then looks similar to the exact Pareto case. Loosely speaking, the slope of the lines are a very useful visual tool for detecting the number of upper order statistics k after which a Pareto tail is feasible. This can also be seen in the Burr (Fig.6.4) and loggamma case (Fig.6.5), where the regime of a Pareto tail is only reached for high quantiles.

The bottom panels of Figures 6.1–6.5 suggest two ways of measuring the aforementioned flatness of the LTH estimator as a function of b . The first one computes the empirical variance of $T_{b,k}$, $b = 1, \dots, k$, while the second one fits a linear model with independent variable $b = 1, \dots, k$ and response variable $T_{b,k}$, and then plots the magnitude of the resulting slope coefficient.

6.3 Regularly varying tails

We now move from the simple Pareto sample to a general Fréchet domain of attraction, with tails of the form (6.1). Denote by Q the quantile function associated to F , and define

$$U(x) = Q(1 - 1/x), \quad x > 1,$$

such that the condition (6.1) is equivalent to

$$\lim_{A \rightarrow \infty} \frac{U(Ax)}{U(A)} = x^{-\xi}.$$

Assumptions on the rate of convergence of the above limit make it possible to obtain explicit results concerning asymptotic properties of the lower-trimmed Hill estimator. Hence, we impose the second order condition

$$\lim_{A \rightarrow \infty} \frac{\log U(Ax) - \log U(A) - \xi \log(x)}{Q_0(A)} = \frac{x^p - 1}{p}, \quad (6.13)$$

for some regularly varying function Q_0 with index $p < 0$.

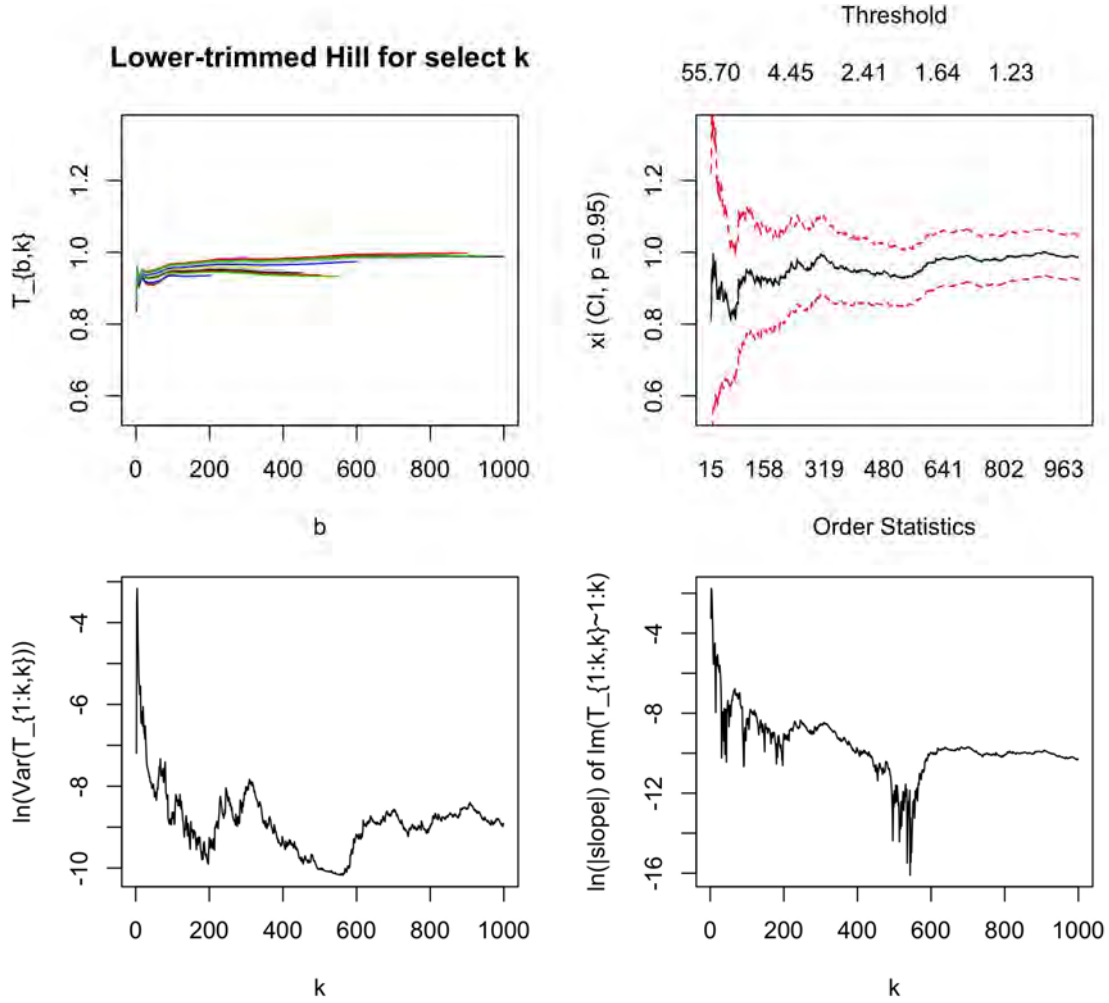


Figure 6.1: Exact Pareto case ($\xi = 1$). Top left: $T_{b,k}$ for varying lower trimming b , for $k = 1, 51, 101, \dots, 1000$. Top right: Hill plot. Bottom left: empirical variance of the LTH as a function of k . Bottom right: slope of a fitted linear model to the LTH as a function of k .

Theorem 6.3.1. *Under the model (6.1) and second order condition (6.13), $T_{b,k}$ as defined in (6.11) satisfies the following asymptotic distributional identity, for $n, k, n/k \rightarrow \infty$,*

$$T_{b,k} \stackrel{d}{=} \xi \frac{\bar{E}_b + \sum_{j=b+1}^k E_j/j}{1 + \sum_{j=b+1}^k j^{-1}} + \frac{Q_0(n/k)}{p} \frac{\frac{((k+1)/b)^p}{1-p} - 1}{1 + \sum_{j=b+1}^k j^{-1}} (1 + o_p(1)), \quad (6.14)$$

where E_1, \dots, E_k are i.i.d. standard exponential random variables, and where we use the notation $\bar{E}_b = b^{-1} \sum_{i=1}^b E_i$.

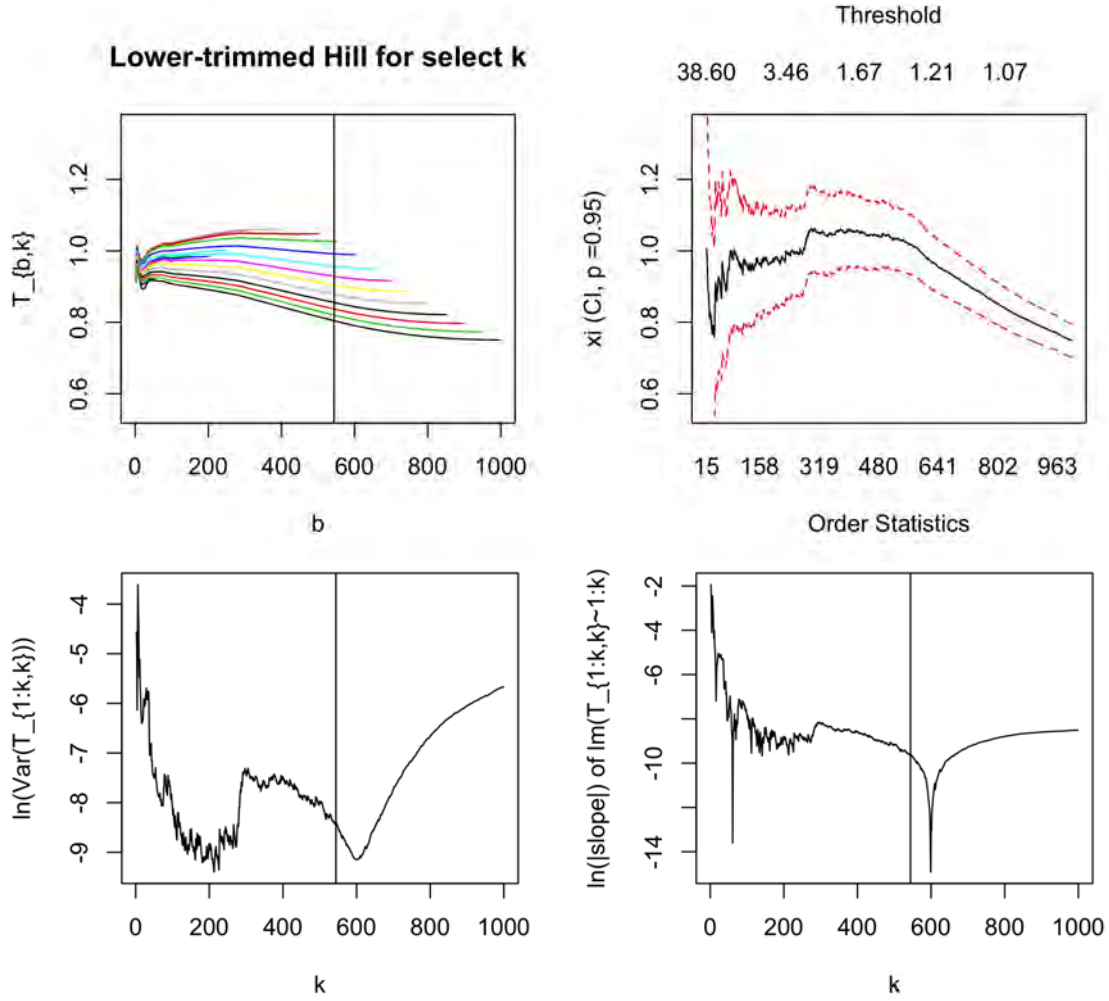


Figure 6.2: Spliced Pareto case (body parameter: 4 and tail parameter: 1). Top left: $T_{b,k}$ for varying lower trimming b , for $k = 1, 51, 101, \dots, 1000$. The vertical line is the splicing location. Top right: Hill plot. Bottom left: empirical variance of the LTH as a function of k . Bottom right: slope of a fitted linear model to the LTH as a function of k .

6.3.1 Distribution of the average

Define the average of the $T_{b,k}$ across b as

$$\bar{T}_k := \frac{1}{k} \sum_{b=1}^k T_{b,k}, \quad (6.15)$$

which by Theorem 3.1 satisfies

$$\bar{T}_k \stackrel{d}{=} \frac{\xi}{k} \sum_{b=1}^k \frac{\bar{E}_b + \sum_{j=b+1}^k E_j/j}{1 + \sum_{j=b+1}^k j^{-1}} + \frac{Q_0(n/k)}{pk} \sum_{b=1}^k \frac{\frac{((k+1)/b)^p - 1}{1-p}}{1 + \sum_{j=b+1}^k j^{-1}} (1 + o_p(1)).$$

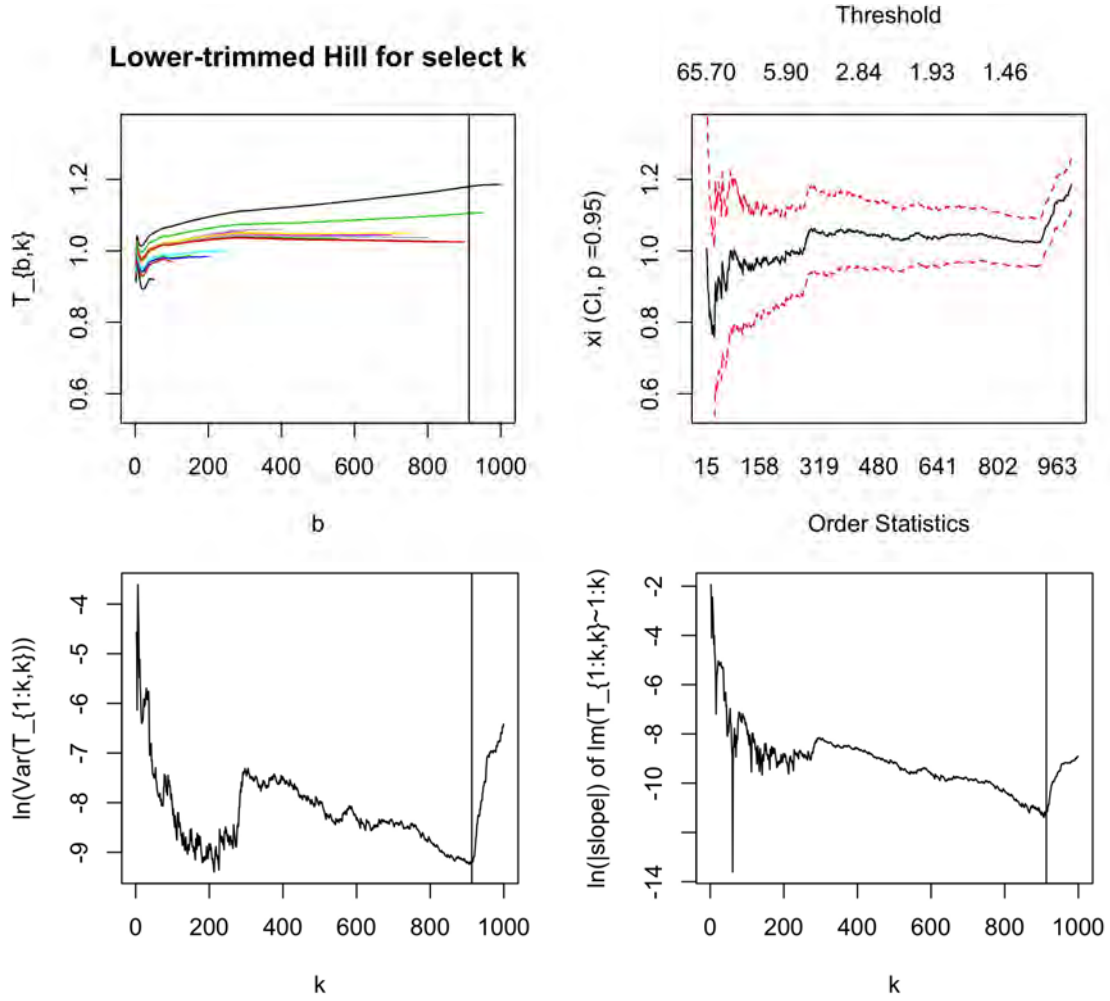


Figure 6.3: Spliced Pareto case (body parameter: 1/4 and tail parameter: 1). Top left: $T_{b,k}$ for varying lower trimming b , for $k = 1, 51, 101, \dots, 1000$. The vertical line is the splicing location. Top right: Hill plot. Bottom left: empirical variance of the LTH as a function of k . Bottom right: slope of a fitted linear model to the LTH as a function of k .

We can immediately see that

$$\mathbb{E}(T_{b,k}) = \xi + \frac{Q_0(n/k)}{p} \frac{\frac{((k+1)/b)^p - 1}{1-p} - 1}{1 + \sum_{j=b+1}^k j^{-1}} (1 + o_p(1)),$$

$$\mathbb{E}(\bar{T}_k) = \xi + \frac{Q_0(n/k)}{pk} \sum_{b=1}^k \frac{\frac{((k+1)/b)^p - 1}{1-p} - 1}{1 + \sum_{j=b+1}^k j^{-1}} (1 + o_p(1)),$$

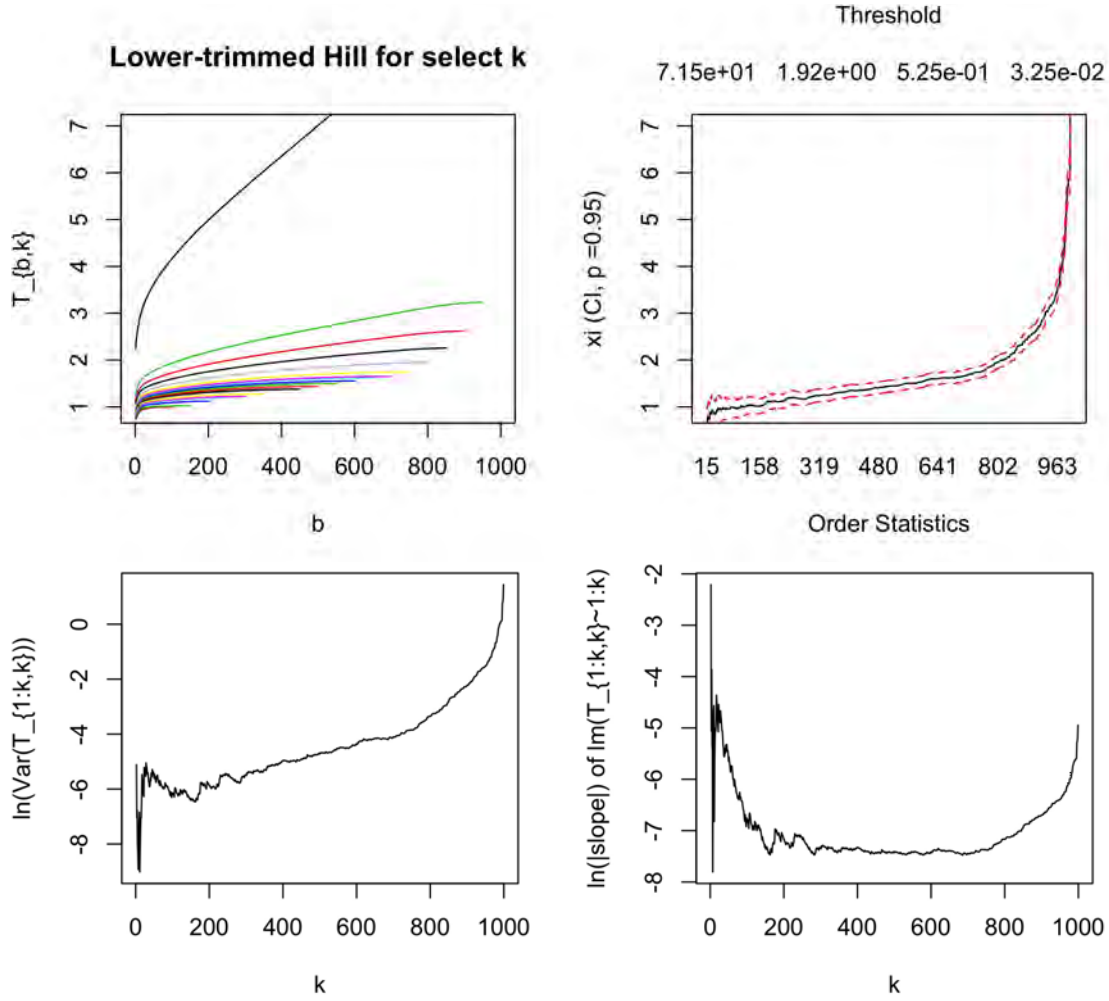


Figure 6.4: Burr case (all parameters set to 1). Top left: $T_{b,k}$ for varying lower trimming b , for $k = 1, 51, 101, \dots, 1000$. Top right: Hill plot. Bottom left: empirical variance of the LTH as a function of k . Bottom right: slope of a fitted linear model to the LTH as a function of k .

so that the asymptotic bias terms can be recognized directly. To ease notation, let us introduce the constants

$$\begin{aligned}
 c_{b,k,p} &:= \frac{1}{p} \cdot \frac{\frac{((k+1)/b)^p - 1}{1-p} - 1}{1 + \sum_{j=b+1}^k j^{-1}} \approx \frac{1}{p} \cdot \frac{\frac{((k+1)/b)^p - 1}{1-p} - 1}{1 + \log((k+1)/b)} \\
 \bar{c}_{k,p} &:= \frac{1}{pk} \sum_{b=1}^k \frac{\frac{((k+1)/b)^p - 1}{1-p} - 1}{1 + \sum_{j=b+1}^k j^{-1}} \approx \frac{1}{pk} \sum_{b=1}^k \frac{\frac{((k+1)/b)^p - 1}{1-p} - 1}{1 + \log((k+1)/b)}.
 \end{aligned} \tag{6.16}$$

Theorem 6.3.2. *The average \bar{T}_k as defined in (6.15), under model (6.1) and second order condition (6.13) satisfies the following asymptotic distributional identity, for*

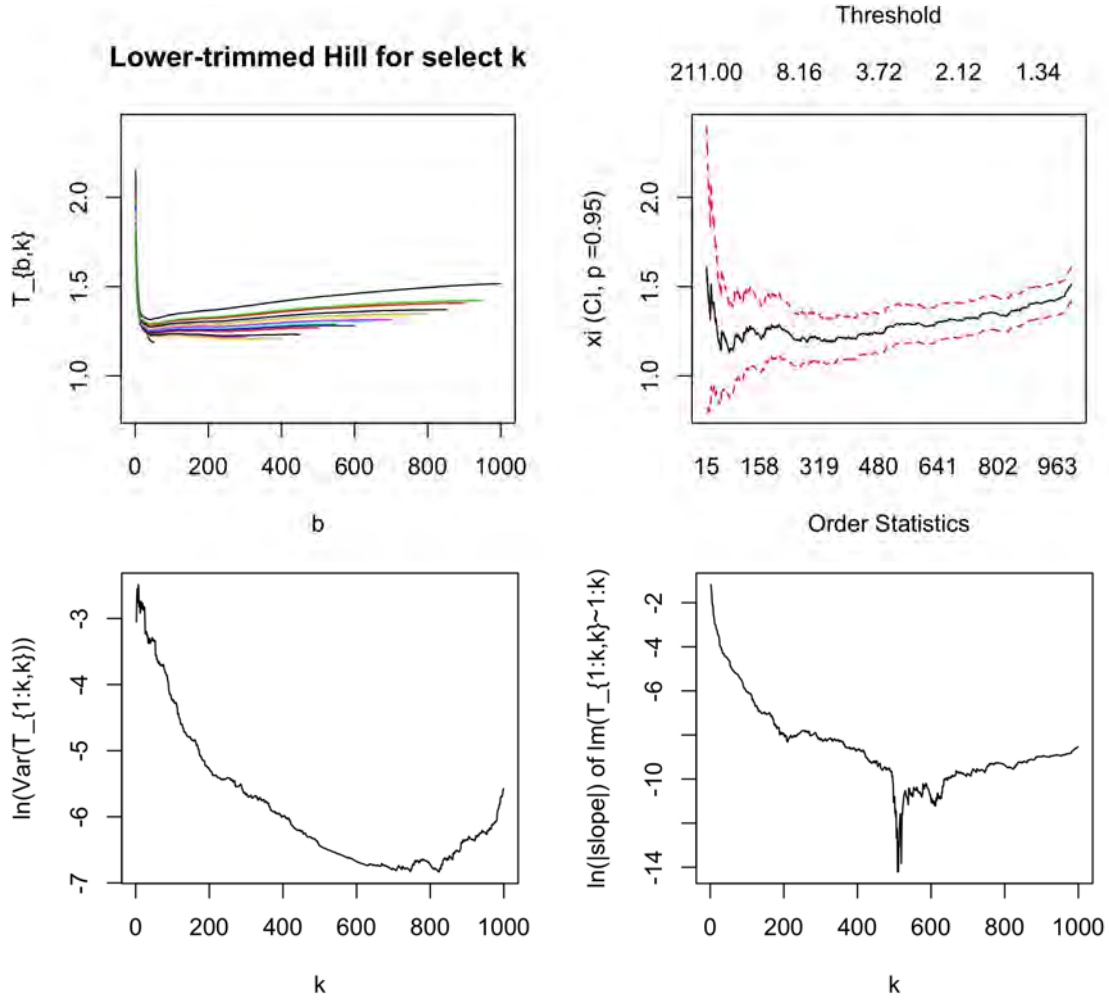


Figure 6.5: Loggamma case (logshape parameter: $3/2$, lograte parameter: 1). Top left: $T_{b,k}$ for varying lower trimming b , for $k = 1, 51, 101, \dots, 1000$. Top right: Hill plot. Bottom left: empirical variance of the LTH as a function of k . Bottom right: slope of a fitted linear model to the LTH as a function of k .

$n, k, n/k \rightarrow \infty$,

$$\begin{aligned} \bar{T}_k \stackrel{d}{=} & \frac{\xi}{k-1} \sum_{j=1}^k E_j \left[\log(1 + \log(k/j)) + \frac{ek}{j} E(1 + \log(k/j)) \right] (1 + o(1)) \quad (6.17) \\ & + Q_0(n/k) \left[\frac{e^{1-p}}{p(1-p)} E(1-p) - \frac{e}{p} E(1) \right] (1 + o_p(1)), \end{aligned}$$

where

$$E(x) := \int_x^\infty e^{-v}/v \, dv,$$

is the exponential integral.

Equipped with the representations in terms of exponential variables that we obtained in Theorems 6.3.1 and 6.3.2, we set on to analyze the mean of the empirical variance of $T_{b,k}$ as a function of b .

Theorem 6.3.3. *The mean of the empirical variance of $\{T_{b,k}; 1 \leq b \leq k\}$, under model (6.1) and second order condition (6.13) satisfies the following asymptotic identity, for $n, k, n/k \rightarrow \infty$,*

$$\mathbb{E} \left[\frac{1}{k} \sum_{b=1}^k (T_{b,k} - \bar{T}_k)^2 \right] = \frac{C}{k} \xi^2 (1 + o(1)) + Q_0^2(n/k) f(p) (1 + o_p(1))$$

where $C = 0.502727$ and

$$\begin{aligned} f(p) := & \frac{1 - e^{1-2p}(1-2p) \mathbb{E}(1-2p) - e^{2-2p} \mathbb{E}^2(1-p)}{p^2(1-p)^2} \\ & + 2 \frac{e^{2-p} \mathbb{E}(1-p) \mathbb{E}(1) - 1 + e^{1-p}(1-p) \mathbb{E}(1-p)}{p^2(1-p)} \\ & + \frac{1 - e \mathbb{E}(1) - e^2 \mathbb{E}^2(1)}{p^2} > 0. \end{aligned} \quad (6.18)$$

6.3.2 Optimal k in the Hall class

We now make a further assumption on the regularly varying class, in order to get an explicit form of Q_0 . Concretely, we assume the Hall class (Hall (1982)), which satisfies the property

$$U(x) = Ax^\xi(1 + Dx^p(1 + o(1))), \quad x \rightarrow \infty. \quad (6.19)$$

An immediate consequence then is the explicit expression

$$Q_0(x) = -pDx^p(1 + o(1)).$$

Hence,

$$\mathbb{E} \left[\frac{1}{k} \sum_{b=1}^k (T_{b,k} - \bar{T}_k)^2 \right] = \frac{C}{k} \xi^2 (1 + o(1)) + p^2 D^2 (n/k)^{2p} f(p) (1 + o_p(1)). \quad (6.20)$$

Recall that the classical Hill estimator for this class has AMSE given by

$$\frac{\xi^2}{k} + \left(\frac{Q_0(n/k)}{1-p} \right)^2,$$

which is minimized for

$$\begin{aligned} k_0^* & \sim (Q_0^2(n))^{-1/(1-2p)} \left(\frac{\xi^2(1-p)^2}{-2p} \right)^{1/(1-2p)} \\ & = \left(\frac{n^{-2p} \xi^2 (1-p)^2}{-2p^3 D^2} \right)^{1/(1-2p)}, \end{aligned} \quad (6.21)$$

see e.g. (Beirlant et al., 2004, p.125)). In a similar way, the minimizer of (6.20) is simply

$$\begin{aligned} k^* &\sim (Q_0^2(n))^{-1/(1-2p)} \left(\frac{C\xi^2}{-2pf(p)} \right)^{1/(1-2p)} \\ &= \left(\frac{n^{-2p}C\xi^2}{-2p^3D^2f(p)} \right)^{1/(1-2p)}. \end{aligned} \quad (6.22)$$

Hence from (6.21) and (6.22) we obtain a simple expression of the optimal threshold k_0^* of the Hill estimator in terms of k^* :

$$k_0^* = k^* \left(\frac{C}{(1-p)^2f(p)} \right)^{-1/(1-2p)}. \quad (6.23)$$

6.3.3 Interpretation of \bar{T}_k as a weighted Hill estimator

Observe that, for fixed k ,

$$\bar{T}_k = \frac{1}{k} \sum_{b=1}^k T_{b,k} = \frac{1}{k} \sum_{i=1}^k \theta_i \log(X_{n-i+1,n}/X_{n-k,n}), \quad (6.24)$$

with

$$\theta_i := \sum_{b=i}^k \frac{1}{b(1 + \sum_{j=b+1}^k j^{-1})},$$

so that one can interpret the estimator \bar{T}_k as a modification of the classical Hill estimator that uses different weights for different order statistics. It is not hard to see that asymptotically the correction factors behave like

$$\theta_i \sim \log \left(\frac{\log(i/k) - 1}{\log(1 - 1/k) - 1} \right), \quad k \rightarrow \infty. \quad (6.25)$$

Figure 6.6(left) highlights the accuracy of this approximation for $k = 100$ across different values of i , and also illustrates the fact that the largest data point receives a weight of almost 2 in this case, whereas on from the 20th-largest observation the weight is lower than for the classical Hill estimator, and the weight diminishes for smaller data points. Note that, as k increases, the weight of the largest observation grows above any bound, but extremely slowly, namely

$$\theta_1 = \log(\log(k) + 1) - 1/k + O(1/k^3).$$

Figure 6.6(right) illustrates that even for a value as large as $k = 10000$, θ_1 is still below 2.4.

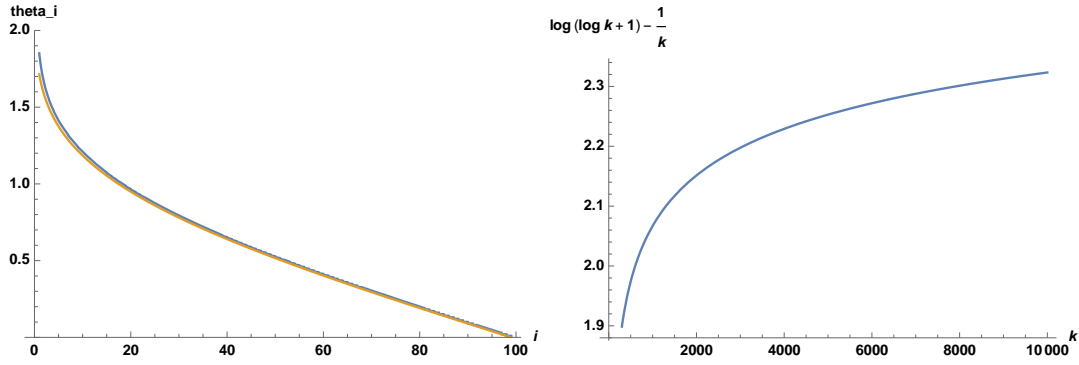


Figure 6.6: Left panel: for $k = 100$, the true θ_i (blue) and the asymptotic approximation (6.25) (orange) as a function of i . Right panel: Leading terms of the series expansion of θ_1 with respect to k .

6.4 A ratio statistic

Once a k^* has been selected, it is important to be able to statistically assess whether the remaining upper tail differs significantly from the one of a pure Pareto. In order to recognize whether a Pareto tail has been achieved or not, we have seen that flatness of the lower-trimmed Hill estimator is desirable. Inspired by the T-statistic introduced in Bhattacharya et al. (2017), we introduce the ratio statistics

$$R_{b,k} = \frac{T_{b+1,k}}{T_{b,k}}, \quad b = 1, \dots, k-1,$$

quantities which we expect to be close to one. Although these statistics do not have the property of being i.i.d. and hence test sizes have to be calibrated using Monte Carlo simulation, an advantage which carries over to the present setting is that they do not depend on ξ . Indeed,

$$R_{b,k} \stackrel{d}{=} \frac{\omega(b,k)}{\omega(b+1,k)} \left(1 + \frac{\log(\Gamma_{b+1}/\Gamma_{k+1})}{\sum_{i=1}^b \log(\Gamma_i/\Gamma_{k+1})} \right),$$

by the order statistics property of the Poisson process, where $\Gamma_m = \sum_{i=1}^m E_i$, and E_i , $i = 1, 2, \dots$, is an i.i.d. sequence of independent unit-rate exponential random variables. This invariance with respect to the ξ parameter permits to assess the goodness of selection of a threshold k^* as follows:

1. Simulate the R_{b,k^*} statistics N_{MC} times, and call them

$$R_{b,k^*}^m, \quad m = 1, \dots, N_{MC}, \quad b = 2, \dots, k^* - 1.$$

2. For fixed $\alpha \in (0, 1)$, find the empirical $\alpha/2$ and $1-\alpha/2$ quantiles corresponding to each of the $b = 2, \dots, k^* - 1$ samples,

$$R_{b,k^*}^m, \quad m = 1, \dots, N_{MC},$$

and call them $(q_1, q_2)_2, \dots, (q_1, q_2)_{k^*-1}$.

- Count the proportion of the the N_{MC} trajectories

$$R_{b,k^*}^m, \quad b = 2, \dots, k^* - 1,$$

which fall outside of their confidence interval $(q_1, q_2)_b$ for some $2 \leq b \leq k^* - 1$. Call this proportion α_r .

- If α_r is, up to some tolerance level, too large (too small), go to step (2) and decrease (increase) α to a value within its two last values.
- Plot the R_{b,k^*} , $b = 2, \dots, k^* - 1$, from the data, together with the last set of quantiles $(q_1, q_2)_1, \dots, (q_1, q_2)_{k^*}$. It is also a good idea, for visualization, to plot the standardized version

$$\frac{R_{b,k^*} - q_{1,b}}{q_{2,b} - q_{1,b}}, \quad b = 2, \dots, k^* - 1,$$

which for a pure Pareto tail is expected by construction to lie between 0 and 1 in $100(1 - \alpha)\%$ of the cases.

Example 6.4.1. For the Burr sample of Figure 6.4, we compare taking $k^* = 326$ and $k^* = 600$ in the plots of Figure 6.7. The first number, $k^* = 326$ is precisely the one that minimizes the expected empirical variance, according to the parameters of the Burr sample and to formula (6.22), with p chosen to be -1 . The number of Monte Carlo simulations was in each case $N_{MC} = 10000$, and the significance level is $\alpha = 0.05$. Observe how the fit is good for $k = 326$, but is outside the bands for $k = 600$.

Remark 6.4.1. This approach can only be considered as a selection procedure itself if the corresponding sequential testing is adjusted to have the correct size. In other words, if the above algorithm is used multiple times to choose k , the rejection probability will exceed the desired α level. An alternative is to take sequential values of k into the algorithm, which makes the routine highly computationally intensive. Hence, we presently recommend it solely as a goodness of selection evaluation.

6.5 Simulations

We perform a simulation study based on three different and common distributions which belong to the Hall class (6.19). We consider simulating $N_{sim} = 1000$ times from the following three distributions, with four sub-cases for each distribution, for varying sample size and parameters:

- The Burr distribution, with tail given by

$$\bar{F}(x) = \left(\frac{\eta}{\eta + x^\tau} \right)^\lambda, \quad x > 0, \quad \eta, \tau, \lambda > 0,$$

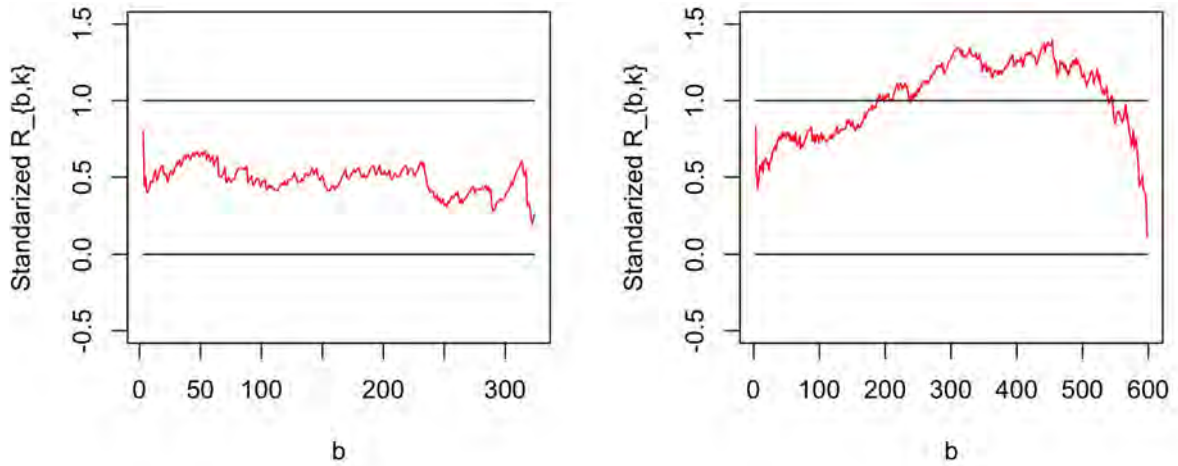


Figure 6.7: Standardized R-statistic for the Burr sample of Figure 6.4 (all parameters set to 1), for two choices of threshold: $k = 326, 600$, respectively. $N_{MC} = 10000$ and $\alpha = 0.05$.

which implies by Taylor expansion that

$$\xi = \frac{1}{\lambda\tau}, \quad A = \eta^{1/\tau}, \quad D = -\frac{1}{\tau}, \quad p = -\frac{1}{\lambda}.$$

We consider for $n = 100, 500$ the two sets of parameters $\eta = 1, \lambda = 2, \tau = 1/2$; and $\eta = 3/2, \lambda = 1/2, \tau = 2$.

- The Fréchet distribution with tail

$$\bar{F}(x) = 1 - \exp(-x^{-\alpha}), \quad \alpha > 0,$$

which implies

$$\xi = \frac{1}{\alpha}, \quad A = 1, \quad D = -\frac{1}{2\alpha}, \quad p = -1.$$

We consider for $n = 100, 500$ the two parameters $\alpha = 1, 1/2$.

- The Generalized Pareto Distribution (GPD) distribution, with tail given by

$$\bar{F}(x) = \left(1 + \frac{\gamma x}{\sigma}\right)^{-1/\gamma}, \quad \gamma, \sigma > 0,$$

which implies

$$\xi = \gamma, \quad A = \frac{\sigma}{\gamma}, \quad D = -1, \quad p = -\gamma.$$

We consider for $n = 100, 500$ the two sets of parameters $\gamma = 1/2, \sigma = 2$; and $\gamma = 5/2, \sigma = 1$.

For each sample we evaluate the Hill estimator

$$H_k = T_{k,k}$$

and the averaged trimmed estimator

$$\bar{T}_k = \frac{1}{k} \sum_{b=1}^k T_{b,k}$$

at three particular choices of k . Note that these threshold choices are designed for the Hill estimator, but will turn out sensible for the latter estimator as well.

- (i) We use the popular procedure of Guillou and Hall (2001) as a benchmark for finding the optimal choice of k , and denote the resulting tail estimators by $H_{\hat{k}_{GH}}, \bar{T}_{\hat{k}_{GH}}$. Such a threshold selection procedure has been subject to comparisons (both in Guillou and Hall (2001) itself and in Beirlant et al. (2002)) to other alternatives like Danielsson et al. (2001) and Drees and Kaufmann (1998), and we refer the reader to these papers for further details.
- (ii) An estimator of k_0^* from (6.21) obtained as follows. Motivated by (6.20), we compute k^* as the minimizer of the empirical variance (the search beginning at $1/5$ of the sample size, to avoid degeneracies) of the trimmed Hill estimator, as a function of b , and using (6.23) to set

$$k_0^* := k^* \left(\frac{C}{(1-p)^2 f(p)} \right)^{-1/(1-2p)}.$$

Observe that while we still have to input p , here prior knowledge of ξ, D is no longer needed. We choose $p = -1$ as the canonical choice.

- (iii) As in (ii), but using the true parameter of p , in order to quantify how the removal of a potential misspecification of p by the canonical choice $p = -1$ affects the estimators (this complements Beirlant et al. (2002), where it was concluded from simulation studies for various estimators that this potential misspecification does not seem to be of major importance).

We then plot the bias, variance and MSE of each resulting estimator as a function of k .

The results are given in Figures 6.8, 6.9 for the Burr case; Figures 6.10, 6.11 for the Fréchet case; and Figures 6.12, 6.13 for the GPD. We observe that the behaviour is very similar for the three families (which is not uncommon in this context, cf. (Beirlant et al., 2002, p.178)).

For the Hill estimator, we notice that our method fares very favourably against the benchmark, and the misspecification of the second order parameter p does

not play a substantial role. The same behaviour is observed within the three \bar{T} -estimators. When comparing Hill against \bar{T} -estimators, the latter improve the bias and MSE for nearly all k , and in most cases also the variance (except for very heavy-tails ($\xi \geq 1$) and small values of k).

Remarkably, the estimator $\bar{T}_{k_0, p=-1}$, where the canonical $p = -1$ is used, is highly competitive against the Hill estimator, especially so for $\xi \leq 1$. This is not a contradiction, since the optimality of the Hill estimator refers to choices for k within the class of H_k , whereas the \bar{T}_k estimators span a different class (visible in the weighting interpretation of Section 6.3.3), and when k is optimized w.r.t. AMSE in that class, even better performance can be feasible, which, however, is not the subject of the present paper.

6.6 Insurance data

Let us now consider a real-life insurance data set consisting of 837 motor third party liability (MTPL) insurance claims from the period 1995-2010 that was studied intensively in Albrecher et al. (2017) (where it is referred to as "Company A"). These data are right-censored, and were also analyzed recently combining survival analysis techniques and expert information in Bladt et al. (2020). Here, we focus only on the *ultimates*, see Figure 6.14, which are the actual final claim sizes for the settled claims and an expert prediction of the total payment until closure for all claims that are still open.

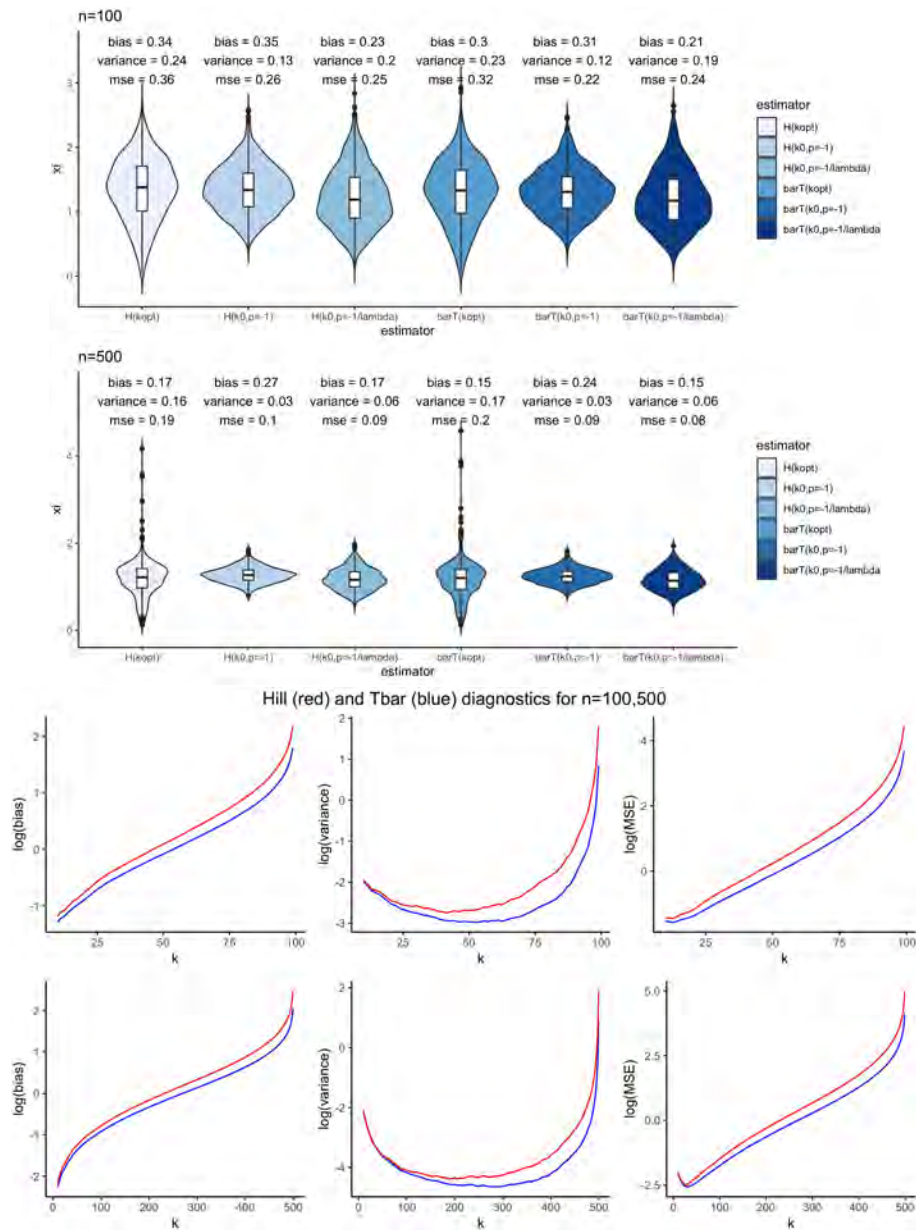


Figure 6.8: Burr distribution, parameters $\eta = 1$, $\lambda = 2$, $\tau = 1/2$. Top: Violin plots for $n = 100, 500$ of the estimators $H_{\hat{k}_{GH}}$, $H_{\hat{k}_0^*, p=-1}$, $H_{\hat{k}_0^*, p=-1/\lambda}$, $\bar{T}_{\hat{k}_{GH}}$, $\bar{T}_{\hat{k}_0^*, p=-1}$, $\bar{T}_{\hat{k}_0^*, p=-1/\lambda}$. Bottom: diagnostics of \bar{T}_k (blue) and H_k (red) as a function of k .

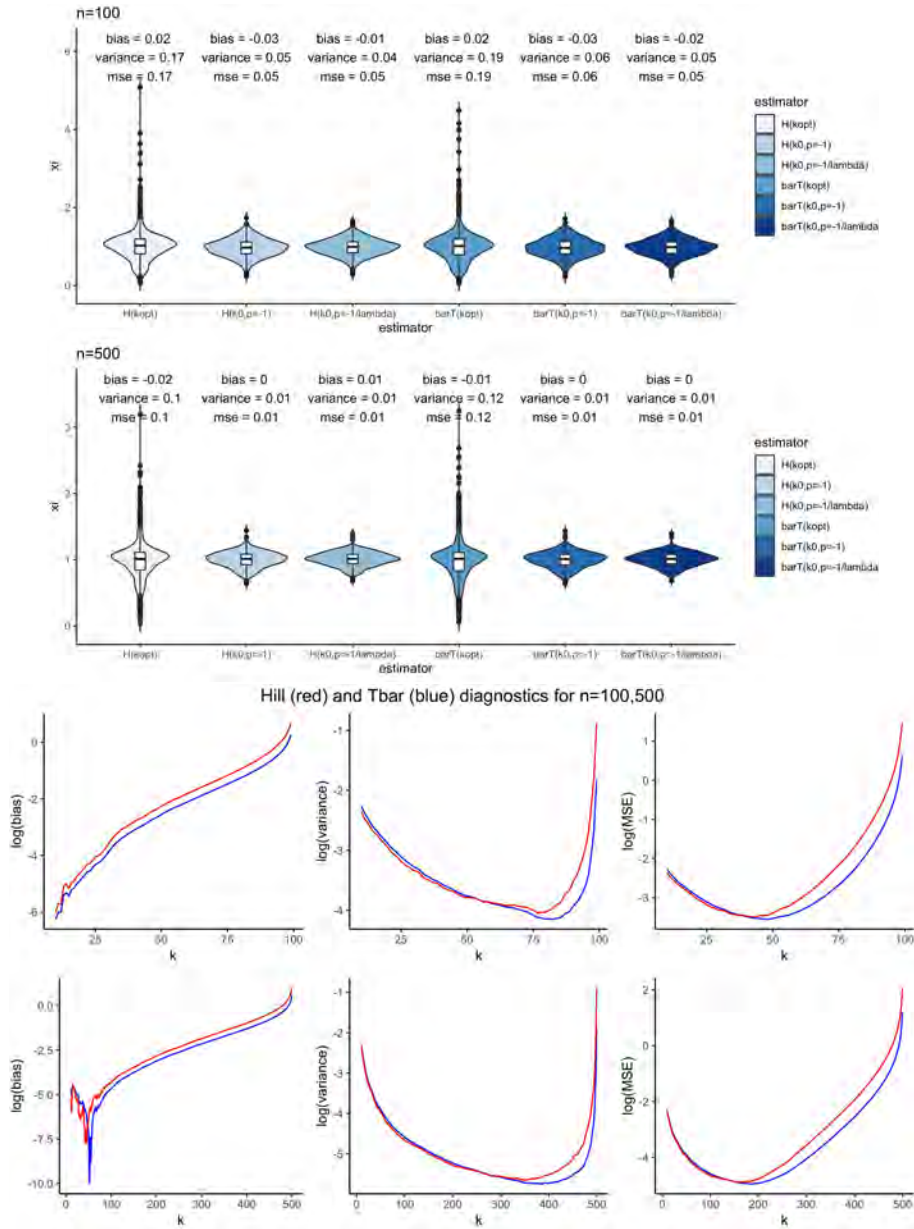


Figure 6.9: Burr distribution, parameters $\eta = 3/2$, $\lambda = 1/2$, $\tau = 2$. Top: Violin plots for $n = 100, 500$ of the estimators $H_{\hat{k}_{GH}}$, $H_{\hat{k}_0^*, p=-1}$, $H_{\hat{k}_0^*, p=-1/\lambda}$, $\bar{T}_{\hat{k}_{GH}}$, $\bar{T}_{\hat{k}_0^*, p=-1}$, $\bar{T}_{\hat{k}_0^*, p=-1/\lambda}$. Bottom: diagnostics of \bar{T}_k (blue) and H_k (red) as a function of k .

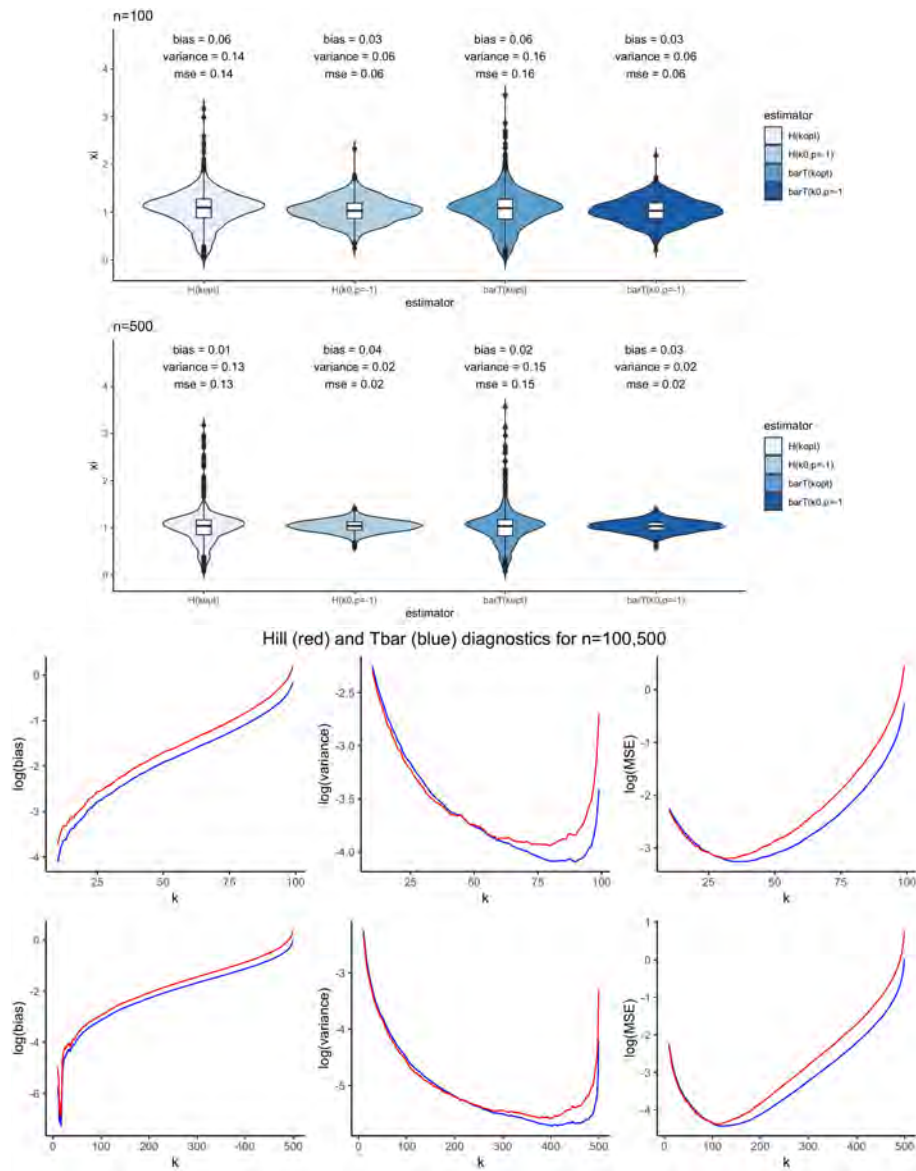


Figure 6.10: Fréchet distribution, parameter $\alpha = 1$. Top: Violin plots for $n = 100, 500$ of the estimators $H_{\hat{k}_{GH}}$, $H_{\hat{k}_{0^*, p=-1}}$, $\bar{T}_{\hat{k}_{GH}}$, $\bar{T}_{\hat{k}_{0^*, p=-1}}$. Bottom: diagnostics of \bar{T}_k (blue) and H_k (red) as a function of k .

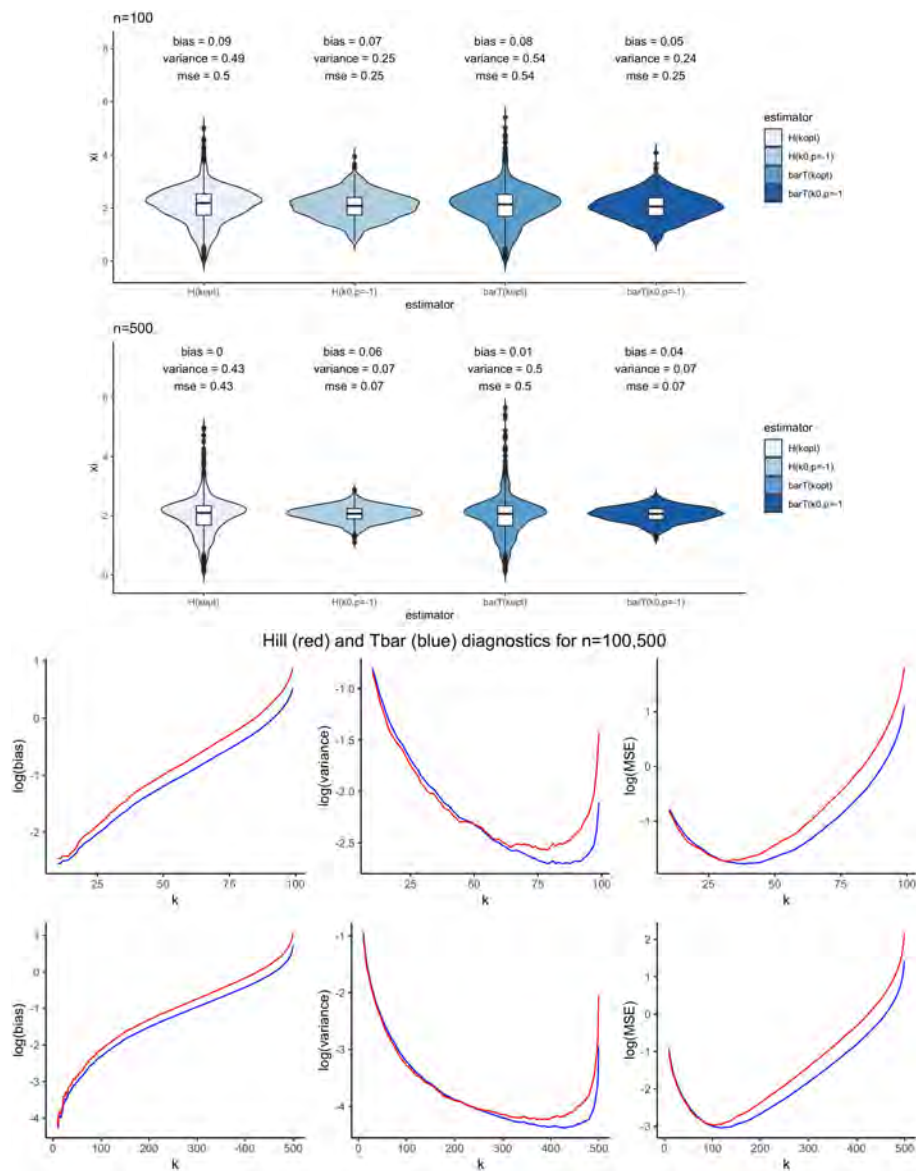


Figure 6.11: Fréchet distribution, parameter $\alpha = 1/2$. Top: Violin plots for $n = 100, 500$ of the estimators $H_{\hat{k}_{GH}}$, $H_{\hat{k}_{0^*, p=-1}}$, $\bar{T}_{\hat{k}_{GH}}$, $\bar{T}_{\hat{k}_{0^*, p=-1}}$. Bottom: diagnostics of \bar{T}_k (blue) and H_k (red) as a function of k .

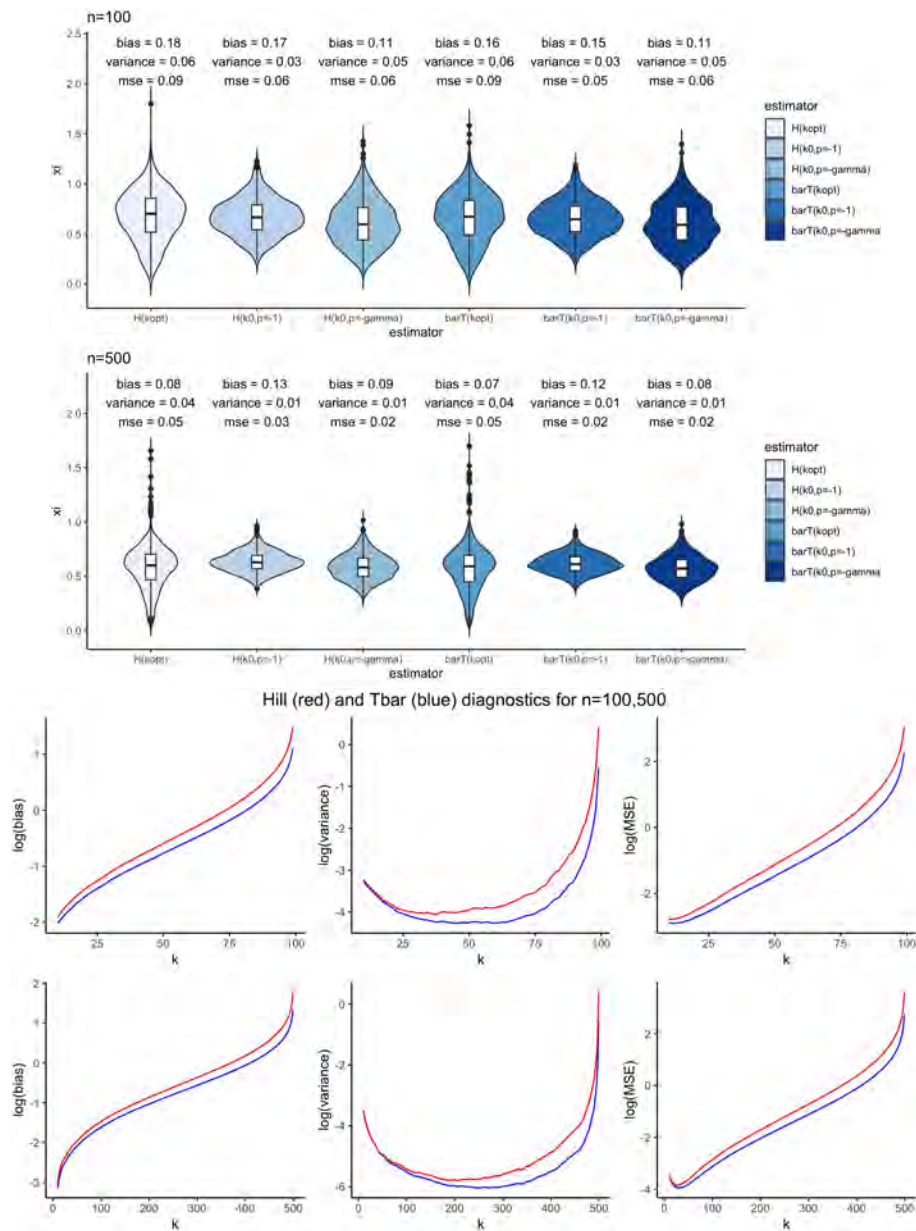


Figure 6.12: GPD distribution, parameters $\gamma = 1/2$, $\sigma = 2$. Top: Violin plots for $n = 100, 500$ of the estimators $H_{\hat{k}_{GH}}$, $H_{\hat{k}_0^*, p=-1}$, $H_{\hat{k}_0^*, p=-\gamma}$, $\bar{T}_{\hat{k}_{GH}}$, $\bar{T}_{\hat{k}_0^*, p=-1}$, $\bar{T}_{\hat{k}_0^*, p=-\gamma}$. Bottom: diagnostics of \bar{T}_k (blue) and H_k (red) as a function of k .

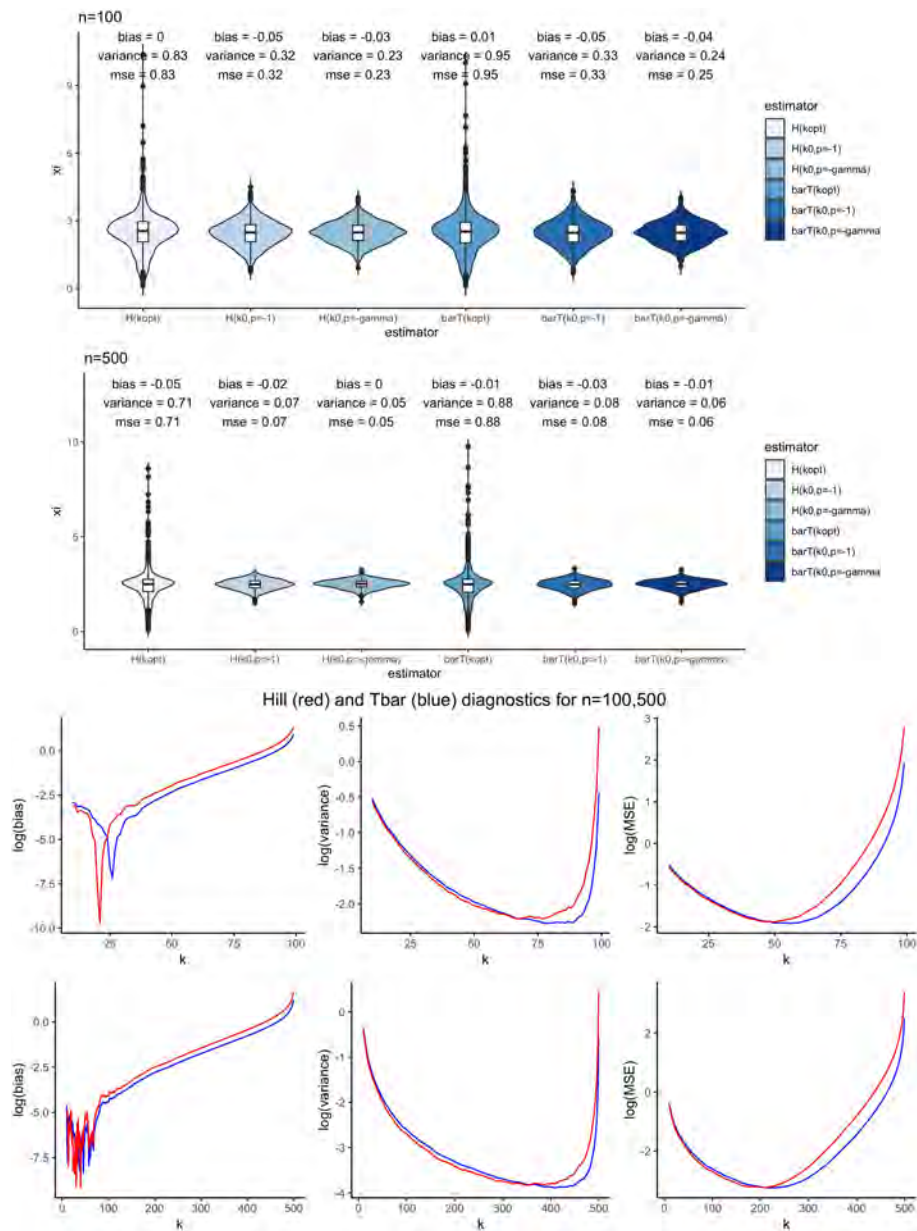


Figure 6.13: GPD distribution, parameters $\gamma = 5/2$, $\sigma = 1$. Top: Violin plots for $n = 100, 500$ of the estimators $H_{\hat{k}_{GH}}$, $H_{\hat{k}_0^*, p=-1}$, $H_{\hat{k}_0^*, p=-\gamma}$, $\bar{T}_{\hat{k}_{GH}}$, $\bar{T}_{\hat{k}_0^*, p=-1}$, $\bar{T}_{\hat{k}_0^*, p=-\gamma}$. Bottom: diagnostics of \bar{T}_k (blue) and H_k (red) as a function of k .

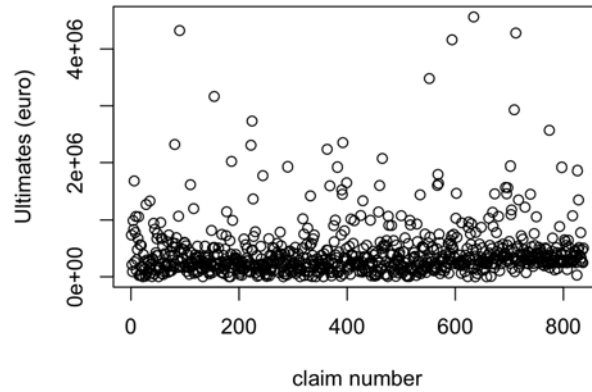


Figure 6.14: Ultimates from an MTPL insurance portfolio.

In Figure 6.15 we depict the lower-trimmed Hill plots and the usual Hill plot, together with the empirical variance. As in the simulation studies in Section 6.5, in order to avoid degeneracies, we only look at candidates for the minimizer to the right of $n/5$, which corresponds to 167 in this case. The minimum empirical variance is then obtained for $k^* = 222$. Using the canonical choice $p = -1$, we have that $k_0 = 222/2.62421 \approx 85$. Note that for the same choice of $p = -1$, and using the prior eyeballed estimate $\xi \approx 0.5$, and consequently $D = -0.5$, we get by (6.21) the suggestion $k_0^* \approx 112$ (which might be considered the classical choice of the threshold in this case).

The corresponding estimates of ξ are given by

$$H_{k_0} = 0.508, \quad H_{k_0^*} = 0.560, \quad \bar{T}_{k_0} = 0.480, \quad T_{k_0^*} = 0.525.$$

The simulation studies of Section 6.5 may suggest the third of the above numbers to be the most reliable estimate here. The ratio statistic test in Figure 6.16 suggests that for both thresholds the sample is Pareto in the tail (with only a slight issue for the two largest observations).

In (Albrecher et al., 2017, p.99), a splicing point was suggested for this data set at around $k = 20$, based on expert opinion. A semi-automated option using our method for detecting this splicing point would be to replace the left limit $k = 167$ by a very small number (in this case $k = 4$ is chosen after visual inspection of the erratic nature of the empirical variance for the first three), and then to apply our method, which leads to the detection of the minimum variance at $k = 38$ (which is clearly visible in Figure 6.15). Under the assumption $p = -1$ this then leads to $k \approx 14$ as a suggested splicing point. Note that in the nature of the present data set, the ultimates for the highest claims have intrinsic uncertainty (as they are just estimates of the final closed claim size), and a more systematic way to approach this particular situation would be to combine the trimming of the Hill estimator from below and above, which is not the focus of the present paper.

6.7 Conclusion

In this paper, we showed that trimming the Hill estimator from the left can lead to favorable properties in connection with the expected empirical variance of the

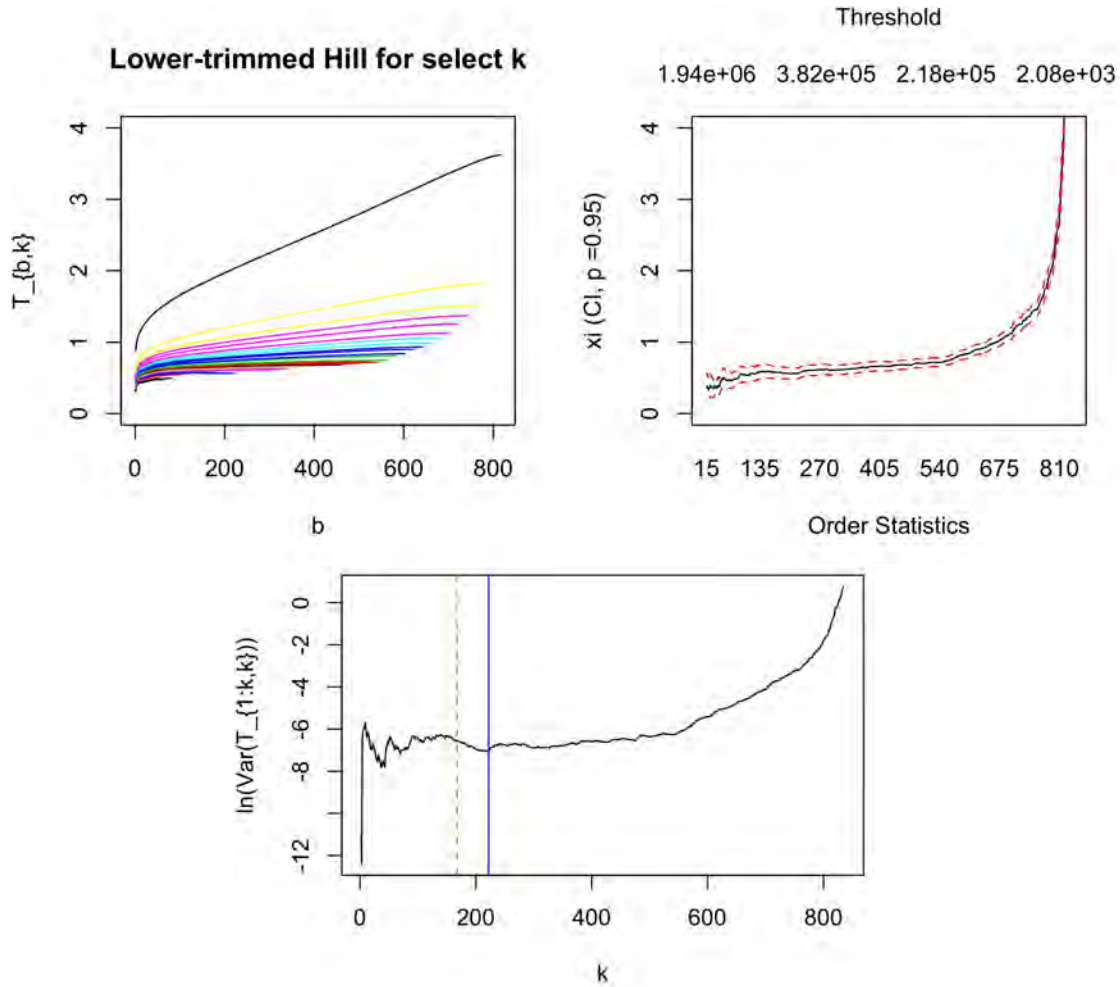


Figure 6.15: MTPL insurance ultimates. Top left: Lower-trimmed Hill estimator (LTH) estimator for varying lower trimming b , for k uniformly spaced from 1 to 837 in units of 20. Top right: Hill plot. Bottom: empirical variance of the LTH as a function of k . The dotted line is the left limit for candidates, and the solid line is the resulting minimum.

tail index estimators in extreme value statistics. For the Hall class, we established asymptotic results on the behavior of this expected empirical variance, which allows to develop a guideline for the choice of the optimal threshold in the tail index estimation problem. It turns out that there is an intrinsic link between this optimal threshold and the classical optimal threshold for the Hill estimator. Since in the trimming context the identification of the optimal threshold is much more insensitive on the tail characteristics (it only depends on the p -parameter in the Hall class, not on D nor on the tail index ξ), this link allows to circumvent the classical problem in threshold selection for the Hill estimator. As a by-product, by suitable averaging we develop a novel tail index estimator which assigns a non-uniform weight to each observation in a natural way, relies on fewer assumptions on the tail characteristics, is simple to implement and outperforms the classical Hill estimator in most cases. The latter is illustrated in extensive simulation stud-

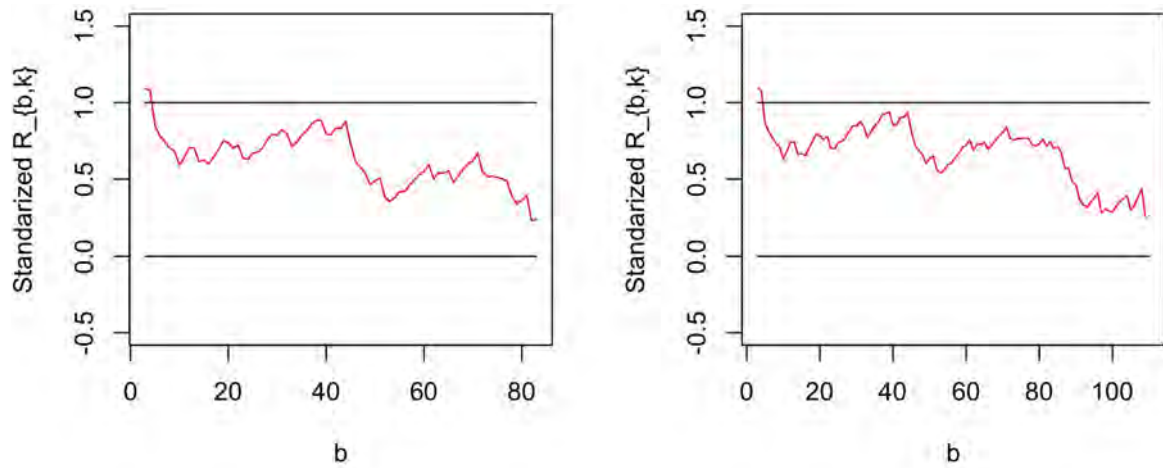


Figure 6.16: Standardized R-statistic for the MTPL ultimates, for the two threshold choices $k = 85$ (left) and $k = 112$ (right). $N_{MC} = 100,000$ and $\alpha = 0.05$.

ies. In addition, the technique is applied to a real-life insurance data set that was previously studied by other techniques. We conclude by noting that the approach taken in this paper is in principle also applicable for the potential improvement of tail index estimators other than the Hill estimator. Further possible directions of future research include the combination of left trimming with right trimming in situations with possible outliers, as well as the consideration of possibly censored data.

6.8 Proofs

Proof of Proposition 2.1. Set $q = k - b + 1$. By the Rényi representation (6.6),

$$\mathbb{V}(T_{b,k}) = \mathbb{V}\left(\sum_{j=1}^k E_j^* \sum_{i=j \vee q}^k \frac{\gamma_i}{k-j+1}\right) = \xi^2 \frac{\sum_{j=1}^k \left(\frac{k-j \vee q+1}{k-j+1}\right)^2}{\left(\sum_{j=1}^k \frac{k-j \vee q+1}{k-j+1}\right)^2}.$$

Plugging in $q = 1$ ($b = k$) gives

$$\mathbb{V}(T_{k,k}) = \frac{\xi^2}{k},$$

which corresponds to the usual Hill estimator $T_{k,k}$ and gives the first identity. In the general case,

$$\mathbb{V}(T_{b,k}) = \xi^2 \frac{\sum_{j=1}^q \left(\frac{k-q+1}{k-j+1}\right)^2 + k - q + 1}{\left(\sum_{j=1}^q \frac{k-q+1}{k-j+1} + k - q + 1\right)^2}.$$

But $j \leq q$ implies $\frac{k-q+1}{k-j+1} \leq 1$, such that

$$\sum_{i=1}^q \frac{1}{k-j+1} \geq \sum_{j=1}^q \frac{k-q+1}{k-j+1} \frac{1}{k-j+1},$$

so

$$\sum_{j=1}^q \frac{k-q+1}{k-j+1} + k - q + 1 \geq \sum_{j=1}^q \left(\frac{k-q+1}{k-j+1}\right)^2 + k - q + 1.$$

Thus

$$\begin{aligned} \mathbb{V}(T_{b,k}) &\leq \xi^2 \frac{\sum_{j=1}^q \left(\frac{k-q+1}{k-j+1}\right)^2 + k - q + 1}{\left(\sum_{j=1}^q \left(\frac{k-q+1}{k-j+1}\right)^2 + k - q + 1\right)^2} \\ &= \frac{\xi^2}{\sum_{j=1}^q \left(\frac{k-q+1}{k-j+1}\right)^2 + k - q + 1}, \end{aligned}$$

which gives the second identity. \square

Proof of Theorem 3.1. We first note that

$$T_{b,k} \stackrel{d}{=} \frac{\sum_{i=1}^b \log(U(Y_{n-i+1,n})/U(Y_{n-k,n}))}{b(1 + \sum_{j=b+1}^k j^{-1})},$$

where $Y_{1,n} < \dots < Y_{n,n}$ are the order statistics of a standard Pareto sample (the $\xi = 1$ case). Then, from the second order condition (6.13) we obtain that for $A = Y_{n-k,n}$ and $x = Y_{n-i+1,n}/Y_{n-k,n}$, as $k, n, n/k \rightarrow \infty$,

$$T_{b,k} \stackrel{d}{=} \frac{\xi \sum_{i=1}^b \log(Y_{n-i+1,n}/Y_{n-k,n}) + \frac{Q_0(Y_{n-k,n})}{p} \sum_{i=1}^b ((Y_{n-i+1,n}/Y_{n-k,n})^p - 1)(1 + o_p(1))}{b(1 + \sum_{j=b+1}^k j^{-1})}.$$

But by the Rényi representation (6.6) of exponential order statistics, the first term is distributed as

$$\sum_{i=1}^b \log(Y_{n-i+1,n}/Y_{n-k,n}) \stackrel{d}{=} \sum_{j=1}^b E_j + b \sum_{j=b+1}^k E_j/j,$$

where E_1, E_2, \dots, E_k are i.i.d. standard exponential random variables. For the second term, by convergence to uniform random variables and a Riemann integral approximation, we get

$$\begin{aligned} \frac{1}{b} \sum_{i=1}^b ((Y_{n-i+1,n}/Y_{n-k,n})^p - 1) &\approx \frac{1}{b} \sum_{i=1}^b (((k+1)/i)^p - 1) \\ &\approx \frac{k+1}{b} \int_0^{b/(k+1)} (u^{-p} - 1) du = \frac{((k+1)/b)^p}{1-p} - 1, \end{aligned}$$

and since $(1 - 1/Y_{n-k,n})$ is a uniform order statistic, we further get that

$$\frac{Q_0(Y_{n-k,n})}{Q_0(n/k)} \xrightarrow{P} 1.$$

Putting the three pieces together then establishes (6.14). \square

Proof of Theorem 3.2. With the shortened notation, we write

$$T_{b,k} \stackrel{d}{=} \xi \frac{\bar{E}_b + \sum_{j=b+1}^k E_j/j}{1 + \sum_{j=b+1}^k j^{-1}} + Q_0(n/k) c_{b,k,p} (1 + o_p(1)), \quad (6.26)$$

and by exchange of the order of summation, we can write

$$\begin{aligned} \bar{T}_k &\stackrel{d}{=} \frac{\xi}{k} \sum_{b=1}^k \frac{\bar{E}_b + \sum_{j=b+1}^k E_j/j}{1 + \sum_{j=b+1}^k j^{-1}} + Q_0(n/k) \bar{c}_{k,p} (1 + o_p(1)) \\ &= \frac{\xi}{k} \left[\sum_{j=1}^k E_j \sum_{b=j}^k \frac{1}{b(1 + \sum_{j=b+1}^k j^{-1})} + \sum_{j=2}^k E_j \sum_{b=1}^{j-1} \frac{1}{j(1 + \sum_{j=b+1}^k j^{-1})} \right] \\ &\quad + Q_0(n/k) \bar{c}_{k,p} (1 + o_p(1)) \\ &= \frac{\xi}{k} \sum_{j=1}^k E_j \left[\sum_{b=j}^k \frac{1}{b(1 + \log(k/b))} + \sum_{b=1}^{j-1} \frac{1}{j(1 + \log(k/b))} \right] (1 + o(1)) \\ &\quad + Q_0(n/k) \bar{c}_{k,p} (1 + o_p(1)). \end{aligned}$$

Again, by Riemann integration we have that

$$\frac{1}{k} \sum_{b=j}^k \frac{1}{(b/k)(1 + \log(k/b))} \approx \int_{j/k}^1 \frac{du}{u(1 - \log(u))} = \log(1 + \log(k/j)),$$

and

$$\sum_{b=1}^{j-1} \frac{1}{j(1 + \log(k/b))} \approx \frac{k}{j} \int_0^{j/k} \frac{du}{1 - \log(u)} = \frac{ek}{j} \mathbf{E}(1 + \log(k/j)).$$

Similarly,

$$\begin{aligned} \bar{c}_{k,p} &\approx \frac{1}{p} \int_0^1 \frac{(1-p)^{-1} u^{-p}}{1 - \log(u)} du - \frac{1}{p} \int_0^1 \frac{du}{1 - \log(u)} \\ &= \frac{e^{1-p}}{p(1-p)} \mathbf{E}(1-p) - \frac{e}{p} \mathbf{E}(1). \end{aligned} \quad (6.27)$$

Putting the pieces together then indeed yields (6.17). \square

Proof of Theorem 3.3. Let us first decompose each summand by writing

$$\mathbb{E}((T_{b,k} - \bar{T}_k)^2) = \mathbb{E}((T_{b,k} - \xi)^2) + \mathbb{E}((\bar{T}_k - \xi)^2) - 2\mathbb{E}((T_{b,k} - \xi)(\bar{T}_k - \xi)),$$

and subsequently consider each term separately. From (6.26) we have that

$$\mathbb{E}((T_{b,k} - \xi)^2) = \mathbb{V}(T_{b,k}) + \text{Bias}^2(T_{b,k}) = \xi^2 \frac{\frac{1}{b} + \sum_{j=b+1}^k 1/j^2}{(1 + \sum_{j=b+1}^k j^{-1})^2} + Q_0^2(n/k) c_{b,k,p}^2 (1 + o_p(1)).$$

On the other hand, (6.17) gives

$$\begin{aligned} \mathbb{E}((\bar{T}_k - \xi)^2) &= \mathbb{V}(\bar{T}_k) + \text{Bias}^2(\bar{T}_k) \\ &= \frac{\xi^2}{k^2} \sum_{j=1}^k \left[\log(1 + \log(k/j)) + \frac{ek}{j} \mathbf{E}(1 + \log(k/j)) \right]^2 (1 + o(1)) \\ &\quad + Q_0^2(n/k) \left[\frac{e^{1-p}}{p(1-p)} \mathbf{E}(1-p) - \frac{e}{p} \mathbf{E}(1) \right]^2 (1 + o_p(1)). \end{aligned}$$

The third term can be analyzed using both (6.17) and (6.26) as follows

$$\begin{aligned} \mathbb{E}((T_{b,k} - \xi)(\bar{T}_k - \xi)) &= \mathbb{E}((T_{b,k} - \mathbb{E}(T_{b,k}))(\bar{T}_k - \mathbb{E}(\bar{T}_k))) + Q_0^2(n/k) c_{b,k,p} \bar{c}_{k,p} \\ &= \xi^2 \mathbb{E} \left[\left(\frac{\frac{1}{b} \sum_{j=1}^b (E_j - 1) + \sum_{j=b+1}^k (E_j - 1)/j}{1 + \sum_{j=b+1}^k j^{-1}} \right) \left(\sum_{i=1}^k \frac{E_i - 1}{k} S(i, k) (1 + o(1)) \right) \right] \\ &\quad + Q_0^2(n/k) c_{b,k,p} \bar{c}_{k,p} \\ &= \xi^2 \frac{\sum_{j=1}^k (j \vee b)^{-1} S(j, k)}{k(1 + \sum_{j=b+1}^k j^{-1})} (1 + o(1)) + Q_0^2(n/k) c_{b,k,p} \bar{c}_{k,p} \end{aligned}$$

where $S(j, k) := \log(1 + \log(k/j)) + \frac{ek}{j} \mathbf{E}(1 + \log(k/j))$.

We now proceed to add the k summands of the expected variance. To this end, some preparatory calculations will be helpful. By (6.16) and Riemann approximation we have

$$\begin{aligned} \frac{1}{k} \sum_{b=1}^k c_{b,k,p}^2 &\approx \frac{1}{k} \sum_{b=1}^k \frac{1}{p^2} \cdot \frac{\frac{((k+1)/b)^{2p}}{(1-p)^2} - 2\frac{((k+1)/b)^p}{1-p} + 1}{(1 + \log((k+1)/b))^2} \\ &\approx \frac{1}{p^2(1-p)^2} [1 - e^{1-2p}(1-2p)\mathbf{E}(1-2p)] \\ &\quad - \frac{2}{p^2(1-p)} [1 - e^{1-p}(1-p)\mathbf{E}(1-p)] \\ &\quad + \frac{1}{p^2} [1 - e\mathbf{E}(1)]. \end{aligned}$$

By virtue of (6.27),

$$\bar{c}_{k,p}^2 \approx \frac{e^{2(1-p)}}{p^2(1-p)^2} \mathbf{E}^2(1-p) - 2\frac{e^{2-p}}{p^2(1-p)} \mathbf{E}(1-p)\mathbf{E}(1) + \frac{e^2}{p^2} \mathbf{E}^2(1),$$

from which we deduce that as $k \rightarrow \infty$,

$$\frac{1}{k} \sum_{b=1}^k c_{b,k,p}^2 - \bar{c}_{k,p}^2 \rightarrow f(p),$$

where $f(p)$ is given by (6.18).

Observe that

$$\begin{aligned} \frac{1}{k} \sum_{b=1}^k \frac{\frac{1}{b} + \sum_{j=b+1}^k 1/j^2}{(1 + \sum_{j=b+1}^k j^{-1})^2} &\approx \frac{2}{k} \int_0^1 \frac{du}{u(1 - \log(u))^2} - \frac{1}{k} \int_0^1 \frac{du}{(1 - \log(u))^2} \\ &= \frac{1 + e\mathbf{E}(1)}{k}. \end{aligned}$$

Next,

$$\begin{aligned} &\frac{1}{e} \frac{1}{k} \sum_{b=1}^k \frac{\sum_{j=1}^b b^{-1}(\log(1 + \log(k/j)) + \frac{ek}{j} E_1(1 + \log(k/j)))}{1 + \log(k/b)} \\ &\approx \int_0^1 \frac{1}{z \log(e/z)} \left(\int_0^z \frac{1}{u} \left(\int_{\log(e/u)}^\infty \log(v) e^{-v} dv \right) du \right) dz \\ &\approx 0.266 =: I_1 \end{aligned}$$

and

$$\begin{aligned} &\frac{1}{e} \frac{1}{k} \sum_{b=1}^k \frac{\sum_{j=b+1}^k j^{-1}(\log(1 + \log(k/j)) + \frac{ek}{j} E_1(1 + \log(k/j)))}{1 + \log(k/b)} \\ &\approx \int_0^1 \frac{1}{\log(e/z)} \left(\int_z^1 \frac{1}{u^2} \left(\int_{\log(e/u)}^\infty \log(v) e^{-v} dv \right) du \right) dz \\ &\approx 0.135746 =: I_2 \end{aligned}$$

Finally,

$$\begin{aligned} & \frac{1}{k} \sum_{j=1}^k (k/j)^2 \left(\int_{\log(e/u)}^{\infty} \log(v) e^{-v} dv \right)^2 \\ & \approx \int_0^1 u^{-2} \left(\int_{\log(e/u)}^{\infty} \log(v) e^{-v} dv \right)^2 du \\ & \approx 0.148005 =: I_3. \end{aligned}$$

Altogether we hence obtain

$$\begin{aligned} & \mathbb{E} \left[\frac{1}{k} \sum_{b=1}^k (T_{b,k} - \bar{T}_k)^2 \right] \\ & = \frac{1}{k} \sum_{b=1}^k \left(\mathbb{E}((T_{b,k} - \xi)^2) + \mathbb{E}((\bar{T}_k - \xi)^2) - 2\mathbb{E}((T_{b,k} - \xi)(\bar{T}_k - \xi)) \right) \\ & = \xi^2 \left(\frac{1 + e \mathbf{E}(1)}{k} \right) (1 + o(1)) + \xi^2 \frac{e^2}{k} I_3 (1 + o(1)) \\ & \quad - 2\xi^2 \frac{e}{k} (I_1 + I_2) (1 + o(1)) + Q_0^2(n/k) \left[\frac{1}{k} \sum_{b=1}^k c_{b,k,p}^2 - \bar{c}_{k,p}^2 \right] (1 + o_p(1)) \\ & = \xi^2 \left(\frac{1 + e \mathbf{E}(1)}{k} \right) (1 + o(1)) + \xi^2 \frac{e^2}{k} I_3 (1 + o(1)) \\ & \quad - 2\xi^2 \frac{e}{k} (I_1 + I_2) (1 + o(1)) + Q_0^2(n/k) f(p) (1 + o_p(1)) \\ & = \frac{C}{k} \xi^2 (1 + o(1)) + Q_0^2(n/k) f(p) (1 + o_p(1)) \end{aligned}$$

with

$$C = 1 + e \mathbf{E}(1) + e^2 I_3 - 2e(I_1 + I_2) \approx 0.502727.$$

□

Chapter 7

Trimmed extreme value estimators for censored heavy-tailed data

This chapter is based on the following manuscript being prepared for submission:

Albrecher, H., Beirlant, J., & Bladt, M. (2020). Trimmed extreme value estimators for censored heavy-tailed data. Preprint, University of Lausanne.

Abstract

We consider estimation of the extreme value index and extreme quantiles for heavy-tailed data that are right-censored. We study a general procedure of removing low importance observations in tail estimators. This trimming procedure is applied to the state-of-the-art estimators for randomly right-censored tail estimators, consequently deriving various families of trimmed estimators. The latter are put into a kernel framework together with one of the existing estimators. Extensive simulation suggests that one of the new considered kernels leads to a highly competitive estimator against virtually any other available alternative in this framework.

7.1 Introduction

In recent years the problem of tail estimation for heavy-tailed distributions when the available data are right-censored has received considerable attention. Several papers on this subject have been motivated by heavy-tailed insurance claim data with long development times of the claims, see e.g. Beirlant et al. (2016, 2018, 2019); Worms and Worms (2014, 2016, 2018); Ndao et al. (2014). The underlying model assumption here is that the random variable of interest X has a Pareto-type distribution function

$$F(x) = \mathbb{P}(X \leq x) = 1 - x^{-1/\xi}\ell(x), \quad \xi > 0, x > 1, \quad (7.1)$$

where ℓ is slowly varying at infinity:

$$\lim_{x \rightarrow \infty} \frac{\ell(tx)}{\ell(x)} = 1, \text{ for every } t > 1.$$

A popular model for modelling incomplete right-censored observations is given by the random right-censoring model, where the independent and identically distributed (i.i.d.) observations X_1, \dots, X_n of X may be preceded by censoring variables C_1, \dots, C_n , and it is known if that happens. One then observes

$$Z_i = \min\{X_i, C_i\}, \quad e_i = 1\{X_i \leq C_i\}, \quad i = 1, \dots, n,$$

where C_1, \dots, C_n is an i.i.d. sequence of censoring random variables, independent of the observations X_i . In order to avoid that the largest X observations would almost surely be censored, one assumes that also the censoring variables are Pareto-type distributed with distribution function

$$G(x) = \mathbb{P}(C_1 \leq x) = 1 - x^{-1/\xi_c} \ell_c(x), \quad \xi_c > 0, x > 1,$$

with ℓ_c another slowly varying function at infinity. Then we have that for $x > 1$,

$$H(x) = \mathbb{P}(Z_1 \leq x) = 1 - x^{-1/\xi_z} \ell_z(x), \quad \xi_z = \frac{\xi \xi_c}{\xi + \xi_c},$$

where $\ell_z(x) = \ell(x)\ell_c(x)$. As explained in Einmahl et al. (2008) the parameter $p = \xi_z/\xi = \frac{\xi_c}{\xi + \xi_c}$ is the limit of $\mathbb{P}(e_1 = 1 | Z_1 = z)$ as $z \rightarrow \infty$, and can be interpreted as the non-censoring probability in the limit, or the tail limiting proportion of non-censored data. In the exact Pareto setting (i.e. ℓ and ℓ_c being constant) the censoring indicators e_1, \dots, e_n turn out to be i.i.d. Bernoulli(p) random variables, independent of Z_1, \dots, Z_n .

Within this censoring and regularly varying context, Beirlant et al. (2007) proposed a first estimator of ξ in the spirit of the classical Hill estimator (cf. Hill (1975)). Concretely, define the order statistics of the observed sample as

$$Z_{1,n} \leq \dots \leq Z_{n,n},$$

and $e_{i,n}$ the corresponding censoring indicators $i = 1, \dots, n$. Then the Hill estimator adapted for censoring is given by

$$H_k = \frac{\sum_{i=1}^k \log(Z_{n-i+1,n}/Z_{n-k,n})}{\sum_{i=1}^k e_{n-i+1,n}}, \quad 1 < k < n. \quad (7.2)$$

Einmahl et al. (2008) showed that, under some regularity assumptions, H_k is consistent and asymptotically normally distributed, whatever the value of $p \in (0, 1)$. In the present paper it will be useful to define

$$p_k = \frac{1}{k} \sum_{i=1}^k e_{n-i+1,n}, \quad H_k^Z = \frac{\sum_{i=1}^k \log(Z_{n-i+1,n}/Z_{n-k,n})}{k}, \quad k < n,$$

and then, under some conditions, as $k, n \rightarrow \infty$ and $k/n \rightarrow 0$

$$H_k = \frac{H_k^Z}{p_k} \rightarrow_p \frac{\xi_z}{p} = \xi.$$

Worms and Worms (2014) proposed an alternative generalization of the Hill estimator based on the fact that

$$\mathbb{E}(\log(Z/t)|Z > t) = \int_1^\infty \frac{\overline{F}(ut)}{\overline{F}(t)} \frac{1}{u} du \rightarrow \xi \text{ as } t \rightarrow \infty, \quad (7.3)$$

where $\overline{F}(x) = 1 - F(x)$. In the exact Pareto case, the above limit is an equality. Replacing \overline{F} with the Kaplan-Meier estimate

$$\widehat{\overline{F}}(x) = \prod_{Z_{i,n} \leq x} \left(\frac{n-i}{n-i+1} \right)^{e_{i,n}}$$

for the tail yields the estimator

$$H_k^W = \sum_{i=1}^k \frac{\widehat{\overline{F}}(Z_{n-i,n})}{\widehat{\overline{F}}(Z_{n-k,n})} \log(Z_{n-i,n}/Z_{n-i+1,n}), \quad (7.4)$$

which was shown to be consistent in Worms and Worms (2014), while Beirlant et al. (2019) derived asymptotic normality under light censoring, i.e. $\xi < \xi_c$ or $p > 1/2$, and some regularity conditions. Observe that both (7.2) and (7.4) reduce to H_k^Z when there is no censoring. Based on simulation studies, see e.g. Beirlant et al. (2018), the estimator H_k^W is known to exhibit superior behaviour in comparison with H_k , especially with respect to bias.

Before introducing the trimming procedure, we first propose a simplified version of H_k^W , that will be more amenable for the approach in the sequel. To this end, note that one can write

$$H_k^W = \sum_{i=1}^k \left[\prod_{j=i+1}^k (1 - 1/j)^{e_{n-i+1,n}} \right] \log(Z_{n-i+1,n}/Z_{n-i,n}), \quad (7.5)$$

where the term in the square bracket has expectation

$$\mathbb{E} \left[\prod_{j=i+1}^k (1 - 1/j)^{e_{n-i+1,n}} \right] = \prod_{j=i+1}^k (1 - p/j) \quad (7.6)$$

for the exact Pareto case, while the second factor in (7.5) satisfies, by the Rényi representation,

$$\mathbb{E}(\log(Z_{n-i+1,n}/Z_{n-i,n})) = \frac{1}{i \sum_{m=i}^k m^{-1}} \mathbb{E}(\log(Z_{n-i+1,n}/Z_{n-k,n})). \quad (7.7)$$

Based on (7.6) and (7.7), and using the approximations $\sum_{m=i}^k m^{-1} \approx \log((k+1)/i)$ and $(1 - p_k/j) \approx \exp(-p_k/j)$, we define a novel simpler estimator of ξ by

$$H_k^A = \frac{1}{k+1} \sum_{i=1}^k \left(\frac{i}{k+1} \right)^{p_k-1} \frac{1}{\log((k+1)/i)} \log(Z_{n-i+1,n}/Z_{n-k,n}), \quad k < n, \quad (7.8)$$

where the log-spacings in the sum are all taken with respect to the same baseline order statistic $Z_{n-k,n}$. The latter will allow to apply the trimming operation of removing low importance observations in the tail estimation, developed in Bladt et al. (2019) for the non-censoring case, to the present situation with censoring.

In this paper, we extend the trimming method proposed in Bladt et al. (2019) to the case of random right censoring, both for H_k and H_k^A . Averaging the trimmed statistics over the amount of trimming then leads to new estimators which belong to a general family of kernel estimators comprising H_k and H_k^A . This family turns out to be closed under the proposed averaging operation after trimming. After studying the basic asymptotic properties of the kernel estimators in Section 7.3, we discuss the optimal choice of k when using the proposed estimators in Section 7.4. In a final section, the merits of the new kernel estimators and the threshold selection method are illustrated through simulations and a case study from insurance.

7.2 Trimmed estimators for ξ

7.2.1 Trimming tail estimators

In Bladt et al. (2019), lower trimming of the classical Hill estimator was shown to be an effective strategy to obtain Hill-type plots with lower variance arising from the changes of the baseline order statistic, which aids in the visual selection of a horizontal part of the trajectory. Here, we extend this approach to the censored case, and consider lower trimming of the estimators H_k and H_k^A , deleting the smallest $k - b$ ($b \leq k$) peaks over thresholds $Z_{n-i+1,n}/Z_{n-k,n}$, $i = b + 1, \dots, k$:

$$H_{b,k} = \frac{1}{1 + \sum_{j=b+1}^k j^{-1}} \cdot \frac{\frac{1}{b} \sum_{i=1}^b \log(Z_{n-i+1,n}/Z_{n-k,n})}{p_k}, \quad b \leq k \leq n - 1, \quad (7.9)$$

and analogously

$$H_{b,k}^A = \frac{1}{b+1} \sum_{i=1}^b \left(\frac{i}{b+1} \right)^{p_k-1} \frac{1}{\log((k+1)/i)} \log(Z_{n-i+1,n}/Z_{n-k,n}), \quad b \leq k \leq n - 1, \quad (7.10)$$

as the trimmed versions of H_k and H_k^A . Note that $H_{k,k} = H_k$ and $H_{k,k}^A = H_k^A$.

7.2.2 Averaging and kernels

The above trimming procedure naturally leads to new estimators when considering the empirical mean of the trimmed estimators across $b = 1, \dots, k$

$$\frac{1}{k} \sum_{b=1}^k H_{b,k}, \quad \frac{1}{k} \sum_{b=1}^k H_{b,k}^A.$$

For instance, in case of $H_{b,k}^A$ this is asymptotically equivalent to

$$\overline{H}_k^A = \frac{1}{k} \sum_{i=1}^k \frac{1}{1-p_k} \left(\left(\frac{k+1}{i} \right)^{1-p_k} - 1 \right) \frac{1}{\log((k+1)/i)} \log(Z_{n-i+1,n}/Z_{n-k,n}), \quad k < n, \quad (7.11)$$

as can be seen by a simple Riemann sum approximation as before.

In fact H_k , H_k^A and \overline{H}_k^A can all be put into a kernel framework, by defining

$$H_k^{\mathcal{K}} = \frac{1}{k} \sum_{i=1}^k \mathcal{K} \left(\frac{i}{k+1}, p_k \right) \frac{1}{\log((k+1)/i)} \log(Z_{n-i+1,n}/Z_{n-k,n}), \quad (7.12)$$

where \mathcal{K} is a positive kernel function satisfying

$$\int_0^1 \mathcal{K}(u; p) du = \frac{1}{p}, \quad \text{for all } p \in (0, 1].$$

In particular, we get

$$\begin{aligned} H_k &= H_k^{\mathcal{K}_0}, \quad \text{with } \mathcal{K}_0(u, p) = \frac{1}{p} \log \left(\frac{1}{u} \right), \\ H_k^A &= H_k^{\mathcal{K}_1}, \quad \text{with } \mathcal{K}_1(u, p) = u^{p-1}, \\ \overline{H}_k^A &= H_k^{\mathcal{K}_2}, \quad \text{with } \mathcal{K}_2(u, p) = \frac{u^{p-1} - 1}{1-p}. \end{aligned}$$

Note that H_k^W does not fall into this framework, but its simplified version H_k^A does.

Also notice that, when trimming any kernel estimator $H_k^{\mathcal{K}}$ to obtain

$$H_{b,k}^{\mathcal{K}} = \frac{1}{b+1} \sum_{i=1}^b \mathcal{K} \left(\frac{i}{b+1}, p_k \right) \frac{1}{\log((k+1)/i)} \log(Z_{n-i+1,n}/Z_{n-k,n}), \quad (7.13)$$

the averaging operation $\frac{1}{k} \sum_{b=1}^k H_{b,k}^{\mathcal{K}}$ leads to an associated kernel estimator

$$H_k^{\overline{\mathcal{K}}} = \frac{1}{k} \sum_{i=1}^k \overline{\mathcal{K}} \left(\frac{i}{k+1}, p_k \right) \frac{1}{\log((k+1)/i)} \log(Z_{n-i+1,n}/Z_{n-k,n}), \quad (7.14)$$

with

$$\bar{\mathcal{K}}(u, p) = \int_u^1 \frac{\mathcal{K}(v, p)}{v} dv,$$

where $\bar{\mathcal{K}}(\frac{i}{k+1}, p)$ is obtained using a Riemann approximation of $\frac{1}{k} \sum_{b=i}^k \frac{k}{b+1} \mathcal{K}(\frac{i}{k+1} \frac{k+1}{b+1}, p)$ as $k \rightarrow \infty$ for fixed b . An interesting question in its own right, which we will not pursue further in the sequel, is whether creating kernels in such an iterative fashion leads to a limit in some appropriate functional space.

7.3 Asymptotic representations

In this section we derive the asymptotic distributions of the kernel estimators and their trimmed counterparts as introduced in the preceding section. In Einmahl et al. (2008) the asymptotics for $H_k = H_k^{\mathcal{K}_0}$ was discussed in detail (and note that Beirlant et al. (2019) provided an asymptotic normality result for H_k^W when $p > 1/2$, but that estimator is not in the current kernel framework). Here we provide asymptotic representations for the class of kernel estimators $H_k^{\mathcal{K}}$ in general. To this end, we make use of second-order assumptions which were first proposed in Hall and Welsh (1985) and since then have gained popularity in the extreme value community:

$$\begin{aligned} \ell(x) &= C(1 + Dx^{-\beta}(1 + o(1))), \\ \ell_c(x) &= C_c(1 + D_c x^{-\beta_c}(1 + o(1))), \quad x \rightarrow \infty, \end{aligned}$$

where β, β_c, C, C_c are positive constants and D, D_c are real constants. It now follows that

$$\ell_z(x) = C_z(1 + D_z x^{-\beta_z}(1 + o(1))),$$

where

$$C_z = CC_c, \quad \beta_z = \min\{\beta, \beta_c\}, \quad D_z = D \cdot 1_{\beta \leq \beta_c} + D_c \cdot 1_{\beta_c \leq \beta}.$$

Concerning the peaks-over-threshold values $Z_{n-i+1,n}/Z_{n-k,n}$, $i = 1, \dots, k$, one then has the following expansion as $n, k \rightarrow \infty$ and $k/n \rightarrow 0$ (see pp.75-76 in de Haan and Ferreira (2007)):

$$\begin{aligned} \log(Z_{n-i+1,n}/Z_{n-k,n}) &= \xi_z \log((k+1)/i) + \frac{\xi_z}{\sqrt{k}} V\left(\frac{i}{k+1}\right) \\ &\quad + Q_{0,z}(n/k) k_{\rho_z}\left(\frac{k+1}{i}\right)(1 + o(1)), \end{aligned} \quad (7.15)$$

where $\rho_z = -\beta_z \xi_z$, $Q_{0,z}(t) = -\xi_z^2 \beta_z D_z C^{\rho_z} t^{\rho_z}$, $k_{\rho_z}(u) = \frac{u^{\rho_z-1}}{\rho_z}$, and $\{V(u)\}_{u>0}$ is a centered Gaussian process with covariance function

$$\mathbb{E}(V(u)V(v)) = \frac{1 - \max\{u, v\}}{\max\{u, v\}}, \quad u, v > 0.$$

Next, from Einmahl et al. (2008) and Beirlant et al. (2016) one obtains that

$$\sqrt{k}(p_k - p) = \sqrt{p(1-p)}Z + \sqrt{k}Q_{0,z}(n/k) \frac{\kappa_z}{1-\rho_z}(1+o(1)), \quad (7.16)$$

where $Z \sim N(0, 1)$ independent of V , and $\kappa_z = -\frac{(D\xi)_z}{D_z\xi\xi_c}$, with $(D\xi)_z = (D\xi)1_{\beta \leq \beta_c} - (D_c\xi_c)1_{\beta_c \leq \beta}$. Based on (7.15) and (7.16) we now derive that

$$\begin{aligned} H_{b,k}^{\mathcal{K}} - \xi &= \frac{1}{b+1} \sum_{i=1}^b \left(\mathcal{K}\left(\frac{i}{b+1}, p_k\right) - \mathcal{K}\left(\frac{i}{b+1}, p\right) \right) \frac{\log(Z_{n-i+1,n}/Z_{n-k,n})}{\log((k+1)/i)} \\ &\quad + \left(\frac{1}{b+1} \sum_{i=1}^b \mathcal{K}\left(\frac{i}{b+1}, p\right) \frac{\log(Z_{n-i+1,n}/Z_{n-k,n})}{\log((k+1)/i)} - \xi \right) \\ &=: T_{1,b,k} + T_{2,b,k}. \end{aligned} \quad (7.17)$$

Using the mean value theorem, we have from (7.15) and (7.16) that

$$\begin{aligned} T_{1,b,k} &\sim_p \xi_z(p_k - p) \alpha_b^{\mathcal{K}} \\ &= \xi_z \alpha_b^{\mathcal{K}} \left(\sqrt{p(1-p)} \frac{Z}{\sqrt{k}} + Q_{0,z}(n/k) \frac{\kappa_z}{1-\rho_z}(1+o(1)) \right), \end{aligned} \quad (7.18)$$

with $\alpha_b^{\mathcal{K}} = \frac{1}{b+1} \sum_{i=1}^b \frac{\partial \mathcal{K}}{\partial p}\left(\frac{i}{b+1}, p\right)$. Next, using (7.15),

$$\begin{aligned} T_{2,b,k} &= \xi_z \left(\frac{1}{b+1} \sum_{i=1}^b \mathcal{K}\left(\frac{i}{b+1}, p\right) - \frac{1}{p} \right) \\ &\quad + \frac{\xi_z}{\sqrt{k}} \frac{1}{b+1} \sum_{i=1}^b \mathcal{K}\left(\frac{i}{b+1}, p\right) \frac{V(i/(k+1))}{\log((k+1)/i)} \\ &\quad + Q_{0,z}(n/k)(1+o(1)) \frac{1}{b+1} \sum_{i=1}^b \mathcal{K}\left(\frac{i}{b+1}, p\right) \frac{k_{\rho_z}((k+1)/i)}{\log((k+1)/i)}. \end{aligned} \quad (7.19)$$

Concerning the non-trimmed kernel estimators, we find the asymptotic expansion

$$\begin{aligned} \sqrt{k}(H_k^{\mathcal{K}} - \xi) &= -\sqrt{p(1-p)} \frac{\xi_z}{p^2} Z + \frac{\xi_z}{k+1} \sum_{i=1}^k \mathcal{K}\left(\frac{i}{k+1}, p\right) \frac{V(i/(k+1))}{\log((k+1)/i)} \\ &\quad + \sqrt{k}Q_{0,z}(n/k) \left\{ \frac{-\kappa_z \xi_z}{p^2(1-\rho_z)} + \frac{1}{k+1} \sum_{i=1}^k \mathcal{K}\left(\frac{i}{k+1}, p\right) \frac{k_{\rho_z}((k+1)/i)}{\log((k+1)/i)} \right\} (1+o(1)), \end{aligned} \quad (7.20)$$

using $\alpha_k^{\mathcal{K}} = -1/p^2 + O(k^{-1})$ and $\frac{1}{k+1} \sum_{i=1}^k \mathcal{K}\left(\frac{i}{k+1}, p\right) - \frac{1}{p} = O(k^{-1})$. Hence the asymptotic mean squared error of $H_k^{\mathcal{K}}$ is given by

$$\text{AMSE}(H_k^{\mathcal{K}}) = \frac{\xi_z^2}{k} v_p + Q_{0,z}^2\left(\frac{n}{k}\right) b_p \quad (7.21)$$

with

$$v_p = \frac{1-p}{p^3} + \frac{2}{k+1} \sum_{i=1}^k \mathcal{K}\left(\frac{i}{k+1}, p\right) \frac{1 - \frac{i}{k+1}}{\frac{i}{k+1} \log\left(\frac{k+1}{i}\right)} \left(\frac{1}{k+1} \sum_{j=1}^i \mathcal{K}\left(\frac{j}{k+1}, p\right) \frac{1}{\log\left(\frac{k+1}{j}\right)} \right),$$

$$b_p = \left\{ \frac{-\kappa_z \xi_z}{p^2(1-\rho_z)} + \frac{1}{k+1} \sum_{i=1}^k \mathcal{K}\left(\frac{i}{k+1}, p\right) \frac{k\rho_z((k+1)/i)}{\log((k+1)/i)} \right\}^2.$$

7.4 Optimal choice of k when estimating ξ

Denoting the trimmed version of the Hill estimator in the fully observed case by $H_{b,k}^Z = H_{b,k} p_k$, it was shown in Bladt et al. (2019) that the value $k_{\text{opt}}(H_k^Z)$ of k minimizing the asymptotic MSE of H_k^Z satisfies

$$k_{\text{opt}}(H_k^Z) = \left(\frac{K}{(1-\rho_z)^2 f(\rho_z)} \right)^{\frac{-1}{1-2\rho_z}} k_{\text{opt}}(H_{b,k}^Z),$$

for a universal constant K and a specific function f . Here $k_{\text{opt}}(H_{b,k}^Z)$ is the optimal sample fraction minimizing the expectation of the empirical variance $S_k^2 = \frac{1}{k} \sum_{b=1}^k \left(H_{b,k}^Z - \frac{1}{k} \sum_{b=1}^k H_{b,k}^Z \right)^2$.

On the other hand, based on (7.20) we find under $\beta < \beta_c$, i.e. when the bias is largest compared with the classical Hill estimator in case of no censoring, that

$$\text{AMSE}(H_k) = \frac{1}{p^4} \left(p \frac{\xi_z^2}{k} + \frac{Q_{0,z}^2(n/k)}{(1-\rho_z)^2} \right), \tag{7.22}$$

which yields that

$$k_{\text{opt}}(H_k) = \left(\frac{p^{-1} K}{(1-\rho_z)^2 f(\rho_z)} \right)^{-\frac{1}{1-2\rho_z}} k_{\text{opt}}(H_{b,k}^Z). \tag{7.23}$$

This means that the optimal k for the estimator H_k with respect to minimization of the AMSE is linked to the optimal k of its trimmed versions for the minimization of the expected empirical variance in the non-censored case. A consequence of the above formula is that

$$k_{\text{opt}}(H_k) = p^{\frac{1}{1-2\rho_z}} k_{\text{opt}}(H_k^Z).$$

That is, a larger percentage of censoring leads to a higher threshold, when compared to the non-censored case. This can already be seen from the expression of the AMSE given in (7.22), where a smaller p leads to more weight being given to the bias term. From an intuitive point of view, when dealing with censored datasets, two sources of bias have to be accounted for, and hence a smaller sample fraction k is needed to control them.

In practice, for a given sample, one finds an estimate $\hat{k}_0 = \hat{k}_{\text{opt}}(H_{b,k}^Z)$ of $k_{\text{opt}}(H_{b,k}^Z)$ through minimization of s_k^2 over k , from which an adaptive choice of k is found through

$$\left(\frac{p_{\hat{k}_0}^{-1} K}{(1-\hat{\rho}_z)^2 f(\hat{\rho}_z)} \right)^{-\frac{1}{1-2\hat{\rho}_z}} \hat{k}_0,$$

using an estimate $\hat{\rho}_z$ of ρ_z and replacing p by $p_{\hat{k}_0}$. The second-order parameter ρ_z is known to be hard to estimate, even in the non-censored case. In the next section we use the choices $\rho_z = -1, -1/2, -3/2$, but the results are not very sensitive to this parameter, which is commonly taken as simply $\rho_z = -1$ in practice.

7.5 Simulations

We performed simulations using the following distributions.

- Burr distribution with survival function $1 - F(x) = \left(\frac{\theta}{\theta + x^\beta}\right)^\lambda$ with (θ, β, λ) taken as (10,2,1) for X and (10,3,1) for C so that $p < 1/2$, next to (10,3,1) for X and (10,2,1) for C so that $p > 1/2$, and (10,2,1) for both X and C with $p = 1/2$.
- Fréchet distribution with $F(x) = \exp(-x^{-1/\xi})$ with ξ taken as 1/2 for X and 1/4 for C and correspondingly $p < 1/2$, as 1/4 for X and 1/2 for C and correspondingly $p > 1/2$, and finally as $\xi = 1/4$ for both X and C , so that $p = 1/2$.
- Log-gamma distribution with density $f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} (\log x)^{\alpha-1} x^{-\lambda-1}$ with (α, λ) taken as (3/2,2) for X and (3/2,4) for C so that $p < 1/2$, as (3/2,4) for X and (3/2,2) for C so that $p > 1/2$, and as (3/2,4) for X and C so that $p = 1/2$.

The results are based on 200 simulations of sample size $n = 200$ each.

In Figures 7.1, 7.2 and 7.3 we plot the bias, variance and mean squared error as a function of k of the various estimators considered above. Note that the MSE characteristics of the estimator $H_k^{\mathcal{K}_2}$ are quite comparable to those of H_k^W in the Burr and Fréchet cases, and are even better for the log-gamma model.

In Figures 7.4, 7.5 and 7.6 we provide violin plots for the optimal threshold for the \mathcal{K}_0 -based estimator, selected according to the automatic procedure given in the previous section. We have taken $\rho_z = -1, -3/2, -1/2$ respectively for the three distributions that we consider. These values were permuted (the resulting plots are omitted) and the results were not very sensitive to the choice of ρ_z . To avoid degeneracies, a cutoff of 1/5 of the size of the data set was used for the empirical variance estimates. Where relevant we also add the results of the parameter estimates when taking k fixed at the theoretical optimal value.

7.6 Insurance Application: censored claims data vs ultimates

We now proceed to analyze an insurance dataset consisting of 837 motor third-party liability (MTPL) insurance claims from 1995 till 2010. This data set has been described and studied in Albrecher et al. (2017), Bladt et al. (2019) (without censoring, using ultimate values instead) and Bladt et al. (2020) (using both censoring and ultimate values).

The data exhibit right-censoring, that is, a claim size is partially observed whenever the development of the claim payment is ongoing and the claim is not yet closed. Observed claim sizes thus are considered as observed data points. In Bladt et al. (2020) it was argued that the assumption of random censoring and heavy-tailedness is adequate.

Using the same mechanism as for the simulation study (and $\rho_z = -1$) we find that $k = 35$ is the optimal threshold for the estimator using the kernel \mathcal{K}_0 (see Figure 7.7). As observed in the simulations, it makes sense to evaluate the other estimators at this value as well. This yields the estimates

$$\hat{\xi} = 0.8781691, 0.8745332, 0.6981923, 0.7003625.$$

The latter two values correspond to the kernel \mathcal{K}_2 and to the Worms estimator H_k^W . They are quite close, and the simulations suggest that they are also the best performing. Previous studies, using the ultimate values (cf. Bladt et al. (2019), with subsequent agreement in Albrecher et al. (2019)), that is, internal projected values of the claim sizes at closure, suggested a tail index of about 0.48. In Bladt et al. (2020), combining this expert information with the estimator corresponding to \mathcal{K}_0 , intermediate values between the purely statistical 0.87 and the purely expert information 0.48 were suggested. The present value of 0.7 is an interesting intermediate value that arises from a purely statistical procedure.

7.7 Conclusion

In this paper we developed novel extreme value estimators under right-censoring in a kernel framework. The latter class is closed (in the asymptotic sense) under the averaging operation of their trimmed versions, by a simple replacement of kernel. The asymptotic behaviour is given for arbitrary kernels, which allows us to compute, for instance, the expression for the MSE as a function of k . The choice of the optimal threshold with respect to MSE is explored in connection with the empirical variance of the trimmed trajectories, which leads to an automated way of selecting a threshold. As for the non-censored case, the idea of selecting a threshold by exploiting this link, circumvents the usual estimation difficulties and instabilities which arise in previous approaches in the literature which typically require the estimation of the second-order parameter D , and of ξ itself. Despite its simplicity, simulation studies suggest that the method is also efficient. In fact, when compared with the theoretically optimal value, the latter sometimes is too small to be of any practical relevance, and then our adaptive estimator is superior. In the other cases, when the theoretically optimal value is sensible, our estimator also performs well against it. We finally apply the procedure to a well-understood insurance dataset, and the simulation studies suggest that the instances where \mathcal{K}_0 has been used in the literature (either alone, or in combination with expert information) to analyze these data could very possibly be improved by considering \mathcal{K}_2 instead. Interesting directions for further research include trimming the kernel estimators from above, to remove outliers from data, and to apply combined tail information using censored data and expert information with the new kernels, improving the previous methods.

Finally, it will be interesting to consider optimality criteria for the choice of k for any kernel, and to work out criteria for the selection of the optimal kernel from a purely mathematical point of view.

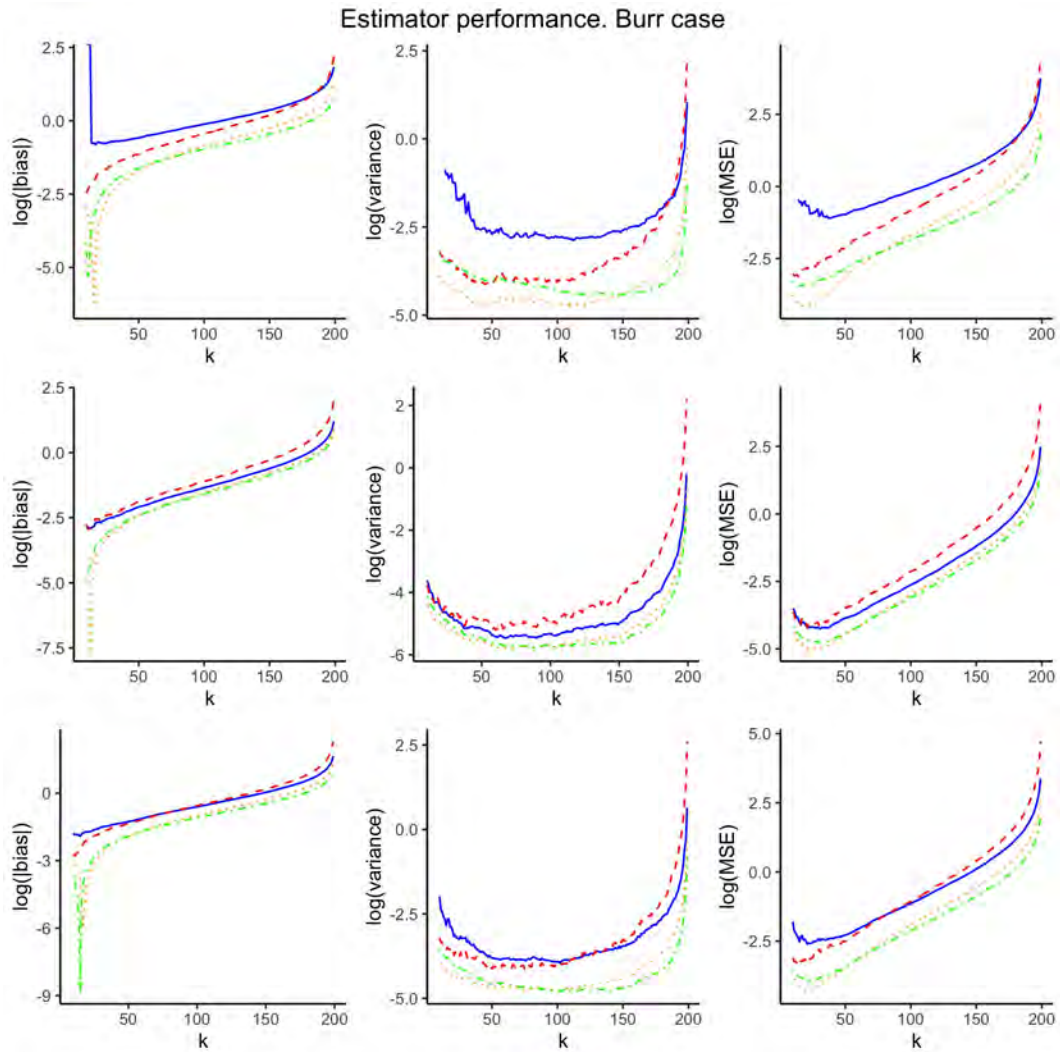


Figure 7.1: Burr distributions: bias, variance and mean square error of the kernel estimator ($H_k = H_k^{K_0}$ in solid blue, $H_k^A = H_k^{K_1}$ in dashed red, $H_k^{K_2}$ in dotted orange) and the Worms estimator H_k^W (dashed and dotted green), as a function of k . The top, middle and bottom levels correspond to $2p < 1$, $2p > 1$ and $2p = 1$, respectively.

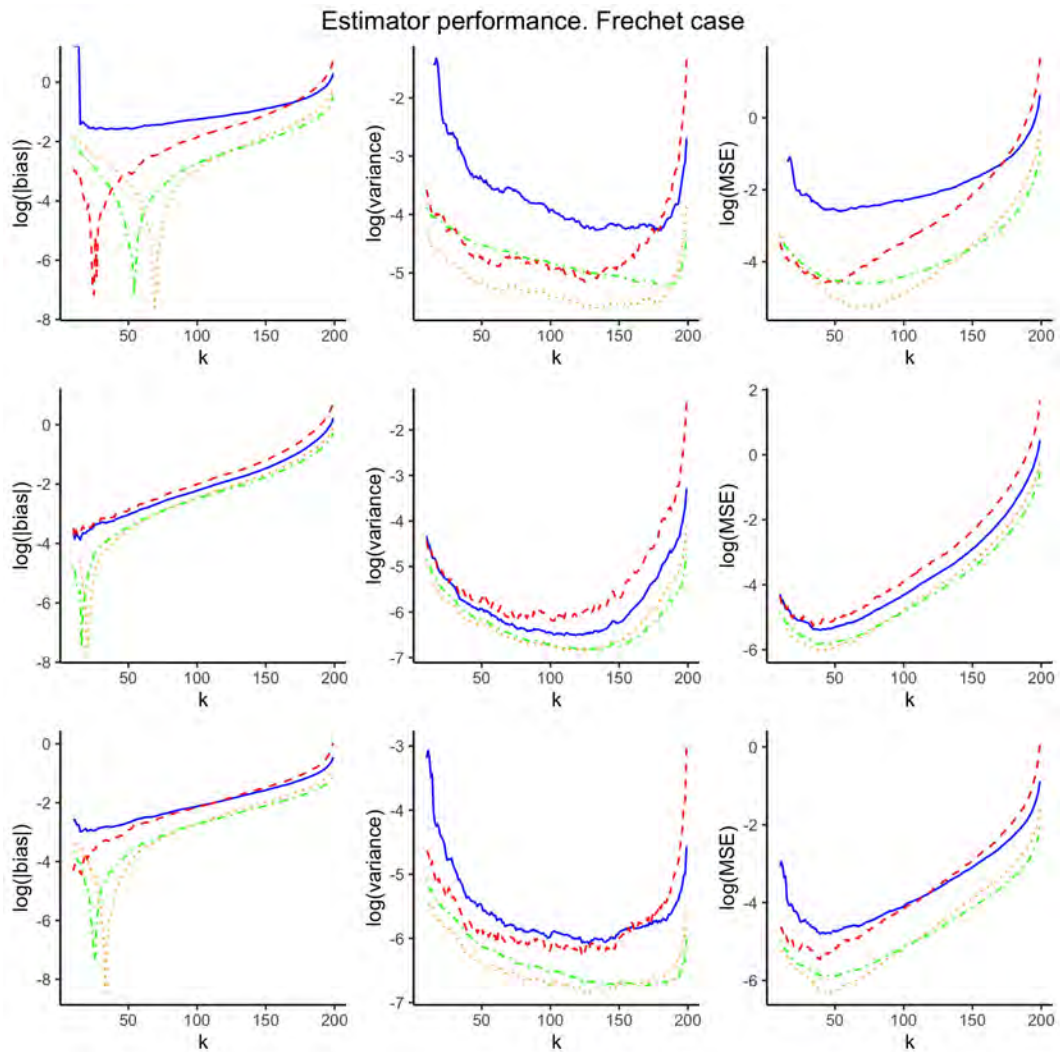


Figure 7.2: Fréchet distributions: bias, variance and mean square error of the kernel estimator ($H_k = H_k^{K_0}$ in solid blue, $H_k^A = H_k^{K_1}$ in dashed red, $H_k^{K_2}$ in dotted orange) and the Worms estimator H_K^W (dashed and dotted green), as a function of k . The top, middle and bottom levels correspond to $2p < 1$, $2p > 1$ and $2p = 1$, respectively.

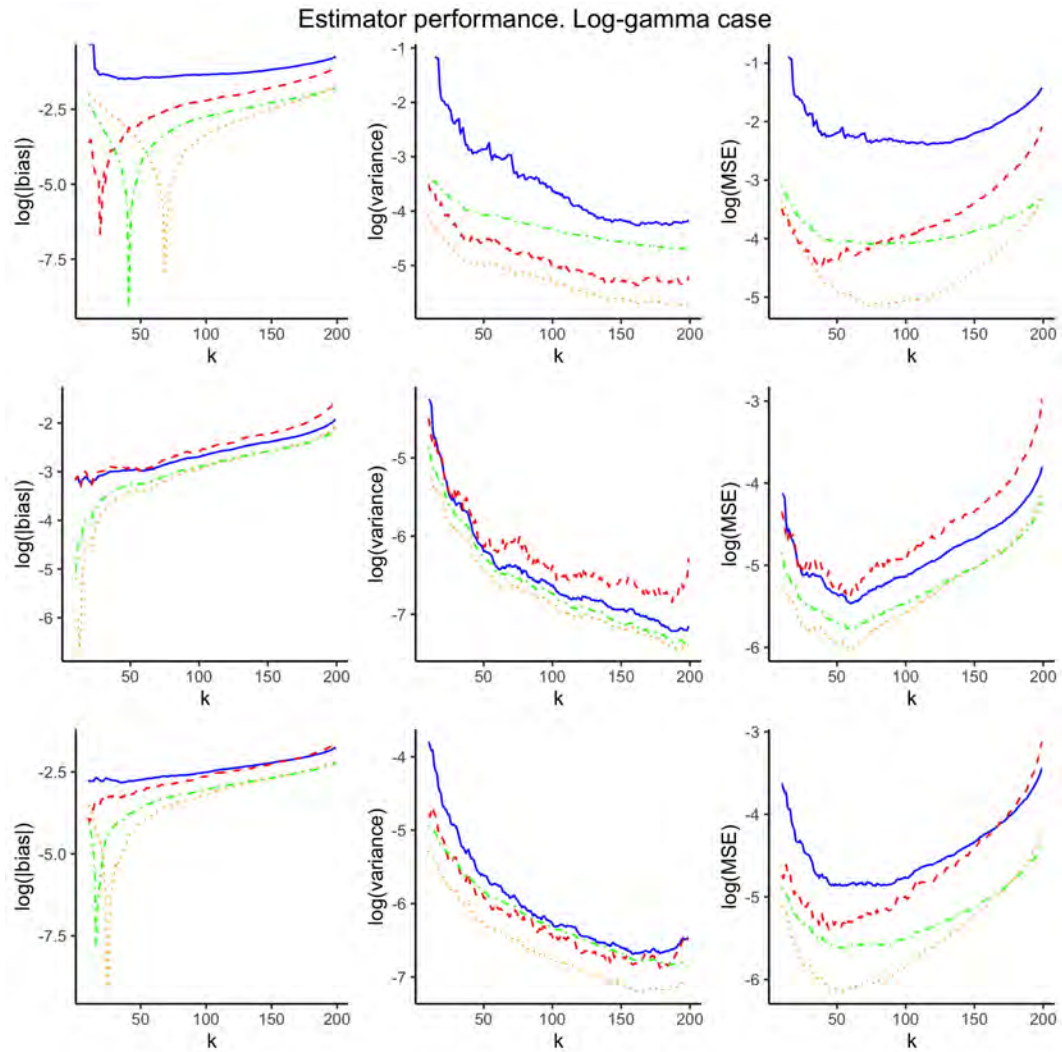


Figure 7.3: Log-gamma distributions: bias, variance and mean square error of the kernel estimator ($H_k = H_k^{K_0}$ in solid blue, $H_k^A = H_k^{K_1}$ in dashed red, $H_k^{K_2}$ in dotted orange) and the Worms estimator K_k^W (dashed and dotted green), as a function of k . The top, middle and bottom levels correspond to $2p < 1$, $2p > 1$ and $2p = 1$, respectively.

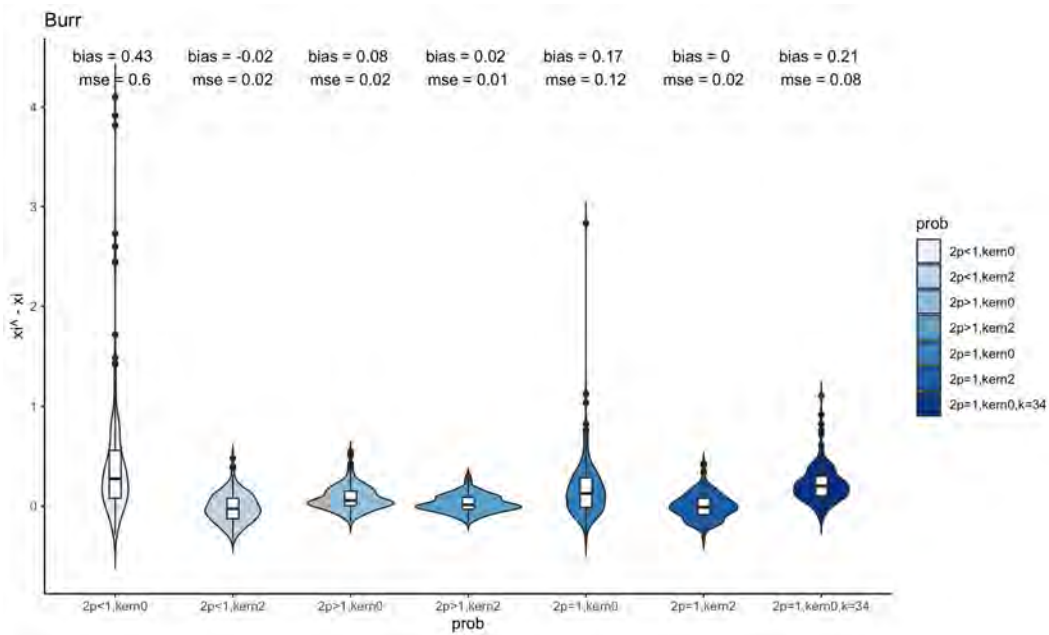


Figure 7.4: Violin plots for the simulation results in case of Burr distributions under different non-censoring asymptotic probabilities.

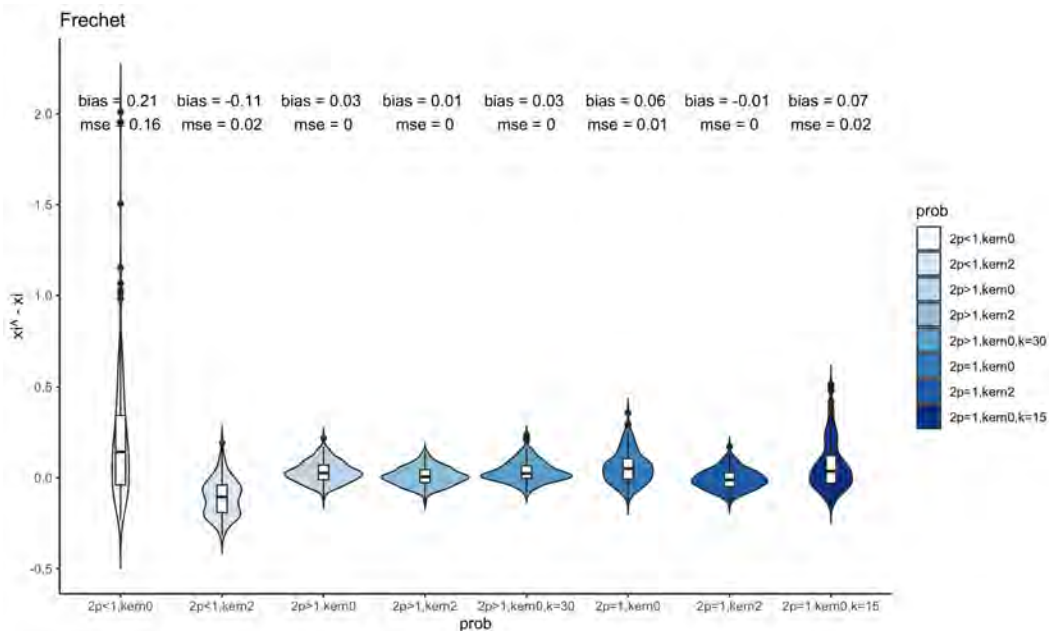


Figure 7.5: Violin plots for the simulation results in case of Fréchet distributions under different non-censoring asymptotic probabilities.

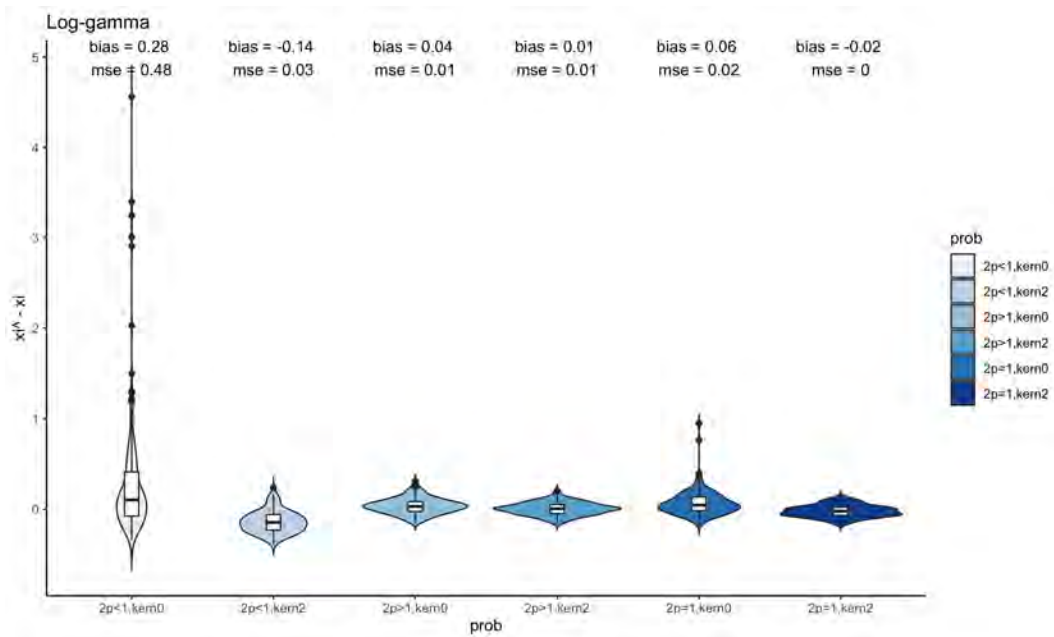


Figure 7.6: Violin plots for the simulation results in case of log-gamma distributions under different non-censoring asymptotic probabilities.

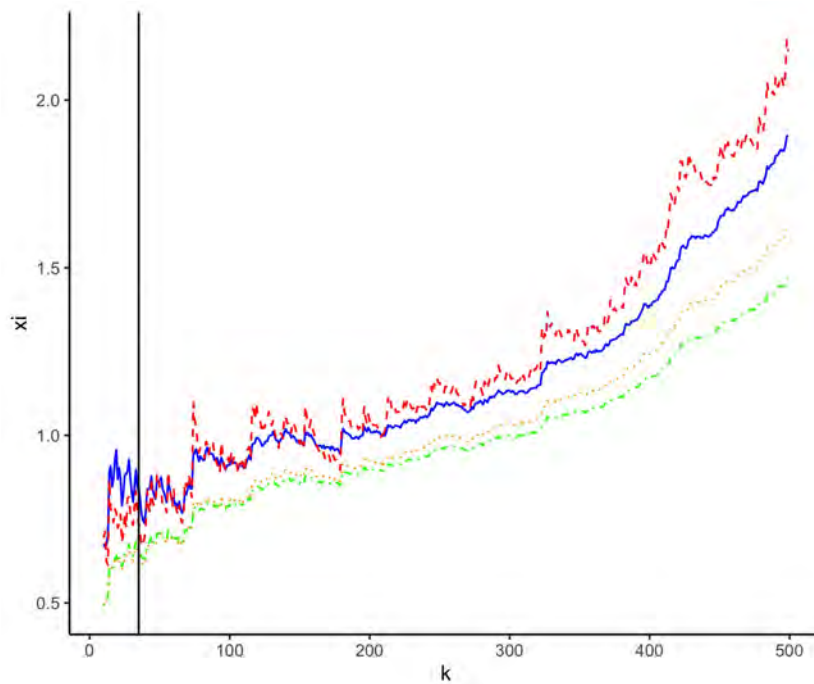


Figure 7.7: Estimates of ξ for the MTPL insurance claim size data: $H_k = H_k^{K_0}$ in solid blue, $H_k^A = H_k^{K_1}$ in dashed red, $H_k^{K_2}$ in dotted orange and the Worms estimator H_k^W in dashed and dotted green. The vertical line is at the estimated optimal k for $H_k = H_k^{K_0}$.

Chapter 8

Novelty detection for heavy-tailed randomly censored data

This chapter is based on the following manuscript being prepared for submission:

Bladt, M (2020). Novelty detection for heavy-tailed randomly censored data. Preprint, University of Lausanne.

Abstract

The problem of determining whether a small sample comes from a specific distribution is considered, the null hypothesis being rejected when there is novelty in the sample. We consider an extension of an existing methodology to the censored case. Specifically, consider a heavy-tailed sample, randomly and independently right-censored by another heavy-tailed sample, and testing in low-density regions, by means of point processes of exceedances. A transformation of the original data lets us write a survival analysis likelihood whose distribution can be approximated analytically, by considering peaks over different thresholds for the censored part and the non-censored part of the likelihood, respectively. We introduce two additional methods along similar lines, and finally exemplify the performance of such approximations using synthetic data.

8.1 Introduction

In pattern recognition or in signal processing, novelty detection is defined as the task of differentiating two samples, where one is very large and the other is small. The problem can be seen as a one-sided classification problem where the question of interest is the determination of whether the small sample comes from the distribution of the larger one or not. The area has applications in IT security, healthcare informatics, industrial or medical monitoring and others, cf. Pimentel et al. (2014) for a more detailed description of the classical applications. A decade and a half ago, the review of novelty detection given in Markou and Singh (2003a,b) distinguished the classification techniques as being statistical or of neural network-type.

In the context of machine learning the problem can be cast as the task of recognising change in a test sample with respect to a training sample. While appropriate at the time, in the more recent review given in Pimentel et al. (2014) it is argued that not only have machine learning and statistics converged as study fields, but many new methods for detecting novelty have arisen in the last decade and a new distinction is in place. More precisely, the proposed categories for novelty detection methods are: probabilistic, distance-based, domain-based, reconstruction-based and information theoretic.

As argued in Luca et al. (2018), the majority of the existing work deals with the point-wise classification and when dealing with more than a single point, multiple hypothesis testing is then adopted, which is prone to have large misclassification rate. Although there is no universally accepted definition, the difference between outlier detection and novelty detection is argued in Pimentel et al. (2014) to be significant: outliers are data which is not desired in a statistical analysis, and which may arise in contaminated samples, while novelty detection aims at measuring patterns that change. In Luca et al. (2018), a likelihood-approximation approach (using the so-called Janossy densities) reduces a multidimensional problem into a unidimensional one. Their setup allows for a second random mechanism, by making the number of points of the small sample a random variable. Such a method is then worked out in detail for the Gaussian distribution, and a simulation sensitivity analysis is also provided for that case. They also suggest a method which can be used for detecting novelty in extreme value theory, i.e. a change of behaviour of incoming data in under-sampled regions.

In this paper we take up the ideas in Luca et al. (2018) and extend them to the case where the data is possibly right-censored. We hence want to classify point patterns $\{\xi_1, \xi_2, \dots, \xi_n\}$, $n \in \mathbb{N}$ but we observe only $\{X_1, \dots, X_n\}$ where the data-points have been randomly right-censored and we also observe whether or not a specific variable was censored:

$$X_k = \min\{\xi_k, C_k\}, \quad e_k = 1\{\xi_k = X_k\}, \quad k = 1, \dots, n,$$

where $\{C_1, \dots, C_n\}$ is a censoring mechanism that is independent of the ξ_k . We will be especially concerned with the classification of Pareto distributed samples, but the theory will be presented in a general point process approach, and the methodology used ensures that deviations from the Pareto family can be sensibly classified using the same test. We will modify an approach found in Luca et al. (2018) to fit our current goal and develop transparent and well-motivated methods which use a survival analysis likelihood or a Bayesian-like posterior, respectively, instead of Janossy densities. Such classifiers fall into the category of parametric probabilistic detection methods.

Notice that existing novelty detection methods could be applied to the X_k or to the binary variables e_i directly, but there are cases when both ξ_k and C_k change in a manner that the X_k remain fairly similar (or even identical) to the case where there is no change, and then the e_i yield additional information and hence play an

important role. The same comment is true if we exchange the roles of X_k and e_k . Hence, if we stick to non-censored data methods, the novelty detection for each of the two samples $\{\xi_1, \xi_2, \dots, \xi_n\}$ and $\{e_1, e_2, \dots, e_n\}$ has to be obtained and suitably combined. In the present work we will use the classical maximum-likelihood classification method for this bivariate problem as a benchmark, combined using the Bonferroni correction (cf. Hochberg and Tamhane (1987)). The latter correction is unreliable for doing multiple hypothesis testing when data are dependent and the number of tests considered is large, neither of which will pose an issue in what follows. Building on the latter method we also show on synthetic data how the inclusion of information on the number of peaks over thresholds can lead to a drastic improvement on the power of such a double test.

Novelty detection or one-class classification for samples that have been randomly censored has, to the best of the author's knowledge, not been considered before. In Eo et al. (2014) some algorithms for recognising an outlier in a randomly censored sample are considered, by means of quantile regression. For non-censored samples, extreme value methods for detecting outliers can be found in Chapter 2 of Aggarwal (2016), where the focus is on multivariate Gaussian samples, and the Mahalanobis distance plays an important role in the incorporation of the inter-vector covariance matrix.

The main tool that facilitates the analysis of a censored sample for low density areas will be the point process of exceedances (cf. Embrechts et al. (2013)). Such point processes are used in Extreme Value Theory (EVT) for modelling extreme events, such as natural disasters, accidents, or the related insurance claims. This kind of data is often prone to be right-censored. A classical problem in EVT deals with the estimation of the tail index, cf. Beirlant et al. (2004) or Embrechts et al. (2013) for an overview, and Bladt et al. (2020) for a recently proposed estimation method for when external information is available for the right-censored data-points. EVT methods have also been used previously for classifying non-censored data. Specifically, in Clifton et al. (2011, 2013, 2014) one-sample classification methods are derived, where only the most extreme point of the small sample is used to make a decision. In Luca et al. (2014), the Poisson point process model is used in novelty detection to introduce "anomaly scores" for number of exceedances and size of exceedances separately and then combined with a third score. In Luca et al. (2016) the Poisson approximation was again used in a more integrated way, where also an analysis on the mean and maximal exceedances is jointly taken into account. In Section 5.4 of Luca et al. (2018) a detection test is derived which uses jointly the size and number of exceedances.

The remainder of the paper is organised as follows. In Section 8.2 we lay out the formulation of the classification problem. In Section 8.3 we give the background needed from point processes and EVT needed for part of the main result, which is then given in Section 8.4, together with two additional methods. A performance study on synthetic data is given in Section 8.5, before concluding in Section 8.6.

8.2 Problem Formulation

We will consider the problem of determining whether an independent and identically distributed (i.i.d.) randomly censored sample $X_k = \min\{\xi_k, C_k\}$, $e_k = 1\{\xi_k = X_k\}$, $k = 1, \dots, n$ comes from the same distribution as a large and well-understood sample. That is to say if

$$\mathbb{P}(\xi_1 \leq x) = F_\xi(x), \quad \mathbb{P}(C_1 \leq x) = F_C(x), \quad x \geq 0$$

are the true cumulative distribution functions of the target sample and the censoring mechanism, respectively, we are interested in devising a test for the following statistical hypothesis:

$$\begin{aligned} H_0 : \mathbb{P}(X_k \leq x) = F_X(x) &= 1 - \overline{F}_\xi(x)\overline{F}_C(x) \\ H_1 : \mathbb{P}(X_k \leq x) = F_X(x) &\neq 1 - \overline{F}_\xi(x)\overline{F}_C(x). \end{aligned} \tag{8.1}$$

Observe that we make the assumption that we know the (large-sample) distribution of both ξ_1 and C_1 , and the resulting test will also be sensitive to changes on the censoring mechanism. The methods developed below also apply without change to the case where the large-sample distribution of ξ_1 is known whereas the small-sample distribution of C_1 is also known, by simply incorporating the latter into H_0 through F_C .

In terms of novelty detection, it is not classically assumed we know a distribution but instead we have a large and well-sampled dataset, so F_ξ and F_C should be replaced with their estimated counterparts, say \widehat{F}_ξ and \widehat{F}_C , respectively, leading to the statistical hypothesis which is common in the novelty detection literature

$$\begin{aligned} H_0 : \mathbb{P}(X_k \leq x) = F_X(x) &= 1 - \widehat{\overline{F}}_\xi(x)\widehat{\overline{F}}_C(x) \\ H_1 : \mathbb{P}(X_k \leq x) = F_X(x) &\neq 1 - \widehat{\overline{F}}_\xi(x)\widehat{\overline{F}}_C(x). \end{aligned}$$

Estimating \widehat{F}_ξ , \widehat{F}_C corresponds to training, and the determination of the above statistical hypotheses corresponds to testing, in machine learning terminology. The difference between the two hypothesis results in very small discrepancies of the testing procedure if the training sample is large, as is the case in the area of novelty detection, but care must be taken if this is not the case, and then a two-sample classification method might be more appropriate. In the remainder of the paper we will study the testing part of the procedure. Namely, we focus on testing the hypothesis (8.1).

8.3 Extreme Value Theory

One of the tools that has proven to be very useful in the analysis of novelty detection for heavy-tailed data in low-density regions is EVT, cf. Clifton et al. (2011, 2013, 2014) and a point process approximation approach in Luca et al. (2014, 2016,

2018) has been evolving in an ever more integrated fashion. In this section we introduce the main notation, concepts and limits which motivate the approximations that we will crucially need at a later stage when devising the novelty detection test. We present first selected classical EVT results and then proceed to introduce point processes and point processes of exceedances, which will play a central role when choosing a parametric form of the test. Results here are stated without any additional effort for extreme value distributions in general, although later we will concentrate exclusively on Pareto and Gumbel laws.

8.3.1 Domains of attraction and GPD

Given an i.i.d. sequence of random variables X_1, X_2, \dots , we say that they follow a max-stable distribution if the following relation holds

$$c_n^{-1}(M_n - d_n) \stackrel{d}{=} X_1$$

for some norming constants c_n, d_n and where $M_n = \max\{X_1, \dots, X_n\}$. The class of max-stable distributions is important because of its connection with limiting distributions of maxima. More precisely, assume that $\mathbb{P}(M_n \leq c_n x + d_n) \rightarrow H(x)$ (non-degenerate). Then also $\mathbb{P}(M_{nk} \leq c_{nk} x + d_{nk}) \rightarrow H(x)$, while on the other hand $\mathbb{P}(M_{nk} \leq c_n x + d_n) \rightarrow H^k(x)$. By the convergence to types theorem (cf. Embrechts et al. (2013)), if $Y_i \stackrel{i.i.d.}{\sim} H$,

$$\max\{Y_1, \dots, Y_k\} \stackrel{d}{=} \tilde{c}_k Y_1 + \tilde{d}_k, \quad \tilde{c}_k = \lim_n \frac{c_{nk}}{c_n}, \quad \tilde{d}_k = \lim_n \frac{d_{nk} - d_n}{c_n}.$$

Conversely, it is clear that any max-stable sequence has a maximum which has a limiting distribution. This characterisation of max-stable laws goes even further, since the following theorem (cf. Fisher and Tippett (1928)) guarantees a fully explicit way of writing the associated limiting distribution:

Theorem 8.3.1. *For X_i i.i.d., if there exist non-degenerate H and norming constants such that $c_n^{-1}(M_n - d_n) \xrightarrow{d} H$, then necessarily H is one of the following:*

Frechet: $\Phi_\alpha(x) = e^{-x^{-\alpha}} 1_{(0, \infty)}(x),$

Weibull: $\Psi_\alpha(x) = e^{-(-x)^\alpha} 1_{(-\infty, 0]}(x) + 1_{(0, \infty)}(x)$

Gumbel: $\Lambda(x) = e^{-e^{-x}}.$

An important remark is that, although distributionally different, mathematically the variables satisfy

$$X \sim \Phi_\alpha \Leftrightarrow -X^{-1} \sim \Psi_\alpha \Leftrightarrow \log X^\alpha \sim \Lambda.$$

This relation is important when transforming data. The above distributions are called the extreme value distributions. It follows that they are max-stable with norming constants $c_n = n^{1/\alpha}, n^{-1/\alpha}, 1$ and $d_n = 0, 0, \ln n$. For instance, in the exponential case, $\mathbb{P}(M_n - \ln n \leq x) = (1 - n^{-1}e^{-x})^n \rightarrow \Lambda(x)$ and in the Cauchy case,

by L'Hopital $\bar{F}(x) \sim \frac{1}{\pi x}$, so $\mathbb{P}(M_n \leq nx/\pi) = (1 - \bar{F}(nx/\pi))^n \rightarrow e^{-x^{-1}} = \Phi_1(x)$. In general, it is easier to verify the membership to a so-called domain of attraction via the following result.

Say that a distribution function $F \in \mathcal{R}_{-\alpha}$, that is, it is regularly varying with index α if $\bar{F}(x) = x^{-\alpha}\ell(x)$ where $\lim_{x \rightarrow \infty} \ell(tx)/\ell(x) = 1$, for all $t > 0$. The mean excess function is defined as

$$s_X(u) = \mathbb{E}(X - u | X > u) = \frac{\int_u^\infty \bar{F}(y) dy}{\bar{F}(u)}.$$

Theorem 8.3.2. (*Characterisation*)

- $F \in MDA(\Phi_\alpha) \Leftrightarrow \bar{F} \in \mathcal{R}_{-\alpha}$. The norming constants can be taken as $d_n = 0$ and $c_n = F^{\leftarrow}(1 - n^{-1})$.
- $F \in MDA(\Psi_\alpha) \Leftrightarrow x_F < \infty$ and $\bar{F}(x_F - x^{-1}) \in \mathcal{R}_{-\alpha}$. The norming constants can be taken as $d_n = x_F$ and $c_n = x_F - F^{\leftarrow}(1 - n^{-1})$.
- $F \in MDA(\Lambda) \Leftrightarrow \bar{F}(x) = c(x) \exp\left(-\int_z^x \frac{g(t)}{a(t)} dt\right)$, with $g \rightarrow 1, c \rightarrow c_0 > 0, a' \rightarrow 0$. The norming constants can be taken as $d_n = F^{\leftarrow}(1 - n^{-1})$ and $c_n = a(d_n)$. One can choose a to be the mean excess function.

The extreme value distributions can be parametrised for simplicity in the so-called Generalised Extreme Value distribution (GEV):

$$H_\xi(x) = \exp(-(1 + \xi x)^{-1/\xi}), \quad 1 + \xi x > 0,$$

corresponding to Frechet $\xi = \alpha^{-1} > 0$, Weibull $\xi = -\alpha^{-1} < 0$ and Gumbel $\xi = 0$. The following result links the distribution functions of normalised maxima and of the normalised exceedances over a high threshold, cf. Balkema and De Haan (1974); Pickands III et al. (1975):

Theorem 8.3.3. *The following are equivalent:*

- a) $F \in MDA(H_\xi)$
- b) *There exists a positive function a such that for $1 + \xi x > 0$,*

$$\lim_{u \uparrow x_F} \frac{\bar{F}(u + xa(u))}{\bar{F}(u)} = -\log H_\xi(x).$$

An interesting discussion arises from the preceding theorem as follows. The limit in b) can be interpreted as the limit distribution of the excess over a high threshold

$$\lim_{u \uparrow x_F} \mathbb{P}((X - u)/a(u) > x | X > u) = -\log H_\xi(x) =: \bar{G}_\xi(x), \quad (8.2)$$

which we define as the Generalised Pareto distribution (GPD). Again, in the Gumbel case we can take $a(u) = s_X(u)$. This suggests that a good approximation for

the normalised excess distribution or the distribution of the so-called *peaks over thresholds* (POT) is given by the corresponding GPD. In fact, $F_u(x) = \mathbb{P}(X - u \leq x | X > u)$ can be approximated almost uniformly by $G_{\xi, \beta(u)} := G_{\xi}(\cdot / \beta(u))$, for some positive function β in the sense that $F \in \text{MDA}(H_{\xi})$ if and only if

$$\lim_{u \uparrow x_F} \sup_{0 < x < x_F - u} |F_u(x) - G_{\xi, \beta(u)}(x)| = 0. \quad (8.3)$$

In conclusion, we have found the asymptotic behaviour for the excess distribution of i.i.d. data over increasingly high thresholds.

In practice, to check when this approximation is reasonable, for $\beta(u) \equiv \beta$, one has to choose u as to make the empirical mean excess function

$$e_n(u) = \frac{\int_u^{\infty} \bar{F}_n(y) dy}{\bar{F}_n(u)} = \frac{\sum_{i \in \Delta_n(u)} (X_i - u)}{\text{card} \Delta_n(u)}, \quad \Delta_n(u) = \{i : X_i > u\}$$

approximately linear, since the mean excess function of a GPD is given by

$$\frac{\beta + u\xi}{1 - \xi}, \quad \beta(u) + u\xi > 0.$$

8.3.2 Point Processes

For an i.i.d. sequence X_1, X_2, \dots , with distribution F , thresholds (u_n) and $\lambda \in (0, \infty)$, it holds that $n\bar{F}(u_n) \rightarrow \lambda$ if and only if $B_n = \sum_{i=1}^n 1_{X_i > u_n} \sim \text{Bin}(n, \bar{F}(u_n))$ converges to a $\text{Pois}(\lambda)$ variable Y . Indeed, the Laplace transform of the binomial variable satisfies the following limit

$$\Psi_{B_n}(u) = (F(u_n) + \bar{F}(u_n)e^{-u})^n = \left(1 - \frac{n\bar{F}(u_n)}{n}(1 - e^{-u})\right)^n \quad (8.4)$$

$$\rightarrow \exp\{-\lambda(1 - e^{-u})\} = \Psi_Y(u) \quad (8.5)$$

if and only if $n\bar{F}(u_n) \rightarrow \lambda$. The point process approach to extreme value theory links this Poisson limit result with the GPD limit in (8.3) in a unified way.

The notion of a point process is that of a measurable map $N : \Omega \rightarrow (M_p(E), \mathcal{M}_p(E))$ where $M_p(E)$ is the space of all locally finite counting measures on the space E . The sigma field $\mathcal{M}_p(E)$ is the one which makes all projections measurable. In short, a point process is a random variable taking values in a function space. The functions of the latter space are used for counting.

The structure $N = \sum_{i=1}^{\infty} \epsilon_{X_i}$ for random variables (X_i) in $E \subset \mathbb{R}$, where $\epsilon_{X_i}(A) := 1\{X_i \in A\}$ for all measurable $A \subset \mathbb{R}$, defines a point process, which counts for each $\omega \in \Omega$ the number of realisations $X_1(\omega), X_2(\omega), \dots$ falling into the set A , the resulting count being $N(A)(\omega)$. The mean measure is defined as $\mu(\cdot) = \mathbb{E}(N(\cdot))$. We require $\mu(K) < \infty$ for any compact $K \subset E$, so then $\mathbb{E}(N(K)) < \infty$ and by non-negativity of N , even $N(K) < \infty$ a.s., so that N is indeed a point process.

Of special importance is the Poisson Point Process $\text{PPP}(\mu)$, defined on the space E as a point process N such that $N(A)$ is distributed $\text{Pois}(\mu(A))$ and for disjoint A_1, \dots, A_k , the variables $N(A_1), \dots, N(A_k)$ are independent. We observe that, analogously to the Poisson random variable, a Poisson point process is characterised by the mean measure μ . In general, the distribution of a point process $N = \sum_{i=1}^{\infty} \epsilon_{X_i}$ can be characterised by the use of the Laplace functional, defined as

$$\Psi_N(g) = \mathbb{E} \left(\exp \left\{ - \int g(x) dN(x) \right\} \right) = \mathbb{E}(\exp \{ - \sum_{i=1}^{\infty} g(X_i) \}),$$

for every non-negative, bounded and measurable function g . We now define the Point Process of exceedances for a sequence of i.i.d. max-stable random variables X_1, X_2, \dots, X_n with norming constants c_n, d_n as

$$N_n(A) = \sum_{i=1}^n \epsilon_{\left(\frac{X_i - d_n}{c_n}, \frac{i}{n+1}\right)}(A), \quad (8.6)$$

for every measurable $A \subset (u, \infty) \times (0, 1)$. Using a Laplace functional version of the convergence in (8.4), one can show that weak convergence holds (cf. Mikosch (2009); Embrechts et al. (2013)) from the point process (8.6) to a PPP with mean measure given by

$$\mu(A) = (t - s) \overline{G}_{\xi}(x), \quad A = (x, \infty) \times (s, t) \subset (u, \infty) \times (0, 1) \quad (8.7)$$

whenever $n\overline{F}(d_n + c_n u) \rightarrow \overline{G}_{\xi}(u) =: \lambda$.

In particular, we can translate the limits (8.2) and (8.4) into the language of point processes by doing the following calculations:

$$\begin{aligned} \mathbb{P}((X_1 - d_n)/c_n > x | X_1 > d_n) &= n \mathbb{P}((X_1 - d_n)/c_n > x) \\ &= \mathbb{E}(N_n((x, \infty) \times (0, 1))) \\ &\rightarrow \mu((x, \infty) \times (0, 1)) = \overline{G}_{\xi}(x), \end{aligned} \quad (8.8)$$

where we have used that $\mathbb{P}(X_1 > d_n) = \mathbb{P}(X_1 > F^{\leftarrow}(1 - n^{-1})) = n^{-1}$, by Theorem 8.3.2, and

$$\mathbb{P}(N_n((u, \infty) \times (0, 1)) = k) \rightarrow \frac{\lambda^k}{k!} e^{-\lambda}, \quad (8.9)$$

respectively. Notice that, apart from characterisations, we did not build upon previous results, meaning that point processes provide alternative ways of proving EVT limits, in a uniquely transparent fashion.

8.4 Novelty detection for randomly censored data

We focus on the classification derived from a censored extreme-value sample

$$X = \{X_1, X_2, \dots, X_n\}, \quad e = \{e_1, e_2, \dots, e_n\}, \quad n \in \mathbb{N},$$

where X_1, X_2, \dots , is an independent and identically distributed (i.i.d.) sample with X_1 having a law in the Fréchet domain of attraction, e_1, e_2, \dots , are binary indicators taking the value 1 if the observation is non-censored and 0 otherwise. The archetype of such a domain of attraction is the Pareto distribution and will be our main focus when investigating the performance of the test. However, the increased generality with which we will present the results comes with little additional effort. Further, we assume random censoring from two distributions from the Fréchet domain of attraction is in place. Concretely, by the regularly varying tail characterization we have that if

$$X_i = \min\{\xi_i, C_i\}, \quad i = 1, 2, \dots,$$

with ξ_i being the target i.i.d. sample to estimate and C_i being an independent i.i.d. censoring sequence, with respective tails given by

$$\mathbb{P}(\xi_1 > x) = x^{-\alpha} \ell_\xi(x), \quad \mathbb{P}(C_1 > x) = x^{-c} \ell_C(x), \quad \alpha, c, x > 0,$$

where ℓ_ξ, ℓ_C are slowly varying functions, then

$$\mathbb{P}(X_1 > x) = x^{-(\alpha+c)} \ell_X(x), \quad x > 0,$$

with ℓ_X again slowly varying. We also have that the censoring indicator can be written as

$$e_1 = 1\{X_1 = \xi_1\}.$$

In the remainder of this section we will study a point process of exceedances that will be useful to determine the asymptotic law of an approximate survival-analysis likelihood. We then proceed to introduce two main methods that can be used for novelty detection for censored data. We also examine some tests based on the Bonferroni correction, both as a benchmark and as a third main method.

8.4.1 Point Process of Exceedances

Assume that a density $f_\xi(x)$ of ξ_1 exists and denote by $\bar{F}_\xi(x) = \int_x^\infty f_\xi(z) dz$.

Further, we require that the negative logarithm of the densities and tails:

$$Z_i = \log(\alpha) - \log f_\xi(X_i), \quad W_i = -\log \bar{F}_\xi(X_i), \quad i = 1, 2, \dots, n,$$

are themselves in the Gumbel domain of attraction. Let their associated cumulative distribution functions be denoted by F_Z and F_W , respectively.

We proceed to introduce a point process of exceedances for the variables Z_i and W_i , $i = 1, \dots$ which will be useful in the next subsection. Define for the respective norming constants a_n, b_n and c_n, d_n , the process

$$N_n(A) = \sum_{i=1}^n \epsilon_{(e_i(\frac{Z_i - b_n}{a_n}) + (1 - e_i)(\frac{W_i - d_n}{c_n}), \frac{i}{n+1})}(A), \quad (8.10)$$

for every measurable $A \subset (0, \infty) \times (0, 1)$. In the exact Pareto case we have that Z_i, W_i and e are independent. For the regularly varying case, we will consider

only those cases where the excesses $Z_i - b_n | Z_i > b_n$ and $W_i - d_n | W_i > d_n$ are asymptotically independent of e , and where furthermore (Einmahl et al. (2008))

$$\lim_{n \rightarrow \infty} \mathbb{P}(e_i = 1 | X_i > n) \rightarrow p = \frac{\alpha}{\alpha + c}. \quad (8.11)$$

We have the following Poisson convergence.

Theorem 8.4.1. *The point process of exceedances (8.10) converges weakly to a PPP, say N , with mean measure given by*

$$\mathbb{E}(N(A)) = \mu(A) = (t - s)e^{-x}, \quad A = (x, \infty) \times (s, t) \subset (0, \infty) \times (0, 1) \quad (8.12)$$

if and only if for $u > 0$

$$\lim_{n \rightarrow \infty} n\bar{F}_Z(b_n + a_n u) = \lim_{n \rightarrow \infty} n\bar{F}_W(d_n + c_n u) = e^{-u}.$$

Proof. We use the Laplace functional convergence. To this end we calculate for any measurable, non-negative and bounded function of the type $f(x, y) = 1\{x > z\}h(y)$,

$$\begin{aligned} & \Psi_{N_n}(f) \\ &= \mathbb{E}(\exp\{-\sum_{i=1}^n 1\{e_i(Z_i - b_n)/a_n + (1 - e_i)(W_i - d_n)/c_n > z\}h(i/(n + 1))\}) \\ &= \prod_{i=1}^n \mathbb{E}(1 - 1\{e_i(Z_i - b_n)/a_n + (1 - e_i)(W_i - d_n)/c_n > z\} \\ &\quad \times (1 - \exp\{-h(i/(n + 1))\})) \\ &= \prod_{i=1}^n (1 - \mathbb{P}(e_i(Z_i - b_n)/a_n + (1 - e_i)(W_i - d_n)/c_n > z) \\ &\quad \times (1 - \exp\{-h(i/(n + 1))\})) \\ &\sim \prod_{i=1}^n \exp\{-n \mathbb{P}(e_i(Z_i - b_n)/a_n + (1 - e_i)(W_i - d_n)/c_n > z) \\ &\quad \times (1 - \exp\{-h(i/(n + 1))\})/n\} \\ &\rightarrow \exp\{-(pe^{-z} + (1 - p)e^{-z}) \int_0^1 (1 - e^{-h(y)}) dy\} \\ &= \exp\{-\int_0^\infty \int_0^1 (1 - e^{-1\{x > z\}h(y)}) dy e^{-x} dx\} = \Psi_N(f). \end{aligned}$$

which holds if and only if for $u > 0$ we have $\bar{F}_Z(b_n + a_n u), n\bar{F}_Z(d_n + c_n u) \rightarrow e^{-u}$. This concludes the proof for such f . For general f it suffices to use the standard approximation by functions of the aforementioned type and an application of the dominated convergence theorem (cf. Kallenberg (2006)). \square

The previous enables us to obtain some quantities of interest with little effort, analogous to (8.2) and (8.4) of the classical case.

Corollary 8.4.2. *Let $\lim_{n \rightarrow \infty} n\bar{F}_Z(a_n x + b_n) = \lim_{n \rightarrow \infty} n\bar{F}_W(c_n x + d_n) = e^{-x} =: \lambda(x)$. Define*

$$M(x, n) = \mathbb{P}(e_1(Z_1 - b_n)/a_n + (1 - e_1)(W_1 - d_n)/c_n > x | e_1(Z_1 - b_n) \quad (8.13)$$

$$+ (1 - e_1)(W_1 - d_n) > 0), \quad (8.14)$$

$$K(x, n) = \sum_{i=1}^n 1\{e_i(Z_i - b_n)/a_n + (1 - e_i)(W_i - d_n)/c_n > x\}, \quad n \in \mathbb{N} \quad (8.15)$$

Then

$$\lim_{n \rightarrow \infty} M(x, n) = \lambda(x), \quad x \geq 0.$$

and

$$\lim_{n \rightarrow \infty} \mathbb{P}(K(x, n) = k) = \frac{\lambda(x)^k}{k!} e^{-\lambda(x)}, \quad k \geq 0$$

Proof. For the second identity simply observe that since the following equality holds

$$K(x, n) = N_n((x, \infty) \times (0, 1))$$

we can use the convergence

$$\mathbb{P}(N_n((x, \infty) \times (0, 1)) = k) \rightarrow \frac{\lambda(x)^k}{k!} e^{-\lambda(x)},$$

by definition of PPP.

For the first identity we note first that by Theorem 8.3.2.

$$\begin{aligned} & \mathbb{P}(e_1(Z_1 - b_n) + (1 - e_1)(W_1 - d_n) > 0) \\ & \sim p \mathbb{P}(Z_1 > F_Z^{\leftarrow}(1 - n^{-1})) + (1 - p) \mathbb{P}(W_1 > F_W^{\leftarrow}(1 - n^{-1})) = n^{-1}. \end{aligned}$$

which implies $M(x, n) \sim \mathbb{E}(K(x, n))$ and we may apply the previous result, since the mean of a Poisson variable is the rate $\lambda(x)$. Alternatively, one can directly calculate:

$$\begin{aligned} & \mathbb{P}(e_i(Z_i - b_n)/a_n + (1 - e_i)(W_i - d_n)/c_n > x | e_i(Z_i - b_n) + (1 - e_i)(W_i - d_n) > 0) \\ & \sim n \mathbb{P}(e_i(Z_i - b_n)/a_n + (1 - e_i)(W_i - d_n)/c_n > x) \\ & = \mathbb{E}(N_n((x, \infty) \times (0, 1))) \\ & \rightarrow \mu((x, \infty) \times (0, 1)) = e^{-x}. \end{aligned}$$

□

8.4.2 Law of the likelihood

We define

$$K_0(u, v) = \sum_{i=1}^n 1\{e_i(Z_i - u)/s_Z(u) + (1 - e_i)(W_i - v)/s_W(v) > 0\},$$

and to simplify notation assume that the observations which contribute to $K_0(u, v)$ are precisely the first $K_0(u, v)$ of the n data points. That is,

$$K_0(u, v) = \sum_{i=1}^{K_0(u, v)} 1\{e_i(Z_i - u)/s_Z(u) + (1 - e_i)(W_i - v)/s_W(v) > 0\},$$

which for varying n can be achieved by re-labeling. Here

$$s_Z(u) = \mathbb{E}(Z_1 - u | Z_1 > u), \quad s_W(v) = \mathbb{E}(W_1 - v | W_1 > v)$$

are the mean excess functions of the transformed variables.

From Corollary 8.4.2 we know that the exceedances $Z_i - u | Z_i > u$ and $W_i - v | W_i > v$ are for large thresholds distributed according to exponential distributions with respective means given by $s_Z(u)$ and $s_W(v)$. We hence consider, conditionally on $K_0(u, v) = k$, the asymptotic likelihood of the transformed exceedances, defined as the random variable

$$\begin{aligned} L_k(X, e, u, v) \\ = \prod_{i=1}^k \exp\left(-e_i \left[\frac{Z_i - u}{s_Z(u)} + \log s_Z(u)\right]\right) \exp\left(-(1 - e_i) \left[\frac{W_i - v}{s_W(v)} + \log s_W(v)\right]\right), \end{aligned}$$

for $1 \leq k \leq n$ and $L_0(X, e, u, v) = 1$. Let $p(k) := \mathbb{P}(K_0(u, v) = k)$, $0 \leq k \leq n$. If n is large and we choose u, v , such that $n\bar{F}_Z(u) = n\bar{F}_W(v) = \lambda > 0$, we have by Corollary 8.4.2 that

$$p(k) \rightarrow \frac{\lambda^k}{k!} e^{-\lambda}.$$

Theorem 8.4.3. *As $u = u(n), v = v(n), n \rightarrow \infty$ such that $n\bar{F}_Z(u), n\bar{F}_W(v) \rightarrow \lambda > 0$, the asymptotic cumulative distribution function of the random variable $L_{K_0(u, v)}(X, e, u, v)$ is given by*

$$G(l) = e^{-\lambda} 1\{1 \leq l\} + \sum_{k=1}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} F_k,$$

where

$$\sum_{i=0}^{k \wedge \kappa} \binom{k}{i} p^i (1 - p)^{k-i} \frac{\gamma_u(k, -\log(s_W(v)^{k-l}) - i \log(s_Z(u)/s_W(v)))}{(k-1)!} \rightarrow F_k.$$

$\kappa = 1\{s_Z(u) > s_W(v)\} \frac{-\log(s_W(v)^{k-l})}{\log(s_Z(u)/s_W(v))} + k 1\{s_Z(u) \leq s_W(v)\}$, $p = \alpha/(\alpha + c)$ and $\gamma_u(\cdot, \cdot)$ is the upper incomplete gamma function.

Proof. Note first that for large u, v, n , the normalised exceedances $(Z_i - u)/s_Z(u)$ and $(W_i - v)/s_W(v)$ are asymptotically standard exponentially distributed, $X \perp e$, and the probability of non-censoring (8.11) stabilises at $p = \alpha/(\alpha + c)$.

We first determine the law of $A_k = -\log(L_k(X, e, u, v))$ for deterministic k . We write

$$\begin{aligned} A_k &= \sum_{i=1}^k e_i \left[\frac{Z_i - u}{s_Z(u)} + \log s_Z(u) \right] + (1 - e_i) \left[\frac{W_i - v}{s_W(v)} + \log s_W(v) \right] \\ &= \sum_{i=1}^k \left(e_i \left[\frac{Z_i - u}{s_Z(u)} \right] + (1 - e_i) \left[\frac{W_i - v}{s_W(v)} \right] \right) \\ &\quad + \log(s_Z(u)/s_W(v)) \sum_{i=1}^k e_i + \log(s_W(v))k \end{aligned}$$

and observe that the first sum has an Erlang($k, 1$) distribution. For the second sum, we have a scaled Binomial(k, p) random variable. We hence get that for $s_Z(u) > s_W(v)$

$$\begin{aligned} &\mathbb{P} \left(\frac{A_k - \log(s_W(v))k}{\log(s_Z(u)/s_W(v))} < x \right) \\ &\sim \sum_{i=0}^{x \wedge k} \binom{k}{i} p^i (1-p)^{k-i} \frac{\gamma_l(k, (x-i) \log(s_Z(u)/s_W(v)))}{(k-1)!}, \end{aligned}$$

where $\gamma_l(\cdot, \cdot)$ denotes the lower incomplete gamma function. Now using

$$\begin{aligned} &\frac{A_k - \log(s_W(v))k}{\log(s_Z(u)/s_W(v))} < x \\ &\Leftrightarrow L_k(X, e, u, v) > \exp(-\log(s_W(v))k - x \log(s_Z(u)/s_W(v))) \end{aligned}$$

we get

$$\begin{aligned} &\mathbb{P}(L_k(X, e, u, v) \leq l) \sim \\ &k \wedge \frac{-\log(s_W(v)kl)}{\log(s_Z(u)/s_W(v))} \sum_{i=0}^k \binom{k}{i} p^i (1-p)^{k-i} \frac{\gamma_u(k, -\log(s_W(v)kl) - i \log(s_Z(u)/s_W(v)))}{(k-1)!}. \end{aligned}$$

If $s_Z(u) < s_W(v)$ we have

$$\begin{aligned} &\mathbb{P} \left(\frac{A_k - \log(s_W(v))k}{-\log(s_Z(u)/s_W(v))} < x \right) \\ &\sim \sum_{i=0}^k \binom{k}{i} p^i (1-p)^{k-i} \frac{\gamma_l(k, (x+i) \log(s_W(v)/s_Z(u)))}{(k-1)!}, \end{aligned}$$

and

$$\begin{aligned} &\frac{A_k - \log(s_W(v))k}{-\log(s_Z(u)/s_W(v))} < x \\ &\Leftrightarrow L_k(X, e, u, v) < \exp(-\log(s_W(v))k - x \log(s_W(v)/s_Z(u))) \end{aligned}$$

so we get

$$\begin{aligned} & \mathbb{P}(L_k(X, e, u, v) \leq l) \\ & \sim \sum_{i=0}^k \binom{k}{i} p^i (1-p)^{k-i} \frac{\gamma_u(k, -\log(s_W(v)^{kl}) - i \log(s_Z(u)/s_W(v)))}{(k-1)!}. \end{aligned}$$

The formula can also be verified to hold for when the mean excess functions agree. Finally, by conditioning and using the Poisson process exceedance probabilities, the main result now follows. \square

Observe that within the last theorem there is an implicit check not only of the distribution of the exceedances, but also of the number of such exceedances, according to the Poisson number of POTs approximation.

Remark 8.4.4. (Method 1) In practice, for a finite sample size n , it is natural to consider the quantity

$$\begin{aligned} \mathcal{G}(l) & := e^{-\lambda} \mathbf{1}\{1 \leq l\} \\ & + \sum_{k=1}^n \frac{\lambda^k}{k!} e^{-\lambda} \sum_{i=0}^{k \wedge \kappa} \binom{k}{i} p^i (1-p)^{k-i} \frac{\gamma_u(k, -\log(s_W(v)^{kl}) - i \log(s_Z(u)/s_W(v)))}{(k-1)!} \end{aligned}$$

as the determining factor to check whether a sample is novel or not: if it is above or below $1 - \alpha/2$ or $\alpha/2$, respectively, we say that the sample is novel, or reject the null hypothesis. For instance, the arbitrary $\alpha = 0.05, 0.01$ agree with most statistical significances used in practice. In Luca et al. (2018) the test for heavy-tails is rejected only when $\mathcal{G}(l) < \alpha$, and while this works well when a novel tail is heavier than the reference one, empirical tests show that for an arbitrary change in the tail, $\mathcal{G} > 1 - \alpha/2$ or $\mathcal{G} < \alpha/2$ is superior.

In contrast, the results in Luca et al. (2018) could be adapted to the sample X_1, \dots, X_n without the use of the e_k . The rationale is as follows. Define

$$\tilde{Z}_i = \log(\alpha + c) - \log(f_X(X_i)), \quad i = 1, \dots, n, \quad (8.16)$$

and

$$\tilde{K}_0(u) = \sum_{i=1}^n \mathbf{1}\{(\tilde{Z}_i - u)/s_{\tilde{Z}}(u) > 0\}, \quad (8.17)$$

and let the law of the exceedances $\tilde{Z}_i - u | \tilde{Z}_i > u$ for large thresholds u converge to an exponential distribution with respective mean given by $s_{\tilde{Z}}(u)$. We hence, conditionally on $K_0(u, v) = k$, could consider

$$\prod_{i=1}^k \exp\left(-\left[\frac{\tilde{Z}_i - u}{s_{\tilde{Z}}(u)} + \log s_{\tilde{Z}}(u)\right]\right).$$

In order to accommodate for the (asymptotically independent) information brought by the e_i , we instead consider the joint asymptotic likelihood of the normalised exceedances and the censoring indicators:

$$\tilde{L}_k(X, e, u) = \prod_{i=1}^k \left(\frac{\alpha}{\alpha + c} \right)^{e_i} \left(\frac{c}{\alpha + c} \right)^{1-e_i} \exp \left(-\frac{\tilde{Z}_i - u}{s_{\tilde{Z}}(u)} \right),$$

for $1 \leq k \leq n$ and $\tilde{L}_0(X, e, u) = 1$. Let $\tilde{p}(k) := \mathbb{P}(\tilde{K}_0(u) = k)$, $0 \leq k \leq n$. If n is large and we choose u , such that $n\bar{F}_{\tilde{Z}}(u) = \lambda > 0$, we have by (8.9) that

$$\tilde{p}(k) \rightarrow \frac{\lambda^k}{k!} e^{-\lambda}.$$

Theorem 8.4.5. *As $u = u(n)$, $n \rightarrow \infty$ such that $n\bar{F}_{\tilde{Z}}(u) \rightarrow \lambda > 0$, the asymptotic cumulative distribution function of the random variable $\tilde{L}_{\tilde{K}_0(u)}(X, e, u)$ is given by*

$$\tilde{G}(l) = e^{-\lambda} 1\{1 \leq l\} + \sum_{k=1}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} \tilde{F}_k$$

where

$$\tilde{F}_k = \sum_{i=0}^{k \wedge \kappa(p)} \binom{k}{i} p^i (1-p)^{k-i} \frac{\gamma_u(k, \log((1-p)^k/l) + i \log(p/(1-p)))}{(k-1)!},$$

$\kappa(p) = 1\{p < 1/2\} \frac{k \log((1-p)^k/l)}{\log((1-p)/p)} + k1\{p \geq 1/2\}$, $p = \alpha/(\alpha + c)$ and $\gamma_u(\cdot, \cdot)$ is the upper incomplete gamma function.

Proof. As noted above, for large u, v, n , the normalised exceedances $(\tilde{Z}_i - u)/s_{\tilde{Z}}(u)$ are asymptotically standard exponentially distributed, $X \perp e$, and the probability of non-censoring (8.11) stabilises at $p = \alpha/(\alpha + c)$.

We define $\tilde{A}_k = -\log(\tilde{L}_k(X, e, u))$ for deterministic k . We have

$$\begin{aligned} \tilde{A}_k &= \sum_{i=1}^k \frac{\tilde{Z}_i - u}{s_{\tilde{Z}}(u)} - \log(p) \sum_{i=1}^k e_i - \log(1-p) \sum_{i=1}^k (1-e_i) \\ &= \sum_{i=1}^k \frac{\tilde{Z}_i - u}{s_{\tilde{Z}}(u)} - \log(p/(1-p)) \sum_{i=1}^k e_i - \log(1-p)k. \end{aligned}$$

The first sum has an Erlang($k, 1$) distribution. For the second sum, we have a scaled Binomial(k, p) random variable. We hence get that for $p < 1/2$

$$\begin{aligned} &\mathbb{P} \left(\frac{\tilde{A}_k + \log(1-p)k}{\log((1-p)/p)} < x \right) \\ &\sim \sum_{i=0}^{x \wedge k} \binom{k}{i} p^i (1-p)^{k-i} \frac{\gamma_l(k, (x-i) \log((1-p)/p))}{(k-1)!}, \end{aligned}$$

where as before γ_l denotes the lower incomplete gamma function. Since

$$\begin{aligned} \frac{\tilde{A}_k + \log(1-p)k}{\log((1-p)/p)} &< x \\ \Leftrightarrow \tilde{L}_k(X, e, u) &> \exp(\log(1-p)k - x \log((1-p)/p)) \end{aligned}$$

we get

$$\begin{aligned} \mathbb{P}(\tilde{L}_k(X, e, u) \leq l) &\sim \\ k \wedge \frac{k \log((1-p)^k/l)}{\log((1-p)/p)} & \\ \sum_{i=0}^{k \wedge \frac{k \log((1-p)^k/l)}{\log((1-p)/p)}} \binom{k}{i} p^i (1-p)^{k-i} &\frac{\gamma_u(k, \log((1-p)^k/l) - i \log((1-p)/p))}{(k-1)!}. \end{aligned}$$

If $p > 1/2$ then

$$\begin{aligned} \mathbb{P}\left(\frac{\tilde{A}_k + \log(1-p)k}{\log(p/(1-p))} < x\right) \\ \sim \sum_{i=0}^k \binom{k}{i} p^i (1-p)^{k-i} \frac{\gamma_l(k, (x+i) \log(p/(1-p)))}{(k-1)!}, \end{aligned}$$

And since

$$\frac{\tilde{A}_k + \log(1-p)k}{\log(p/(1-p))} < x \Leftrightarrow \tilde{L}_k(X, e, u) > \exp(\log(1-p)k - x \log(p/(1-p)))$$

we get

$$\begin{aligned} \mathbb{P}(\tilde{L}_k(X, e, u) \leq l) \\ \sim \sum_{i=0}^k \binom{k}{i} p^i (1-p)^{k-i} \frac{\gamma_u(k, \log((1-p)^k/l) - i \log(1-p)/p)}{(k-1)!}. \end{aligned}$$

It remains to check that the formula remains valid for $p = 0$ and the proof is finished off by conditioning and invoking the classical Poisson process exceedance probabilities. \square

Remark 8.4.6. (Method 2) In practice, for a finite sample size n , we consider

$$\begin{aligned} \tilde{\mathcal{G}}(l) &= e^{-\lambda} 1\{1 \leq l\} \\ &+ \sum_{k=1}^n \frac{\lambda^k}{k!} e^{-\lambda} \sum_{i=0}^{k \wedge \kappa(p)} \binom{k}{i} p^i (1-p)^{k-i} \frac{\gamma_u(k, \log((1-p)^k/l) + i \log(p/(1-p)))}{(k-1)!}, \end{aligned}$$

as the determining factor to check whether a sample is novel or not: if it is above or below $1 - \alpha/2$ or $\alpha/2$, respectively, we say that the sample is novel, or reject the null hypothesis.

8.4.3 The Bonferroni correction

As a benchmark to the bivariate and asymptotically independent hypothesis testing at hand, it is natural to consider hypothesis testing based on the sufficient statistics of the exceedances $(\log(X_i^{\alpha+c}) - u)$ and corresponding e_i , $i = 1, \dots, k$, which are approximately independent and i.i.d. exponential and binomial, respectively, when X_i are Pareto in the tail.

Remark 8.4.7. (Benchmark) When X_i have a Pareto tail, the classic Bonferroni test for censored data is as follows. Since for high thresholds

$$b_1 := \sum_{i=1}^k \frac{\log(X_i^{\alpha+c}) - u}{s_{\log(X_i^{\alpha+c})}(u)} \sim \text{Gamma}(k, 1)$$

and (near) independently

$$b_2 := \sum_{i=1}^k e_i \sim \text{Binom}\left(k, p = \frac{\alpha}{\alpha + c}\right)$$

we set

$$\mathcal{B}_1(b_1) = \frac{\gamma_1(k, b_1)}{(k-1)!}$$

and

$$\mathcal{B}_2(b_2) = \sum_{i=0}^{b_2} \binom{k}{i} p^i (1-p)^{k-i}$$

and we reject the null hypothesis whenever any of the following holds:

$$\mathcal{B}_1(b_1) < \alpha/4, \quad \mathcal{B}_1(b_1) > 1 - \alpha/4, \quad \mathcal{B}_2(b_2) < \alpha/4, \quad \mathcal{B}_2(b_2) > 1 - \alpha/4.$$

Notice that for the binomial test \mathcal{B}_2 above it is crucial to use the same k as the number of exceedances used to calculate b_2 , making it unfeasible to apply a Poisson conditional argument as before.

In the same spirit of Method 1 and Method 2 we may apply a Bonferroni correction to a double hypothesis test which also takes into account the conditional distribution of the exceedances given the sample size, which is asymptotically Poisson. More precisely, using the notation of (8.16), (8.17) we consider the asymptotic likelihood of the normalised exceedances without the censoring indicators:

$$L_k^B(X, e, u) = \prod_{i=1}^k \exp\left(-\frac{\tilde{Z}_i - u}{s_{\tilde{Z}}(u)}\right), \quad (8.18)$$

for $1 \leq k \leq n$ and $L_0^B(X, e, u) = 1$. With $p(k) = \mathbb{P}(\tilde{K}_0(u) = k)$, $0 \leq k \leq n$, if n is large and we choose u , such that $n\bar{F}_{\tilde{Z}}(u) = \lambda > 0$, we have by (8.9) that

$$\tilde{p}(k) \rightarrow \frac{\lambda^k}{k!} e^{-\lambda}.$$

It is then easy to prove in the same manner as Theorem 8.4.6 the following result

Theorem 8.4.8. *As $u = u(n)$, $n \rightarrow \infty$ such that $n\bar{F}_{\bar{z}}(u) \rightarrow \lambda > 0$, the asymptotic cumulative distribution function of the random variable $L_{\bar{K}_0(u)}^B(X, e, u)$ is given by*

$$G^B(l) = e^{-\lambda}1\{1 \leq l\} + \sum_{k=1}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} F_k^B$$

where

$$F_k^B = \frac{\gamma_u(k, -\log(l))}{(k-1)!},$$

and $\gamma_u(\cdot, \cdot)$ is the upper incomplete gamma function.

We also consider the the asymptotic likelihood of the censoring indicators:

$$L_k^C(X, e, u) = \prod_{i=1}^k \left(\frac{\alpha}{\alpha + c} \right)^{e_i} \left(\frac{c}{\alpha + c} \right)^{1-e_i}, \quad (8.19)$$

for $1 \leq k \leq n$ and $L_0^C(X, e, u) = 1$. Using the same proof method as before, we get

Theorem 8.4.9. *As $u = u(n)$, $n \rightarrow \infty$ such that $n\bar{F}_{\bar{z}}(u) \rightarrow \lambda > 0$, the asymptotic cumulative distribution function of the random variable $L_{\bar{K}_0(u)}^B(X, e, u)$ is given by*

$$G^C(l) = e^{-\lambda}1\{1 \leq l\} + \sum_{k=1}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} F_k^C$$

where

$$F_k^C = \sum_{i \in \Sigma(p, l)} \binom{k}{i} p^i (1-p)^{k-i},$$

and

$$\Sigma(p, l) = \begin{cases} \{0, \dots, \min \{k, \lfloor k \log((1-p)/l) / \log((1-p)/p) \rfloor\}\}, & p > 1/2 \\ \{\min \{k, \lfloor k \log((1-p)/l) / \log((1-p)/p) \rfloor\}, \dots, k\}, & p \leq 1/2. \end{cases}$$

With the two last results we can create an enhanced bivariate Bonferroni correction method that takes into account the conditional distribution of the exceedances given the sample size.

Remark 8.4.10. (Method 3) In practice, for a finite sample size n , we consider realisations l_1, l_2 of the (8.18) and (8.19), respectively and set

$$\mathcal{G}^B(l_1) = e^{-\lambda}1\{1 \leq l_1\} + \sum_{k=1}^n \frac{\lambda^k}{k!} e^{-\lambda} \frac{\gamma_u(k, -\log(l_1))}{(k-1)!},$$

and

$$\mathcal{G}^C(l_2) = e^{-\lambda}1\{1 \leq l_2\} + \sum_{k=1}^n \frac{\lambda^k}{k!} e^{-\lambda} \sum_{i \in \Sigma(p, l_2)} \binom{k}{i} p^i (1-p)^{k-i},$$

and we reject the null hypothesis whenever any of the following holds:

$$\mathcal{G}^B(l_1) < \alpha/4, \quad \mathcal{G}^B(l_1) > 1 - \alpha/4, \quad \mathcal{G}^C(l_2) < \alpha/4, \quad \mathcal{G}^C(l_2) > 1 - \alpha/4.$$

8.4.4 A special class of regularly varying distributions

We have made the assumption that the variables X_i have regularly varying tails, $\mathbb{P}(X_1 > x) = \ell_X(x)x^{-(\alpha+c)}$, $\alpha + c > 0$, and at the same time that the variables $Z_i = -\log f(X_i)$, $W_i = -\log \bar{F}(X_i)$, are in the Gumbel domain of attraction. For this to hold true we require that

$$\mathbb{P}(-\log(\ell_\xi(X_1)X_1^{-\alpha}) > x)$$

is tail equivalent with a Von Mises function, and likewise for the Z_i . We initially examine a simple case.

Assume that $\ell_\xi = l_\xi$, a constant, so $\ell_X(x) = l_\xi \ell_C(x)$, and we further make the requirement that $\ell_C(x)$ converges to a constant l_C . We have

$$\mathbb{P}(W_1 > x) = \mathbb{P}(X_1^\alpha > l_\xi \exp(x)) = l_\xi^{-c/\alpha} l_C \left(l_\xi^{1/\alpha} \exp(x/\alpha) \right) \exp(-x(\alpha+c)/\alpha),$$

which is tail equivalent to a Von Mises function. In a similar fashion the expression

$$\begin{aligned} \mathbb{P}(Z_1 > x) &= \mathbb{P}(X_1^{\alpha+1} > \alpha l_\xi \exp(x)) \\ &= \alpha^{-(\alpha+c)/(\alpha+1)} l_\xi^{-(c+1)/(\alpha+1)} l_C \left((\alpha l_\xi)^{1/(\alpha+1)} \exp(x/(\alpha+1)) \right) \\ &\quad \times \exp(-x(\alpha+c)/(\alpha+1)), \end{aligned}$$

demonstrates the belonging to the Gumbel domain of attraction. By convergence, the same is true for the case $\ell_\xi \rightarrow l_\xi$. A subclass of such distributions which was first suggested by Hall et al. (1985), and subsequently accepted in the Extreme Value Theory community is the following:

$$\begin{aligned} \mathbb{P}(\xi_i > u) &= C_1 u^{-\alpha} (1 + K_1 u^{-\beta_1} (1 + o(1))) \quad \text{for } u \rightarrow \infty, \\ \mathbb{P}(C_i > u) &= C_2 u^{-c} (1 + K_2 u^{-\beta_2} (1 + o(1))) \quad \text{for } u \rightarrow \infty, \end{aligned} \tag{8.20}$$

where $\beta_1, \beta_2, C_1, C_2$ are positive constants and K_1, K_2 real constants.

8.5 Performance

We consider a simulation study where the variables ξ and C follow a regularly varying tail given by the following special case of (8.20):

$$\begin{aligned} \mathbb{P}(\xi_i > u) &= u^{-\alpha} (1 + K u^{-\beta}) \quad \text{for } u \rightarrow \infty, \\ \mathbb{P}(C_i > u) &= u^{-c} (1 + K u^{-\beta}) \quad \text{for } u \rightarrow \infty. \end{aligned}$$

We will study the successful classification proportion, or power of the test, given the significance level of 0.05. The following table summarises the different settings that we cover, where α_0 and c_0 are the hypothesis (test) parameters:

The mean excess function is taken to be as in the exact Pareto case, as are the thresholds, defined by λ through the equations

$$n\bar{F}_Z(u) = \lambda > 0, \quad n\bar{F}_W(v) = \lambda > 0.$$

	α_0	c_0	α	c	β	K	λ
Study 1	2	1.2	2	2.2	2	0.2	$\in [5, 30]$
Study 2	1.5	1.2	1	1.2	2	0.5	$\in [5, 30]$
Study 3	2	1	1.8	1	2	$\in [0, 2]$	15
Study 4	2	1.2	2.2	2.2	2	$\in [0, 2]$	15

For the Method 1 (Remark 8.4.4), Method 2 (Remark 8.4.6), Method 3 (Remark 8.4.10), and Baseline (Remark 8.4.7) the results based on a sample of size $n = 200$ and averaged over 1000 simulations (for each value of the varying parameter) are given in Figure 8.1.

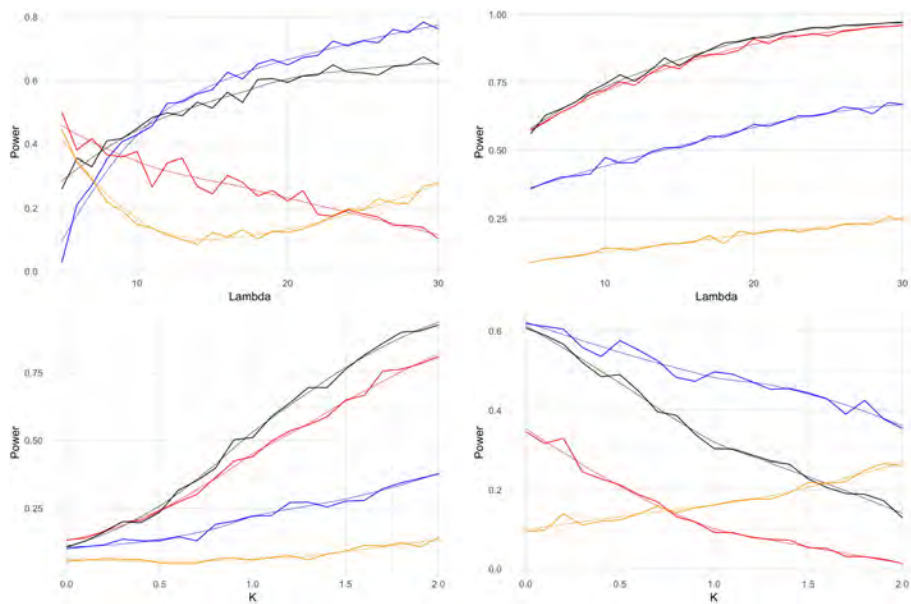


Figure 8.1: Power of various tests introduced in the text, as a function of a test parameter. Method 1 in blue; Method 2 in red; Method 3 in black; Baseline in yellow.

We observe that the three methods introduced in this paper which make use of the conditional information of the exceedances given the sample size drastically improve novelty detection for censored data. A Bonferroni correction on two such tests (Method 3) is seen to be competitive against the univariate tests (Method 1 and 2). When the tail of the new (testing) sample becomes heavier, the usual MLE bivariate approach is always outperformed in the cases that we consider (Study 2 and 3). From Study 1, where the testing sample has lighter tail, we observe that too few exceedances (too high thresholds) or too many exceedances (too low thresholds) can sometimes be detrimental to the methods which rely on the Point Processes of Exceedances. Finally, from Study 4 we can observe that large departures from Pareto behaviour when the testing sample has lighter tail can be detrimental to the PPP methods. The latter is not surprising, since when tails are lighter than expected, too few observations fall above a pre-set threshold, and the statistics using those very few points become less reliable.

8.6 Conclusion

In this paper we have developed novelty detection tests for the situation when datasets have right-censored observations, building and generalising on existing methodology. The focus was on detecting a change in either tail of the underlying sample or of the censoring mechanism. We have shown how a PPP arises from a point process of exceedances in the censored case, exploiting the fact that for Pareto-like tails, the peaks over high thresholds and the censoring labels are virtually independent. We have used this tool, and the already existing Poisson convergence theorem to develop tests which jointly include information on not only the censoring mechanism and the size of the exceedances but also on the number of exceedances.

Since we could also see the problem as a bivariate hypothesis testing problem, the loss of power expected from, say, a Bonferroni correction, was expected to not be large, given that the tests are virtually independent for high thresholds. Nonetheless, the inclusion of PPP methods allowed us to develop a probabilistically sound test which takes into account the conditional distribution of the number of exceedances given a sample size (Poisson approximation) which yields high sensitivity, especially when the tail of the new sample is heavier.

With this contribution we aim at filling a gap in the literature which is classically of interest primarily to the health, finance and insurance sector. The methods we developed are designed and perform best in under-sampled regions of heavy-tailed distributions, i.e. in the tail of the distribution.

There are various ways of how to generalise the results we have obtained. For instance, the introduction of dependence between the censoring mechanism and the underlying sample could be interesting to model, i.e. the novelty detection study of data which is censored, but not completely at random. We would also like to study the novelty detection problem for the case of several dependent samples which have been randomly censored by a possibly dependent multivariate censoring mechanism.

Chapter 9

Matrix Mittag–Leffler distributions and modeling heavy-tailed risks

This chapter is based on the following article:

Albrecher, H., Bladt, M., & Bladt, M. (2019). Matrix Mittag–Leffler distributions and modeling heavy-tailed risks. *Extremes*, to appear. arXiv:1906.05316.

Abstract

In this paper we define the class of matrix Mittag-Leffler distributions and study some of its properties. We show that it can be interpreted as a particular case of an inhomogeneous phase-type distribution with random scaling factor, and alternatively also as the absorption time of a semi-Markov process with Mittag-Leffler distributed interarrival times. We then identify this class and its power transforms as a remarkably parsimonious and versatile family for the modelling of heavy-tailed risks, which overcomes some disadvantages of other approaches like the problem of threshold selection in extreme value theory. We illustrate this point both on simulated data as well as on a set of real-life MTPL insurance data that were modeled differently in the past.

9.1 Introduction

The modeling of heavy-tailed risks is a classical topic in probability, statistics and its applications to understand and interpret data, see e.g. Embrechts et al. (1997), Beirlant et al. (2004) and Klugman et al. (2012). The folklore heavy-tailed distributions like Pareto, (heavy-tailed) Weibull and lognormal distributions can often serve as very useful benchmark models, particularly when only a few data points are available. In addition, the Pareto distribution is simple to work with and has an intuitive justification in terms of limit properties of extremes. However, in situations with more (but not an abundance of) data points, one often empirically

observes that a simple Pareto distribution does not serve as a good model across the entire range of the distribution. This is also the case for more general parametric families like Burr or Benktander distributions. Traditionally, and according to a main paradigm of the extreme value statistics approach, this is handled by only using the largest available data points to estimate the tail behavior, and model the bulk of the distribution separately by another distribution, finally splicing together the respective parts (see e.g. (Albrecher et al., 2017, Ch.4) for details). In insurance practice one often refers to this separate modeling as the modeling of *attritional* and *large* claims, and the resulting models are frequently referred to as *composite models* Pigeon and Denuit (2011). A natural problem in this context is how to choose the threshold between the separate regions, often boiling down to the compromise of not leaving too few data points for the tail modeling. At the same time, the consequences of that choice can be considerable, for instance for the determination of solvency capital requirements in insurance (cf. (Albrecher et al., 2017, Ch.6) for illustrations). A considerable effort has therefore been made to develop techniques and criteria for an appropriate choice of such thresholds, see e.g. Beirlant et al. (2004) for an overview and Bladt et al. (2019) for a recent contribution in that direction.

If the confidence in the relevance of available data points for the description of the (future) risk is sufficiently high, another possible approach is to use a tractable, but much larger family of distributions and identify a good fit. A particularly popular candidate for such an approach is the class of phase-type (PH) distributions, see e.g. Asmussen et al. (1996). The class of PH distributions is dense (in the sense of weak convergence) in the class of distributions on the positive real line, meaning that they can approximate any positive distribution arbitrarily well. They are, however, light-tailed, which may be a problem in applications which require a heavier tail and where the quantities of interest heavily depend on the tail behavior (as e.g. for ruin probabilities, cf. Asmussen and Albrecher (2010)). Fitting heavy-tailed distributions with a PH distribution can then lead to requiring many phases (rendering its use computationally cumbersome), and the resulting model will still not capture the tail behavior in a satisfactory manner. Two approaches to remedy this problem are Bladt and Rojas-Nandayapa (2018) and Bladt et al. (2015). In Albrecher and Bladt (2019) recently another direction was suggested, namely to transform time in the construction of PH distributions (as absorption times of Markov jump processes), leading to inhomogeneous phase-type (IPH) distributions. For suitable transformations, this approach allows to transport the versatility of PH distributions into the domain of heavy-tailed distributions, by introducing dense classes of genuinely heavy-tailed distributions. As a by-product, it was shown in Albrecher and Bladt (2019) that a class of matrix-Pareto distributions can be identified, where the scalar parameter of a classical Pareto distribution is replaced by a matrix, providing an intuitive and somewhat natural extension of the Pareto distribution, much as the matrix-exponential distribution, which is a powerful extension of the classical exponential distribution, see e.g. Bladt and Nielsen (2017).

In this paper we establish another matrix version of a distribution, namely the Mittag-Leffler distribution (first studied by Pillai (1990)). While the identifica-

tion of matrix versions of distributions is of mathematical interest in its own right, we will show that the resulting matrix Mittag-Leffler distributions (and its power transforms) have favorable properties for the modeling of heavy-tails, and it can outperform some other modeling approaches in a remarkable way. Furthermore, we will identify this class of distributions as a particular extension of the IPH class, where the scaling is random. In addition, we will establish the matrix Mittag-Leffler distribution as the absorption time of a semi-Markov process with (scalar) Mittag-Leffler distributed inter-arrival times, extending the role of the exponential distribution for the inter-arrival in continuous-time Markov chains.

The Mittag-Leffler function was first introduced in Mittag-Leffler (1904) and over the years turned out to be a crucial object in fractional calculus. It can be seen as playing the same role for fractional differential equations as the exponential function does for ordinary differential equations, see e.g. Gorenflo et al. (2014) for a recent overview. A recent application of fractional calculus for a particular risk model in insurance can be found in Constantinescu et al. (ress). Mittag-Leffler functions with matrix argument were first introduced in Chikrii and Eidel'man (2000) and play a prominent role for identifying solutions of systems of fractional differential equations, see e.g. Garrappa and Popolizio (2018). Here we will use them to define the class of matrix Mittag-Leffler (MML) distributions, which enjoy some attractive mathematical properties and are heavy-tailed with a regularly varying tail with index $\alpha < 1$. While such extremely heavy-tails with resulting infinite mean can be relevant in the modeling of operational risk Nešlehová et al. (2006) and possibly insurance losses due to natural catastrophes Albrecher et al. (2017), in most applications of interest the tails are slightly less heavy. We therefore enlarge the class of matrix Mittag-Leffler distributions by also including its power transforms and estimate the corresponding power together with the other parameters from the data in the fitting procedure. The index of regular variation of this larger class of distributions can now be any positive number. Whereas the number of needed phases for a PH or IPH fit can be very large also due to multi-modality or other irregularities in the shape of the main body of the distribution, we will see that the class of matrix Mittag-Leffler distribution and its power transforms (which we call *power matrix Mittag-Leffler (PMML) distributions*) offers a significant reduction in the number of phases needed to obtaining adequate fits. For this reason, it can even be worthwhile to scale light-tailed data points to heavy-tailed ones first, then apply a matrix Mittag-Leffler fit to the latter and transform the fit back to the original light-tailed scaling. This procedure is to some extent the reverse direction of the philosophy that underlied the PH fitting of heavy-tailed distributions. We will illustrate the potential advantage of this alternative approach in the numerical section at the end of the paper.

The remainder of the paper is organized as follows. Section 9.2 recollects some useful definitions and properties of Mittag-Leffler functions, Mittag-Leffler distributions and PH distributions. Section 9.3 then defines matrix Mittag-Leffler distributions and derives a number of its properties. We also give three explicit examples. Section 9.4 establishes MML distributions as IPH distributions under a particu-

lar random scaling, which allows to intuitively understand the additional flexibility gained from using MML distributions for the modeling of heavy-tailed risks. We then also establish MML distributions as the absorption times of a semi-Markov process with ML distributed interarrival times, which is yet another perspective on the potential of the MML class as a modeling tool. Finally, Section 9.5 is devoted to the modeling of data using (power-transformed) MML distributions. We first illustrate the convincing performance of the numerical fitting procedure to something as involved as tri-modal data. Secondly, we consider an MTPL data set taken from Albrecher et al. (2017) and already studied by various other means in the literature. We show that a plain maximum-likelihood fit to this data set gives a convincing fit to the entire range of the data with remarkably few parameters, and even identifies the tail index with striking accuracy when compared to recent extreme value techniques as in Bladt et al. (2019), without having to choose a threshold for the tail modeling at all. We then also provide an example where transforming light-tailed data into heavy-tailed ones, fitting with a PMML distribution and transforming back can lead to a much better fit for the same number of parameters than a classical phase-type distribution. We then also discuss the signature of MML distributions in the tail in terms of the behavior of the Hill plot, which allows to develop an intuitive guess as to when MML distributions are particularly adequate for a fitting procedure of heavy-tails. Finally, Section 9.6 concludes.

9.2 Some relevant background

9.2.1 Mittag–Leffler functions

The Mittag–Leffler (ML) function is defined by

$$E_{\alpha,\beta}(z) = \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\alpha k + \beta)}, \quad z \in \mathbb{C},$$

where $\beta \in \mathbb{R}$ and $\alpha > 0$. The ML function is an entire function if $\beta > 0$, and it satisfies (see e.g. (Erdélyi et al., 1955, p.210))

$$\frac{d^m}{dz^m} [z^{\beta-1} E_{\alpha,\beta}(z^\alpha)] = z^{\beta-m-1} E_{\alpha,\beta-m}(z^\alpha).$$

This implies that (see (Garrappa and Popolizio, 2018, Prop.2))

$$E_{\alpha,\beta}^{(k)}(z) = \frac{d^k}{dz^k} E_{\alpha,\beta}(z) = \frac{1}{\alpha^k z^k} \sum_{j=0}^k c_j^{(k)} E_{\alpha,\beta-j}(z) \quad (9.1)$$

where

$$c_j^{(k)} = \begin{cases} (1 - \beta - \alpha(k-1))c_0^{(k-1)}, & j = 0, \\ c_{j-1}^{(k-1)} + (1 - \beta - \alpha(k-1) + j)c_j^{(k-1)}, & j = 1, \dots, k-1, \\ 1, & j = k. \end{cases}$$

For a matrix \mathbf{A} , we may define its ML function as

$$E_{\alpha,\beta}(\mathbf{A}) = \sum_{k=0}^{\infty} \frac{\mathbf{A}^k}{\Gamma(\alpha k + \beta)}.$$

If $\beta > 0$, one can then express the entire ML function of a matrix \mathbf{A} by Cauchy's formula

$$E_{\alpha,\beta}(\mathbf{A}) = \frac{1}{2\pi i} \int_{\gamma} E_{\alpha,\beta}(z)(z\mathbf{I} - \mathbf{A})^{-1} dz,$$

where γ is a simple path enclosing the eigenvalues of \mathbf{A} . If \mathbf{A} has a Jordan normal form $\mathbf{A} = \mathbf{P} \text{diag}(\mathbf{J}_1, \dots, \mathbf{J}_r) \mathbf{P}^{-1}$ with

$$\mathbf{J}_i = \begin{pmatrix} \lambda_i & 1 & 0 & \cdots & 0 \\ 0 & \lambda_i & 1 & \cdots & 0 \\ 0 & 0 & \lambda_i & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_i \end{pmatrix},$$

then we may equivalently express $E_{\alpha,\beta}(\mathbf{A})$ by

$$E_{\alpha,\beta}(\mathbf{A}) = \mathbf{P} \text{diag}(E_{\alpha,\beta}(\mathbf{J}_1), \dots, E_{\alpha,\beta}(\mathbf{J}_r)) \mathbf{P}^{-1},$$

where

$$E_{\alpha,\beta}(\mathbf{J}_i) = \begin{pmatrix} E_{\alpha,\beta}(\lambda_i) & E_{\alpha,\beta}^{(1)}(\lambda_i) & \frac{E_{\alpha,\beta}^{(2)}(\lambda_i)}{2!} & \cdots & \frac{E_{\alpha,\beta}^{(m_i-1)}(\lambda_i)}{(m_i-1)!} \\ 0 & E_{\alpha,\beta}(\lambda_i) & E_{\alpha,\beta}^{(1)}(\lambda_i) & \cdots & \frac{E_{\alpha,\beta}^{(m_i-2)}(\lambda_i)}{(m_i-2)!} \\ 0 & 0 & E_{\alpha,\beta}(\lambda_i) & \cdots & \frac{E_{\alpha,\beta}^{(m_i-3)}(\lambda_i)}{(m_i-3)!} \\ \vdots & \vdots & \cdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & E_{\alpha,\beta}(\lambda_i) \end{pmatrix},$$

and m_i is the dimension of \mathbf{J}_i .

In either case, we shall need to evaluate the derivatives of ML functions at the eigenvalues, which by (9.1) implies the evaluation of ML functions with possibly negative indices $\beta - j$. This is not a problem, but it is important that initially $\beta > 0$ to ensure that the ML function is entire and thereby the existence of the Cauchy integral formula is guaranteed.

For further properties on the ML function we refer e.g. to Erdélyi et al. (1955), Garrappa and Popolizio (2018), Matychyn and Onyshchenko (2018) and Haubold et al. (2011).

9.2.2 Mittag–Leffler distributions

A random variable having Mittag-Leffler (ML) distribution was defined in Pillai (1990) through the cumulative distribution function and consequently density given by

$$F_{\delta,\alpha}(x) = 1 - E_{\alpha,1}(-(x/\delta)^\alpha), \quad x > 0, \quad 0 < \alpha \leq 1,$$

$$f_{\delta,\alpha}(x) = \frac{x^{\alpha-1}}{\delta^\alpha} E_{\alpha,\alpha}(-(x/\delta)^\alpha), \quad x > 0, \quad 0 < \alpha \leq 1,$$

with Laplace transform

$$\frac{1}{1 + (\delta u)^\alpha}. \quad (9.2)$$

A convenient representation, due to Kozubowski Kozubowski (2001), for a ML random variable X is

$$X \stackrel{d}{=} \delta Z R^{1/\alpha},$$

where Z is standard exponential and R has cumulative distribution function

$$F_R(x) = \frac{2}{\pi\alpha} \left[\arctan \left(\frac{x}{\sin(\alpha\pi/2)} + \cot(\alpha\pi/2) \right) - \frac{\pi}{2} \right] + 1.$$

The tail behaviour of R is equivalent to that of a Cauchy random variable, and hence X is regularly varying with parameter α in the tail (see e.g. (Mikosch, 1999, Prop.1.3.9)).

The following extension (for the case $\delta = 1$) will also play a role in the sequel: a random variable X is said to follow a generalized Mittag-Leffler (GML) distribution with parameters α ($0 < \alpha \leq 1$) and $\beta > 0$, if its Laplace transform is given by

$$\mathbb{E}(e^{-uX}) = (1 + u^\alpha)^{-\beta}.$$

The corresponding cumulative distribution function then is

$$F_{\alpha,\beta}(x) = \sum_{k=0}^{\infty} \frac{(-1)^k \Gamma(k + \beta) x^{\alpha(k+\beta)}}{\Gamma(\beta) k! \Gamma(1 + \alpha(k + \beta))} = \sum_{k=0}^{\infty} \frac{(-1)^k x^{\alpha(k+\beta)}}{\mathfrak{B}(\beta, k) k \Gamma(1 + \alpha(k + \beta))},$$

where $\mathfrak{B}(x, y)$ is the Beta function (see Jose et al. (2010)). The analogous representation for a GML variable X is

$$X \stackrel{d}{=} W^{1/\alpha} S_\alpha, \quad (9.3)$$

where W is Gamma with scale parameter 1 and shape parameter β , and S_α is a random variable with Laplace transform given by

$$\mathbb{E}(e^{-uS_\alpha}) = \exp(-u^\alpha).$$

9.2.3 Phase–type distributions

A random variable τ is said to be phase–type distributed with generator (or representation) $(\boldsymbol{\pi}, \mathbf{T})$, and we write $\tau \sim \text{PH}(\boldsymbol{\pi}, \mathbf{T})$, if it is the time until absorption of a (time–homogeneous) Markov jump process $\{X_t\}_{t \geq 0}$ with state–space $E = \{1, 2, \dots, p, p + 1\}$ where states $1, \dots, p$ are transient and state $p + 1$ is absorbing. The row vector $\boldsymbol{\pi} = (\pi_1, \dots, \pi_p)$ is the initial distribution, $\pi_i = \mathbb{P}(X_0 = i)$, and $\mathbf{T} = \{t_{ij}\}_{i,j=1,\dots,p}$ where t_{ij} denotes the transition rates of jumps between transient states i and j . We assume that $\pi_1 + \dots + \pi_p + 1$, i.e. X_0 cannot start in the absorbing

state which would have caused an atom at zero. The intensity matrix for $\{X_t\}_{t \geq 0}$ can be written as

$$\Lambda = \begin{pmatrix} \mathbf{T} & \mathbf{t} \\ \mathbf{0} & 0 \end{pmatrix},$$

where $\mathbf{t} = (t_1, \dots, t_p)'$ is a column vector of exit rates, i.e. t_i is the rate of transition from state i to the absorbing state $p + 1$. We notice that $-\mathbf{T}\mathbf{e} = \mathbf{t}$, where $\mathbf{e} = (1, 1, \dots, 1)'$ is the p -dimensional column vector of ones, since the row sums in Λ must all be zero. Hence $(\boldsymbol{\pi}, \mathbf{T})$ fully parametrises the Markov process.

The class of Phase-type distributions is dense in the class of distributions on the positive reals, meaning that they may approximate any positive distribution arbitrarily well. On the other hand, they constitute a class of probabilistically tractable distributions, which often allows for exact solutions to complex stochastic problems, and frequently in a closed form. The theory is well developed with numerous applications in insurance risk and queueing theory (see e.g. Bladt and Nielsen (2017) and reference therein). Phase-type distributions are light-tailed (i.e., their tail has an exponential decay), which makes them inadequate for modelling certain phenomena like insurance risks with heavy-tailed claims. Recently, Albrecher and Bladt (2019) proposed an extension of the PH construction principle to time-inhomogeneous Markov processes, in which case the absorption times can also be heavy-tailed with a wide spectrum of possible tail shapes.

If $\tau \sim \text{PH}(\boldsymbol{\pi}, \mathbf{T})$, then its density function is given by $f_\tau(x) = \boldsymbol{\pi}e^{\mathbf{T}x}\mathbf{t}$, its distribution function by $F_\tau(x) = 1 - \boldsymbol{\pi}e^{\mathbf{T}x}\mathbf{e}$ and its Laplace transform by $L_\tau(s) = \boldsymbol{\pi}(s\mathbf{I} - \mathbf{T})^{-1}\mathbf{t}$, where \mathbf{I} denotes the identity matrix. The (fractional) moments are $\mathbb{E}(\tau^\alpha) = \Gamma(\alpha + 1)\boldsymbol{\pi}(-\mathbf{T})^{-\alpha}\mathbf{e}$. For further details on Phase-type distributions we refer to Bladt and Nielsen (2017).

9.3 Matrix Mittag–Leffler distributions

Let us now derive a matrix version of the Mittag-Leffler distribution by defining its Laplace transform and identifying the distribution associated to it. To this end, in view of (9.2) consider the function

$$\phi(u) = \boldsymbol{\pi}(u^\alpha\mathbf{I} - \mathbf{T})^{-1}\mathbf{t}, \quad 0 < \alpha \leq 1, \quad (9.4)$$

where $(\boldsymbol{\pi}, \mathbf{T})$ is a PH generator.

Theorem 9.3.1. *$\phi(u)$ is the Laplace transform of a probability distribution.*

Proof. Let $g(u) = u^\alpha$ and

$$f(x) = \boldsymbol{\pi}(x\mathbf{I} - \mathbf{T})^{-1}\mathbf{t}.$$

Then $\phi(u) = f(g(u))$. Now $\mathbf{T} - u^\alpha\mathbf{I}$ is a sub-intensity matrix for all $u \geq 0$ and therefore $(u^\alpha\mathbf{I} - \mathbf{T})^{-1}$ is a non-negative matrix (Green matrix, see (Bladt and Nielsen, 2017, p.134)). Thus

$$f^{(n)}(g(u)) = (-1)^n n! \boldsymbol{\pi} (u^\alpha\mathbf{I} - \mathbf{T})^{-n-1} \mathbf{t},$$

which has sign $(-1)^n$. Concerning g ,

$$g^{(j)}(u) = \alpha(\alpha - 1) \cdots (\alpha - j + 1)u^{\alpha-j}$$

has sign $(-1)^{j+1}$. We shall employ Faà di Bruno's formula,

$$\frac{d^n}{dx^n} f(g(x)) = \sum \frac{n!}{m_1! m_2! \cdots m_n!} \cdot f^{(m_1+\cdots+m_n)}(g(x)) \cdot \prod_{j=1}^n \left(\frac{g^{(j)}(x)}{j!} \right)^{m_j},$$

where the summation is over n -tuples for which

$$1 \cdot m_1 + 2 \cdot m_2 + 3 \cdot m_3 + \cdots + n \cdot m_n = n,$$

to determine the sign of $\phi^{(n)}(u)$. Notice that

$$\begin{aligned} \text{sign} \left(f^{(m_1+m_2+\cdots+m_n)}(g(u)) \prod_{i=1}^n g^{(i)}(u)^{m_i} \right) &= (-1)^{m_1+\cdots+m_n} \prod_{i=1}^n (-1)^{m_i(i+1)} \\ &= (-1)^{2 \sum_i m_i} (-1)^{\sum_i i m_i} \\ &= (-1)^n. \end{aligned}$$

Hence all terms in the summation have the same sign $(-1)^n$, and we conclude that also the sum itself has sign $(-1)^n$, i.e. $\text{sign}(\phi^{(n)}(u)) = (-1)^n$. Since $\phi(0) = \boldsymbol{\pi}(-\mathbf{T})\mathbf{t} = \boldsymbol{\pi}(-\mathbf{T})(-\mathbf{T}\mathbf{e}) = \boldsymbol{\pi}\mathbf{e} = 1$, the result then follows with Bernstein's theorem (see (Feller, 1971, p.439)). \square

Remark 9.3.2. Note that the proof does not rely on the special form of u^α , and it follows that if g is any function with $g(0) = 0$ and $-g$ completely monotone, then

$$\phi(u) = \boldsymbol{\pi}(g(u)\mathbf{I} - \mathbf{T})^{-1}\mathbf{t}$$

is the Laplace transform of a probability distribution, which may also be useful in other contexts.

Theorem 9.3.3. *Let X be a random variable with Laplace transform (9.4). Then the density function of X is given by*

$$f(x) = x^{\alpha-1} \boldsymbol{\pi} E_{\alpha,\alpha}(\mathbf{T}x^\alpha) \mathbf{t}.$$

Proof. We show that f has the required Laplace transform. For u sufficiently large

one has

$$\begin{aligned}
 \int_0^\infty e^{-ux} f(x) \, dx &= \pi \int_0^\infty e^{-ux} \sum_{n=0}^\infty \frac{\mathbf{T}^n x^{\alpha n}}{\Gamma((n+1)\alpha)} x^{\alpha-1} \, dx \, \mathbf{t} \\
 &= \pi \sum_{n=0}^\infty \frac{\mathbf{T}^n}{\Gamma((n+1)\alpha)} \int_0^\infty x^{\alpha(n+1)-1} e^{-ux} \, dx \, \mathbf{t} \\
 &= \pi \sum_{n=0}^\infty \frac{\mathbf{T}^n}{\Gamma((n+1)\alpha)} \Gamma((n+1)\alpha) u^{-(n+1)\alpha} \mathbf{t} \\
 &= -\pi \sum_{n=0}^\infty \mathbf{T}^{n+1} u^{-(n+1)\alpha} \mathbf{e} \\
 &= -\pi \sum_{n=1}^\infty (-\mathbf{T} u^{-\alpha})^n \mathbf{e} \\
 &= -\pi (\mathbf{I} - \mathbf{T} u^{-\alpha})^{-1} (\mathbf{T} u^{-\alpha}) \mathbf{e} \\
 &= \pi (u^\alpha \mathbf{I} - \mathbf{T})^{-1} \mathbf{t}.
 \end{aligned}$$

The result for all $u \in [0, \infty)$ then follows by analytic continuation. □

Definition 9.3.4. Let $(\boldsymbol{\pi}, \mathbf{T})$ be a PH generator and let $0 < \alpha \leq 1$. A random variable X is said to have a matrix Mittag–Leffler distribution, if its Laplace transform is given by

$$\mathbb{E}(e^{-uX}) = \boldsymbol{\pi} (u^\alpha \mathbf{I} - \mathbf{T})^{-1} \mathbf{t}.$$

In this case we write $X \sim \text{MML}(\alpha, \boldsymbol{\pi}, \mathbf{T})$.

Remark 9.3.5. The proof Theorem 9.3.3 shows that for any triplet $(\boldsymbol{\pi}, \mathbf{T}, \mathbf{t})$ for which

$$f(x) = \boldsymbol{\pi} e^{\mathbf{T}x} \mathbf{t} \tag{9.5}$$

is a density function, the function $\phi(u) = \boldsymbol{\pi} (u^\alpha \mathbf{I} - \mathbf{T})^{-1} \mathbf{t}$ is indeed a Laplace transform of a probability distribution. Distributions with density (9.5) are referred to as matrix–exponential distributions, and contain the class of phase–type distributions as a strict subset. We have stated the above definition in terms of a phase–type generator, but it is hence clear that the construction also applies to any matrix–exponential distribution. While most results in the following could be stated in terms of a matrix–exponential triplet, for the sake of simplicity and notation we restrict the parameters to phase–type generators only.

Corollary 9.3.6. Let $X \sim \text{MML}(\alpha, \boldsymbol{\pi}, \mathbf{T})$. Then the cumulative distribution function for X is given by

$$F(x) = 1 - \boldsymbol{\pi} E_{\alpha,1}(\mathbf{T}x^\alpha) \mathbf{e}.$$

Proof. The derivative of this function is

$$\begin{aligned}
 F'(x) &= -\boldsymbol{\pi} \sum_{n=1}^\infty \frac{\mathbf{T}^n n x^{\alpha(n-1)} \alpha x^{\alpha-1}}{\Gamma(1 + \alpha n)} \mathbf{e} \\
 &= -x^{\alpha-1} \boldsymbol{\pi} \sum_{n=1}^\infty \frac{(\mathbf{T}x^\alpha)^{n-1}}{\Gamma(\alpha n)} \mathbf{T} \mathbf{e} \\
 &= x^{\alpha-1} \boldsymbol{\pi} E_{\alpha,\alpha}(\mathbf{T}x^\alpha) \mathbf{t},
 \end{aligned}$$

which indeed coincides with the density. It remains to note that F satisfies the boundary condition $F(0) = 0$. \square

One can now realize that the matrix ML distribution provides an extension of representation (9.3) in the following way:

Theorem 9.3.7. *Let $X \sim MML(\alpha, \boldsymbol{\pi}, \mathbf{T})$. Then*

$$X \stackrel{d}{=} W^{1/\alpha} S_\alpha, \quad (9.6)$$

where $W \sim PH(\boldsymbol{\pi}, \mathbf{T})$, and S_α is an independent (positive stable) random variable with Laplace transform given by $\exp(-u^\alpha)$.

Proof. Simply note that

$$\begin{aligned} \mathbb{E}(\exp(-uW^{1/\alpha}S_\alpha)) &= \int_0^\infty \mathbb{E}(\exp(-ux^{1/\alpha}S_\alpha)) \boldsymbol{\pi} e^{\mathbf{T}x} \mathbf{t} \, dx \\ &= \boldsymbol{\pi} \int_0^\infty e^{-u^\alpha x} e^{\mathbf{T}x} \, dx \, \mathbf{t} \\ &= \boldsymbol{\pi} \int_0^\infty e^{-x(u^\alpha \mathbf{I} - \mathbf{T})} \, dx \, \mathbf{t} \\ &= \boldsymbol{\pi} (u^\alpha \mathbf{I} - \mathbf{T})^{-1} \mathbf{t}. \end{aligned}$$

\square

Corollary 9.3.8. *Let $X \sim MML(\alpha, \boldsymbol{\pi}, \mathbf{T})$. The fractional moments of order $\rho < \alpha \leq 1$ are given by*

$$\mathbb{E}(X^\rho) = \frac{\Gamma(1 - \rho/\alpha) \Gamma(1 + \rho/\alpha) \boldsymbol{\pi} (-\mathbf{T})^{-\rho/\alpha} \mathbf{e}}{\Gamma(1 - \rho)}.$$

Proof. It is known Wolfe (1975) that the fractional moments of a random variable S_α with Laplace transform $\exp(-u^\alpha)$ are given by

$$\mathbb{E}(S_\alpha^\rho) = \frac{\Gamma(1 - \rho/\alpha)}{\Gamma(1 - \rho)}.$$

From Bladt and Nielsen (2017) we know that the ν th fractional moment of a random variable W with $PH(\boldsymbol{\pi}, \mathbf{T})$ distribution is given by

$$\mathbb{E}(W^\nu) = \Gamma(\nu + 1) \boldsymbol{\pi} (-\mathbf{T})^{-\nu} \mathbf{e}.$$

By Theorem 9.3.7, and setting $\nu = \rho/\alpha$, X will have the ρ th fractional moment given by

$$\mathbb{E}(X^\rho) = \mathbb{E}(S_\alpha^\rho) \mathbb{E}(W^{\rho/\alpha}) = \frac{\Gamma(1 - \rho/\alpha) \Gamma(1 + \rho/\alpha) \boldsymbol{\pi} (-\mathbf{T})^{-\rho/\alpha} \mathbf{e}}{\Gamma(1 - \rho)}.$$

\square

Remark 9.3.9. Representation (9.6) is not only useful for establishing closed-form formulas, but it also suggests a simple and efficient simulation technique for X . Simulation algorithms for PH and stable distributions are for instance available in the statistical software R via the packages `actuar` and `stabledist`, respectively.

Example 9.3.1. Consider $X \sim \text{MML}(\alpha, \boldsymbol{\pi}, \mathbf{T})$ where $(\boldsymbol{\pi}, \mathbf{T})$ is the PH representation of an Erlang distribution with p ($p \in \mathbb{N}$) stages and intensity λ , i.e. $\boldsymbol{\pi} = (1, 0, \dots, 0)$ and

$$\mathbf{T} = \begin{pmatrix} -\lambda & \lambda & 0 & \dots & 0 & 0 \\ 0 & -\lambda & \lambda & \dots & 0 & 0 \\ 0 & 0 & -\lambda & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -\lambda & \lambda \\ 0 & 0 & 0 & \dots & 0 & -\lambda \end{pmatrix}.$$

In this case

$$(s\mathbf{I} - \mathbf{T})^{-1} = \begin{pmatrix} \frac{1}{s+\lambda} & \frac{\lambda}{(s+\lambda)^2} & \frac{\lambda^2}{(s+\lambda)^3} & \dots & \frac{\lambda^{p-1}}{(s+\lambda)^p} \\ 0 & \frac{1}{s+\lambda} & \frac{\lambda}{(s+\lambda)^2} & \dots & \frac{\lambda^{p-2}}{(s+\lambda)^{p-1}} \\ 0 & 0 & \frac{1}{s+\lambda} & \dots & \frac{\lambda^{p-3}}{(s+\lambda)^{p-2}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \frac{1}{s+\lambda} \end{pmatrix},$$

so

$$\begin{aligned} f(x) &= x^{\alpha-1} \boldsymbol{\pi} E_{\alpha, \alpha}(\mathbf{T} x^\alpha) \mathbf{t} \\ &= x^{\alpha-1} \frac{1}{2\pi i} \int_{\gamma} E_{\alpha, \alpha}(s) \boldsymbol{\pi} (s\mathbf{I} - x^\alpha \mathbf{T})^{-1} \mathbf{t} \, ds \\ &= x^{\alpha-1} \frac{1}{2\pi i} \int_{\gamma} E_{\alpha, \alpha}(s) \lambda \frac{(\lambda x^\alpha)^{p-1}}{(s + x^\alpha \lambda)^p} \, ds \\ &= \frac{\lambda^p x^{\alpha p-1}}{(p-1)!} E_{\alpha, \alpha}^{(p-1)}(-\lambda x^\alpha), \end{aligned}$$

where γ is a simple path enclosing $-\lambda x^\alpha$, and where in the last step we used the residue theorem. Note that this corresponds to the GML random variable given already for general shape parameter $p = \beta \in \mathbb{R}_+$ in (9.3) (although this explicit form of the density was not given in Jose et al. (2010)). Figure 9.1 depicts the density for several choices of parameters. The parameter α controls the heaviness of the tail (and more generally the deviation from exponentiality of the Mittag-Leffler function), the parameter p determines the shape of the body of the distribution (since larger p implies a more pronounced Erlang component), and λ is a scaling parameter. \square

Example 9.3.2. Mixture of Erlang distributions form the simplest sub-class of PH distributions which are dense in the class of distributions on the positive real line

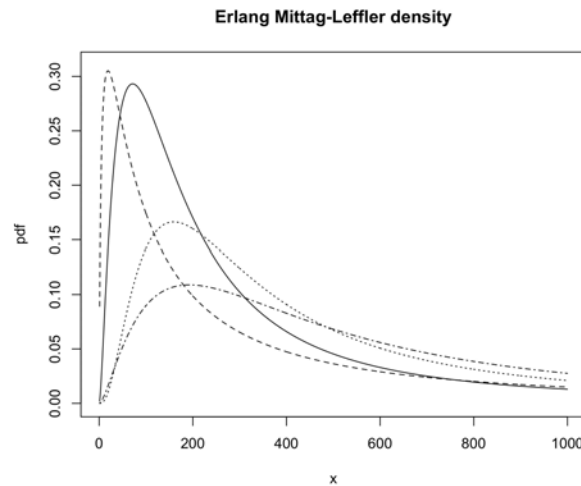


Figure 9.1: Density of a $\text{MML}(\alpha, \boldsymbol{\pi}, \mathbf{T})$, for $\alpha = 0.7$ and Erlang phase-type component with $p = 4$ and $\lambda = 2$ (solid). The remaining curves have a change in one of the parameters: $\alpha = 0.5$ (dashed), $p = 6$ (dotted), and $\lambda = 1$ (dashed and dotted).

(in the sense of weak convergence). Let h be the density

$$h(x) = \sum_{i=1}^m \theta_i f_{\text{Erl}}(x; p_i, \lambda_i),$$

where $f_{\text{Erl}}(x; p_i, \lambda_i) = \lambda_i^{p_i} x^{p_i-1} \exp(-\lambda_i x) / (p_i - 1)!$ denotes the density of an Erlang distribution and $\theta_i \geq 0$ are weights with $\sum \theta_i = 1$. Then it follows immediately from Example 9.3.1 that the corresponding MML distribution has density

$$f(x) = \sum_{i=1}^m \theta_i \frac{\lambda_i^{p_i} x^{\alpha p_i - 1}}{(p_i - 1)!} E_{\alpha, \alpha}^{(p_i-1)}(-\lambda_i x^\alpha).$$

In Figure 9.2, a trimodal distribution is considered, corresponding to $X \sim \text{MML}(\alpha, \boldsymbol{\pi}, \mathbf{T})$, for a mixture of three Erlang PH components. The densities of $\log(X)$ and X are both depicted. \square

Example 9.3.3. A Coxian phase-type distribution has a representation of the form

$$\boldsymbol{\pi} = (\pi_1, \dots, \pi_p), \quad \mathbf{T} = \begin{pmatrix} -\lambda_1 & \lambda_1 & 0 & \dots & 0 \\ 0 & -\lambda_2 & \lambda_2 & \dots & 0 \\ 0 & 0 & -\lambda_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & -\lambda_p \end{pmatrix},$$

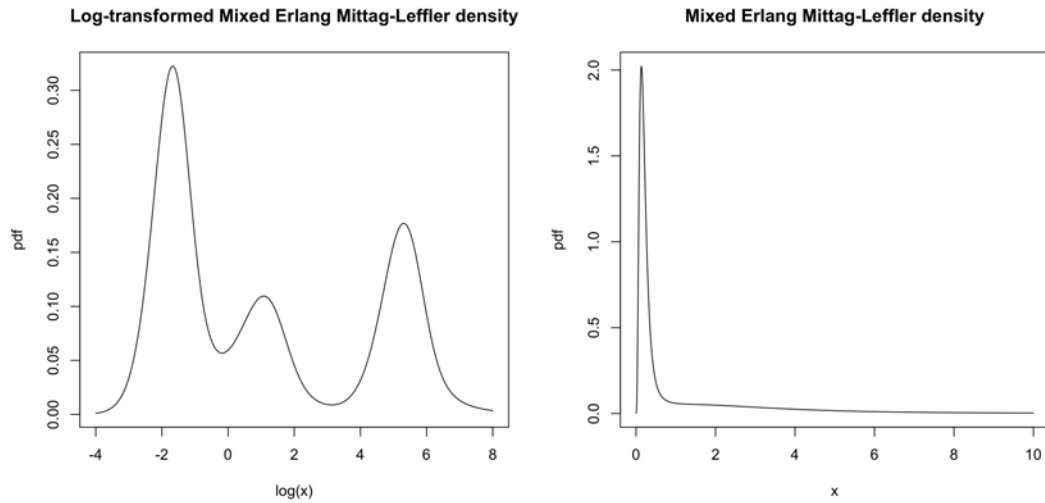


Figure 9.2: Densities of $\log(X)$ and of X , where $X \sim \text{MML}(\alpha, \boldsymbol{\pi}, \mathbf{T})$ is a mixture of $m = 3$ Erlang PH components with parameters $\alpha = 0.9$, $p_1 = 5$, $p_2 = 3$, $p_3 = 4$, $\lambda_1 = 20$, $\lambda_2 = 1$, $\lambda_3 = 0.03$, $\theta_1 = 0.5$, $\theta_2 = 0.2$, $\theta_3 = 0.3$.

where all λ_i , $i = 1, \dots, p$, are distinct. In this case

$$(z\mathbf{I} - \mathbf{T})^{-1} = \begin{pmatrix} \frac{1}{z+\lambda_1} & \frac{\lambda_1}{(z+\lambda_1)(z+\lambda_2)} & \frac{\lambda_1\lambda_2}{(z+\lambda_1)(z+\lambda_2)(z+\lambda_3)} & \cdots & \frac{\lambda_1\cdots\lambda_{p-1}}{(z+\lambda_1)(z+\lambda_2)\cdots(z+\lambda_p)} \\ 0 & \frac{1}{z+\lambda_2} & \frac{\lambda_2}{(z+\lambda_2)(z+\lambda_3)} & \cdots & \frac{\lambda_2\cdots\lambda_{p-1}}{(z+\lambda_2)(z+\lambda_3)\cdots(z+\lambda_p)} \\ 0 & 0 & \frac{1}{z+\lambda_3} & \cdots & \frac{\lambda_3\cdots\lambda_{p-1}}{(z+\lambda_3)(z+\lambda_4)\cdots(z+\lambda_p)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{1}{z+\lambda_p} \end{pmatrix},$$

so

$$\begin{aligned} f(x) &= x^{\alpha-1} \frac{1}{2\pi i} \int_{\gamma} E_{\alpha,\alpha}(s) \boldsymbol{\pi} (s\mathbf{I} - x^\alpha \mathbf{T})^{-1} \mathbf{t} ds \\ &= x^{\alpha-1} \sum_{j=1}^p \pi_j \frac{1}{2\pi i} \int_{\gamma} E_{\alpha,\alpha}(s) \frac{(\lambda_j x^\alpha)(\lambda_{j+1} x^\alpha) \cdots (\lambda_{p-1} x^\alpha) \lambda_p}{(s + \lambda_j x^\alpha)(s + \lambda_{j+1} x^\alpha) \cdots (s + \lambda_p x^\alpha)} ds \\ &= \sum_{j=1}^p \pi_j x^{\alpha(p-j+1)-1} \left(\prod_{k=j}^p \lambda_k \right) \frac{1}{2\pi i} \int_{\gamma} \frac{E_{\alpha,\alpha}(s)}{(s + x^\alpha \lambda_j) \cdots (s + x^\alpha \lambda_p)} ds \\ &= \sum_{j=1}^p \pi_j x^{\alpha(p-j+1)-1} \left(\prod_{k=j}^p \lambda_k \right) \sum_{m=j}^p \frac{E_{\alpha,\alpha}(-\lambda_m x^\alpha)}{\prod_{\substack{n=j \\ n \neq m}}^p (-x^\alpha \lambda_m + x^\alpha \lambda_n)} \\ &= x^{\alpha-1} \sum_{j=1}^p \pi_j \left(\prod_{k=j}^p \lambda_k \right) \sum_{m=j}^p \frac{E_{\alpha,\alpha}(-\lambda_m x^\alpha)}{\prod_{\substack{n=j \\ n \neq m}}^p (\lambda_n - \lambda_m)}. \end{aligned}$$

In Figure 9.3 four such densities are plotted. □

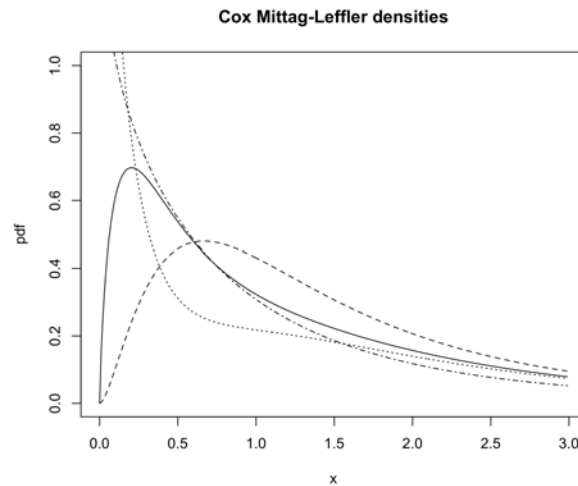


Figure 9.3: Density of $X \sim \text{MML}(\alpha, \boldsymbol{\pi}, \mathbf{T})$, with Coxian PH component ($p = 4$, $\alpha = 0.9$, $\lambda_1 = 1$, $\lambda_2 = 2$, $\lambda_3 = 3$, $\lambda_4 = 4$, and $\boldsymbol{\pi} = (0.5, 0, 0.5, 0)$ (solid), $\boldsymbol{\pi} = (0.5, 0.5, 0, 0)$ (dashed), $\boldsymbol{\pi} = (0.5, 0, 0, 0.5)$ (dotted), and $\boldsymbol{\pi} = (0.25, 0.25, 0.25, 0.25)$ (dashed–dotted)).

9.4 Sample path representations

We now provide two different representations of the MML distribution as sample path properties of a stochastic process. The first one will be as an absorption time of a randomly scaled time-inhomogeneous Markov jump process, and the second one as an absorption time of a particular semi-Markov process (where random time scaling is not needed).

9.4.1 Random time-inhomogeneous phase–type distributions

We recall from Section 9.2.3 that a random variable is PH distributed if it is the time until absorption of a time-homogeneous Markov jump process on a finite state-space, where one state is absorbing and the remaining states are transient. In this section we show that a MML distribution can be interpreted as a time-inhomogeneous phase-type distribution with random intensity matrix.

Let us define a random time-inhomogeneous Markov jump process X_t as a jump process with state space $E = \{1, 2, \dots, p, p+1\}$, where $p+1$ is an absorbing state, the remaining states being transient, and the intensity matrix given by

$$\mathbf{\Lambda}(t) = \frac{1}{Y} \begin{pmatrix} \mathbf{T}(t) & \mathbf{t}(t) \\ \mathbf{0} & 0 \end{pmatrix}, \quad (9.7)$$

where $\mathbf{t}(t) = -\mathbf{T}(t)\mathbf{e}$, $\mathbf{e} = (1, 1, \dots, 1)^T$, $\mathbf{0} = (0, 0, \dots, 0)$, $\boldsymbol{\pi} = (\pi_1, \dots, \pi_p)$,

$$\mathbb{P}(X_0 = p+1) = 0, \quad \mathbb{P}(X_0 = k) = \pi_k, \quad k = 1, \dots, p,$$

and the independent, positive random variable Y is a random scaling factor. For the resulting absorption time

$$\tau = \inf\{t \geq 0 : X_t = p+1\},$$

we write $\tau \sim \text{RIPH}(Y, \boldsymbol{\pi}, \mathbf{T}(t))$. Note that the special case $Y \equiv 1$ corresponds to the IPH class in Albrecher and Bladt (2019).

If furthermore we can write $\mathbf{T}(t) = \lambda(t)\mathbf{T}$, the intensity matrix (9.7) takes the form

$$\boldsymbol{\Lambda}(t) = \frac{\lambda(t)}{Y} \begin{pmatrix} \mathbf{T} & \mathbf{t} \\ \mathbf{0} & 0 \end{pmatrix},$$

in which case we write $\tau \sim \text{RIPH}(Y, \boldsymbol{\pi}, \mathbf{T}, \lambda)$. The following result is then immediate.

Theorem 9.4.1. *Let $\tau \sim \text{RIPH}(Y, \boldsymbol{\pi}, \mathbf{T}, \lambda)$ for a random variable Y with density g . Then the density f and distribution function F of τ are given by*

$$f(x) = \int_0^\infty \lambda(x/v) \boldsymbol{\pi} \exp\left(\int_0^{x/v} \lambda(u) du \mathbf{T}\right) \mathbf{t} \frac{g(v)}{v} dv,$$

$$F(x) = 1 - \int_0^\infty \boldsymbol{\pi} \exp\left(\int_0^{x/v} \lambda(u) du \mathbf{T}\right) \mathbf{e} g(v) dv.$$

Furthermore, if $\lambda(t)$ is a strictly positive function and we define h by

$$h^{-1}(x) = \int_0^x \lambda(t) dt,$$

then

$$\tau \stackrel{d}{=} h(\tau_0) \cdot Y, \tag{9.8}$$

where $\tau_0 \sim \text{PH}(\boldsymbol{\pi}, \mathbf{T})$.

Combining Theorem 9.3.7 and Theorem 9.4.1, the matrix Mittag-Leffler random variable $X \sim \text{MML}(\alpha, \boldsymbol{\pi}, \mathbf{T})$ can hence be interpreted as a particular random scaling of a time-inhomogeneous phase-type distribution $\tau_0 \sim \text{PH}(\boldsymbol{\pi}, \mathbf{T})$ with $h(x) = x^{1/\alpha}$ (translating into $\lambda(t) = \alpha t^{\alpha-1}$) and heavy-tailed random scaling factor $Y = S_\alpha$. As we will illustrate in Section 9.5, this represents a particularly versatile yet simple class of random variables for fitting real data.

Remark 9.4.2. For any random variable W with cumulative distribution function F_W , it is possible to write

$$W \stackrel{d}{=} F_W^{-1}(1 - \exp(-E)) =: h_W(E),$$

where E is a unit mean exponential random variable. For modeling purposes, the rationale in Albrecher and Bladt (2019) can be interpreted as approximating the transformation function h_W before-hand by some function h (in absence of the knowledge of W) and then replacing E with a general PH distribution, providing flexibility for the fit with often explicit formulas for the resulting random variable. The matrix-Pareto distributions defined in Albrecher and Bladt (2019) are then the special case $Y \equiv 1$ and (up to a constant) $h(x) = e^x - 1$ in (9.8), which

entails $\lambda(t) = 1/(1+t)$. The fitting in that case is particularly parsimonious for distributions 'close' to a Pareto distribution (where the distance concept here is then inherited from the distance in the PH domain after the log-transform). The general representation (9.8), in contrast, allows to introduce a potential heavy-tail behavior also through the random scaling factor Y , providing more flexibility for the shape of the function $h(x) = x^{1/\alpha}$ (through the choice of α) in a fitting procedure while keeping the resulting expressions tractable.

Remark 9.4.3. The form (9.8) may also suggest to consider – for modelling purposes – the somewhat simpler case of a PH variable multiplied by a standard Pareto variable with tail index $\beta > 0$, that is $h(x) = x$ and $f_Y(y) = \beta x^{-\beta-1}$, $x \geq 1$. For general $\tau_0 \sim PH(\boldsymbol{\pi}, \mathbf{T})$ it is straightforward to see that then

$$\begin{aligned} f_\tau(t) &= \beta z^{-\beta-1} \int_0^t w^\beta \boldsymbol{\pi} \exp(\mathbf{T}w) \mathbf{t} \, dw \\ &= \beta z^{-\beta-1} m_\beta F_{m_\beta}(t), \end{aligned}$$

where m_β is the β -th moment of τ_0 , and F_{m_β} is its β -th moment distribution.

For $\tau_0 \sim \text{Exp}(\lambda)$ this simplifies to

$$f_\tau(t) = \frac{\beta(\lambda t)^{-\beta/2} \exp(-\lambda t/2) \mathcal{W}_M\left(\frac{\beta}{2}, \frac{\beta}{2} + \frac{1}{2}, \lambda t\right)}{\lambda t(\beta + 1)},$$

where

$$\mathcal{W}_M(k, m, z) = z^{m+1/2} e^{-z/2} \sum_{n=0}^{\infty} \frac{(m-k+\frac{1}{2})_n}{n!(2m+1)_n} z^n$$

is the Whittaker M function and $(x)_n$ is the Pochhammer symbol. Inserting a matrix into the third argument of this function, one may now proceed again with the matrix version of Cauchy's formula and by Jordan decomposition, potentially giving rise to a theory similar to the one for Mittag-Leffler distributions. However, this direction is not the focus of the present paper. In addition, a key difference between the above product construction and the MML distribution will be discussed in Section 9.4.2 below in the context of a non-random path representation, for which the fine properties of the ML distribution play a crucial role.

9.4.2 Semi–Markov framework

Let $E^* = \{1, 2, \dots, p\}$ be a state space and let $\mathbf{Q} = \{q_{ij}\}_{i,j \in E}$ denote a transition matrix for some Markov chain $\{Y_n\}_{n \in \mathbb{N}}$ defined on E^* . We assume that $q_{ii} = 0$ for all i . $\{Y_n\}_{n \in \mathbb{N}}$ will serve as an embedded Markov chain in a Markov renewal process.

Let $\alpha \in (0, 1]$ and $\lambda_i > 0$. For $i = 1, \dots, p$, let T_n^i be i.i.d. random variables with a Mittag–Leffler distribution $\text{ML}(\alpha, \lambda_i)$, where we use the parametrisation such that the density of a generic T^i is given by

$$f_i(x) = \lambda_i x^{\alpha-1} E_{\alpha, \alpha}(-\lambda_i x^\alpha). \quad (9.9)$$

An alternative common parametrisation is obtained in terms of the parameter ρ_i that satisfies $\rho_i^{-\alpha} = \lambda_i$.

We now construct a semi-Markov process $\{X_t\}_{t \geq 0}$ as follows. Let $S_0 = 0$ and

$$S_n = \sum_{i=1}^n T_i^{Y_i}, \quad n \geq 1.$$

Define

$$X_t = \sum_{n=1}^{\infty} Y_{n-1} 1_{\{S_{n-1} \leq t < S_n\}}. \tag{9.10}$$

Then $\{X_t\}_{t \geq 0}$ changes states according to the Markov chain Y_n , S_n denotes the time of the n 'th jump, and the sojourn times in states i are Mittag-Leffler distributed with parameters (α, λ_i) . The construction is illustrated in Figure 9.4.

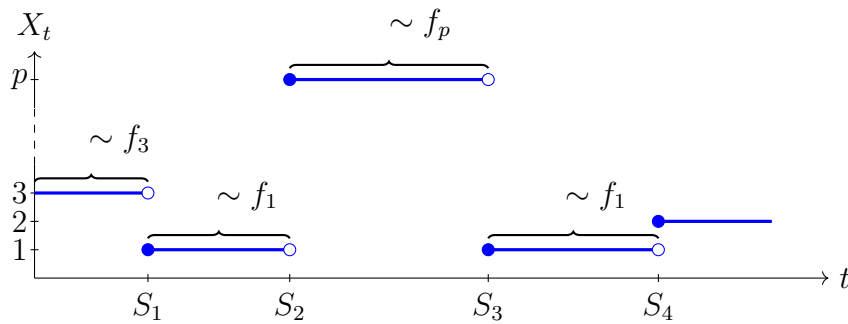


Figure 9.4: Construction of a semi-Markov process based on Mittag-Leffler distributed interarrivals.

Define the intensity matrix $\mathbf{\Lambda} = \{\lambda_{ij}\}_{i=1, \dots, p}$ by

$$\lambda_{ij} = \lambda_i q_{ij}, \quad i \neq j, \quad \text{and} \quad \lambda_{ii} = -\lambda_i = \sum_{k \neq i} \lambda_{ik},$$

and let

$$p_{ij}(t) = \mathbb{P}(X_t = j | X_0 = i), \quad \mathbf{P}(t) = \{p_{ij}(t)\}_{i,j=1, \dots, p}.$$

Theorem 9.4.4. *We have*

$$\mathbf{P}(t) = E_{\alpha,1}(\mathbf{\Lambda}t^\alpha).$$

Proof. Conditioning on the time of the first jump, we get that

$$\begin{aligned} p_{ij}(t) &= \delta_{ij} \mathbb{P}(T_1^i > t) + \int_0^t f_i(s) \sum_{k \neq i} q_{ik} p_{kj}(t-s) \, ds \\ &= \delta_{ij} E_{\alpha,1}(-\lambda_i t^\alpha) + \sum_{k \neq i} q_{ik} \int_0^t \lambda_i s^{\alpha-1} E_{\alpha,\alpha}(-\lambda_i s^\alpha) p_{kj}(t-s) \, ds. \end{aligned}$$

Taking Laplace transforms, and using that

$$\mathcal{L} [x^{\beta-1} E_{\alpha,\beta}(ax^\alpha)](s) = s^{-\beta} (1 - as^{-\alpha})^{-1},$$

together with $\lambda_{ik} = \lambda_i q_{ik}$, we get that

$$\hat{p}_{ij}(s) := \mathbb{E}(e^{-sp_{ij}(t)}) = \delta_{ij} \frac{1}{1 + \lambda_i s^{-\alpha}} + \sum_{k \neq i} \lambda_{ik} \frac{s^{-\alpha}}{1 + \lambda_i s^{-\alpha}} \cdot \hat{p}_{kj}(s)$$

or

$$(1 + \lambda_i s^{-\alpha}) \hat{p}_{ij}(s) = \delta_{ij} + s^{-\alpha} \sum_{k \neq i} \lambda_{ik} \hat{p}_{kj}(s).$$

Now using $\lambda_{ii} = -\lambda_i$, we get

$$\hat{p}_{ij}(s) = \delta_{ij} + s^{-\alpha} \sum_{k=1}^p \lambda_{ik} \hat{p}_{kj}(s). \quad (9.11)$$

In matrix form this amounts to

$$\hat{\mathbf{P}}(s) = \mathbf{I} + s^{-\alpha} \mathbf{\Lambda} \hat{\mathbf{P}}(s)$$

which has the solution

$$\hat{\mathbf{P}}(s) = (\mathbf{I} - s^{-\alpha} \mathbf{\Lambda})^{-1}.$$

The right-hand side is the Laplace transform of $E_{\alpha,1}(\mathbf{\Lambda}t^\alpha)$, establishing the result. \square

Next we consider the case where $E = \{1, 2, \dots, p, p+1\}$ and where the states $1, \dots, p$ are transient and state $p+1$ is absorbing (with respect to the Markov chain $\{Y_n\}_{n \in \mathbb{N}}$). This means that $\{Y_n\}_{n \in \mathbb{N}}$ has a transition matrix of the form

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}^1 & \mathbf{q}^1 \\ \mathbf{0} & 1 \end{pmatrix},$$

and regarding the intensities we set $\lambda_{p+1} = 0$. The matrix $\mathbf{\Lambda}$ then is of the form

$$\mathbf{\Lambda} = \begin{pmatrix} \mathbf{T} & \mathbf{t} \\ \mathbf{0} & 0 \end{pmatrix}. \quad (9.12)$$

We notice the following useful result.

Lemma 9.4.5.

$$E_{\alpha,1} \left(\begin{pmatrix} \mathbf{T} & \mathbf{t} \\ \mathbf{0} & 0 \end{pmatrix} x^\alpha \right) = \begin{pmatrix} E_{\alpha,1}(\mathbf{T}x^\alpha) & \mathbf{e} - E_{\alpha,1}(\mathbf{T}x^\alpha)\mathbf{e} \\ \mathbf{0} & 1 \end{pmatrix},$$

where $\mathbf{t} = -\mathbf{T}\mathbf{e}$ (rows sum to zero).

Proof. By definition,

$$\begin{aligned}
 E_{\alpha,1} \left(\begin{pmatrix} \mathbf{T} & \mathbf{t} \\ \mathbf{0} & 0 \end{pmatrix} x^\alpha \right) &= \sum_{n=0}^{\infty} \begin{pmatrix} \mathbf{T} & \mathbf{t} \\ \mathbf{0} & 0 \end{pmatrix}^n \frac{x^{\alpha n}}{\Gamma(\alpha n + 1)} \\
 &= \mathbf{I} + \sum_{n=1}^{\infty} \begin{pmatrix} \mathbf{T} & \mathbf{t} \\ \mathbf{0} & 0 \end{pmatrix}^n \frac{x^{\alpha n}}{\Gamma(\alpha n + 1)} \\
 &= \mathbf{I} + \sum_{n=1}^{\infty} \begin{pmatrix} \mathbf{T}^n & -\mathbf{T}^n \mathbf{e} \\ \mathbf{0} & 0 \end{pmatrix} \frac{x^{\alpha n}}{\Gamma(\alpha n + 1)} \\
 &= \begin{pmatrix} \mathbf{I} + \sum_{n=1}^{\infty} \mathbf{T}^n \frac{x^{\alpha n}}{\Gamma(\alpha n + 1)} & - \left(\sum_{n=1}^{\infty} \mathbf{T}^n \frac{x^{\alpha n}}{\Gamma(\alpha n + 1)} \mathbf{e} \right) \\ \mathbf{0} & 1 \end{pmatrix} \\
 &= \begin{pmatrix} E_{\alpha,1}(\mathbf{T}x^\alpha) & \mathbf{e} - E_{\alpha,1}(\mathbf{T}x^\alpha)\mathbf{e} \\ \mathbf{0} & 1 \end{pmatrix}.
 \end{aligned}$$

□

Thus the restriction of $E_{\alpha,1}(\mathbf{\Lambda}x^\alpha)$ to the transient states $1, \dots, p$ equals $E_{\alpha,1}(\mathbf{T}x^\alpha)$ and is hence the sub-transition matrix between the transient states.

Theorem 9.4.6. *Let $\{X_t\}_{t \geq 0}$ be a semi-Markov process, where the matrix $\mathbf{\Lambda}$ has the form*

$$\mathbf{\Lambda} = \begin{pmatrix} \mathbf{T} & \mathbf{t} \\ \mathbf{0} & 0 \end{pmatrix}.$$

Let $\tau = \inf\{t \geq 0 : X_t = p + 1\}$ denote the time until absorption. Then τ has a MML($\alpha, \boldsymbol{\pi}, \mathbf{T}$) distribution, with cumulative distribution function given by

$$F_\tau(u) = 1 - \boldsymbol{\pi} E_{\alpha,1}(\mathbf{T}u^\alpha)\mathbf{e}.$$

Proof. For $E^* = \{1, 2, \dots, p\}$, the events $\{\tau > u\}$ and $\{X_u \in E^*\}$ coincide. Thus,

$$\begin{aligned}
 1 - F_\tau(u) &= \mathbb{P}(\tau > u) = \mathbb{P}(X_u \in E^*) = \sum_{j=1}^p \mathbb{P}(X_u = j) \\
 &= \sum_{i,j=1}^p \mathbb{P}(X_u = j | X_0 = i) \mathbb{P}(X_0 = i) = \sum_{i,j=1}^p \pi_i \mathbf{P}_{ij}(u) = \boldsymbol{\pi} E_{\alpha,1}(\mathbf{T}u^\alpha)\mathbf{e}.
 \end{aligned}$$

□

Remark 9.4.7. The proofs above heavily depend on the form of the Laplace transform of the ML distribution, and its similarity with the exponential function. Note that the above construction naturally extends the definition of PH distributions as absorption times of continuous-time Markov chains, the latter being the limit case $\alpha \rightarrow 1$. In general, such a semi-Markov representation will hence not be available for other product distributions.

9.5 Statistical modeling using MML distributions

In this section we present some examples of MML distribution fitting to data. Let us start with an illustration of the maximum likelihood fitting performance to simulated data.

Example 9.5.1. We simulate 300 observations from $X \sim \text{MML}(\alpha, \boldsymbol{\pi}, \mathbf{T})$, with a mixture of Erlang PH component, with parameters chosen in such a way that the log-data is trimodal. The corresponding maximum likelihood fit is depicted in Figure 9.5 (for visualization purposes the scale of the x -axis is logarithmic). The true parameters are $m = 3$, $\alpha = 0.9$, $p_1 = p_2 = p_3 = 3$, $\lambda_1 = 10$, $\lambda_2 = 1$, $\lambda_3 = 0.1$, $\theta_1 = 0.3$, $\theta_2 = 0.3$, $\theta_3 = 0.4$, whereas the maximum likelihood estimator is found to be

$$\begin{aligned} \hat{\alpha} &= 0.905532, & \hat{\theta}_1 &= 0.3133867, & \hat{\theta}_2 &= 0.3138739, & \hat{\theta}_3 &= 0.3727393, \\ & & \hat{\lambda}_1 &= 9.193643, & \hat{\lambda}_2 &= 1.137208, & \hat{\lambda}_3 &= 0.08746225. \end{aligned}$$

The negative log-likelihood at the fitted parameters was 1020.102, compared to 1023.963 at the true parameters (i.e., in the likelihood sense, the fitted model even outperforms the true model for the simulated data points). Note that the tri-modal shape of the underlying density is nicely identified here (which in pure PH fitting would typically not work as smoothly). We also provide a Hill plot of the simulated data, where the potentially heavy-tailed behavior can be clearly appreciated. \square

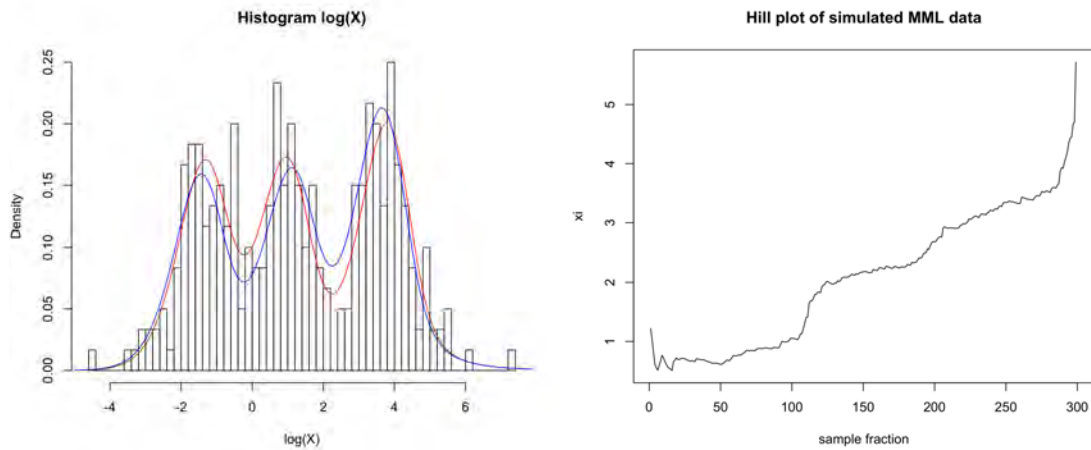


Figure 9.5: Left panel: maximum likelihood fit (red) to simulated MML data with mixture of Erlang PH component and parameters $m = 3$, $\alpha = 0.9$, $p_1 = p_2 = p_3 = 3$, $\lambda_1 = 10$, $\lambda_2 = 1$, $\lambda_3 = 0.1$, $\theta_1 = 0.3$, $\theta_2 = 0.3$, $\theta_3 = 0.4$. The true density is plotted in blue. Right panel: Hill plot of the untransformed data.

The fact that MML distributions behave in a Pareto manner with parameter $\alpha \in (0, 1]$ in the tail can be seen from (9.6). Since such a tail will be too heavy for most applications, we introduce a simple power transformation to gain flexibility for the tail behavior, which particularly allows lighter tails as well.

Definition 9.5.1. Let $X \sim \text{MML}(\alpha, \boldsymbol{\pi}, \mathbf{T})$. For $\nu > 0$, we define

$$X^{1/\nu} \sim \text{PMML}(\alpha, \boldsymbol{\pi}, \mathbf{T}, \nu),$$

and refer to it as the class of Power-MML (PMML) distributions.

The density function of a $\text{PMML}(\alpha, \boldsymbol{\pi}, \mathbf{T}, \nu)$ distribution is given by

$$f(x) = \nu x^{\nu\alpha-1} \boldsymbol{\pi} E_{\alpha, \alpha}(\mathbf{T}x^{\nu\alpha}) \mathbf{t},$$

which will be needed for the maximum likelihood procedure below.

Remark 9.5.2. The introduction of the PMML class allows for an adaptive transformation of the data during the fitting procedure. The interpretation of ν is then as the power to which the data should be taken in order for the latter to be most adequately fit by a pure MML distribution. As the number of MML components grows, the product $\alpha\nu$ is expected to estimate the tail index. However, this estimate might be far off when the matrix \mathbf{T} is not large enough in order for the global fit to be adequate. In those cases, the power transform will tend to improve the fit of the body of the distribution, rather than the tail. When compared to the approach taken in Albrecher and Bladt (2019) (fitting a PH density to log-transformed heavy-tailed variables) one can consider the present procedure as adaptive selection of the transformation function, as opposed to fixing it to be the logarithm.

Example 9.5.2. We consider a real-life motor third party liability (MTPL) insurance data set which was thoroughly studied in Albrecher et al. (2017), mainly from a heavy-tailed perspective (referred to as "Company A" there). The data set originally consists of 837 observations, having the interpretation of claim sizes reported to the company during the time frame 1995-2010. The data are right-censored, and were analyzed recently in Bladt et al. (2020) using perturbed likelihood with censoring techniques. For the present purpose, we solely focus on the *ultimates*, which consists of imputing an expert prediction of the final claim amount for all claims which are still open, i.e. right-censored. We restrict our analysis here to the largest 800 observations, since the inclusion of the 37 smallest claims lead to a sub-optimal fit, but are somehow irrelevant for modeling purposes. For convenience, we divided the claim sizes by 100,000.

The heavy-tailed nature of the data suggests that using MML distributions to model the claim sizes is appropriate. Recently, in Bladt et al. (2019), a tail index of $\alpha^{-1} = 0.48$ was suggested through an automated threshold selection procedure, using a novel trimming approach for the Hill estimator. Since this (or also other much rougher pre-analysis techniques like Pareto QQ-plots) suggests a finite mean, we employ the PMML distributions for the present purpose. This is in fact advised as a general procedure, since in situations where a pure MML fit is appropriate, the fitting procedure will suggest a value for ν close to 1 anyway. The maximum likelihood procedure identifies here a surprisingly simple PMML distribution as adequate, namely with a PH component just being a simple exponential random variable:

$$\hat{\alpha} = 0.3025553, \quad \hat{\mathbf{T}} = -0.08293046, \quad \hat{\nu} = 6.941576.$$

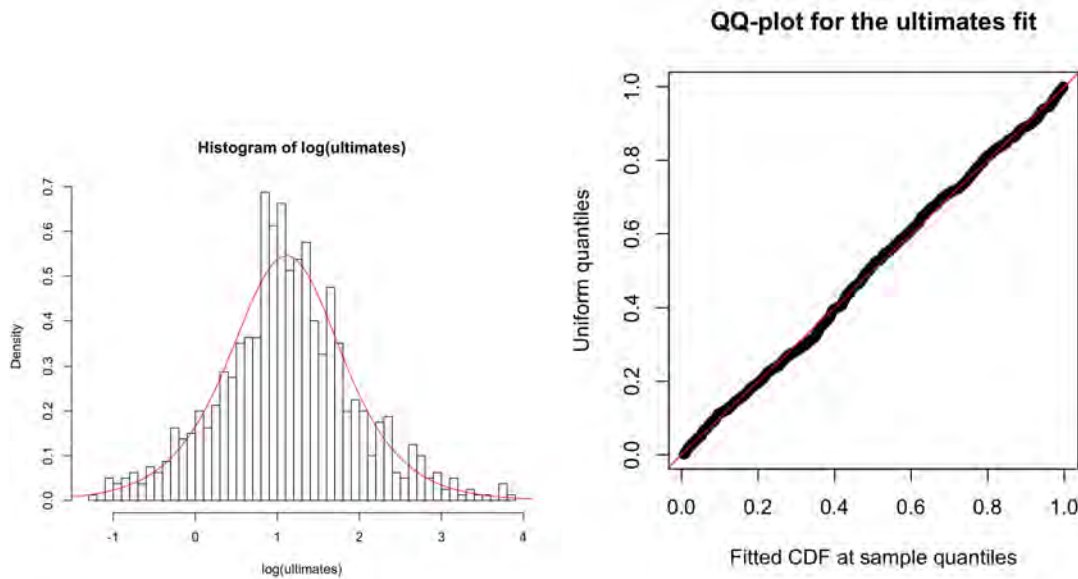


Figure 9.6: Liability insurance ultimates. Left panel: maximum likelihood fit using a PMML with exponential PH component. Right panel: QQ-plot of the fitted distribution function evaluated at the sample quantiles, against theoretical uniform quantiles.

More complex PH components turn out to indeed numerically degenerate into this simple model again. The resulting model density is hence given by

$$f(x) = 0.56 x^{1.1} E_{0.30,0.30}(-0.08x^{2.1}).$$

The adequate fit can be appreciated in Figure 9.6. Observe that the maximum likelihood approach is concerned not only with the tail behaviour but also with adequately fitting the body of the distribution. Nonetheless, the tail index of the PMML fit is given by $(\hat{\alpha} \cdot \hat{\nu})^{-1} = 0.4761427$, which is strikingly(!) close to the 0.48 suggested in Bladt et al. (2019).

A previous approach to describe the entire data set by one model was given in (Albrecher et al., 2017, p.99), where a splicing point was suggested for this data set at around the 20th largest order statistic, based purely on expert opinion. A semi-automated approach in Bladt et al. (2019) suggested splicing at the 14th largest data point. Notice that not only does our model fit the data well and is much more parsimonious, but it also circumvents threshold or splicing point selection completely. \square

Phase-type distributions are weakly dense in the set of all probability distributions on the positive real line. However, often a very large dimension of the PH distribution is needed to get a decent fit to data. Here, we show an example of how the class of PMML distributions can be used to reduce the dimension of a PH fit, thanks to the increased flexibility that the randomization with an α -stable distribution and the power function $(\cdot)^{1/\nu}$ provide.

Example 9.5.3. We consider $n = 500$ simulated data points X_1, \dots, X_n following a mixture of two Erlang(40) distributions. The PH representation is of dimension 80, with parameters

$$\pi_1^0 = \pi_{41}^0 = 0.5, \quad \lambda_1^0 = 100, \quad \lambda_2^0 = 50.$$

Since the data is light-tailed, and PMML are heavy-tailed, we consider the transformed observations

$$Y_i = \exp(X_i) - 1, \quad i = 1, \dots, n, \quad (9.13)$$

which are Pareto in the tail. We then proceed to fit a PMML distribution to the transformed data, but with a much lower matrix dimension. Concretely, we consider a mixture of two Erlang distributions of three phases each for the PH component of the PMML representation. In this way we are led to the maximum likelihood estimates

$$\begin{aligned} \hat{\alpha} &= 0.8649503, & \hat{\pi}_1 &= 0.5386982, & \hat{\pi}_2 &= 0.4613018, \\ \hat{\lambda}_1 &= 25.47413, & \hat{\lambda}_2 &= 1.298168, & \hat{\nu} &= 3.871273. \end{aligned}$$

The (back-transformed) fitted density is plotted in Figure 9.7, along with a histogram of the original PH data points. We also include a fitted density using a pure PH distribution of the same dimension and kind: a mixture of two Erlangs of three phases each. We observe how transforming the data into the heavy-tail domain, fitting a PMML and then back-transforming adds only two extra parameters and improves the estimation dramatically. Additionally, a Hill plot of the transformed data is provided, which shows that the tail index empirically could correspond to $\alpha^{-1} \approx 0.1$, such that it is necessary to use the PMML class, as opposed to only the MML. For reference, the resulting tail is $(\hat{\alpha} \cdot \hat{\nu})^{-1} = 0.223427$, but here the quantification of the tail behaviour is not the main focus of the estimation, and used only qualitatively. In fact, a quick calculation shows that the true index is $\alpha^{-1} = 1/50 = 0.02$, and it is well-known that it is a hard task to estimate tail indices in the transition area between Fréchet and Gumbel domains of attraction. The additional Hill plots of simulated paths of the estimated model in Figure 9.7 show how the hump at the middle of the Hill plot is not a random fluctuation, but rather a systematic feature.

Notice that we took the exponential transformation (9.13) of the simulated data because in this case we know that their tail is exponential and hence the transforms will be regularly varying, in accordance with the PMML distribution. In general, the exponentiation of light-tailed data does not imply that the resulting underlying distribution is regularly varying in the tail. Hence, for real data, a preliminary assessment of the tail behavior and the one of their exponential transforms is recommended.

Remark 9.5.3. When fitting a MML or PMML distribution to data, one has to decide upon the dimension of the underlying phase-type representation. This problem arises similarly when fitting phase-type distributions to light-tailed data, and there are no generally accepted and well established methods for model selection,

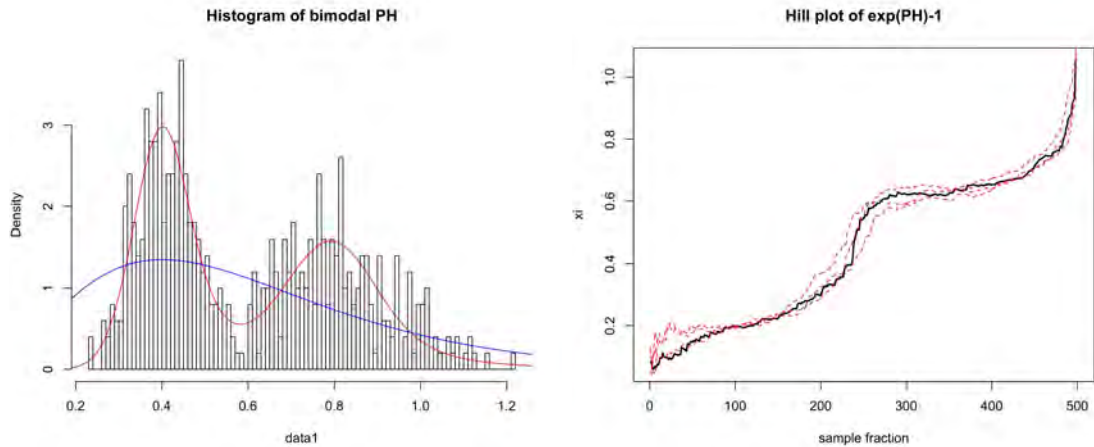


Figure 9.7: Left panel: a back-transformed PMML fit (red) using a 6-dimensional matrix representation, and a pure PH fit (blue) of the same dimension, to an 80-dimensional PH simulated data set. Right panel: Hill plot of the transformed data (black, solid), together with Hill plots for simulated data from the resulting estimated model (red, dashed).

since PH distributions may be well overparametrised so that penalized methods such AIC or BIC indices will not work in general. The order of the PMML (or phase-type distribution) is therefore usually chosen by fitting a range of models of different dimension and then comparing the fit and likelihood values (which, as opposed to information indices, are comparable).

9.6 Conclusion

In this paper we define the class of matrix Mittag-Leffler distributions and derive some of its properties. We identify this class as a particular case of inhomogeneous phase-type distributions under random scaling with a stable law, which together with its power transforms is surprisingly versatile for modeling purposes. In addition, the class is shown to correspond to absorption times of semi-Markov processes with Mittag-Leffler distributed interarrival times, providing a natural extension of the phase-type construction. We illustrate with several examples that this class can simultaneously fit the main body and the tail of a distribution with remarkable accuracy in a parsimonious manner. It turns out that the flexibility of this heavy-tailed class of distributions can even make it worthwhile to transform data into the heavy-tailed domain, fitting the resulting data points and then transforming them back. It will be an interesting direction for future research to further explore the potential of such fitting procedures, both from a theoretical and practical perspective.

Chapter 10

Multivariate Matrix Mittag–Leffler distributions

This chapter is based on the following article (in press):

Albrecher, H., Bladt, M., & Bladt, M. (2020). Multivariate Matrix Mittag-Leffler distributions. *Annals of the Institute of Statistical Mathematics*, to appear.

Abstract

We extend the construction principle of multivariate phase-type distributions to establish an analytically tractable class of heavy-tailed multivariate random variables whose marginal distributions are of Mittag-Leffler type with arbitrary index of regular variation. The construction can essentially be seen as allowing a scalar parameter to become matrix-valued. The class of distributions is shown to be dense among all multivariate positive random variables and hence provides a versatile candidate for the modelling of heavy-tailed, but tail-independent, risks in various fields of application.

10.1 Introduction

The joint modelling of dependent risks is a crucial task in many areas of applied probability and quantitative risk management, see e.g. McNeil et al. (2015). While in many situations there is a reasonable amount of data available for the fitting procedure of univariate risks, the identification of multivariate models is much more delicate. A frequent approach proposed in applications is to use the available data for univariate fitting, and choose a parametric copula to combine the margins, where the parameters of that copula are then either assumed a priori or estimated from the joint data. The choice of such a copula is of course crucial for the resulting joint distribution and the conclusions one draws from it, cf. Mai and Scherer (2017); Mikosch (2006). In multivariate extremes, which is currently a very active research topic, one typically uses less restrictive assumptions for the quantification of joint exceedances, see e.g. Falk et al. (2019); Kiriliouk et al. (2019). Some specific

families, like multivariate regular variation, are considered particularly attractive in this context, as they have a natural interpretation in terms of how to extend univariate behaviour into higher dimensions Ho and Dombry (2019); Joe and Li (2011); Resnick (2002). These results focus, however, on the asymptotic behaviour, so that for a concrete application with an available data set one typically has to choose thresholds above which this respective behaviour is assumed Wan and Davis (2019), and the bulk of the distribution is then to be modelled by a different distribution (see e.g. Beirlant et al. (2004) and (Albrecher et al., 2017, Ch.IV.5)).

In this paper we would like to establish a family of multivariate distributions that can be applied for modeling across the entire positive orthant, so that no threshold selection is needed. In particular, we are interested in a family that leads to explicit and tractable expressions for the model fitting and interpretation. While such a family already exists for marginally light (exponentially bounded) tails in the form of multivariate phase-type (MVPH) distributions, our goal here is to develop a related family with heavy-tailed marginal distributions. The univariate starting point for this procedure is the matrix Mittag-Leffler (MML) distribution, which is a heavy-tailed distribution that was recently studied in Albrecher et al. (2019), and which proved to be very tractable, with excellent fitting properties. While in principle there are many possible ways of defining a vector of random variables with given marginals, we want to consider here the natural concept of multivariate families that can be characterized by the property that any linear combination of the components of such a vector is again of the same marginal type. This is exactly one possible definition of MVPH distributions (so any linear combination of the coordinates of a random vector are again (univariate) phase-type), and it is also a characterizing property of multivariate regular variation of a random vector, namely that any linear combination of the coordinates of such a vector is again (univariate) regularly varying, see Basrak et al. (2002).

The goal is hence to study the class of multivariate random vectors for which such a property applies with MML marginal distributions. It will turn out that for this approach to work, we first need to consider a slightly more general class, which we will refer to as generalized MML distributions. We will show that the analysis developed for the MVPH case can then be extended to our more general situation. In particular, we will establish some properties of this class and work out explicit expressions for a number of concrete cases. The analysis is considerably simpler for the symmetric situation where all marginal distributions share the same index of regular variation, but the general case can be handled as well. The resulting multivariate MML distribution is asymptotically independent, i.e. there is tail-independence for each bivariate pair of components. In the case of multivariate regular variation, the subclass of random vectors with asymptotic independence was studied and characterized in terms of second order conditions in Resnick (2002), where also concrete application areas for such heavy-tailed, but asymptotically independent risks are given. In a sense, the multivariate MML family of distributions we introduce here is another candidate for models in this domain, with the advantage of being explicit and tractable across the entire range \mathbb{R}_+^n . In that respect, this family is also an in-

interesting alternative to multivariate Linnik distributions (see e.g. Anderson (1992) and Lim and Teo (2010a)), which can be conveniently defined in terms of their characteristic function, have the range \mathbb{R}^n (rather than \mathbb{R}_+^n) and also have heavy-tailed marginals, but which do not lead to explicit expressions for the multivariate density.

The remainder of the paper is organized as follows. Section 10.2 recapitulates the construction principle of univariate and multivariate PH distributions and provides the available background on MML distributions. Section 10.3 introduces generalized MML distributions. In Section 10.4 we then develop the necessary theoretical background for our definition of the multivariate MML family and establish some of its properties. We also consider power transforms, which will provide useful flexibility for modeling applications, and we derive denseness properties of the resulting multivariate family. In Section 10.5 we work out a concrete simple example in detail and illustrate resulting dependence properties for this case. Section 10.6 concludes.

10.2 Phase-type distributions

10.2.1 Notation

We shall apply a common convention from phase-type theory that matrices are expressed in bold capital letters (e.g. \mathbf{T} , $\mathbf{\Lambda}$), row vectors are bold minuscule greek letters (e.g. $\boldsymbol{\pi}$, $\boldsymbol{\alpha}$) while column vectors are bold minuscule roman letters (e.g. \mathbf{t} , \mathbf{x}). Elements of matrices and vectors are denoted by their corresponding minuscule unbold letters with indices, e.g. $\mathbf{A} = \{a_{ij}\}$ and $\mathbf{a} = (a_i)$. If $\mathbf{a} = (a_1, \dots, a_n)$ is a vector, then by $\mathbf{\Delta}(\mathbf{a})$ we shall denote the diagonal matrix with \mathbf{a} as diagonal.

10.2.2 Univariate phase-type distributions

Phase-type distributions are defined as the distribution of the time until absorption of a finite state-space Markov jump process with one absorbing state and the other states being transient.

Let p be a positive integer, and $\{X_t\}_{t \geq 0}$ denote a Markov jump process on $E = \{1, \dots, p, p+1\}$, where states $1, 2, \dots, p$ are transient and state $p+1$ is absorbing. Let $\pi_i = \mathbb{P}(X_0 = i)$ and assume that $\pi_1 + \dots + \pi_p = 1$, i.e. initiation in the absorbing state is not possible. The intensity matrix of $\{X_t\}_{t \geq 0}$ can be written as

$$\mathbf{\Lambda} = \begin{pmatrix} \mathbf{T} & \mathbf{t} \\ \mathbf{0} & 0 \end{pmatrix}, \quad (10.1)$$

where \mathbf{T} is the $p \times p$ *sub-intensity* matrix whose off diagonal elements consist of transition rates between the transient states, \mathbf{t} is a p -dimensional column vector $\mathbf{0}$ is a p -dimensional row vector. The diagonal elements of \mathbf{T} are given by $t_{ii} = -\sum_{j \neq i} t_{ij} + t_i$, since the row sums of $\mathbf{\Lambda}$ must be zero.

Let \mathbf{e} denote the vector of ones and $\boldsymbol{\pi} = (\pi_1, \dots, \pi_p)$. Dimensions are usually suppressed and \mathbf{e} may then have any adequate dimension depending on the context.

Then the time until absorption,

$$\tau = \inf\{t \geq 0 : X_t = p+1\},$$

is said to have a phase-type (PH) distribution with representation $(\boldsymbol{\pi}, \mathbf{T})$ and we write $\text{PH}_p(\boldsymbol{\pi}, \mathbf{T})$. Since rows of \mathbf{A} sum to zero, we get $\mathbf{t} = -\mathbf{T}\mathbf{e}$. Note that the case $p = 1$ leads to an exponential distribution.

If $\tau \sim \text{PH}_p(\boldsymbol{\pi}, \mathbf{T})$, then a number of relevant formulas can be written compactly in matrix notation, like e.g.

$$\begin{aligned} f(x; \boldsymbol{\pi}, \mathbf{T}) &= \boldsymbol{\pi} e^{\mathbf{T}x\mathbf{t}}, \quad x > 0, \\ F(x; \boldsymbol{\pi}, \mathbf{T}) &= 1 - \boldsymbol{\pi} e^{\mathbf{T}x\mathbf{t}}, \quad x > 0, \\ L(s; \boldsymbol{\pi}, \mathbf{T}) &= \boldsymbol{\pi} (s\mathbf{I} - \mathbf{T})^{-1}\mathbf{t}, \quad s > \text{Re}(\eta_{\max}), \\ \mathbb{E}(\tau^\alpha) &= \Gamma(\alpha + 1)\boldsymbol{\pi}(-\mathbf{T})^{-\alpha}\mathbf{e}, \quad \alpha > 0, \end{aligned}$$

for the density, c.d.f., Laplace transform and (fractional) moments, respectively. Here η_{\max} denotes the eigenvalue with maximum real part of \mathbf{T} , and this real part is strictly negative. In particular, the Laplace transform is well defined for all $s \geq 0$ and in a neighbourhood around zero.

Remark 10.2.1. Representations $(\boldsymbol{\pi}, \mathbf{T})$ of phase-type distributions are not unique. In fact, one can construct an infinite number of different representations, which may even be of different orders p . Hence phase-type representations may also suffer from over-parametrisation, and it is not possible to attach a specific significance to individual elements of an intensity matrix. While one can typically construct a certain behaviour by means of structuring the sub-intensity matrix \mathbf{T} , the opposite task of deducing such a behaviour from a given matrix is typically not possible. Some simple cases, however, may be described. For instance, $p = 1$ means one phase and the resulting distribution is exponential, hence unimodal. For $p = 2$, bimodality cannot be achieved either, as one could at most aim for a mixture of exponentials. For $p = 3$ it is possible to have a mixture of an exponential with an Erlang(2) which is bimodal.

For further details on phase-type expressions, we refer to Albrecher et al. (2019) and Bladt and Nielsen (2017).

10.2.3 Multivariate phase-type distributions

A non-negative random vector $\mathbf{X} = (X_1, \dots, X_n)$ is phase-type distributed (MVPH) if all non-negative, non-vanishing linear combinations of its coordinates X_i , $i = 1, \dots, n$ have a (univariate) phase-type distribution. This is the most general definition of a multivariate phase-type distribution which, however, lacks practicality since it does not suggest how to construct such distributions. It contains a subclass of multivariate distributions, MPH*, which have multidimensional Laplace transforms of the form

$$L_{\mathbf{X}}(\mathbf{u}; \boldsymbol{\pi}, \mathbf{T}, \mathbf{R}) = \mathbb{E}(e^{-\langle \mathbf{u}, \mathbf{X} \rangle}) = \boldsymbol{\pi} (\boldsymbol{\Delta}(\mathbf{R}\mathbf{u}) - \mathbf{T})^{-1}\mathbf{t}. \quad (10.2)$$

and we write that $\mathbf{X} \sim \text{MPH}^*(\boldsymbol{\pi}, \mathbf{T}, \mathbf{R})$. Here $(\boldsymbol{\pi}, \mathbf{T})$ is a phase-type representation of dimension p , say, \mathbf{R} is a $p \times n$ matrix and $\mathbf{u} = (u_1, \dots, u_n) \in \mathbb{R}_+^n$. Furthermore, the joint Laplace transform exists in a neighbourhood around zero ((Bladt and Nielsen, 2017, Thm.8.1.2)).

The form (10.2) is established from the following probabilistic construction (cf. Kulkarni (1989)). Consider the Markov jump process $\{X_t\}_{t \geq 0}$ underlying the phase-type distribution with representation $(\boldsymbol{\pi}, \mathbf{T})$. The n columns of $\mathbf{R} = \{r_{ik}\}$ are p -dimensional vectors which contain non-negative numbers. These numbers are “rewards” to be earned during sojourns in state i . If τ denotes the time until absorption of the underlying Markov jump process, then

$$X_k = \int_0^\tau \sum_{i=1}^p 1\{X_t = i\} r_{ik} dt, \quad k = 1, \dots, n \tag{10.3}$$

is the total reward earned according to column k of \mathbf{R} until absorption. The structure matrix \mathbf{R} hence picks scaled sojourns out of the underlying Markov jump process. Correlation between different total rewards, X_i and X_j say, will then depend on the structure of \mathbf{R} and on the underlying stochastic process. If there are common states in which reward is earned for both X_i and X_j , then this will contribute to a positive correlation between them. If there are no common states, the correlation will be entirely generated by the structure of the \mathbf{T} matrix. Negative correlation between X_i and X_j is achieved if large rewards earned in one reduces the one earned in the other and vice versa. Specific constructions of dependencies between Phase-type distributed random variables with given marginals is non-trivial, see. e.g. Bladt and Nielsen (2010) for an example with exponentially distributed marginals.

The random variables X_k defined in (10.3) are again phase-type distributed and in general dependent since different variables may be generated through earning positive rewards on certain common states (while in other states there may be zero reward for one variable whenever the other has positive reward). If all $r_{ik} > 0$, $i = 1, \dots, p$, then X_k is phase-type distributed with initial distribution $\boldsymbol{\pi}$ and sub-intensity matrix $\boldsymbol{\Delta}^{-1}(\mathbf{r}_k)\mathbf{T}$. This follows easily from a sample path argument: if reward r_{ik} is earned during a sojourn in state i , then the distribution of the reward during a sojourn is exponentially distributed with intensity $-t_{ii}/r_{ik}$.

If some $r_{ik} = 0$, then finding a representation for X_k is more involved. Let $\mathbf{w} \geq \mathbf{0}$ denote a non-zero vector. For obtaining the k 'th marginal distribution we would choose $\mathbf{w} = \mathbf{e}'_k$, the k 'th Euclidean unit vector, while for a more general projection we may choose $\mathbf{w} = c_1 \mathbf{e}_1 + \dots + c_n \mathbf{e}_k$ for some constants c_i , $i = 1, \dots, n$. For this given \mathbf{w} , decompose the set of transient state $E = \{1, \dots, p\}$ into $E = E_+ \cup E_0$, where E_+ denotes states $i \in E$ for which $(\mathbf{R}\mathbf{w})_i > 0$ and E_0 states $i \in E$ for which $(\mathbf{R}\mathbf{w})_i = 0$. Decompose $\boldsymbol{\pi} = (\boldsymbol{\pi}_+, \boldsymbol{\pi}_0)$ and

$$\mathbf{T} = \begin{pmatrix} \mathbf{T}_{++} & \mathbf{T}_{+0} \\ \mathbf{T}_{0+} & \mathbf{T}_{00} \end{pmatrix} \tag{10.4}$$

accordingly. Then we have the following theorem which is proved in (Bladt and Nielsen, 2017, p.441).

Theorem 10.2.2. *The distribution of $\langle \mathbf{X}, \mathbf{w} \rangle$ is given by an atom at zero of size $q = \boldsymbol{\pi}_0 (\mathbf{I} - (-\mathbf{T}_{00})^{-1} \mathbf{T}_{0+}) \mathbf{e}$ and an absolute continuous part given by a possibly defective phase-type distribution with representation $(\boldsymbol{\pi}_w, \mathbf{T}_w)$, where*

$$\boldsymbol{\pi}_w = \boldsymbol{\pi}_+ + \boldsymbol{\pi}_0 (-\mathbf{T}_{00})^{-1} \mathbf{T}_{0+} \text{ and } \mathbf{T}_w = \boldsymbol{\Delta} ((\mathbf{R}\mathbf{w})_+)^{-1} (\mathbf{T}_{++} + \mathbf{T}_{+0} (-\mathbf{T}_{00})^{-1} \mathbf{T}_{0+})$$

This means that

$$\begin{aligned}\pi(\Delta(\mathbf{R}u\mathbf{w}) - \mathbf{T})^{-1}\mathbf{t} &= \mathbb{E}(e^{-\langle \mathbf{X}, u\mathbf{w} \rangle}) \\ &= \mathbb{E}(e^{-u\langle \mathbf{X}, \mathbf{w} \rangle}) \\ &= q + \pi_{\mathbf{w}}(u\mathbf{I} - \mathbf{T}_{\mathbf{w}})^{-1}\mathbf{t}_{\mathbf{w}},\end{aligned}\quad (10.5)$$

where $\mathbf{t}_{\mathbf{w}} = -\mathbf{T}_{\mathbf{w}}\mathbf{e}$.

Remark 10.2.3. It is still an open question whether $\text{MPH}^* \subset \text{MVPH}$ or whether $\text{MPH}^* = \text{MVPH}$.

Remark 10.2.4. As for univariate phase-type distributions, representations $(\pi, \mathbf{T}, \mathbf{R})$ of MPH^* are not uniquely determined by their distributions, and they may be over-parametrised as well. In particular, the interplay between \mathbf{T} and \mathbf{R} introduces further ambiguity.

While both MPH^* and MPVH distributions lack explicit formulas for distribution and density functions, there is a sub-class of MPH^* distributions that does allow explicit forms. The latter is the one where the structure of the underlying Markov chain is of so-called *feed-forward* type.

Let $\mathbf{C}_1, \dots, \mathbf{C}_n$ be sub-intensity matrices and let $\mathbf{D}_1, \dots, \mathbf{D}_n$ denote non-negative matrices such that $-\mathbf{C}_i\mathbf{e} = \mathbf{D}_i\mathbf{e}$. The matrices \mathbf{D}_i are not necessarily square matrices, with the number of rows being equal to the number of rows in \mathbf{C}_i and the number of columns equal to the number of rows (and columns) of \mathbf{C}_{i+1} . Define

$$\beta = (\pi, \mathbf{0}, \dots, \mathbf{0}) \quad \text{and} \quad \mathbf{T} = \begin{pmatrix} \mathbf{C}_1 & \mathbf{D}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_2 & \mathbf{D}_2 & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{C}_3 & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{C}_n \end{pmatrix} \quad (10.6)$$

and let the reward matrix be

$$\mathbf{R} = \begin{pmatrix} \mathbf{e} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{e} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{e} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{e} \end{pmatrix}. \quad (10.7)$$

The structure of the \mathbf{R} matrix implies that the i 'th total reward, X_i , then equals the inter-arrival time between arrivals $i-1$ and i . Positive correlation between two consecutive inter-arrivals $i-1$ and i can then be obtained by choosing the matrix \mathbf{D}_i in such a way that a long (short) duration of the Markov chain in block $i-1$ will imply a long (short) duration in block i as well. For a negative correlation we have to choose the matrix \mathbf{D}_1 such that the implications are reversed.

Then the joint density of the MPH^* distribution is then given by

$$f(x_1, \dots, x_n; \beta, \mathbf{T}, \mathbf{R}) = \pi e^{\mathbf{C}_1 x_1} \mathbf{D}_1 e^{\mathbf{C}_2 x_2} \mathbf{D}_2 \cdots \mathbf{D}_{n-1} e^{\mathbf{C}_n x_n} \mathbf{D}_n \mathbf{e}. \quad (10.8)$$

Remark 10.2.5. The matrices \mathbf{C}_i are sub-intensity matrices, providing a phase-type distributed time until arrival i . The matrices \mathbf{D}_i are non-negative matrices containing intensities for initiating a new inter-arrival time for arrival $i + 1$ at the time of the arrival i . Hence the matrices \mathbf{D}_i create the dependence between the inter-arrival. In particular, if $\mathbf{D}_i = \mathbf{c}_i \boldsymbol{\pi}_{i+1}$, where $\mathbf{c}_i = -\mathbf{C}_i \mathbf{e}$ is the exit rate (column) vector corresponding to \mathbf{C}_i and $\boldsymbol{\pi}_{i+1}$ is some probability (row) vector on $\{1, 2, \dots, p_i\}$, then the inter-arrivals are independent.

Remark 10.2.6. The (full) matrix \mathbf{D}_n is not really needed for our purposes, but only the exit vector $\mathbf{c}_n = -\mathbf{C}_n \mathbf{e} = \mathbf{D}_n \mathbf{e}$. Thus we may rewrite (10.8) in the form

$$f(x_1, \dots, x_n; \boldsymbol{\beta}, \mathbf{T}, \mathbf{R}) = \boldsymbol{\pi} e^{\mathbf{C}_1 x_1} \mathbf{D}_1 e^{\mathbf{C}_2 x_2} \mathbf{D}_2 \cdots \mathbf{D}_{n-1} e^{\mathbf{C}_n x_n} \mathbf{c}_n, \quad (10.9)$$

We shall, however, maintain the notation with \mathbf{D}_n for notational reasons. Since $-\mathbf{C}_i \mathbf{e} = \mathbf{D}_i \mathbf{e}$ for all i , this also implies the exit vector

$$\mathbf{t} = -\mathbf{T} \mathbf{e} = (0, 0, \dots, 0, \mathbf{c}_n)',$$

so $\mathbf{D}_n \mathbf{e}$, which is not part of \mathbf{T} , is part of \mathbf{t} (see (10.1)).

Remark 10.2.7. Note that the restriction $-\mathbf{C}_i \mathbf{e} = \mathbf{D}_i \mathbf{e}$ reduces the effective number of parameters contributed from those matrices from $2p_i^2$ to $2p_i^2 - p_i$. In particular, the model of (10.9), and therefore also (10.8), has $p_1 - 1 + \sum_{i=1}^{n-1} p_i(2p_i - 1) + p_n^2$ effective degrees of freedom.

Remark 10.2.8. If $\mathbf{C}_i = \mathbf{C}$ and $\mathbf{D}_i = \mathbf{D}$ for all i , then (10.8) is the joint density function for the first n inter-arrival times of a Markovian Arrival Process (MAP) (see e.g. Neuts (1979), Bladt and Nielsen (2017)). This class of point processes is dense in class of point process on \mathbb{R}_+ (see Asmussen and Koole (1993)), and therefore the class of distributions given by (10.8) is also dense – in the sense of weak convergence and with flexible dimension of the matrices \mathbf{C} and \mathbf{D} – in the class of multivariate distributions on \mathbb{R}_+^n .

Later we shall need the joint fractional moments for such distributions, which are given in the following lemma.

Lemma 10.2.9. *Suppose that $\mathbf{X} = (X_1, X_2, \dots, X_n)$ has a joint phase-type distribution with density (10.8). Then for $\theta_i > 0$, $i = 1, \dots, n$,*

$$\mathbb{E}(X_1^{\theta_1} X_2^{\theta_2} \cdots X_n^{\theta_n}) = \left(\prod_{i=1}^n \Gamma(\theta_i + 1) \right) \boldsymbol{\pi} \left(\prod_{i=1}^n (-\mathbf{C}_i)^{-\theta_i - 1} \mathbf{D}_i \right) \mathbf{e}$$

Proof. It is sufficient to prove the lemma for $n = 2$.

$$\begin{aligned} \mathbb{E}(Z_1^{\theta_1} Z_2^{\theta_2}) &= \int_0^\infty \int_0^\infty z_1^{\theta_1} z_2^{\theta_2} \boldsymbol{\pi} e^{\mathbf{C}_1 z_1} \mathbf{D}_1 e^{\mathbf{C}_2 z_2} \mathbf{D}_2 \mathbf{e} \, dz_1 \, dz_2 \\ &= \boldsymbol{\pi} \int_0^\infty z_1^{\theta_1} e^{\mathbf{C}_1 z_1} \, dz_1 \mathbf{D}_1 \int_0^\infty z_2^{\theta_2} e^{\mathbf{C}_2 z_2} \, dz_2 \mathbf{D}_2 \mathbf{e} \\ &= \boldsymbol{\pi} L_{z^{\theta_1}}(-\mathbf{C}_1) \mathbf{D}_1 L_{z^{\theta_2}}(-\mathbf{C}_2) \mathbf{D}_2 \mathbf{e}, \end{aligned}$$

where $L_{z^\theta}(u) = \Gamma(u + 1)/u^{\theta+1}$ is the Laplace transform for $z \rightarrow z^\theta$. Since the Laplace transforms are analytic (where they are defined), the result follows by a functional calculus argument (see Theorem 3.4.4 of Bladt and Nielsen (2017)). \square

10.2.4 Matrix Mittag–Leffler distributions

Let $(\boldsymbol{\pi}, \mathbf{T})$ be a phase–type representation. Then a random variable X has a matrix Mittag–Leffler (MML) distribution with representation $(\alpha, \boldsymbol{\pi}, \mathbf{T})$, if it has Laplace transform

$$L_X(u; \alpha, \boldsymbol{\pi}, \mathbf{T}) = \boldsymbol{\pi} (u^\alpha \mathbf{I} - \mathbf{T})^{-1} \mathbf{t}, \quad u \geq 0,$$

where $0 < \alpha \leq 1$. We write $X \sim \text{MML}(\alpha, \boldsymbol{\pi}, \mathbf{T})$. Let

$$E_{\alpha, \beta}(z) = \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\alpha k + \beta)}, \quad z \in \mathbb{C},$$

denote the Mittag–Leffler (ML) function. Then (see Albrecher et al. (2019)) the density of X is given by

$$f(x; \alpha, \boldsymbol{\pi}, \mathbf{T}) = x^{\alpha-1} \boldsymbol{\pi} E_{\alpha, \alpha}(\mathbf{T}x^\alpha) \mathbf{t}, \quad x > 0,$$

and the corresponding c.d.f. is

$$F(x; \alpha, \boldsymbol{\pi}, \mathbf{T}) = 1 - \boldsymbol{\pi} E_{\alpha, 1}(\mathbf{T}x^\alpha) \mathbf{e}, \quad x > 0.$$

The ML function with (complex) matrix argument \mathbf{A} is defined as

$$E_{\alpha, \beta}(\mathbf{A}) = \sum_{k=0}^{\infty} \frac{\mathbf{A}^k}{\Gamma(\alpha k + \beta)}.$$

For $\beta > 0$, one can express the (then entire) ML function of a matrix \mathbf{A} by Cauchy’s formula

$$E_{\alpha, \beta}(\mathbf{A}) = \frac{1}{2\pi i} \int_{\gamma} E_{\alpha, \beta}(z) (z\mathbf{I} - \mathbf{A})^{-1} dz,$$

where γ is a simple path enclosing the eigenvalues of \mathbf{A} . Invoking the residue theorem, for each entry of the matrix $E_{\alpha, \beta}(z) (z\mathbf{I} - \mathbf{A})^{-1}$, then provides a simple method for calculating $E_{\alpha, \beta}(\mathbf{A})$.

As outlined in Albrecher et al. (2019), MML distributions with $0 < \alpha < 1$ are heavy-tailed with tail indices less than one, so that their mean does not exist. This may be too restrictive in many situations, and one way to obtain a closely related class of distributions is by considering power transformations of the original MML distributed random variables. Indeed, if $X \sim \text{MML}(\alpha, \boldsymbol{\pi}, \mathbf{T})$, then $X^{1/\nu}$ has density

$$f(x; \nu, \alpha, \boldsymbol{\pi}, \mathbf{T}) = \nu x^{\nu\alpha-1} \boldsymbol{\pi} E_{\alpha, \alpha}(\mathbf{T}x^{\nu\alpha}) \mathbf{t}, \quad x > 0,$$

and distribution function

$$F(x; \nu, \alpha, \boldsymbol{\pi}, \mathbf{T}) = 1 - \boldsymbol{\pi} E_{\alpha, 1}(\mathbf{T}x^{\alpha\nu}) \mathbf{e}, \quad x > 0,$$

for $\nu > 0$ (cf. Albrecher et al. (2019)). Rewriting $\beta = \nu\alpha$ leads to the reparametrization

$$f(x; \beta, \alpha, \boldsymbol{\pi}, \mathbf{T}) = \frac{\beta}{\alpha} x^{\beta-1} \boldsymbol{\pi} E_{\alpha, \alpha}(\mathbf{T}x^\beta) \mathbf{t}, \quad x > 0, \quad (10.10)$$

and

$$F(x; \beta, \alpha, \boldsymbol{\pi}, \mathbf{T}) = 1 - \boldsymbol{\pi} E_{\alpha,1}(\mathbf{T}x^\beta)\mathbf{e}, \quad x > 0. \quad (10.11)$$

Thus, for any $0 < \alpha \leq 1$ and $\beta > 0$, (10.10) and (10.11) define densities and their corresponding distribution functions, with tail index β instead of α . We shall refer to distributions with densities of the form (10.10) as power MML and write $X \sim \text{MML}^{1/\nu}(\alpha, \boldsymbol{\pi}, \mathbf{T})$. Their Laplace transforms are somewhat more involved. Indeed, the Laplace transform for $X \sim \text{MML}^{1/\nu}(\alpha, \boldsymbol{\pi}, \mathbf{T})$ is given by (see formula (5.1.30) in Gorenflo et al. (2014) and compare to (Gorenflo et al., 2014, p.364))

$$L_X(s; \nu, \alpha, \boldsymbol{\pi}, \mathbf{T}) = s^{-\nu\alpha} \boldsymbol{\pi} \left(\sum_{k=0}^{\infty} \frac{\Gamma(\nu\alpha(k+1))}{\Gamma(\alpha(k+1))} (s^{-\nu\alpha}\mathbf{T})^k \right) \mathbf{t}, \quad s \geq 0, \quad (10.12)$$

where the series expansion relates to a generalized Wright hypergeometric function (cf. with (Gorenflo et al., 2014, p.364) for further details). The similarity with the Laplace transform for $Y \sim \text{MML}(\alpha, \boldsymbol{\pi}, \mathbf{T})$ may be appreciated by rewriting

$$L_Y(s; \alpha, \boldsymbol{\pi}, \mathbf{T}) = \boldsymbol{\pi}(s^\alpha\mathbf{I} - \mathbf{T})^{-1}\mathbf{t} = s^{-\alpha}\boldsymbol{\pi}(\mathbf{I} - s^{-\alpha}\mathbf{T})^{-1}\mathbf{t}, \quad s \geq 0, \quad (10.13)$$

where we also notice that (10.12) reduces to (10.13) for $\nu = 1$.

10.3 Generalized matrix Mittag–Leffler distributions

The convolution of Mittag–Leffler distributions is not a Mittag–Leffler distribution. However, if the components in the convolution have the same tail index, then the resulting distribution is a MML.

Theorem 10.3.1. *Suppose that $X \sim \text{MML}(\alpha, \boldsymbol{\pi}_1, \mathbf{T}_1)$ and $Y \sim \text{MML}(\alpha, \boldsymbol{\pi}_2, \mathbf{T}_2)$. Then*

$$X + Y \sim \text{MML}(\alpha, \boldsymbol{\pi}, \mathbf{T}),$$

with

$$\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \mathbf{0}) \quad \text{and} \quad \mathbf{T} = \begin{pmatrix} \mathbf{T}_1 & \mathbf{t}_1\boldsymbol{\pi}_2 \\ \mathbf{0} & \mathbf{T}_2 \end{pmatrix}.$$

Proof. This result follows from the Laplace transform of $X + Y$ being

$$\begin{aligned} L_{X+Y}(u; \alpha, \boldsymbol{\pi}, \mathbf{T}) &= \boldsymbol{\pi}_1(u^\alpha\mathbf{I} - \mathbf{T}_1)^{-1}\mathbf{t}_1\boldsymbol{\pi}_2(u^\alpha\mathbf{I} - \mathbf{T}_2)^{-1}\mathbf{t}_2 \\ &= (\boldsymbol{\pi}_1, \mathbf{0}) \begin{pmatrix} (u^\alpha\mathbf{I} - \mathbf{T}_1)^{-1} & -(u^\alpha\mathbf{I} - \mathbf{T}_1)^{-1}(-\mathbf{t}_1\boldsymbol{\pi}_2)(u^\alpha\mathbf{I} - \mathbf{T}_2)^{-1} \\ \mathbf{0} & (u^\alpha\mathbf{I} - \mathbf{T}_2)^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{0} \\ \mathbf{t}_2 \end{pmatrix} \\ &= (\boldsymbol{\pi}_1, \mathbf{0}) \left(u^\alpha\mathbf{I} - \begin{pmatrix} \mathbf{T}_1 & \mathbf{t}_1\boldsymbol{\pi}_2 \\ \mathbf{0} & \mathbf{T}_2 \end{pmatrix} \right)^{-1} \begin{pmatrix} \mathbf{0} \\ \mathbf{t}_2 \end{pmatrix}. \end{aligned}$$

□

Since $X \sim \text{MML}(\alpha, \boldsymbol{\pi}_1, \mathbf{T}_1)$ implies that $cX \sim \text{MML}(\alpha, \boldsymbol{\pi}, \mathbf{T})$ for any constant $c > 0$, where

$$\boldsymbol{\pi} = \boldsymbol{\pi}_1 \quad \text{and} \quad \mathbf{T} = c^{-\alpha} \mathbf{T}_1,$$

we conclude that if X_1, X_2, \dots, X_n are independent MML with the *same* tail index α , then any linear combination $c_1 X_1 + \dots + c_n X_n$ with $c_1, c_2, \dots, c_n \geq 0$ is again MML with tail index α .

The convolution of MML distributions with different tail indices are not MML distributions, but naturally lead to an extended class of MML distributions which we refer to as *Generalized MML*, as we will define in the sequel. If $X \sim \text{MML}(\alpha, \boldsymbol{\pi}_1, \mathbf{T}_1)$ and $Y \sim \text{MML}(\beta, \boldsymbol{\pi}_2, \mathbf{T}_2)$ with $\alpha \neq \beta$, then calculations similar to the proof of Theorem 10.3.1 lead to $X + Y$ having Laplace transform

$$L_{X+Y}(u) = (\boldsymbol{\pi}_1, \mathbf{0}) \left(\boldsymbol{\Delta}(u^\alpha \mathbf{I}, u^\beta \mathbf{I}) - \begin{pmatrix} \mathbf{T}_1 & \mathbf{t}_1 \boldsymbol{\pi}_2 \\ \mathbf{0} & \mathbf{T}_2 \end{pmatrix} \right)^{-1} \begin{pmatrix} \mathbf{0} \\ \mathbf{t}_2 \end{pmatrix}, \quad (10.14)$$

where $\boldsymbol{\Delta}(\mathbf{A}, \mathbf{B})$ denotes the block diagonal matrix

$$\boldsymbol{\Delta}(\mathbf{A}, \mathbf{B}) = \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{pmatrix}$$

for square matrices \mathbf{A} and \mathbf{B} . The linear combination $c_1 X + c_2 Y$ will then have a Laplace transform on the form,

$$L_{c_1 X + c_2 Y}(u) = (\boldsymbol{\pi}_1, \mathbf{0}) \left(\boldsymbol{\Delta}(u^\alpha \mathbf{I}, u^\beta \mathbf{I}) - \begin{pmatrix} c_1^{-\alpha} \mathbf{T}_1 & c_1^{-\alpha} \mathbf{t}_1 \boldsymbol{\pi}_2 \\ \mathbf{0} & c_2^{-\beta} \mathbf{T}_2 \end{pmatrix} \right)^{-1} \begin{pmatrix} \mathbf{0} \\ c_2^{-\beta} \mathbf{t}_2 \end{pmatrix}.$$

This motivates the following definition.

Definition 10.3.2. *A random variable X is said to have a (univariate) generalized matrix Mittag–Leffler distribution, if there exist $\alpha_1, \dots, \alpha_n$ with $0 < \alpha_i \leq 1$, and a phase-type representation $(\boldsymbol{\pi}, \mathbf{T})$ for which the absolutely continuous part of its Laplace transform is given by*

$$L_X^{\text{cont}}(u; \boldsymbol{\alpha}, \boldsymbol{\pi}, \mathbf{T}) = \boldsymbol{\pi} (\boldsymbol{\Delta}(u^{\alpha_1} \mathbf{I}_1, \dots, u^{\alpha_n} \mathbf{I}_n) - \mathbf{T})^{-1} \mathbf{t}, \quad u \geq 0,$$

where \mathbf{I}_k are identity matrices and $\dim(\mathbf{I}_1) + \dots + \dim(\mathbf{I}_n) = \dim(\mathbf{T})$. We write

$$X \sim \text{GMML}(\boldsymbol{\alpha}, \boldsymbol{\pi}, \mathbf{T}),$$

where $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_n) \in \mathbb{R}_+^n$.

Then, if X_1, \dots, X_n are independent with

$$X_i \sim \text{GMML}(\alpha_i, \boldsymbol{\pi}_i, \mathbf{T}_i),$$

we get

$$X_1 + \dots + X_n \sim \text{GMML}(\boldsymbol{\alpha}, \boldsymbol{\pi}, \mathbf{T})$$

where

$$\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_n),$$

$$\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \mathbf{0}, \dots, \mathbf{0}),$$

and

$$\mathbf{T} = \begin{pmatrix} \mathbf{T}_1 & t_1 \boldsymbol{\pi}_2 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_2 & t_2 \boldsymbol{\pi}_3 & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{T}_3 & \dots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ddots & \mathbf{T}_n \end{pmatrix}.$$

By scaling, any non–negative non–zero linear combination of GMML distributed random variables will again follow a GMML distribution.

10.4 The multivariate matrix Mittag–Leffler distribution

Motivated by Section 10.3, we proceed now to define the multivariate MML in a similar way as their underlying multivariate phase–type distributions.

Definition 10.4.1. *A random vector $\mathbf{X} = (X_1, \dots, X_n)$ has a multivariate GMML distribution in the wide sense, if all non–negative non–vanishing linear combinations $c_1 X_1 + \dots + c_n X_n$ have a GMML distribution.*

As for MVPH distributions, this definition is not very practical from a constructive point of view, and we shall introduce a subclass inspired by (10.2). To this end we first notice the following result.

Lemma 10.4.2. *Let $\phi(s_1, \dots, s_k)$ be a multidimensional Laplace transform and let $g_1(x), \dots, g_k(x)$ denote functions for which $-g_i$ are completely monotone. Then it follows that*

$$L(s_1, \dots, s_k) = \phi(g_1(s_1), \dots, g_k(s_k))$$

is again a Laplace transform.

Proof. This follows immediately from the multidimensional Bernstein–Widder theorem, see (Bochner, 2005, p.87), which states that a multivariate function $\phi(s_1, \dots, s_k)$ is a multidimensional Laplace transform if and only if it is infinitely often differentiable and

$$(-1)^{n_1 + \dots + n_k} \frac{\partial^{n_1 + \dots + n_k} \phi}{\partial s_1^{n_1} \dots \partial s_k^{n_k}} \geq 0$$

for all $n_1 \geq 0, \dots, n_k \geq 0$. □

From this we immediately get the following important result.

Theorem 10.4.3. *Let $(\boldsymbol{\pi}, \mathbf{T}, \mathbf{R})$ be a representation for a multivariate PH distribution (10.2). Then the multidimensional function*

$$\phi(\mathbf{u}) = \boldsymbol{\pi} (\boldsymbol{\Delta}(\mathbf{R}\mathbf{u}^\alpha) - \mathbf{T})^{-1} \mathbf{t}, \quad \mathbf{u} \in \mathbb{R}_+^n, \tag{10.15}$$

with $\mathbf{u}^\alpha = (u_1^{\alpha_1}, \dots, u_n^{\alpha_n})$, is a multidimensional Laplace transform.

From Theorem 10.2.2 we now obtain the following.

Theorem 10.4.4. *Let $\mathbf{w} \geq \mathbf{0}$ denote a non-zero vector and let $\mathbf{X} = (X_1, \dots, X_n)$ have a distribution given by the joint Laplace transform (10.15) with all $\alpha_i = \alpha$. Decompose $(\boldsymbol{\pi}, \mathbf{T})$ as in (10.4) according to $\mathbf{R}\mathbf{w}^\alpha$. Then the distribution of $\langle \mathbf{X}, \mathbf{w} \rangle$ has an atom at zero of size $q = \boldsymbol{\pi}_0 (\mathbf{I} - (-\mathbf{T}_{00})^{-1} \mathbf{T}_{0+}) \mathbf{e}$, and a possibly defective absolute continuous part which is $\text{MML}(\alpha, \boldsymbol{\pi}_{\mathbf{w}^\alpha}, \mathbf{T}_{\mathbf{w}^\alpha})$, where $(\boldsymbol{\pi}_{\mathbf{w}^\alpha}, \mathbf{T}_{\mathbf{w}^\alpha})$ is given in Theorem 10.2.2.*

Proof. The result follows from

$$\begin{aligned} \mathbb{E}(e^{-u\langle \mathbf{X}, \mathbf{w} \rangle}) &= \mathbb{E}(e^{-\langle \mathbf{X}, u\mathbf{w} \rangle}) \\ &\stackrel{(10.15)}{=} \boldsymbol{\pi} (\boldsymbol{\Delta}(\mathbf{R}u^\alpha \mathbf{w}^\alpha) - \mathbf{T})^{-1} \mathbf{t} \\ &\stackrel{(10.5)}{=} q + \boldsymbol{\pi}_{\mathbf{w}^\alpha} (u^\alpha \mathbf{I} - \mathbf{T}_{\mathbf{w}^\alpha})^{-1} \mathbf{t}_{\mathbf{w}^\alpha}. \end{aligned}$$

□

For possibly distinct α_i , we proceed as follows.

Theorem 10.4.5. *Let $\mathbf{w} \geq \mathbf{0}$ denote a non-zero vector and let $\mathbf{X} = (X_1, \dots, X_n)$ be a random vector with joint Laplace transform (10.15). Decompose $(\boldsymbol{\pi}, \mathbf{T})$ as in (10.4) according to $\mathbf{R}\mathbf{w}^\alpha$. Then the distribution of $\langle \mathbf{X}, \mathbf{w} \rangle$ has an atom at zero of size $p = \boldsymbol{\pi}_0 (\mathbf{I} - (-\mathbf{T}_{00})^{-1} \mathbf{T}_{0+}) \mathbf{e}$ and a possibly defective absolute continuous part which is $\text{GMML}(\boldsymbol{\alpha}, \boldsymbol{\pi}_{\mathbf{w}^\alpha}, \mathbf{T}_{\mathbf{w}^\alpha})$, where $(\boldsymbol{\pi}_{\mathbf{w}^\alpha}, \mathbf{T}_{\mathbf{w}^\alpha})$ is given in Theorem 10.2.2.*

Proof. We have that

$$\begin{aligned} \mathbb{E}(e^{-u\langle \mathbf{X}, \mathbf{w} \rangle}) &= \mathbb{E}(e^{-\langle \mathbf{X}, u\mathbf{w} \rangle}) \\ &= \boldsymbol{\pi} (\boldsymbol{\Delta}(\mathbf{R}(u\mathbf{w})^\alpha) - \mathbf{T})^{-1} \mathbf{t} \\ &= \boldsymbol{\pi} (\boldsymbol{\Delta}(\mathbf{R}\mathbf{w}^\alpha) \boldsymbol{\Delta}(u^\alpha) - \mathbf{T})^{-1} \mathbf{t}, \end{aligned}$$

where $\boldsymbol{\Delta}(u^\alpha) = \text{diag}(u^{\alpha_1}, \dots, u^{\alpha_n})$. Now splitting into blocks according to E_+ and E_0 , we see that

$$\begin{aligned} \boldsymbol{\pi} (\boldsymbol{\Delta}(\mathbf{R}\mathbf{w}^\alpha) \boldsymbol{\Delta}(u^\alpha) - \mathbf{T})^{-1} \mathbf{t} &= \boldsymbol{\pi} \begin{pmatrix} \boldsymbol{\Delta}(\mathbf{R}\mathbf{w}^\alpha)_+ \boldsymbol{\Delta}(u^\alpha)_+ - \mathbf{T}_{++} & -\mathbf{T}_{+0} \\ -\mathbf{T}_{0+} & -\mathbf{T}_{00} \end{pmatrix}^{-1} \mathbf{t} \\ &= (\boldsymbol{\pi}_+, \boldsymbol{\pi}_0) \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{t}_+ \\ \mathbf{t}_0 \end{pmatrix}, \end{aligned}$$

where

$$\begin{aligned} \mathbf{A}_{11} &= (\boldsymbol{\Delta}(\mathbf{R}\mathbf{w}^\alpha)_+ \boldsymbol{\Delta}(u^\alpha)_+ - \mathbf{T}_{++} - \mathbf{T}_{+0} (-\mathbf{T}_{00})^{-1} \mathbf{T}_{0+})^{-1} \\ &= (\boldsymbol{\Delta}(u^\alpha)_+ - (\boldsymbol{\Delta}(\mathbf{R}\mathbf{w}^\alpha)_+)^{-1} [\mathbf{T}_{++} + \mathbf{T}_{+0} (-\mathbf{T}_{00})^{-1} \mathbf{T}_{0+}])^{-1} \boldsymbol{\Delta}(\mathbf{R}\mathbf{w}^\alpha)_+^{-1} \\ &= (\boldsymbol{\Delta}(u^\alpha)_+ - \mathbf{T}_{\mathbf{w}^\alpha})^{-1} \boldsymbol{\Delta}(\mathbf{R}\mathbf{w}^\alpha)_+^{-1}, \end{aligned}$$

$$\mathbf{A}_{12} = (\boldsymbol{\Delta}(u^\alpha)_+ - \mathbf{T}_{\mathbf{w}^\alpha})^{-1} \boldsymbol{\Delta}(\mathbf{R}\mathbf{w}^\alpha)_+^{-1} \mathbf{T}_{+0} (-\mathbf{T}_{00})^{-1},$$

$$\mathbf{A}_{21} = (-\mathbf{T}_{00})^{-1} \mathbf{T}_{0+} (\boldsymbol{\Delta}(u^\alpha)_+ - \mathbf{T}_{\mathbf{w}^\alpha})^{-1} \boldsymbol{\Delta}(\mathbf{R}\mathbf{w}^\alpha)_+^{-1},$$

$$\mathbf{A}_{22} = (-\mathbf{T}_{00})^{-1} (\mathbf{I} + \mathbf{T}_{0+} (\boldsymbol{\Delta}(u^\alpha)_+ - \mathbf{T}_{\mathbf{w}^\alpha})^{-1} \boldsymbol{\Delta}(\mathbf{R}\mathbf{w}^\alpha)_+^{-1} \mathbf{T}_{+0} (-\mathbf{T}_{00})^{-1}).$$

Then

$$\begin{aligned}\pi_+ \mathbf{A}_{11} + \pi_0 \mathbf{A}_{21} &= \pi_{w^\alpha} (\Delta(u^\alpha)_+ - \mathbf{T}_{w^\alpha})^{-1} \Delta(\mathbf{R}w^\alpha)_+^{-1}, \\ \pi_+ \mathbf{A}_{12} + \pi_0 \mathbf{A}_{22} &= \pi_0 (-\mathbf{T}_{00})^{-1} + \pi_{w^\alpha} (\Delta(u^\alpha)_+ - \mathbf{T}_{w^\alpha})^{-1} \Delta(\mathbf{R}w^\alpha)_+^{-1} \mathbf{T}_{+0} (-\mathbf{T}_{00})^{-1}.\end{aligned}$$

Now inserting

$$\begin{pmatrix} \mathbf{t}_+ \\ \mathbf{t}_0 \end{pmatrix} = -\mathbf{T}\mathbf{e} = \begin{pmatrix} -\mathbf{T}_{++}\mathbf{e} - \mathbf{T}_{+0}\mathbf{e} \\ -\mathbf{T}_{0+}\mathbf{e} - \mathbf{T}_{00}\mathbf{e} \end{pmatrix},$$

we get

$$\begin{aligned}(\pi_+ \mathbf{A}_{11} + \pi_0 \mathbf{A}_{21}) \mathbf{t}_+ + (\pi_+ \mathbf{A}_{12} + \pi_0 \mathbf{A}_{22}) \mathbf{t}_0 \\ &= \pi_0 (\mathbf{I} - (-\mathbf{T}_{00})^{-1} \mathbf{T}_{0+}) \mathbf{e} + \pi_{w^\alpha} (\Delta(u^\alpha)_+ - \mathbf{T}_{w^\alpha})^{-1} \mathbf{t}_{w^\alpha} \\ &= p + \pi_{w^\alpha} (\Delta(u^\alpha)_+ - \mathbf{T}_{w^\alpha})^{-1} \mathbf{t}_{w^\alpha}\end{aligned}$$

with

$$\mathbf{t}_{w^\alpha} = -\mathbf{T}_{w^\alpha} \mathbf{e}.$$

□

From the previous results we see that we have found a sub-class of multivariate matrix Mittag–Leffler distributions with explicit Laplace transform. This allows us to concentrate on this class, and to make the following definition.

Definition 10.4.6. *Let $\mathbf{X} = (X_1, \dots, X_n)$ be a random vector. Then we say that \mathbf{X} has a multivariate matrix generalized Mittag–Leffler distribution if it has joint Laplace transform given by (10.15), and write*

$$\mathbf{X} \sim \text{GMML}(\boldsymbol{\alpha}, \boldsymbol{\pi}, \mathbf{T}, \mathbf{R}).$$

The following result generalizes Theorem 3.6 of Albrecher et al. (2019) to the multivariate case. In particular, it gives the probabilistic interpretation of the GMML class as a family of random vectors whose marginals are absorption times of randomly-scaled, time-inhomogeneous Markov processes. The dependence of the corresponding Markov processes arises from the fact that they are all generated according to a reward structure on an underlying common Markov jump process.

Theorem 10.4.7. *Let $\mathbf{X} = (X_1, \dots, X_n) \sim \text{GMML}(\boldsymbol{\alpha}, \boldsymbol{\pi}, \mathbf{T}, \mathbf{R})$. Then*

$$\mathbf{X} \stackrel{d}{=} \mathbf{W}^{1/\alpha} \bullet \mathbf{S}_\alpha, \tag{10.16}$$

where $\mathbf{W}^{1/\alpha} = (W_1^{1/\alpha_1}, \dots, W_n^{1/\alpha_n})$ with $\mathbf{W} = (W_1, \dots, W_n) \sim \text{MPH}^*(\boldsymbol{\pi}, \mathbf{T}, \mathbf{R})$ (see (10.2)), and where $\mathbf{S}_\alpha = (S_{\alpha_1}, \dots, S_{\alpha_n})$ is a vector of independent stable random variables, each with Laplace transform $\exp(-u^{\alpha_i})$. Here, \bullet refers to component-wise multiplication of vectors.

Proof. We observe that

$$\begin{aligned}
\mathbb{E}(\exp(-\langle \mathbf{u}, \mathbf{W}^{1/\alpha} \bullet \mathbf{S}_\alpha \rangle)) &= \int_{\mathbb{R}_+^n} \mathbb{E}(\exp(-\langle \mathbf{u}, \mathbf{w}^{1/\alpha} \bullet \mathbf{S}_\alpha \rangle)) dF_{\mathbf{W}}(\mathbf{w}) \\
&= \int_{\mathbb{R}_+^n} \exp(-[u_1^{\alpha_1} w_1 + \cdots + u_n^{\alpha_n} w_n]) dF_{\mathbf{W}}(\mathbf{w}) \\
&= \int_{\mathbb{R}_+^n} \exp(-\langle \mathbf{u}^\alpha, \mathbf{w} \rangle) dF_{\mathbf{W}}(\mathbf{w}) \\
&= \boldsymbol{\pi} (\boldsymbol{\Delta}(\mathbf{R}\mathbf{u}^\alpha) - \mathbf{T})^{-1} \mathbf{t},
\end{aligned}$$

which implies the desired representation. \square

Remark 10.4.8. From representation (10.16), we have that the marginals of any multivariate GMML distribution are regularly varying with indices $\alpha_1, \dots, \alpha_n$, all smaller than 1. Moreover, by the multivariate version of Breiman's lemma (cf. Basrak et al. (2002)) and the fact that multivariate phase-type distributions have moments of all orders, it follows that the tail independence structure of the vector \mathbf{S}_α carries over to \mathbf{X} . That is, the multivariate GMML family introduced in this paper has (very) heavy-tailed GMML marginals, but is tail-independent. As mentioned in the introduction, application areas for such models are e.g. given in Resnick (2002).

A consequence of $\alpha_i < 1$ is that the mean does not exist. To alleviate this potential practical drawback, it was proposed in Albrecher et al. (2019) to consider power-transformed variables in the univariate case. In the same way, we propose the following definition.

Definition 10.4.9. Let $\mathbf{X} \sim \text{GMML}(\boldsymbol{\alpha}, \boldsymbol{\pi}, \mathbf{T}, \mathbf{R})$. For $\boldsymbol{\nu} > \mathbf{0}$, we define

$$\mathbf{Y} = \mathbf{X}^{1/\boldsymbol{\nu}} \sim \text{GMML}^{1/\boldsymbol{\nu}}(\boldsymbol{\alpha}, \boldsymbol{\pi}, \mathbf{T}, \mathbf{R}),$$

and refer to it as the class of power multivariate MML distributions.

Under the power transform, the class is in general no longer closed under linear combinations. For fixed $\boldsymbol{\alpha}$, however, it possesses the following denseness property (in contrast to distributions with Laplace transform (10.15)). Here 'dense on \mathbb{R}_+^n ' means dense in the sense of weak convergence among all distributions on \mathbb{R}_+^n .

Theorem 10.4.10. (i) The class of $\text{GMML}(\boldsymbol{\alpha}, \boldsymbol{\pi}, \mathbf{T}, \mathbf{R})$ variables is dense on \mathbb{R}_+^n .

(ii) For any fixed $\boldsymbol{\alpha}$, the class of $\text{GMML}^{1/\boldsymbol{\nu}}(\boldsymbol{\alpha}, \boldsymbol{\pi}, \mathbf{T}, \mathbf{R})$ variables is dense on \mathbb{R}_+^n .

(iii) For any fixed marginal tail indices $\boldsymbol{\alpha} \bullet \boldsymbol{\nu} = \boldsymbol{\gamma}^{-1} > \mathbf{0}$, the class of $\text{GMML}^{1/\boldsymbol{\nu}}(\boldsymbol{\alpha}, \boldsymbol{\pi}, \mathbf{T}, \mathbf{R})$ variables is dense on \mathbb{R}_+^n .

Proof. (i) The statement is evident by noticing that we may choose $\boldsymbol{\alpha} \equiv \mathbf{1}$ and recalling that the class of variables with Laplace transform (10.2) is dense on \mathbb{R}_+^n .

(ii) Let $\mathbf{0} < \boldsymbol{\nu}_1 < \boldsymbol{\nu}_2 < \cdots$ be any increasing and (entry-wise) diverging sequence of vectors and Y be an arbitrary random vector on \mathbb{R}_+^n . Let \mathbf{S}_α be as in Theorem 10.4.7 and notice that $\mathbf{S}_\alpha^{1/\boldsymbol{\nu}_n} \rightarrow \mathbf{1}$. In particular $\mathbf{S}_\alpha^{1/\boldsymbol{\nu}_n} \xrightarrow{d} \mathbf{1}$. Moreover, we may

choose an independent sequence of vectors \mathbf{W}_n with Laplace transforms of the form (10.2) such that $\mathbf{W}_n^{1/\nu_n} \xrightarrow{d} Y$. Applying the continuous mapping theorem, and by the characterization of Theorem 10.4.7, the statement follows.

(iii) Similar to the previous case, let $\mathbf{0} < \boldsymbol{\alpha}_1 < \boldsymbol{\alpha}_2 < \dots$ be an increasing sequence of vectors, converging to $\mathbf{1}$, and set $\boldsymbol{\nu}_n = (\boldsymbol{\gamma} \bullet \boldsymbol{\alpha}_n)^{-1}$. With \mathbf{S}_α as in Theorem 10.4.7 we have that $\mathbf{S}_{\boldsymbol{\alpha}_n}^{1/\nu_n} \xrightarrow{d} \mathbf{1}$. Choosing an independent sequence of vectors \mathbf{W}_n with Laplace transforms of the form (10.2) and with $\mathbf{W}_n^{1/\nu_n} \xrightarrow{d} Y$, the proof is finished as before. \square

Remark 10.4.11. The above result shows how several classes of multivariate Mittag-Leffler distributions and their power transforms are dense in the set of all distributions of the n -dimensional positive orthant. However, since we are dealing with a tail-independent model, the number of phases increases drastically when faced with the need to capture dependence above high thresholds. Heuristically, the tail dependence is only correctly modelled in the limit. This is in some way analogous to the fact that phase-type distributions are dense on all distributions on the positive real line, but they are all light-tailed (of exponential decay), and very large dimensions are needed for approximations of heavy-tailed distributions, cf. Bladt and Nielsen (2017).

10.5 Special structures and examples

From the previous sections, it becomes clear that the tail behavior of the GMML class is determined by the parameters α_i (cf. Remark 10.4.8) and the dependence structure is mainly triggered by the parameters of the reward matrix \mathbf{R} , as these determine joint contributions to the size of each component. The marginal behavior and overall shape in the body of the distribution is then finally implied by the structure of the phase-type components $(\boldsymbol{\pi}, \mathbf{T})$. In particular, the dimension p of the latter also determines the potential for possible multimodalities of the components. In fact, Theorem 10.4.10 on the denseness of $\text{GMML}^{1/\nu}$ distributions on \mathbb{R}_+^n relies (implicitly in part (i)) on the possibility of having arbitrarily large dimension p , a flexibility that is needed for modelling multiple modes, as the latter can require many phases. However, due to the possibly complex interaction of all parameters, one can not uniquely assign the role of each of the parameters to achieve a particular distributional behavior or shape. Moreover, for arbitrary combinations of parameters it is not always possible to get an explicit expression for the density of a GMML distribution (a complication inherited from the phase-type distributions).

We now proceed to give an example of a subclass that, however, does allow an explicit form. To that end, consider the special structure (10.6) and (10.7) for $(\boldsymbol{\pi}, \mathbf{T}, \mathbf{R})$, which in the exponential case led to the density (10.8),

$$f(x_1, \dots, x_n; \boldsymbol{\pi}, \mathbf{T}, \mathbf{R}) = \boldsymbol{\pi} e^{\mathbf{C}_1 x_1} \mathbf{D}_1 e^{\mathbf{C}_2 x_2} \mathbf{D}_2 \cdots \mathbf{D}_{n-1} e^{\mathbf{C}_n x_n} \mathbf{D}_n \mathbf{e}.$$

This choice of $(\boldsymbol{\pi}, \mathbf{T}, \mathbf{R})$, when plugged into (10.15), results in the joint Laplace

transform of $\mathbf{X} \sim \text{GMML}(\boldsymbol{\alpha}, \boldsymbol{\pi}, \mathbf{T}, \mathbf{R})$

$$L_X(\mathbf{u}; \boldsymbol{\theta}) = \beta \begin{pmatrix} u_1^{\alpha_1} \mathbf{I} - \mathbf{C}_1 & -\mathbf{D}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & u_2^{\alpha_2} \mathbf{I} - \mathbf{C}_2 & -\mathbf{D}_2 & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & u_3^{\alpha_3} \mathbf{I} - \mathbf{C}_3 & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & u_n^{\alpha_n} \mathbf{I} - \mathbf{C}_n \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{D}_n \mathbf{e} \end{pmatrix}, \quad (10.17)$$

where we now use the shorthand notation $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\pi}, \mathbf{T}, \mathbf{R})$. For the resulting class of GMML distributions we can derive joint and marginal density functions, but first we notice the following lemma.

Lemma 10.5.1.

$$\int_0^\infty x^{\alpha-1} E_{\alpha, \alpha}(\mathbf{T}x^\alpha) dx = -\mathbf{T}^{-1}.$$

Proof. Since $\lambda \rightarrow \lambda x^{\alpha-1} E_{\alpha, \alpha}(-\lambda x^\alpha)$ is an analytic function, and a density as a function of x , we get that

$$\begin{aligned} \int_0^\infty x^{\alpha-1} E_{\alpha, \alpha}(\mathbf{T}x^\alpha) dx &= \int_0^\infty x^{\alpha-1} \frac{1}{2\pi i} \int_\gamma E_{\alpha, \alpha}(sx^\alpha) (s\mathbf{I} - \mathbf{T})^{-1} ds dx \\ &= \frac{1}{2\pi i} \int_\gamma \left(\int_0^\infty x^{\alpha-1} E_{\alpha, \alpha}(sx^\alpha) dx \right) (s\mathbf{I} - \mathbf{T})^{-1} ds \\ &= \frac{1}{2\pi i} (-s^{-1}) (s\mathbf{I} - \mathbf{T})^{-1} ds \\ &= -\mathbf{T}^{-1}. \end{aligned}$$

□

Remark 10.5.2. The matrix $\mathbf{U} = -\mathbf{T}^{-1}$ is the so-called Green matrix which has the following probabilistic interpretation: The element (i, j) of \mathbf{U} is the expected time that the Markov jump process underlying a phase-type distribution with generator \mathbf{T} spends in state j (prior to absorption) given that it starts in state i .

The main result of this section is as follows.

Theorem 10.5.3. *The Laplace transform (10.17) can equivalently be written as*

$$L_X(\mathbf{u}; \boldsymbol{\theta}) = \pi \left(\prod_{i=1}^n (u_i^{\alpha_i} \mathbf{I} - \mathbf{C}_i)^{-1} \mathbf{D}_i \right) \mathbf{e}, \quad \mathbf{u} \in \mathbb{R}_+^n. \quad (10.18)$$

The corresponding joint density is given by

$$f_X(x_1, \dots, x_n; \boldsymbol{\theta}) = \pi \left(\prod_{i=1}^n x_i^{\alpha_i-1} E_{\alpha_i, \alpha_i}(\mathbf{C}_i x_i^{\alpha_i}) \mathbf{D}_i \right) \mathbf{e}, \quad x_i > 0, \quad i = 1, \dots, n. \quad (10.19)$$

For the i 'th marginal distribution of X_i we have

$$X_i \sim \text{MML}(\alpha_i, \boldsymbol{\beta}_i, \mathbf{C}_i)$$

where

$$\beta_i = \pi \prod_{j=1}^{i-1} (-\mathbf{C}_j)^{-1} \mathbf{D}_j.$$

Proof. It is sufficient to prove the result for $n = 2$. (10.18) follows from the general block diagonal inversion formula

$$\begin{pmatrix} \mathbf{A} & -\mathbf{B} \\ \mathbf{0} & \mathbf{C} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A}^{-1} & \mathbf{A}^{-1}\mathbf{B}\mathbf{C}^{-1} \\ \mathbf{0} & \mathbf{C}^{-1} \end{pmatrix}.$$

Concerning (10.19), we have that

$$\begin{aligned} & \int_0^\infty \int_0^\infty e^{-s_1x_1 - s_2x_2} \pi x_1^{\alpha_1} E_{\alpha_1, \alpha_1}(\mathbf{C}_1 x_1^{\alpha_1}) \mathbf{D}_1 x_2^{\alpha_2} E_{\alpha_2, \alpha_2}(\mathbf{C}_2 x_2^{\alpha_2}) \mathbf{D}_2 \mathbf{e} \, dx_1 \, dx_2 \\ &= \int_0^\infty e^{-s_1x_1} x_1^{\alpha_1} \pi E_{\alpha_1, \alpha_1}(\mathbf{C}_1 x_1^{\alpha_1}) \, dx_1 \mathbf{D}_1 \int_0^\infty e^{-s_2x_2} x_2^{\alpha_2} E_{\alpha_2, \alpha_2}(\mathbf{C}_2 x_2^{\alpha_2}) \mathbf{D}_2 \mathbf{e} \, dx_2 \\ &= \pi (u_1^{\alpha_1} \mathbf{I} - \mathbf{C}_1)^{-1} \mathbf{D}_1 (u_2^{\alpha_2} \mathbf{I} - \mathbf{C}_2)^{-1} \mathbf{D}_2 \mathbf{e} \\ &= (\pi, \mathbf{0}) \begin{pmatrix} u_1^{\alpha_1} \mathbf{I} - \mathbf{C}_1 & -\mathbf{D}_1 \\ \mathbf{0} & u_2^{\alpha_2} \mathbf{I} - \mathbf{C}_2 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{0} \\ \mathbf{D}_2 \mathbf{e} \end{pmatrix}, \end{aligned}$$

which is of the form (10.15).

The result on the marginal distributions follow from Lemma 10.5.1 and by using that $(\mathbf{C}_i + \mathbf{D}_i)\mathbf{e} = \mathbf{0}$, implying that $(-\mathbf{C}_i)^{-1}\mathbf{D}_i\mathbf{e} = \mathbf{e}$. \square

The previous result can be used in the construction of bivariate (or multivariate) Mittag–Leffler distributions of a reasonably general type.

Example 10.5.1 (Bivariate Mittag–Leffler distribution).

In this example we construct a class of bivariate distributions with Mittag–Leffler distributed marginals. The starting point is the construction of a bivariate exponential distribution underlying the MML. For details on this construction we refer to Section 8.3.2 of Bladt and Nielsen (2017). Let m be a positive integer and

$$\mathbf{S} = \begin{pmatrix} -m\lambda & (m-1)\lambda & 0 & \dots & 0 & 0 \\ 0 & -(m-1)\lambda & (m-2)\lambda & \dots & 0 & 0 \\ 0 & 0 & -(m-2)\lambda & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -2\lambda & \lambda \\ 0 & 0 & 0 & \dots & 0 & -\lambda \end{pmatrix}.$$

Then for any initial distribution $\pi = (\pi_1, \dots, \pi_m)$, the phase–type distribution $\text{PH}(\pi, \mathbf{S})$ is simply an exponential distribution with intensity λ . Similarly, if we let

$$\tilde{\mathbf{S}} = \begin{pmatrix} -\mu & \mu & 0 & \dots & 0 & 0 \\ 0 & -2\mu & 2\mu & \dots & 0 & 0 \\ 0 & 0 & -3\mu & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -(m-1)\mu & (m-1)\mu \\ 0 & 0 & 0 & \dots & 0 & -m\mu \end{pmatrix}$$

and $\tilde{\boldsymbol{\pi}} = \frac{1}{m}\mathbf{e} = (\frac{1}{m}, \dots, \frac{1}{m})$, then $\text{PH}(\tilde{\boldsymbol{\pi}}, \tilde{\mathbf{S}})$ is again exponentially distributed with intensity μ . Let \mathbf{P} be a doubly stochastic matrix, i.e. its elements are non-negative and

$$\mathbf{P}\mathbf{e} = \mathbf{e} \quad \text{and} \quad \mathbf{e}'\mathbf{P} = \mathbf{e}',$$

and define

$$\mathbf{T} = \begin{pmatrix} \mathbf{S} & \lambda\mathbf{P} \\ \mathbf{0} & \tilde{\mathbf{S}} \end{pmatrix}.$$

Consider the reward matrix

$$\mathbf{R} = \begin{pmatrix} \mathbf{e} & \mathbf{0} \\ \mathbf{0} & \mathbf{e} \end{pmatrix}.$$

Then $\text{MPH}^*(\mathbf{e}'_1, \mathbf{T}, \mathbf{R})$ is a bivariate exponential distribution. This class of bivariate exponential distributions is capable of achieving any feasible correlation (ranging from $1 - \pi^2/6$ to 1) by choosing m sufficiently large and \mathbf{P} adequately (see Bladt and Nielsen (2010)). Independence is achieved for

$$\mathbf{P} = \frac{1}{m}\mathbf{E},$$

where $\mathbf{E} = \{1\}_{i,j=1,\dots,m}$ is the matrix of ones, maximum negative (minimum) correlation (up to order m) by

$$\mathbf{P} = \mathbf{I}$$

and maximum positive correlation for order up to m by

$$\mathbf{P} = \{\delta_{i,m-i+1}\},$$

which is the anti-diagonal unit matrix, cf. He et al. (2012).

The corresponding GMML($\boldsymbol{\alpha}, \boldsymbol{\pi}, \mathbf{T}, \mathbf{R}$) then has a density f of the form

$$f(x_1, x_2; \boldsymbol{\theta}) = m\lambda\mu x_1^{\alpha_1-1} x_2^{\alpha_2-1} \mathbf{e}'_1 E_{\alpha_1, \alpha_1}(\mathbf{S}x_1^{\alpha_1}) \mathbf{P} E_{\alpha_2, \alpha_2}(\tilde{\mathbf{S}}x_2^{\alpha_2}) \mathbf{e}_n, \quad x_1, x_2 > 0, \quad (10.20)$$

where as usual \mathbf{e}_i denotes the i 'th Euclidian unit vector. The marginals are Mittag-Leffler distributions with densities

$$f_{X_1}(x; \alpha_1, \lambda) = \lambda x^{\alpha_1-1} E_{\alpha_1, \alpha_1}(-\lambda x^{\alpha_1-1}) \quad \text{and} \quad f_{X_2}(x; \alpha_2, \mu) = \mu x^{\alpha_2-1} E_{\alpha_2, \alpha_2}(-\mu x^{\alpha_2-1}),$$

for $x > 0$, which follows directly from the invariance under different representations (parametrisations), or by simple integration and using Lemma 10.5.1. Note that the present dependence structure has a very natural interpretation as a copula constructed in terms of combining marginal order statistics, cf. Baker (2008) and (Bladt and Nielsen, 2017, Sec.8.3.2), here for Mittag-Leffler marginals.

We can write the expression (10.20) slightly more explicit. The eigenvalues of \mathbf{S} are $-m\lambda, -(m-1)\lambda, \dots, -\lambda$. To the eigenvalue $-\lambda k$ there corresponds an eigenvector $\mathbf{v}^{(k)} = (v_1^{(k)}, \dots, v_n^{(k)})$ with

$$\begin{aligned} v_1^{(k)} &= 1 \\ v_{i+1}^{(k)} &= \left(1 - \frac{k-1}{m-i}\right) v_i^{(k)}, \quad i = 1, \dots, m-1. \end{aligned}$$

Similarly, $\tilde{\mathbf{S}}$ has eigenvalues $-\mu m, -\mu(m-1), \dots, -\mu$ and to the eigenvalue $-k\mu$ there corresponds an eigenvector $\mathbf{w}^{(k)}$ with

$$\begin{aligned} w_1^{(k)} &= 1 \\ w_{i+1}^{(k)} &= \left(1 - \frac{k}{i}\right) w_i^{(k)}, \quad i = 1, \dots, m-1. \end{aligned}$$

Considering $\mathbf{v}^{(k)}$ and $\mathbf{w}^{(k)}$ as column vectors, we form the matrices $\mathbf{V} = (\mathbf{v}^{(1)}, \dots, \mathbf{v}^{(m)})$ and $\mathbf{W} = (\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(m)})$. Then we may write

$$\begin{aligned} E_{\alpha_1, \alpha_1}(\mathbf{S}x^{\alpha_1}) &= \mathbf{V} \Delta (E_{\alpha_1, \alpha_1}(-m\lambda x^{\lambda_1}), \dots, E_{\alpha_1, \alpha_1}(-\lambda x^{\alpha_1})) \mathbf{V}^{-1}, \\ E_{\alpha_2, \alpha_2}(\tilde{\mathbf{S}}x^{\alpha_2}) &= \mathbf{W} \Delta (E_{\alpha_2, \alpha_2}(-m\mu x^{\alpha_2}), \dots, E_{\alpha_1, \alpha_1}(-\mu x^{\alpha_2})) \mathbf{W}^{-1}. \end{aligned}$$

Though the correlation between the Mittag-Leffler marginals is not defined (since moments of orders larger than α do not exist), some notion of dependence may be appreciated from the correlation structure of the underlying phase-type distribution.

In Figure 10.1 we depict a bivariate Mittag-Leffler density along with simulated data for the parameters $\boldsymbol{\alpha} = (0.6, 0.7)$, $m = 20$, $\lambda = 1$, $\mu = 2$, and \mathbf{P} the identity matrix.

In Figure 10.2 we use the same parameters but with \mathbf{P} being the counter-identity matrix. As expected, the sign of the log-correlation is determined by the structure of the latter matrix. Notice that the number of effective parameters corresponding to each of the two proposed structures is five.

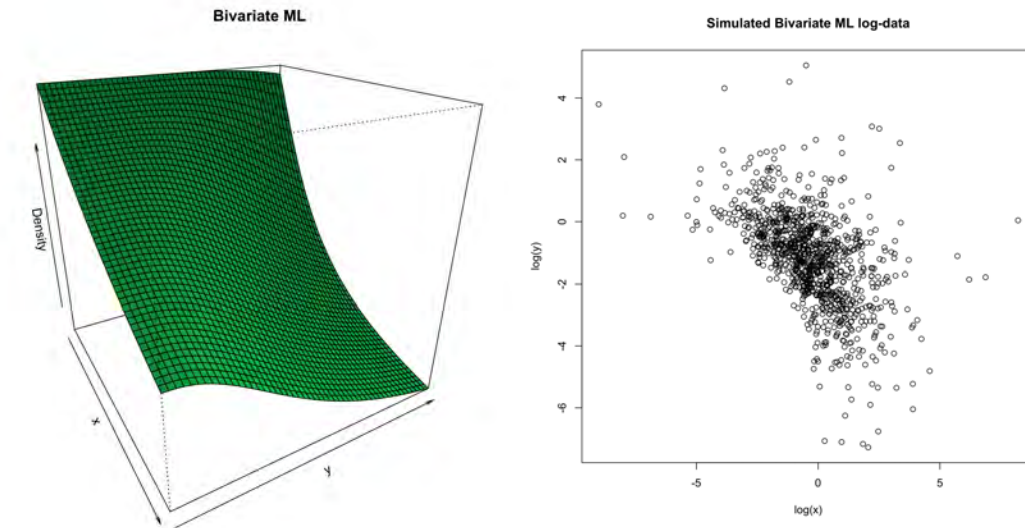


Figure 10.1: Density and 1000 simulated data-points from a bivariate ML distribution with negative log-correlation (empirical correlation of -0.53).

□

Concerning the power MML with this structure we have the following result.

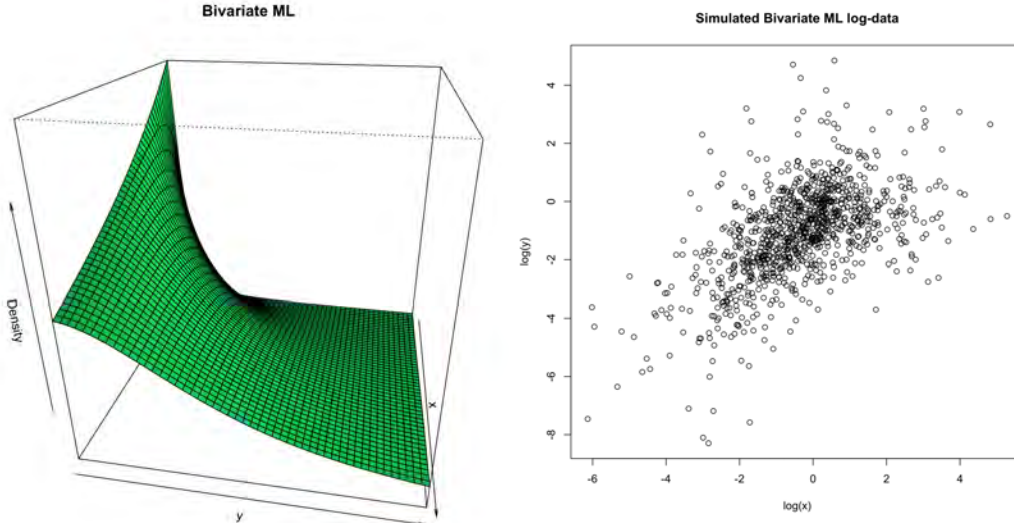


Figure 10.2: Density and 1000 simulated data-points from a bivariate ML distribution with positive correlation (empirical correlation of 0.55).

Theorem 10.5.4. *Assume that \mathbf{X} has joint density (10.19). Then $\mathbf{Y} = \mathbf{X}^{1/\nu}$ has the joint density*

$$f_Y(x_1, \dots, x_n; \boldsymbol{\nu}, \boldsymbol{\theta}) = \pi \left(\prod_{i=1}^n \nu_i x_i^{\alpha_i \nu_i - 1} E_{\alpha_i, \alpha_i}(C_i x_i^{\alpha_i \nu_i}) \mathbf{D}_i \right) \mathbf{e}, \quad x_i > 0, \quad i = 1, \dots, n,$$

and joint moments

$$\begin{aligned} \mathbb{E}(Y_1^{\theta_1} Y_2^{\theta_2} \dots Y_n^{\theta_n}) \\ = \prod_{i=1}^n \left(\frac{\Gamma(1 - \theta_i / (\nu_i \alpha_i)) \Gamma(1 + \theta_i / (\nu_i \alpha_i))}{\Gamma(1 - \theta_i / \nu_i)} \right) \pi \left(\prod_{i=1}^n (-C_i)^{-\theta_i / \nu_i \alpha_i - 1} \mathbf{D}_i \right) \mathbf{e}, \end{aligned}$$

where $\nu_i \alpha_i > \theta_i > 0$, for $i = 1, 2, \dots, n$.

Proof. The form of the joint density is immediate. Concerning the form of the moments, it suffices to consider the case $n = 2$. Using the decomposition (10.4.7), we get

$$\begin{aligned} \mathbb{E}(Y_1^{\theta_1} Y_2^{\theta_2}) &= \mathbb{E} \left(W_1^{\frac{\theta_1}{\alpha_1 \nu_1}} W_2^{\frac{\theta_2}{\alpha_2 \nu_2}} S_{\alpha_1}^{\nu_1} S_{\alpha_2}^{\nu_2} \right) \\ &= \mathbb{E} \left(W_1^{\frac{\theta_1}{\alpha_1 \nu_1}} W_2^{\frac{\theta_2}{\alpha_2 \nu_2}} \right) \mathbb{E} \left(S_{\alpha_1}^{\nu_1} \right) \mathbb{E} \left(S_{\alpha_2}^{\nu_2} \right), \end{aligned}$$

where (W_1, W_2) has a bivariate phase-type distribution with joint density (10.8). Since

$$\mathbb{E} \left(S_{\alpha_i}^{\nu_i} \right) = \frac{\Gamma \left(1 - \frac{\theta_i}{\alpha_i \nu_i} \right)}{\Gamma \left(1 - \frac{\theta_i}{\nu_i} \right)},$$

the result then follows from Lemma 10.2.9. \square

Example 10.5.2. Consider the case of a bivariate MML distribution, $\theta_1 = \theta_2 = 1$, $\nu_i \alpha_i > 1$ and that \mathbf{C}_1 and \mathbf{C}_2 have the same dimension (the latter can always be achieved by augmenting the smaller one). Using the abbreviation

$$c_i = \frac{\Gamma(1 - 1/(\nu_i \alpha_i)) \Gamma(1 + 1/(\nu_i \alpha_i))}{\Gamma(1 - 1/\nu_i)}, \quad i = 1, 2,$$

we get

$$\begin{aligned} \mathbb{E}(Y_1) &= c_1 \boldsymbol{\pi}(-\mathbf{C}_1)^{-1/(\alpha_1 \nu_1)-1} \mathbf{D}_1 \mathbf{e}, \\ \mathbb{E}(Y_2) &= c_2 \boldsymbol{\pi}(-\mathbf{C}_1)^{-1} \mathbf{D}_1 (-\mathbf{C}_2)^{-1/(\alpha_1 \nu_1)-1} \mathbf{D}_2 \mathbf{e}, \\ \mathbb{E}(Y_1 Y_2) &= c_1 c_2 \boldsymbol{\pi}(-\mathbf{C}_1)^{-1/(\alpha_1 \nu_1)-1} \mathbf{D}_1 (-\mathbf{C}_2)^{-1/(\alpha_2 \nu_2)-1} \mathbf{D}_2 \mathbf{e}. \end{aligned}$$

If $\nu_i \alpha_i > 2$ we can calculate variances and correlation. Indeed, with

$$c'_i = \frac{\Gamma(1 - 2/(\nu_i \alpha_i)) \Gamma(1 + 2/(\nu_i \alpha_i))}{\Gamma(1 - 2/\nu_i)}, \quad i = 1, 2,$$

one has

$$\begin{aligned} \mathbb{E}(Y_1^2) &= c'_1 \boldsymbol{\pi}(-\mathbf{C}_1)^{-2/(\alpha_1 \nu_1)-1} \mathbf{D}_1 \mathbf{e} \\ \mathbb{E}(Y_2^2) &= c'_2 \boldsymbol{\pi}(-\mathbf{C}_1^{-1} \mathbf{D}_1) (-\mathbf{C}_1)^{-2/(\alpha_2 \nu_2)-1} \mathbf{D}_2 \mathbf{e} \end{aligned}$$

from which the correlation coefficient is readily calculated.

In Figure 10.3 we depict a bivariate density from a $\text{GMML}^{1/\nu}(\boldsymbol{\alpha}, \boldsymbol{\pi}, \mathbf{T}, \mathbf{R})$ distribution along with simulated data. The parameters are given by

$$\boldsymbol{\alpha} = (0.6, 0.7), \quad \boldsymbol{\beta} = \boldsymbol{\nu} \bullet \boldsymbol{\alpha} = (3, 3),$$

and the phase-type component being of the feed-forward structure (10.6) and (10.7), with $n = 2$, $\boldsymbol{\beta}_1 = (1/3, 1/3, 1/3)$, $\boldsymbol{\beta}_2 = \mathbf{0}$,

$$\mathbf{C}_1 = \mathbf{C}_2 = \begin{pmatrix} -10 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1/10 \end{pmatrix}, \quad \text{and} \quad \mathbf{D}_1 = -\mathbf{C}_1 = \begin{pmatrix} 10 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1/10 \end{pmatrix}.$$

Hence both marginals are mixtures of power Mittag-Leffler distributions. The mixing probabilities of the two distributions are also the same, $(1/3, 1/3, 1/3)$, since the diagonal form of \mathbf{D}_1 ensures that the second mixture draws the same component as the first. The first marginal mixture distribution has a density given by

$$f_1(x) = \frac{5}{3} x^3 \sum_{i=1}^3 \lambda_i E_{0.6, 0.6}(-\lambda_i x^3), \tag{10.21}$$

where $\lambda_1 = 10$, $\lambda_2 = 1$ and $\lambda_3 = 1/10$, while the second marginal density has the form

$$f_2(x) = \frac{10}{7} x^3 \sum_{i=1}^3 \lambda_i E_{0.7, 0.7}(-\lambda_i x^3). \tag{10.22}$$

The reward matrix is

$$\mathbf{R} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}$$

and Y_1 and Y_2 simply correspond to the aforementioned mixtures. The structure of \mathbf{D}_1 implies a strong positive correlation. For example, if Y_1 is picked from the mixture component with rate 10, then Y_2 will be picked from the same component (but then drawn independently).

In Figure 10.4 we use the same parameters, except for

$$\mathbf{D}_1 = \begin{pmatrix} 0 & 0 & 10 \\ 0 & 1 & 0 \\ 1/10 & 0 & 0 \end{pmatrix}.$$

Here the correlation between Y_1 and Y_2 will be negative: if Y_i is drawn from the component with rate 10, then Y_j will be drawn from a component with rate 0.1, $i \neq j$. The marginal distributions are again given by (10.21) and (10.22) since the mixing probabilities are all equal. We observe how the sign of the correlation is affected by the structure of the matrix \mathbf{D}_1 , and the fact that the matrices \mathbf{C}_i are no longer of Erlang structure, the effect is qualitatively opposite to that of the bivariate ML case. One also sees that the class provides quite some flexibility in terms of the shape of the joint density function.

Remark 10.5.5. Dependence may often be constructed by introducing certain structures into the intensity matrices like in Example 10.5.1. More generally, dependence between several random variables of MPH* type may be constructed using the so-called Baker copula (Baker (2008)), where order statistics are used and any feasible correlation structure can be obtained.

10.6 Conclusion

This paper introduces a class GMML of multivariate distributions with matrix Mittag-Leffler distributed marginals. With a construction essentially based on the multivariate phase-type distribution, the GMML class remains a flexible and tractable dense class of distributions maintaining a number of closed form properties. Two important sub-classes are considered, which lead to explicit formulas for distributional properties such as densities and fractional moments. This makes it an attractive candidate for the modelling of both theoretical and practical aspects of multivariate heavy-tailed risks, in situations with tail-independence. The present construction can not be extended to tail-dependent scenarios, so that other approaches will be needed for the latter, which will be an interesting topic for future research.

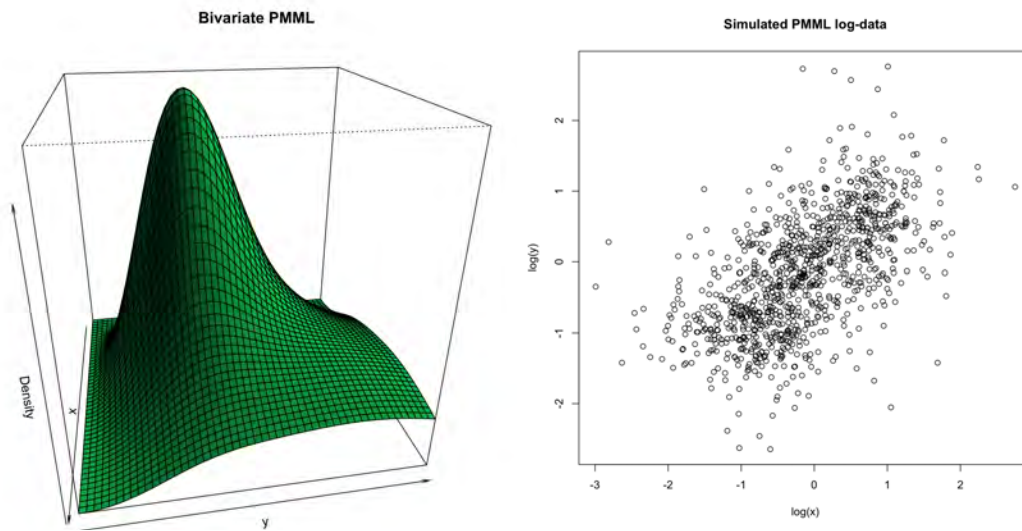


Figure 10.3: Density and 1000 simulated data-points from a power multivariate GMM distribution with positive correlation (true correlation of 0.35 and empirical of 0.37).

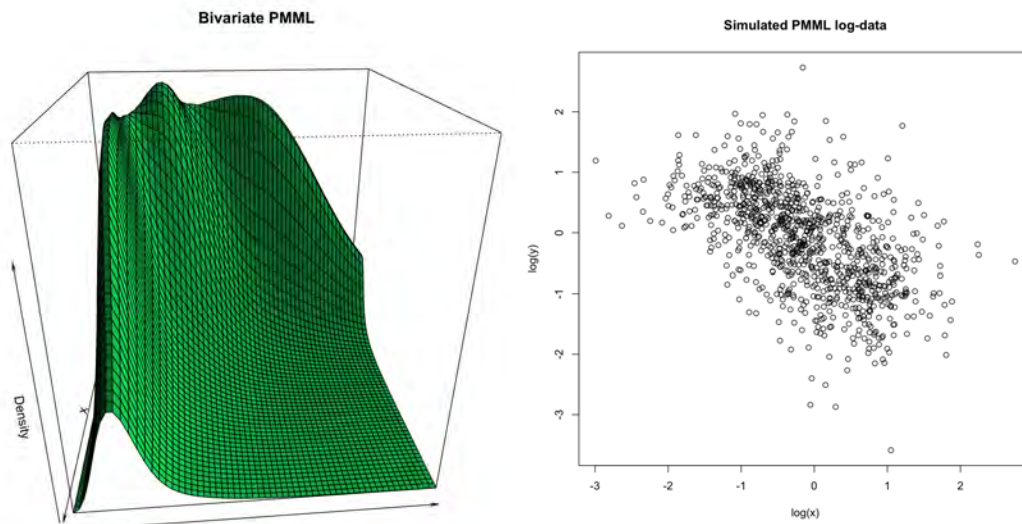


Figure 10.4: Density and 1000 simulated data-points from a power multivariate GMM distribution with negative correlation (true correlation of -0.32 and empirical of -0.33).

Chapter 11

Multivariate fractional phase-type distributions

This chapter is based on the following article, currently submitted for publication:

Albrecher, H., Bladt, M., & Bladt, M. (2020). Multivariate fractional phase-type distributions. arXiv preprint arXiv:2003.11122.

Abstract

We extend the Kulkarni class of multivariate phase-type distributions in a natural time-fractional way to construct a new class of multivariate distributions with heavy-tailed Mittag-Leffler(ML)-distributed marginals. The approach relies on assigning rewards to a non-Markovian jump process with ML sojourn times. This new class complements an earlier multivariate ML construction Albrecher et al. (2020) and in contrast to the former also allows for tail dependence. We derive properties and characterizations of this class, and work out some special cases that lead to explicit density representations.

11.1 Introduction

The formulation of flexible and at the same time parsimonious models for stochastic phenomena is a crucial ingredient in the process of managing risks in various application areas of operations research. On the one hand, a given set of data should be represented reasonably well when putting them into the frame of a calibrated model (and finally replacing them by the latter for the further purposes in the risk analysis). Yet, on the other hand, one needs to avoid overfitting of data and resulting lack of robustness of fitted parameters when applied to updated data sets. In addition, in quite a number of situations (notably in quantitative risk management, see e.g. McNeil et al. (2015)) models are used to extrapolate beyond the range of existing data, and then capturing the main pattern is essential, but overfitting can lead to wrong conclusions about tail properties, particularly in higher dimensions, see also Beirlant et al. (2004). Another important aspect in this context is that it is

quite useful, if models still allow for explicit densities or expressions for the relevant intended measures of risk. This leads to more efficient fitting procedures, and particularly allows to study sensitivities with respect to changes in model parameters in a more explicit way.

In this context, it is quite attractive to have a set of models that a priori is quite general and versatile, but then in the process of fitting the model to actual data reduces to a simpler model in some nested way, if the given data suggest that. One classical example of such a class of models in one dimension are the phase-type distributions (originally introduced by Neuts Neuts (1975)), which builds upon the simplicity of an exponential distribution, but then concatenates exponential ingredients by considering the absorption time of a homogeneous Markov jump process on transient states (phases) into one absorption state with, if needed, many phases and arbitrary intensity matrix (the exponential being the special case of one transient phase only). The gained flexibility is enormous, as the resulting class of phase-type distributions can be shown to be dense (in the sense of weak convergence) in the class of all distributions on the positive half-line (see e.g. Asmussen et al. (1996)). However, the resulting model will only be parsimonious if the underlying risk is close to an exponential structure (e.g. in the tail), as otherwise the number of phases needed for a good fit will be excessive. Yet, on the computational level, the class of phase-type distributions is pleasant, as it can be understood as an (almost exhaustive) subclass of matrix-exponential distributions (that is, an exponential distribution with matrix parameter), for which explicit calculations are available (see e.g. Bladt and Nielsen (2017)). If the underlying risk has a tail heavier than exponential, then it was recently shown in Albrecher and Bladt (2019) that extending the above construction principle to time-inhomogeneous Markov jump processes, adapts the fitting procedure to be built upon other than exponential random variables (namely transforms thereof), and thereby keeps the number of necessary parameters for a good fit very low (essentially leading to matrix-valued parameters of the new base distribution, like Pareto or Weibull). See also Bladt and Rojas-Nandayapa (2018) for another alternative to modelling heavy-tailed data within the phase-type paradigm. Finally, in Albrecher et al. (2019) a random time transformation (based on a stable(α) random variable with $0 < \alpha \leq 1$) in the underlying Markov jump process was considered, which leads to a Mittag-Leffler (ML) distribution as the base distribution, and a resulting flexible family of ML distributions with matrix argument (which later will be referred to as the fractional phase-type class PH_α). The latter is typically heavy-tailed, but contains the phase-type distributions as the limiting special case $\alpha = 1$. Hence the data fitting procedure can decide on which type of model is most suitable for a given data set.

For modelling in more than one dimension, Kulkarni Kulkarni (1989) formulated a multivariate version MPH^* of the phase-type construction by having each component of a random vector collecting different rewards in every state of the (common) Markov jump process, thereby creating possibly dependent phase-type random variables, whose joint Laplace transform is still fully explicit. It could be shown that the resulting family of distributions is again dense in the class of all distributions on the positive orthant. In Albrecher et al. (2020), this multivariate construction was ex-

tended to define a transparent class of multivariate generalized matrix ML (GMML) distributions by applying an independent stable(α_i) random time transformation to each component of the Kulkarni construction. Mathematically, this amounts to a replacement of each argument θ_i in the joint Laplace transform by its power $\theta_i^{\alpha_i}$, leading to explicit expressions for a number of particular cases (see Albrecher et al. (2020) for details). An unfortunate consequence of this procedure is that the resulting multivariate model is necessarily (asymptotically) tail-independent. This can also be seen from an alternative interpretation of the above resulting random vector as the one obtained from stopping each component of a multivariate stable(α_i) Lévy process (with independent components, cf. Kyprianou (2006)) at the (dependent) multivariate phase-type times from the Kulkarni class. However, in many applications one observes possible dependence in the tails, and a proper modelling of that tail dependence is a particular concern in risk management.

In this paper, we propose another way to extend Kulkarni's multivariate phase-type class to formulate a new class MPH_α^* of multivariate Mittag-Leffler distributions that does allow for tail dependence. Concretely, we return to the interpretation of a matrix Mittag-Leffler distributed random variable as the absorption time of a finite state-space semi-Markov process with (state-dependent) ML distributed sojourn times and one absorbing state, see Albrecher et al. (2019). This involves the consideration of Kolmogorov forward equations with fractional derivatives of order α . We then impose the reward structure element of Kulkarni's multivariate construction on this semi-Markov process. Interestingly, the joint Laplace transform of the resulting random vector is again explicit, and on the analytical side differs from the one of the construction in Albrecher et al. (2020) merely by the fact that the power α is applied to the scalar product of each reward vector and the vector of Laplace arguments rather than to the Laplace arguments themselves (with the additional restriction that the value for α in each component now has to be the same). This approach leads to an attractive complement candidate for the modelling of multivariate matrix Mittag-Leffler distributions which allows for dependence in the tail. In a way, the present approach naturally extends Kulkarni's approach onto the appropriate more general semi-Markovian process governed by fractional Kolmogorov forward equations. As compared to the approach in Albrecher et al. (2020), the stretching of time is here applied continuously until absorption, rather than only on the final absorption times, allowing for a different degree of flexibility in the fine structure of the dependence modelling across the different random components. Figure 11.1 depicts the relation between the respective models in the literature, and highlights the fact that the MPH_α^* class proposed here is a natural next step from a conceptual point of view.

The remainder of the paper is organized as follows. In Section 11.2 we review some relevant background on phase-type and (matrix) Mittag-Leffler distributions. Section 11.3 develops the class MPH_α^* of multivariate fractional phase-type distributions as a reward-based multivariate construction using a time-fractional sample path approach with matrix ML distributed marginals. It is shown that this new class (as well as its extension to powers) is itself dense among all distributions on the positive orthant in several ways, and a characterization in terms of a product

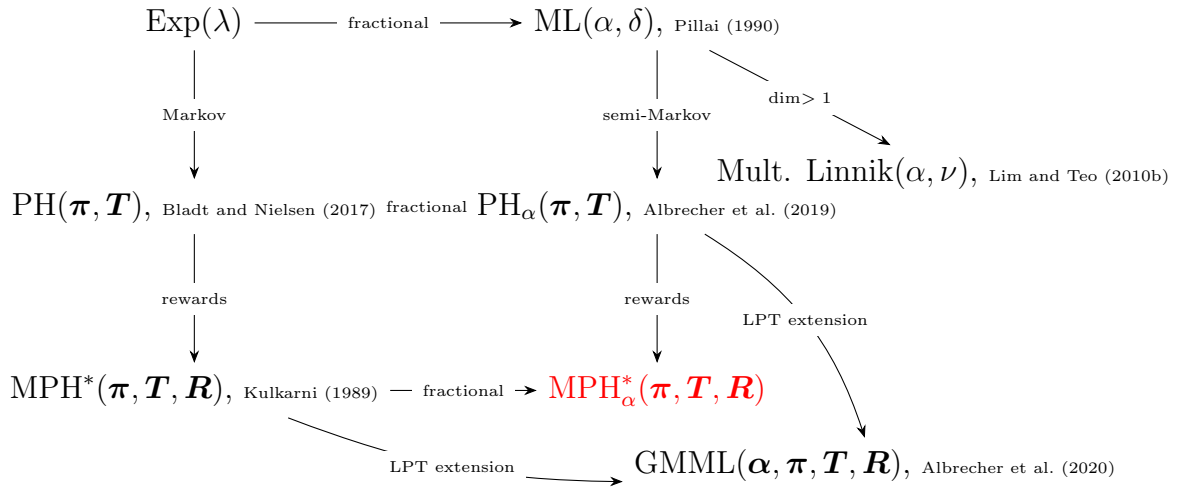


Figure 11.1: Schematic representation of distributions related to multivariate fractional phase-type distributions. Each arrow indicates a generalization.

representation is provided. Finally, it is shown that any linear combination of the random components is again matrix ML distributed (possibly with an additional atom at zero). In Section 11.4 we illustrate two particular cases that lead to explicit density representations. Section 11.5 concludes the paper.

11.2 Background

11.2.1 Phase-type distributions (PH)

Consider a state space $E = \{1, 2, \dots, p, p + 1\}$, and a Markov jump process $\{X_t\}_{t \geq 0}$ evolving on E such that the first p states are transient and the state $p + 1$ is absorbing. The intensity matrix of such a process has the form

$$\Lambda = \begin{pmatrix} \mathbf{T} & \mathbf{t} \\ \mathbf{0} & 0 \end{pmatrix},$$

where \mathbf{T} is a sub-intensity matrix of dimension $p \times p$, consisting of jump rates between the transient states rates. We further specify an initial distribution, concentrated on the transient states $1, \dots, p$, by $\pi_k = \mathbb{P}(X_0 = k)$ for $k = 1, \dots, p$. Thus, if we write $\boldsymbol{\pi} = (\pi_1, \dots, \pi_p)$, we have that $\boldsymbol{\pi} \mathbf{e} = 1$, where \mathbf{e} is the p -dimensional column vector of 1's. We also write by convention

$$\mathbf{t} = -\mathbf{T} \mathbf{e},$$

which is a column vector whose elements are the intensities of jumping to the absorbing state. A phase-type distribution is defined as the absorption time of X_t , that is, if we let

$$\tau = \inf\{t > 0 | X_t = p + 1\},$$

we say that τ follows a phase-type distribution with parameters $\boldsymbol{\pi}, \mathbf{T}$, and write $\tau \sim \text{PH}(\boldsymbol{\pi}, \mathbf{T})$. In general, the parametrization is non-identifiable, in the sense that several initial vectors and sub-intensity matrices can result in the same distribution.

The density and distribution function of $\tau \sim \text{PH}(\boldsymbol{\pi}, \mathbf{T})$ are given by

$$\begin{aligned} f(x) &= \boldsymbol{\pi} e^{\mathbf{T}x} \mathbf{t}, \quad x > 0, \\ F(x) &= 1 - \boldsymbol{\pi} e^{\mathbf{T}x} \mathbf{e}, \quad x > 0, \end{aligned}$$

where the exponential of a matrix M is defined by the formula

$$\exp(\mathbf{M}) = \sum_{n=0}^{\infty} \frac{\mathbf{M}^n}{n!}.$$

The Laplace transform is given by

$$L(u) = \boldsymbol{\pi} (u\mathbf{I} - \mathbf{T})^{-1} \mathbf{t}, \quad (11.1)$$

and is always a rational function, well defined for $u > \text{Re}(\lambda_m)$, where Re denotes the real part and where λ_m is the eigenvalue of \mathbf{T} with largest real part, and \mathbf{I} denotes the identity matrix.

The class of phase-type distributions is closed both under mixing and convolution, which means that also Erlang distributions, Coxian distribution and mixtures thereof are PH distributions. The class is also dense in the class of all distributions on the positive real line (in the sense of weak convergence). This means that any distribution with support on \mathbb{R}_+ may be approximated arbitrarily well by a phase-type distribution (of sufficiently high dimension).

11.2.2 Multivariate phase-type distributions (MPH*)

The class of MPH* was originally introduced in Kulkarni (1989) and is constructed as follows. Let $\tau \sim \text{PH}(\boldsymbol{\pi}, \mathbf{T})$ and let $\{X_t\}_{t \geq 0}$ be the underlying Markov jump. For $i = 1, \dots, n$, let $\mathbf{r}_k = (r_{1k}, r_{2k}, \dots, r_{pk})'$ (column vector) and define

$$Y_k = \int_0^\tau \sum_{i=1}^p r_{ik} 1\{X_t = i\} dt, \quad k = 1, \dots, n.$$

If we interpret r_{ik} as the reward rate earned by the process X_t when it is in state i , then Y_k is the total amount of reward earned according to \mathbf{r}_k prior to absorption. Let \mathbf{R} denote the $p \times n$ matrix

$$\mathbf{R} = (\mathbf{r}_1, \dots, \mathbf{r}_n).$$

whose the columns consist of the different reward rates leading to the variables Y_1, \dots, Y_n . Then we say that $\mathbf{Y} = (Y_1, \dots, Y_n)$ has a multivariate distribution of the MPH* type and we write $\mathbf{Y} \sim \text{MPH}^*(\boldsymbol{\pi}, \mathbf{T}, \mathbf{R})$. The multivariate Laplace transform of $\mathbf{Y} \sim \text{MPH}^*(\boldsymbol{\pi}, \mathbf{T}, \mathbf{R})$ is given by

$$\mathbb{E}(e^{-\langle \mathbf{Y}, \boldsymbol{\theta} \rangle}) = \boldsymbol{\pi} (\boldsymbol{\Delta}(\mathbf{R}\boldsymbol{\theta}) - \mathbf{T})^{-1} \mathbf{t}, \quad (11.2)$$

where $\boldsymbol{\Delta}(\mathbf{v})$ denotes the diagonal matrix which has \mathbf{v} as diagonal.

Multivariate phase-type distributions are dense on \mathbb{R}_+^n , and the marginals and their linear combinations are univariate phase-type distributions, which make them a very flexible and attractive class of distributions for statistical as well as non-statistical applications. However, statistical fitting of this class is still in an experimental stage, since the main dimensionality difficulties of the univariate case are exacerbated with the introduction of the additional parameters of \mathbf{R} .

We refer the reader to Bladt and Nielsen (2017) for a recent comprehensive text on phase-type distributions, both in the uni- and multivariate cases.

11.2.3 Univariate fractional phase-type distributions (PH_α)

A Mittag-Leffler (ML) distribution Pillai (1990) has a density of the form

$$f_{\lambda,\alpha}(x) = \lambda x^{\alpha-1} E_{\alpha,\alpha}(-\lambda x^\alpha), \quad \lambda > 0, \quad 0 < \alpha \leq 1, \quad (11.3)$$

where

$$E_{\alpha,\beta}(z) = \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\alpha k + \beta)}, \quad \beta \in \mathbb{R}, \quad \alpha > 0$$

is the so-called Mittag-Leffler function, and we denote the corresponding class by $\text{ML}(\alpha, \lambda)$. Note that Pillai's definition Pillai (1990) of the ML distribution is recovered with $\rho_i^{-\alpha} = \lambda_i$. For $\alpha = 1$, (11.3) reduces to the density of an exponential random variable. Recently, in Albrecher et al. (2019), a matrix version of the ML distribution with Laplace transform

$$\boldsymbol{\pi}(u^\alpha \mathbf{I} - \mathbf{T})^{-1} \mathbf{t}, \quad 0 < \alpha \leq 1, \quad (11.4)$$

was introduced, which for $\alpha = 1$ reduces to the one of a phase-type distribution (cf. (11.1)). For scalar \mathbf{I} and \mathbf{T} one recovers the classical ML distribution. While the class of distributions with Laplace transform (11.4) was referred to as a *matrix ML* distribution in Albrecher et al. (2019), we suggest to assign to it the additional name *fractional phase-type distribution* ($\text{PH}_\alpha(\boldsymbol{\pi}, \mathbf{T})$), as this will lead to a simple and somewhat more consistent nomenclature in the sequel. As shown in Albrecher et al. (2019), the density and distribution function are given by

$$\begin{aligned} f(x) &= x^{\alpha-1} \boldsymbol{\pi} E_{\alpha,\alpha}(\mathbf{T}x^\alpha) \mathbf{t}, \\ F(x) &= 1 - \boldsymbol{\pi} E_{\alpha,1}(\mathbf{T}x^\alpha) \mathbf{e}, \end{aligned}$$

where

$$E_{\alpha,\beta}(\mathbf{T}x^\alpha) = \sum_{k=0}^{\infty} \frac{\mathbf{T}^k x^{\alpha k}}{\Gamma(\alpha k + \beta)} = \frac{1}{2\pi i} \oint_{\gamma} E_{\alpha,\beta}(zx^\alpha) (z\mathbf{I} - \mathbf{T})^{-1} dz, \quad (11.5)$$

with γ denoting a simple path enclosing the eigenvalues of \mathbf{T} . For $X \sim \text{PH}_\alpha(\boldsymbol{\pi}, \mathbf{T})$ we have the product representation

$$X \stackrel{d}{=} W^{1/\alpha} S_\alpha, \quad (11.6)$$

where $W \sim \text{PH}(\boldsymbol{\pi}, \mathbf{T})$, and S_α is an independent positive stable random variable, cf. Albrecher et al. (2019). Note again that for $\alpha = 1$ we obtain the PH distributions as a special case.

11.3 Multivariate fractional phase-type distributions

11.3.1 The construction

Following Albrecher et al. (2019), we begin by constructing a semi-Markov process which has an absorption time given by a PH_α distribution. Let $E = \{1, 2, \dots, p, p+1\}$ be the state space and let $\mathbf{Q} = \{q_{ij}\}_{i,j \in E}$ denote the transition matrix of a Markov chain $\{Y_n\}_{n \in \mathbb{N}}$ on E , where the first p states are transient and state $p+1$ is absorbing. This means that $\{Y_n\}_{n \in \mathbb{N}}$ has a transition matrix of the form

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}^1 & \mathbf{q}^1 \\ \mathbf{0} & 1 \end{pmatrix}.$$

We assume that $q_{ii} = 0$ for all $i \neq p+1$. This chain will be the embedded Markov chain in a Markov renewal process with Mittag-Leffler distributed holding times defined below. Let $\alpha \in (0, 1]$ and $\lambda_i > 0$. For the states $i = 1, \dots, p$, let T_n^i , $n = 1, 2, \dots$ be independent $\text{ML}(\alpha, \lambda_i)$ -distributed random variables. Let

$$S_n = \sum_{i=1}^n T_i^{Y_i}, \quad n \geq 1,$$

and $S_0 = 0$. Define then the semi-Markov process

$$X_t = \sum_{n=1}^{\infty} Y_{n-1} 1\{S_{n-1} \leq t < S_n\}, \quad t \geq 0. \quad (11.7)$$

The interpretation is that $\{X_t\}_{t \geq 0}$ jumps between states according to the dynamics of the Markov chain Y_n , and S_n denotes the time of the n 'th jump. The holding time in state $i < p+1$ is $\text{ML}(\alpha, \lambda_i)$. The construction is schematically shown in Figure 11.2.

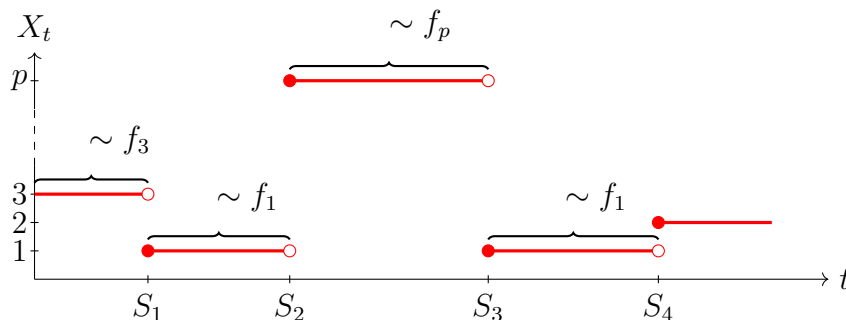


Figure 11.2: Construction of a semi-Markov process based on Mittag-Leffler distributed interarrivals.

Define the intensity matrix $\mathbf{\Lambda} = \{\lambda_{ij}\}_{i=1, \dots, p+1}$ by

$$\lambda_{ij} = \lambda_i q_{ij}, \quad i \neq j, \quad \text{and} \quad \lambda_{ii} = -\lambda_i = \sum_{k \neq i} \lambda_{ik}, \quad i \leq p,$$

and $\lambda_{p+1,i} = 0$, and let

$$p_{ij}(t) = \mathbb{P}(X_t = j | X_0 = i), \quad \mathbf{P}(t) = \{p_{ij}(t)\}_{i,j=1,\dots,p},$$

be the probabilities that describe the dynamics of the process over the transient states. Then we may also write the matrix $\mathbf{\Lambda}$ in the following way

$$\mathbf{\Lambda} = \begin{pmatrix} \mathbf{T} & \mathbf{t} \\ \mathbf{0} & 0 \end{pmatrix}. \tag{11.8}$$

The matrix $\mathbf{\Lambda}$ can be associated with the intensity matrix for some Markov jump process. However, it is important to make the distinction that here we instead consider the semi-Markov process X_t , for which the dynamics on the transient states are not based on the exponential function but rather the Mittag-Leffler function, as is shown in the following result:

Theorem 11.3.1. Albrecher et al. (2019) *Let $\{X_t\}_{t \geq 0}$ be the semi-Markov process constructed above. Then*

$$\mathbf{P}(t) = E_{\alpha,1}(\mathbf{T}t^\alpha).$$

Define the Caputo derivative as the following fractional generalization of the ordinary differentiation operator,

$${}_0^c D_t^\alpha x(t) = \frac{1}{\Gamma(n - \alpha)} \int_0^t (t - \tau)^{n-\alpha-1} x^{(n)}(\tau) d\tau.$$

Then Theorem 11.3.1 yields the following forward and backward type of Kolmogorov fractional differential equations:

Corollary 11.3.2. ${}_0^c D_t^\alpha \mathbf{P}(t) = \mathbf{T}\mathbf{P}(t) = \mathbf{P}(t)\mathbf{T}$.

Proof. It is well-known that the unique solution to the scalar fractional differential equation

$${}_0^c D_t^\alpha x(t) = ax(t), \quad t \geq 0,$$

is given in terms of the Mittag-Leffler function

$$x(t) = E_{\alpha,1}(at^\alpha).$$

The extension to the matrix case now follows from the representation (11.5):

$$\begin{aligned} {}_0^c D_t^\alpha \mathbf{P}(t) &= \frac{1}{2\pi i} \oint_\gamma {}_0^c D_t^\alpha E_{\alpha,\beta}(zt^\alpha)(z\mathbf{I} - \mathbf{T})^{-1} dz \\ &= \frac{1}{2\pi i} \oint_\gamma z E_{\alpha,\beta}(zt^\alpha)(z\mathbf{I} - \mathbf{T})^{-1} dz, \end{aligned}$$

but the latter equals both $\mathbf{T}\mathbf{P}(t)$ and $\mathbf{P}(t)\mathbf{T}$. □

Theorem 11.3.3. Albrecher et al. (2019) *Let $\{X_t\}_{t \geq 0}$ be a semi-Markov process constructed as above, with $\mathbf{\Lambda}$ given by (11.8). Let $\tau = \inf\{t \geq 0 : X_t = p+1\}$ denote the time until absorption. Then τ has a $PH_\alpha(\boldsymbol{\pi}, \mathbf{T})$ distribution, with cumulative distribution function given by*

$$F_\tau(u) = 1 - \boldsymbol{\pi} E_{\alpha,1}(\mathbf{T}u^\alpha)\mathbf{e}.$$

With this representation we are now ready to impose a reward structure on the different states of the process, thereby creating a dependent random vector in a way that extends the MPH* naturally.

For the absorption time τ as defined in Theorem 11.3.3, let now r_{ik} , $i = 1, \dots, p$, $k = 1, \dots, n$ be non-negative numbers and define

$$Y_k = \int_0^\tau \sum_{i=1}^p r_{ik} 1\{X_t = i\} dt, \quad k = 1, \dots, n.$$

Form the column vectors $\mathbf{r}_k = (r_{1k}, r_{2k}, \dots, r_{pk})$, $k = 1, \dots, n$, and matrix

$$\mathbf{R} = (\mathbf{r}_1, \dots, \mathbf{r}_n).$$

The random variable Y_k is interpreted as the total reward earned until absorption of $\{X_t\}$, where r_{ik} is the reward earned during sojourns in state i of the variable k . Hence column k of \mathbf{R} defines a reward structure which defines variable Y_k . See Figure 11.3 for a visual representation of the construction.

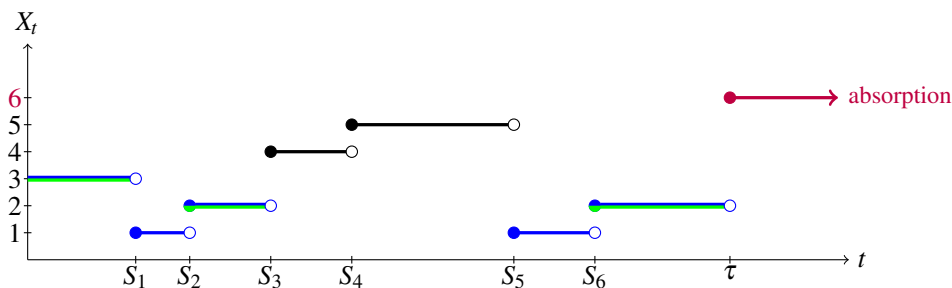


Figure 11.3: Visual representation of the construction of the PH_α^* class. Here, three dimensions are considered: the first collects rewards during the blue holding times, corresponding to the first three states; the second during black holding times (independent of the first, in this case); and the third during green holding times (independent of the second, but not independent of the first).

We are interested in studying the joint distribution of $\mathbf{Y} = (Y_1, \dots, Y_n)$. To this end, let $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)$ and

$$H_i(\boldsymbol{\theta}) = \mathbb{E} (e^{-\langle \mathbf{Y}, \boldsymbol{\theta} \rangle} | X_0 = i).$$

Condition on the first sojourn time Z_{i1} in state i , which has a Mittag-Leffler distribution with parameters (λ_i, α) . Let \mathbf{Y}_{i1} denote the corresponding vector of rewards earned during $[0, Z_{i1})$ and let \mathbf{Y}_r denote the remaining rewards earned during $[Z_{i1}, \tau)$. Then $\mathbf{Y} = \mathbf{Y}_{i1} + \mathbf{Y}_r$ and $\mathbf{Y}_{i1} = Z_{i1} \mathbf{r}_i$. By the Markov renewal property,

$$\begin{aligned} \mathbb{E} (e^{-\langle \mathbf{Y}, \boldsymbol{\theta} \rangle} | X_0 = i) &= \mathbb{E} (e^{-\langle \mathbf{Y}_{i1}, \boldsymbol{\theta} \rangle} | X_0 = i) \mathbb{E} (e^{-\langle \mathbf{Y}_r, \boldsymbol{\theta} \rangle} | X_0 = i, X_{Z_{i1}}) \\ &= \mathbb{E} (e^{-Z_{i1} \langle \mathbf{r}_i, \boldsymbol{\theta} \rangle} | X_0 = i) \mathbb{E} (e^{-\langle \mathbf{Y}_r, \boldsymbol{\theta} \rangle} | X_0 = i, X_{Z_{i1}}) \end{aligned}$$

Since Z_{i1} is Mittag–Leffler distributed with parameters (λ_i, α) , one gets

$$\mathbb{E} \left(e^{-Z_{i1} \langle \mathbf{r}_i, \boldsymbol{\theta} \rangle} \mid X_0 = i \right) = \frac{1}{1 + \langle \mathbf{r}_i, \boldsymbol{\theta} \rangle^\alpha \lambda_i^{-1}}.$$

Recalling that $\mathbf{Q} = \{q_{ij}\}$ contains the transition probabilities for the embedded Markov chain, we then have by a first step argument that

$$H_i(\boldsymbol{\theta}) = \frac{1}{1 + \langle \mathbf{r}_i, \boldsymbol{\theta} \rangle^\alpha \lambda_i^{-1}} \left(q_{i,p+1} + \sum_{j \neq i} q_{ij} H_j(\boldsymbol{\theta}) \right).$$

Using that $t_{ij} = \lambda_i q_{ij}$, $t_i = q_{i,p+1} \lambda_i$ we get that

$$\lambda_i H_i(\boldsymbol{\theta}) + \langle \mathbf{r}_i, \boldsymbol{\theta} \rangle^\alpha H_i(\boldsymbol{\theta}) = t_i + \sum_{j \neq i} t_{ij} H_j(\boldsymbol{\theta})$$

which implies that

$$\langle \mathbf{r}_i, \boldsymbol{\theta} \rangle^\alpha H_i(\boldsymbol{\theta}) = \sum_{j=1}^p t_{ij} H_j(\boldsymbol{\theta}) + t_i.$$

In vector notation, with $\boldsymbol{\Delta}(\mathbf{R}\boldsymbol{\theta})^\alpha$ denoting the diagonal matrix which has $\langle \mathbf{r}_i, \boldsymbol{\theta} \rangle^\alpha$, $i = 1, \dots, p$, on its diagonal, we then write

$$\boldsymbol{\Delta}(\mathbf{R}\boldsymbol{\theta})^\alpha \mathbf{H}(\boldsymbol{\theta}) = \mathbf{T} \mathbf{H}(\boldsymbol{\theta}) + \mathbf{t}$$

or

$$\mathbf{H}(\boldsymbol{\theta}) = (\boldsymbol{\Delta}(\mathbf{R}\boldsymbol{\theta})^\alpha - \mathbf{T})^{-1} \mathbf{t}.$$

If $X_0 \sim \boldsymbol{\pi}$, we then get that the joint Laplace transform for \mathbf{Y} is given by

$$L_{\mathbf{Y}}(\boldsymbol{\theta}) = \boldsymbol{\pi} (\boldsymbol{\Delta}(\mathbf{R}\boldsymbol{\theta})^\alpha - \mathbf{T})^{-1} \mathbf{t}. \quad (11.9)$$

Definition 11.3.4. *The joint distribution of rewards $\mathbf{Y} = (Y_1, \dots, Y_n)$, characterized by its Laplace transform (11.9), is said to have a multivariate fractional phase–type distribution, and we shall denote it by*

$$\mathbf{Y} \sim MPH_\alpha^*(\boldsymbol{\pi}, \mathbf{T}, \mathbf{R}).$$

Remark 11.3.5. Note that the only (yet subtle) difference between the Laplace transform of the GMML distribution introduced in (Albrecher et al., 2020, Eq.15) and the corresponding expression (11.9) above is that the power α is applied after and not before the left-multiplication with the reward matrix. One consequence is that the scalar parameter α represents the regular variation index for all marginals alike, in contrast to the GMML construction in Albrecher et al. (2020), where different values were possible for each component. However, the extension to powers as described in Section 11.3.2 allows to alleviate that issue when desirable.

11.3.2 Denseness properties of the MPH_α^* class and an extension

As members of PH_α , the marginals of the MPH_α^* class all have regularly varying tails with (the same) index $\alpha < 1$ (which in particular entails an infinite mean). In order to allow for more flexibility, a simple extension is to consider (possibly different) powers of each random component, which leads to arbitrary positive index of regular variation for each component.

Let $\mathbf{X} \sim \text{MPH}_\alpha^*(\boldsymbol{\pi}, \mathbf{T}, \mathbf{R})$ with density $f_{\mathbf{X}}(x_1, \dots, x_n)$. Let $\boldsymbol{\nu} = (\nu_1, \dots, \nu_n)$ for $\nu_i > 0$, $i = 1, \dots, n$ and consider the transformed random vector

$$\mathbf{Y} = \mathbf{X}^{1/\boldsymbol{\nu}} = (X_1^{1/\nu_1}, \dots, X_n^{1/\nu_n}),$$

for which the joint density is given by

$$f_{\mathbf{Y}}(y_1, \dots, y_n) = \left(\prod_{i=1}^n \nu_i y_i^{\nu_i-1} \right) f_{\mathbf{X}}(y_1^{\nu_1}, \dots, y_n^{\nu_n}).$$

We refer to this enlarged class as the $\text{MPH}_\alpha^{*1/\boldsymbol{\nu}}$ class. Then we have the following result:

Theorem 11.3.6.

[(i)]

1. The class $\text{MPH}_\alpha^*(\boldsymbol{\pi}, \mathbf{T}, \mathbf{R})$ is dense in the class of distributions on \mathbb{R}_+^n .
2. For any fixed α , the class $\text{MPH}_\alpha^{*1/\boldsymbol{\nu}}(\boldsymbol{\pi}, \mathbf{T}, \mathbf{R})$ is dense in the class of distributions on \mathbb{R}_+^n .
3. For any fixed vector of positive tail indices $(\alpha/\nu_1, \dots, \alpha/\nu_n)$, the class $\text{MPH}_\alpha^{*1/\boldsymbol{\nu}}(\boldsymbol{\pi}, \mathbf{T}, \mathbf{R})$ is dense in the class of distributions on \mathbb{R}_+^n .

Proof. In Section 11.4.1, it will be shown that for the particular subclass of feed-forward type with transition matrix (11.11) and reward matrix (11.12) the identity (11.13) holds and therefore the GMLL and PH_α^* classes agree in that particular case. For this particular structure, the phase-type case $\alpha = 1$ is still dense on \mathbb{R}_+^n , but then the proof of all three items above follows along the same lines as Theorem 4.10 in Albrecher et al. (2020). \square

Notice that (iii) in particular shows that we can approximate any distribution on \mathbb{R}_+^n arbitrarily closely through distributions in $\text{MPH}_\alpha^{*1/\boldsymbol{\nu}}(\boldsymbol{\pi}, \mathbf{T}, \mathbf{R})$ with a pre-specified regularly varying index for each marginal.

11.3.3 A product representation

We proceed to show a representation theorem which sheds some light on the dependence structure of the MPH_α^* class.

Theorem 11.3.7. *Let \mathbf{Y} have Laplace transform (11.9). Then*

$$\mathbf{Y} \stackrel{d}{=} \mathbf{R}^T \mathbf{W}^{1/\alpha} \bullet \mathbf{S}_\alpha, \quad (11.10)$$

where $\mathbf{W}^{1/\alpha} = (W_1^{1/\alpha}, \dots, W_n^{1/\alpha})$ with $\mathbf{W} = (W_1, \dots, W_n) \sim \text{MPH}^*(\boldsymbol{\pi}, \mathbf{T}, \mathbf{I})$ (see (11.2)), and where $\mathbf{S}_\alpha = (S_\alpha^1, \dots, S_\alpha^n)$ is a vector of independent stable random variables, each with Laplace transform $\exp(-u^\alpha)$. Here, \bullet refers to Schur (or entry-wise) multiplication of vectors.

Proof. We first recall that for generic vectors \mathbf{u}, \mathbf{v} we have

$$\langle \mathbf{R}\mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{u}, \mathbf{R}^T \mathbf{v} \rangle,$$

from which

$$\begin{aligned} \mathbb{E}(\exp(-\langle \mathbf{u}, \mathbf{R}^T \mathbf{W}^{1/\alpha} \bullet \mathbf{S}_\alpha \rangle)) &= \int_{\mathbb{R}_+^n} \mathbb{E}(\exp(-\langle \mathbf{R}\mathbf{u}, \mathbf{w}^{1/\alpha} \bullet \mathbf{S}_\alpha \rangle)) dF_{\mathbf{W}}(\mathbf{w}) \\ &= \int_{\mathbb{R}_+^n} \exp(-[(\mathbf{R}\mathbf{u})_1^\alpha w_1 + \dots + (\mathbf{R}\mathbf{u})_n^\alpha w_n]) dF_{\mathbf{W}}(\mathbf{w}) \\ &= \int_{\mathbb{R}_+^n} \exp(-\langle (\mathbf{R}\mathbf{u})^\alpha, \mathbf{w} \rangle) dF_{\mathbf{W}}(\mathbf{w}) \\ &= \boldsymbol{\pi} (\boldsymbol{\Delta} (\mathbf{R}\mathbf{u})^\alpha - \mathbf{T})^{-1} \mathbf{t}. \end{aligned}$$

□

The above result gives insight into how tail dependence is created (in contrast to the analogous Theorem 6 in Albrecher et al. (2020)): the reward matrix \mathbf{R} determines how the a priori independent stable components S_α^i are combined towards tail-dependent components Y_i , ($i = 1, \dots, n$), with tail dependence asymptotically being concentrated on lines with slopes governed by \mathbf{R} .

11.3.4 Distribution of projections

Consider $\mathbf{Y} \sim \text{MPH}_\alpha^*(\boldsymbol{\pi}, \mathbf{T}, \mathbf{R})$ with Laplace transform (11.9). We are interested in the distribution of the linear combination $\langle \mathbf{Y}, \mathbf{w} \rangle$ of the components for some non-zero, non-negative vector \mathbf{w} . Split the state space of E in E_+ and E_0 according to whether $(\mathbf{R}\mathbf{w})_i$ is positive or zero, respectively, and decompose $\boldsymbol{\pi} = (\boldsymbol{\pi}_+, \boldsymbol{\pi}_0)$ and

$$\mathbf{T} = \begin{pmatrix} \mathbf{T}_{++} & \mathbf{T}_{+0} \\ \mathbf{T}_{0+} & \mathbf{T}_{00} \end{pmatrix}$$

accordingly. Then consider the Laplace transform for $\langle \mathbf{Y}, \mathbf{w} \rangle$, which is

$$\begin{aligned}
 \mathbb{E} (e^{-u\langle \mathbf{Y}, \mathbf{w} \rangle}) &= \mathbb{E} (e^{-\langle \mathbf{Y}, u\mathbf{w} \rangle}) \\
 &= \boldsymbol{\pi} (\boldsymbol{\Delta}((\mathbf{R}u\mathbf{w})^\alpha) - \mathbf{T})^{-1} \mathbf{t} \\
 &= \boldsymbol{\pi} (u^\alpha \boldsymbol{\Delta}((\mathbf{R}\mathbf{w})^\alpha) - \mathbf{T})^{-1} \mathbf{t} \\
 &= \boldsymbol{\pi} \begin{pmatrix} u^\alpha \boldsymbol{\Delta}((\mathbf{R}\mathbf{w})^\alpha)_+ - \mathbf{T}_{++} & -\mathbf{T}_{+0} \\ -\mathbf{T}_{0+} & -\mathbf{T}_{00} \end{pmatrix}^{-1} \mathbf{t} \\
 &= (\boldsymbol{\pi}_+, \boldsymbol{\pi}_0) \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{t}_+ \\ \mathbf{t}_0 \end{pmatrix},
 \end{aligned}$$

where

$$\begin{aligned}
 \mathbf{A}_{11} &= (u^\alpha \boldsymbol{\Delta}((\mathbf{R}\mathbf{w})^\alpha)_+ - \mathbf{T}_{++} - \mathbf{T}_{+0}(-\mathbf{T}_{00})^{-1}\mathbf{T}_{0+})^{-1} \\
 &= \left(u^\alpha \mathbf{I} - \boldsymbol{\Delta}((\mathbf{R}\mathbf{w})^\alpha)_+^{-1} [\mathbf{T}_{++} + \mathbf{T}_{+0}(-\mathbf{T}_{00})^{-1}\mathbf{T}_{0+}] \right)^{-1} \boldsymbol{\Delta}((\mathbf{R}\mathbf{w})^\alpha)_+^{-1} \\
 &= (u^\alpha \mathbf{I} - \mathbf{T}_{\mathbf{w}^\alpha})^{-1} \boldsymbol{\Delta}((\mathbf{R}\mathbf{w})^\alpha)_+^{-1},
 \end{aligned}$$

$$\mathbf{A}_{12} = (u^\alpha \mathbf{I} - \mathbf{T}_{\mathbf{w}^\alpha})^{-1} \boldsymbol{\Delta}(\mathbf{R}\mathbf{w}^\alpha)_+^{-1} \boldsymbol{\Delta}((\mathbf{R}\mathbf{w})^\alpha)_+^{-1},$$

$$\mathbf{A}_{21} = (-\mathbf{T}_{00})^{-1}\mathbf{T}_{0+} (\boldsymbol{\Delta}(u^\alpha \mathbf{I} - \mathbf{T}_{\mathbf{w}^\alpha})^{-1} \boldsymbol{\Delta}((\mathbf{R}\mathbf{w})^\alpha)_+^{-1}),$$

$$\mathbf{A}_{22} = (-\mathbf{T}_{00})^{-1} (\mathbf{I} + \mathbf{T}_{0+} (u^\alpha \mathbf{I} - \mathbf{T}_{\mathbf{w}^\alpha})^{-1} \boldsymbol{\Delta}((\mathbf{R}\mathbf{w})^\alpha)_+^{-1} \mathbf{T}_{+0} (-\mathbf{T}_{00})^{-1})$$

and

$$\mathbf{T}_{\mathbf{w}} = \boldsymbol{\Delta}((\mathbf{R}\mathbf{w})^\alpha)_+^{-1} (\mathbf{T}_{++} + \mathbf{T}_{+0} (-\mathbf{T}_{00})^{-1} \mathbf{T}_{0+}).$$

Let

$$\boldsymbol{\pi}_{\mathbf{w}} = \boldsymbol{\pi}_+ + \boldsymbol{\pi}_0 (-\mathbf{T}_{00})^{-1} \mathbf{T}_{0+}.$$

Then

$$\begin{aligned}
 \boldsymbol{\pi}_+ \mathbf{A}_{11} + \boldsymbol{\pi}_0 \mathbf{A}_{21} &= \boldsymbol{\pi}_{\mathbf{w}} (u^\alpha \mathbf{I} - \mathbf{T}_{\mathbf{w}})^{-1} \boldsymbol{\Delta}((\mathbf{R}\mathbf{w})^\alpha)_+^{-1}, \\
 \boldsymbol{\pi}_+ \mathbf{A}_{12} + \boldsymbol{\pi}_0 \mathbf{A}_{22} &= \boldsymbol{\pi}_0 (-\mathbf{T}_{00})^{-1} + \boldsymbol{\pi}_{\mathbf{w}} (u^\alpha \mathbf{I} - \mathbf{T}_{\mathbf{w}})^{-1} \boldsymbol{\Delta}((\mathbf{R}\mathbf{w})^\alpha)_+^{-1} \mathbf{T}_{+0} (-\mathbf{T}_{00})^{-1}.
 \end{aligned}$$

Now inserting

$$\begin{pmatrix} \mathbf{t}_+ \\ \mathbf{t}_0 \end{pmatrix} = -\mathbf{T}\mathbf{e} = \begin{pmatrix} -\mathbf{T}_{++}\mathbf{e} - \mathbf{T}_{+0}\mathbf{e} \\ -\mathbf{T}_{0+}\mathbf{e} - \mathbf{T}_{00}\mathbf{e} \end{pmatrix},$$

we get

$$\begin{aligned}
 &(\boldsymbol{\pi}_+ \mathbf{A}_{11} + \boldsymbol{\pi}_0 \mathbf{A}_{21}) \mathbf{t}_+ + (\boldsymbol{\pi}_+ \mathbf{A}_{12} + \boldsymbol{\pi}_0 \mathbf{A}_{22}) \mathbf{t}_0 \\
 &= \boldsymbol{\pi}_0 (\mathbf{I} - (-\mathbf{T}_{00})^{-1} \mathbf{T}_{0+}) \mathbf{e} + \boldsymbol{\pi}_{\mathbf{w}} (u^\alpha \mathbf{I} - \mathbf{T}_{\mathbf{w}})^{-1} \mathbf{t}_{\mathbf{w}} \\
 &= \mathbf{1} - \boldsymbol{\pi}_{\mathbf{w}} \mathbf{e} + \boldsymbol{\pi}_{\mathbf{w}} (u^\alpha \mathbf{I} - \mathbf{T}_{\mathbf{w}})^{-1} \mathbf{t}_{\mathbf{w}}
 \end{aligned}$$

with

$$\mathbf{t}_{\mathbf{w}} = -\mathbf{T}_{\mathbf{w}} \mathbf{e}.$$

Thus we have proved the following result.

Theorem 11.3.8. Let $\mathbf{Y} \sim \text{MPH}_\alpha^*(\boldsymbol{\pi}, \mathbf{T}, \mathbf{R})$ and $\mathbf{w} \geq \mathbf{0}$ be a non-zero vector. Then $\langle \mathbf{Y}, \mathbf{w} \rangle$ has a distribution with an absolutely continuous part being $\text{PH}_\alpha(\boldsymbol{\pi}_\mathbf{w}, \mathbf{T}_\mathbf{w})$ distributed, where

$$\begin{aligned}\boldsymbol{\pi}_\mathbf{w} &= \boldsymbol{\pi}_+ + \boldsymbol{\pi}_0 (-\mathbf{T}_{00})^{-1} \mathbf{T}_{0+} \\ \mathbf{T}_\mathbf{w} &= \boldsymbol{\Delta} ((\mathbf{R}\mathbf{w})_+^\alpha)^{-1} (\mathbf{T}_{++} + \mathbf{T}_{+0} (-\mathbf{T}_{00})^{-1} \mathbf{T}_{0+})\end{aligned}$$

and an atom at zero of size $1 - \boldsymbol{\pi}_\mathbf{w}\mathbf{e}$.

As a simple consequence of the above result, one can retrieve the form of the marginal distributions for any choice of \mathbf{T} and \mathbf{R} :

Corollary 11.3.9. Let $\tau \sim \text{PH}_\alpha(\boldsymbol{\pi}, \mathbf{T})$. Let $\mathbf{r} = (r_1, \dots, r_p)$ be a non-zero non-negative vector of rewards. Let $\{X_t\}_{t \geq 0}$ denote the semi-Markov process which generates τ and define

$$Y = \int_0^\tau \sum_{i=1}^p r_i 1\{X_t = i\} dt$$

which is the total reward earned up to time τ . Then Y has a distribution with an absolutely continuous part having a $\text{PH}_\alpha(\tilde{\boldsymbol{\pi}}, \tilde{\mathbf{T}})$ form, where

$$\begin{aligned}\tilde{\boldsymbol{\pi}} &= \boldsymbol{\pi}_+ + \boldsymbol{\pi}_0 (-\mathbf{T}_{00})^{-1} \mathbf{T}_{0+} \\ \tilde{\mathbf{T}} &= \boldsymbol{\Delta} (\mathbf{r}_+^\alpha)^{-1} (\mathbf{T}_{++} + \mathbf{T}_{+0} (-\mathbf{T}_{00})^{-1} \mathbf{T}_{0+})\end{aligned}$$

and an atom at zero of size $1 - \tilde{\boldsymbol{\pi}}\mathbf{e}$.

Remark 11.3.10. In case all rewards are strictly positive, the translation between PH_α distributions with and without rewards is even simpler: Consider the process $\{X_t\}_{t \geq 0}$ defined in (11.7) underlying a $\text{PH}_\alpha(\boldsymbol{\pi}, \mathbf{T})$ distribution, and assume that a reward $r_i > 0$ is earned when the process is in state i , $i = 1, 2, \dots, p$. Then the total reward earned up to the time of absorption is $\text{PH}_\alpha(\boldsymbol{\pi}, \mathbf{S})$ distributed with

$$\mathbf{S} = \boldsymbol{\Delta}(\mathbf{r}^{-\alpha})\mathbf{T},$$

where $\mathbf{r}^{-\alpha} = (r_1^{-\alpha}, \dots, r_p^{-\alpha})$. Hence a reward of rate r_i in state i may be achieved by dividing row i of \mathbf{T} by r_i^α . This can also be seen directly from the construction of the semi-Markov process, since the λ_i are scale parameters.

11.4 Two specific examples

11.4.1 The feed-forward case

The MPH_α^* class shares an important sub-class of distributions with the GMMML class introduced in Albrecher et al. (2020). The so-called *feed-forward* sub-class is based on a special structure of the matrix components given as follows.

Let $\mathbf{C}_1, \dots, \mathbf{C}_n$ be sub-intensity matrices and let $\mathbf{D}_1, \dots, \mathbf{D}_n$ be non-negative matrices such that $-\mathbf{C}_i \mathbf{e} = \mathbf{D}_i \mathbf{e}$. Define the initial vector as $\boldsymbol{\beta} = (\boldsymbol{\pi}, \mathbf{0}, \dots, \mathbf{0})$ and the matrix

$$\mathbf{T} = \begin{pmatrix} \mathbf{C}_1 & \mathbf{D}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_2 & \mathbf{D}_2 & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{C}_3 & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{C}_n \end{pmatrix}. \quad (11.11)$$

The reward matrix consists of

$$\mathbf{R} = \begin{pmatrix} \mathbf{e} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{e} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{e} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{e} \end{pmatrix}. \quad (11.12)$$

In this case it is immediate that

$$\Delta(\mathbf{R}\boldsymbol{\theta})^\alpha = \Delta(\mathbf{R}\boldsymbol{\theta}^\alpha) \quad (11.13)$$

in which case the respective distributions in the GMML and MPH_α^* classes coincide (cf. Remark 11.3.5). Correspondingly, the explicit forms of the Laplace transform and density can be found in Theorem 8 of Albrecher et al. (2020) (choosing $\alpha_1 = \dots = \alpha_n = \alpha$ for the present context).

Note, however, that in general we do not have $\text{GMML} \subset \text{MPH}_\alpha^*$ nor that $\text{GMML} \supset \text{MPH}_\alpha^*$ (keeping in mind that the GMML class contains distributions with possibly different tail index in each marginal and no possible tail dependence, whereas the MPH_α^* class contains distributions with the same tail index for the marginals, but possible tail dependence), see also Figure 11.1.

11.4.2 A two-dimensional explicit example with tail dependence

Suppose that $\mathbf{X} = (X_1, X_2) \sim \text{MPH}_\alpha^*(\boldsymbol{\pi}, \mathbf{T}, \mathbf{R})$, where

$$\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \boldsymbol{\pi}_3), \quad \mathbf{T} = \begin{pmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} & \mathbf{T}_{13} \\ \mathbf{0} & \mathbf{T}_{22} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{T}_{33} \end{pmatrix}, \quad \text{and} \quad \mathbf{R} = \begin{pmatrix} \mathbf{e} & \mathbf{e} \\ \mathbf{e} & \mathbf{0} \\ \mathbf{0} & \mathbf{e} \end{pmatrix},$$

$\boldsymbol{\pi}_i$ are p_i -dimensional vectors and \mathbf{T}_{ij} are $p_i \times p_j$ -dimensional matrices for $i = 1, 2, 3$. As usual we let

$$\mathbf{t} = (\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3)' = -\mathbf{T}\mathbf{e}.$$

Hence \mathbf{t}_i is the vector of rates for jumping to the absorbing state from block $i = 1, 2, 3$. Denote the set of transient states by $E = \{1, 2, \dots, p_1 + p_2 + p_3\}$ and let $E_1 = \{1, 2, \dots, p_1\}$ denote the states corresponding to the first block, $E_2 = \{p_1 +$

$1, \dots, p_1 + p_2\}$ the states of the second block and $E_3 = \{p_1 + p_2 + 1, \dots, p_1 + p_2 + p_3\}$ the states of the third block.

The joint density function of the underlying multivariate phase-type distribution $\mathbf{X} = (X_1, X_2) \sim \text{MPH}^*(\boldsymbol{\pi}, \mathbf{T}, \mathbf{R})$ is given by (see (Bladt and Nielsen, 2017, p.448))

$$f(x_1, x_2) = \begin{cases} \boldsymbol{\pi}_1 e^{\mathbf{T}_{11}x_2} \mathbf{T}_{12} e^{\mathbf{T}_{22}(x_1-x_2)} \mathbf{t}_2, & 0 < x_2 < x_1 \\ \boldsymbol{\pi}_1 e^{\mathbf{T}_{11}x_1} \mathbf{T}_{13} e^{\mathbf{T}_{33}(x_2-x_1)} \mathbf{t}_3, & 0 < x_1 < x_2 \\ \boldsymbol{\pi}_1 e^{\mathbf{T}_{11}x_1} \mathbf{t}_1, & x_1 = x_2 \\ \boldsymbol{\pi}_2 e^{\mathbf{T}_{22}x_1} \mathbf{t}_2, & x_1 > 0, x_2 = 0 \\ \boldsymbol{\pi}_3 e^{\mathbf{T}_{33}x_2} \mathbf{t}_3, & x_1 = 0, x_2 > 0. \end{cases}$$

There is a component of sharing rewards in this structure. If the Markov jump process is started in a state in E_1 , then reward is earned for both variables X_1 and X_2 , and if $\mathbf{t}_1 \neq \mathbf{0}$, then there is a positive probability that the underlying process will exit to the absorbing state directly from the Block 1, in which case $X_1 = X_2$.

We shall now consider the distribution of \mathbf{Y} with Laplace transform (11.9). First we notice that

$$\Delta(\mathbf{R}\boldsymbol{\theta})^\alpha = \begin{pmatrix} (\theta_1 + \theta_2)^\alpha \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \theta_1^\alpha \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \theta_2^\alpha \mathbf{I} \end{pmatrix},$$

where the dimensions of the identity matrices \mathbf{I} are p_1 , p_2 and p_3 , respectively. Let

$$\begin{aligned} \mathbf{A}_{11} &= ((\theta_1 + \theta_2)^\alpha \mathbf{I} - \mathbf{T}_{11})^{-1} = \int_0^\infty e^{-(\theta_1 + \theta_2)x} x^{\alpha-1} E_{\alpha, \alpha}(\mathbf{T}_{11}x^\alpha) dx \\ \mathbf{A}_{22} &= (\theta_1^\alpha \mathbf{I} - \mathbf{T}_{22})^{-1} = \int_0^\infty e^{-\theta_1 y} y^{\alpha-1} E_{\alpha, \alpha}(\mathbf{T}_{22}y^{\alpha-1}) dy \\ \mathbf{A}_{33} &= (\theta_2^\alpha \mathbf{I} - \mathbf{T}_{33})^{-1} = \int_0^\infty e^{-\theta_2 y} y^{\alpha-1} x E_{\alpha, \alpha}(\mathbf{T}_{33}y^{\alpha-1}) dy. \end{aligned}$$

Then

$$(\Delta(\mathbf{R}\boldsymbol{\theta})^\alpha - \mathbf{T})^{-1} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{11}\mathbf{T}_{12}\mathbf{A}_{22} & \mathbf{A}_{11}\mathbf{T}_{13}\mathbf{A}_{33} \\ \mathbf{0} & \mathbf{A}_{22} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_{33} \end{pmatrix}.$$

Hence

$$\begin{aligned} L_{\mathbf{X}}(\boldsymbol{\theta}) &= \boldsymbol{\pi} (\Delta(\mathbf{R}\boldsymbol{\theta})^\alpha - \mathbf{T})^{-1} \mathbf{t} \\ &= \boldsymbol{\pi}_1 \mathbf{A}_{11} \mathbf{t}_1 + \boldsymbol{\pi}_2 \mathbf{A}_{22} \mathbf{t}_2 + \boldsymbol{\pi}_3 \mathbf{A}_{33} \mathbf{t}_3 + \boldsymbol{\pi}_1 \mathbf{A}_{11} \mathbf{T}_{12} \mathbf{A}_{22} \mathbf{t}_2 + \boldsymbol{\pi}_1 \mathbf{A}_{11} \mathbf{T}_{13} \mathbf{A}_{33} \mathbf{t}_3. \end{aligned}$$

The term

$$\begin{aligned}
& \boldsymbol{\pi}_1 \mathbf{A}_{11} \mathbf{T}_{12} \mathbf{A}_{22} \mathbf{t}_2 \\
&= \boldsymbol{\pi}_1 \int_0^\infty e^{-(\theta_1+\theta_2)x} x^{\alpha-1} E_{\alpha,\alpha}(\mathbf{T}_{11}x^\alpha) dx \mathbf{T}_{12} \int_0^\infty e^{-\theta_1 y} y^{\alpha-1} E_{\alpha,\alpha}(\mathbf{T}_{22}y^{\alpha-1}) dy \mathbf{t}_2 \\
&= \boldsymbol{\pi}_1 \int_0^\infty \int_0^\infty e^{-(\theta_1+\theta_2)x-\theta_1 y} x^{\alpha-1} y^{\alpha-1} E_{\alpha,\alpha}(\mathbf{T}_{11}x^\alpha) \mathbf{T}_{12} E_{\alpha,\alpha}(\mathbf{T}_{22}y^\alpha) dx dy \mathbf{t}_2 \\
&= \boldsymbol{\pi}_1 \int_0^\infty \int_0^\infty e^{-\theta_1(x+y)-\theta_2 x} x^{\alpha-1} y^{\alpha-1} E_{\alpha,\alpha}(\mathbf{T}_{11}x^\alpha) \mathbf{T}_{12} E_{\alpha,\alpha}(\mathbf{T}_{22}y^\alpha) dy dx \mathbf{t}_2 \\
&= \boldsymbol{\pi}_1 \int_0^\infty \int_x^\infty e^{-\theta_1 z - \theta_2 x} x^{\alpha-1} (z-x)^{\alpha-1} E_{\alpha,\alpha}(\mathbf{T}_{11}x^\alpha) \mathbf{T}_{12} E_{\alpha,\alpha}(\mathbf{T}_{22}(z-x)^\alpha) dz dx \mathbf{t}_2 \\
&= \boldsymbol{\pi}_1 \int_0^\infty e^{-\theta_2 x} x^{\alpha-1} E_{\alpha,\alpha}(\mathbf{T}_{11}x^\alpha) \mathbf{T}_{12} \int_x^\infty e^{-\theta_1 z} (z-x)^{\alpha-1} E_{\alpha,\alpha}(\mathbf{T}_{22}(z-x)^\alpha) dz dx \mathbf{t}_2,
\end{aligned}$$

which is hence the Laplace transform for the joint density of the form

$$\boldsymbol{\pi}_1 x^{\alpha-1} E_{\alpha,\alpha}(\mathbf{T}_{11}x^\alpha) \mathbf{T}_{12} (y-x)^{\alpha-1} E_{\alpha,\alpha}(\mathbf{T}_{22}(y-x)^\alpha) \mathbf{t}_2$$

when $Y_1 > Y_2$. A similar argument applies to $\boldsymbol{\pi}_1 \mathbf{A}_{11} \mathbf{T}_{13} \mathbf{A}_{33} \mathbf{t}_3$. The terms $\boldsymbol{\pi}_2 \mathbf{A}_{22} \mathbf{t}_2$ and $\boldsymbol{\pi}_3 \mathbf{A}_{33} \mathbf{t}_3$ correspond to the Laplace transform where one of the variables is equal to zero, while the term $\boldsymbol{\pi}_1 \mathbf{A}_{11} \mathbf{t}_1$ corresponds to the joint Laplace transform when $Y_1 = Y_2$. In conclusion,

$$f_{\mathbf{Y}}(x, y) = \begin{cases} \boldsymbol{\pi}_1 y^{\alpha-1} E_{\alpha,\alpha}(\mathbf{T}_{11}y^\alpha) \mathbf{T}_{12} (x-y)^{\alpha-1} E_{\alpha,\alpha}(\mathbf{T}_{22}(x-y)^\alpha) \mathbf{t}_2, & 0 < y < x \\ \boldsymbol{\pi}_1 x^{\alpha-1} E_{\alpha,\alpha}(\mathbf{T}_{11}x^\alpha) \mathbf{T}_{13} (y-x)^{\alpha-1} E_{\alpha,\alpha}(\mathbf{T}_{33}(y-x)^\alpha) \mathbf{t}_2, & 0 < x < y \\ x^{\alpha-1} \boldsymbol{\pi}_1 E_{\alpha,\alpha}(\mathbf{T}_{11}x^\alpha) \mathbf{t}_1, & x = y \\ x^{\alpha-1} \boldsymbol{\pi}_2 E_{\alpha,\alpha}(\mathbf{T}_{22}x^\alpha) \mathbf{t}_2, & x > 0, y = 0 \\ x^{\alpha-1} \boldsymbol{\pi}_3 E_{\alpha,\alpha}(\mathbf{T}_{33}x^\alpha) \mathbf{t}_3, & x = 0, y > 0. \end{cases}$$

An atom at zero (with point mass $1 - \boldsymbol{\pi} \mathbf{e}$) could also have been achieved for both cases by letting $\boldsymbol{\pi} \mathbf{e} < 1$. Figure 11.4 depicts a corresponding density, along with simulated data from the same distribution. The parameters are chosen to be $\alpha = 0.9$, $\boldsymbol{\pi}_1 = (1/2, 1/2)$, $\boldsymbol{\pi}_2 = \boldsymbol{\pi}_3 = \mathbf{0}$, and

$$\mathbf{T}_{11} = \begin{pmatrix} -3 & 2 \\ 0 & -4 \end{pmatrix}, \quad \mathbf{T}_{12} = \mathbf{T}_{13} = \begin{pmatrix} 0 & 1/2 \\ 1 & 1 \end{pmatrix}, \quad \mathbf{T}_{22} = \mathbf{T}_{33} = \begin{pmatrix} -1 & 1 \\ 0 & -2 \end{pmatrix},$$

which implies that there is no mass at $x = 0$, $y = 0$, or $x = y$. One clearly observes the resulting tail dependence across the respective slopes.

11.5 Conclusion

In this paper we propose an extension of Kulkarni's construction method to define a new class of multivariate distributions with matrix Mittag-Leffler distributed marginals. Based on a time-fractional sample path approach of an underlying semi-Markov jump process, this new class allows for dependence in the tails, yet still a

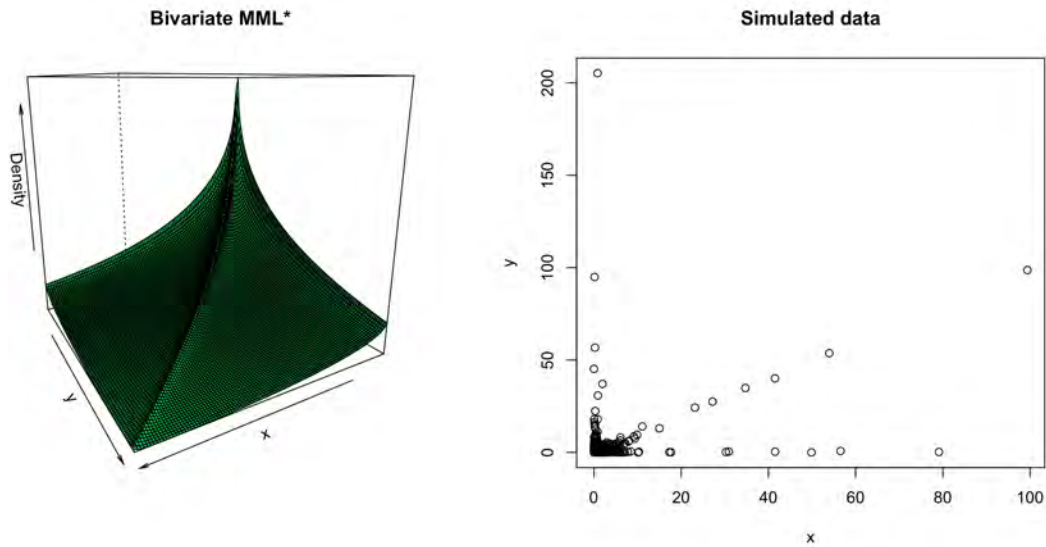


Figure 11.4: Density and 1500 simulated data from a bivariate MPH_α^* distribution.

rather explicit representation. We work out in detail how this class complements an earlier construction of a multivariate Mittag-Leffler distribution in Albrecher et al. (2020). The main contribution of this paper is on the conceptual and mathematical side. It will be interesting in future research to complement the present contribution by developing fitting procedures for real multivariate data sets in applications, which exploit the explicit expressions obtained for this new class and study its versatility in more detail. It will also be challenging to study procedures that decide about the appropriate dimensions of the underlying matrices in concrete applications.

Bibliography

- Aggarwal, C. C. (2016). *Outlier Analysis*. Springer Publishing Company, Incorporated, 2nd edition.
- Albrecher, H., Bäuerle, N., and Thonhauser, S. (2011). Optimal dividend-payout in random discrete time. *Statistics & Risk Modeling*, 28(3):251–276.
- Albrecher, H., Beirlant, J., and Teugels, J. L. (2017). *Reinsurance: Actuarial and Statistical Aspects*. John Wiley & Sons, Chichester.
- Albrecher, H. and Bladt, M. (2019). Inhomogeneous phase-type distributions and heavy tails. *Journal of Applied Probability*, 56(4):to appear.
- Albrecher, H., Bladt, M., and Bladt, M. (2019). Matrix Mittag–Leffler distributions and modeling heavy-tailed risks. *arXiv preprint arXiv:1906.05316*.
- Albrecher, H., Bladt, M., and Bladt, M. (2020). Multivariate Matrix Mittag–Leffler distributions. *Ann. Inst. Statist. Math.* In Press, doi: 10.1007/s10463-020-00750-7.
- Albrecher, H. and Cani, A. (2017). Risk theory with affine dividend payment strategies. In *Number Theory–Diophantine Problems, Uniform Distribution and Applications*, pages 25–60. Springer.
- Albrecher, H. and Thonhauser, S. (2009). Optimality results for dividend problems in insurance. *RACSAM-Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales. Serie A. Matemáticas*, 103(2):295–320.
- Ameraoui, A., Boukhetala, K., and Dupuy, J.-F. (2016). Bayesian estimation of the tail index of a heavy tailed distribution under random censoring. *Computational Statistics & Data Analysis*, 104:148–168.
- Anderson, D. N. (1992). A multivariate Linnik distribution. *Statist. Probab. Lett.*, 14(4):333–336.
- Arnaud, F., Révillon, S., Debret, M., Revel, M., Chapron, E., Jacob, J., Giguet-Covex, C., Poulénard, J., and Magny, M. (2012). Lake Bourget regional erosion patterns reconstruction reveals holocene NW European Alps soil evolution and paleohydrology. *Quaternary Science Reviews*, 51:81–92.
- Asmussen, S. (2003). *Applied probability and queues*, volume 51. Springer-Verlag, New York, second edition edition.

- Asmussen, S. and Albrecher, H. (2010). *Ruin Probabilities*. Advanced Series on Statistical Science & Applied Probability, 14. World Scientific, Second edition edition.
- Asmussen, S. and Glynn, P. W. (2007). *Stochastic Simulation: Algorithms and Analysis*, volume 57. Springer Science & Business Media.
- Asmussen, S. and Koole, G. (1993). Marked point processes as limits of Markovian arrival streams. *J. Appl. Prob.*, 30(2):365–372.
- Asmussen, S. and Kortschak, D. (2012). On error rates in rare event simulation with heavy tails. In *Proceedings of the Winter Simulation Conference*, page 38. Winter Simulation Conference.
- Asmussen, S. and Kortschak, D. (2015). Error rates and improved algorithms for rare event simulation with heavy Weibull tails. *Methodol. Comput. Appl. Probab.*, 17(2):441–461.
- Asmussen, S. and Kroese, D. P. (2006). Improved algorithms for rare event simulation with heavy tails. *Advances in Applied Probability*, 38(2):545–558.
- Asmussen, S., Nerman, O., and Olsson, M. (1996). Fitting phase-type distributions via the em algorithm. *Scandinavian Journal of Statistics*, pages 419–441.
- Asmussen, S. and Taksar, M. (1997). Controlled diffusion models for optimal dividend pay-out. *Insurance Math. Econom.*, 20(1):1–15.
- Avanzi, B. (2009). Strategies for dividend distribution: a review. *N. Am. Actuar. J.*, 13(2):217–251.
- Avanzi, B. and Wong, B. (2012). On a mean reverting dividend strategy with Brownian motion. *Insurance Math. Econom.*, 51(2):229–238.
- Azcue, P. and Muler, N. (2014). *Stochastic optimization in insurance*. SpringerBriefs in Quantitative Finance. Springer, New York.
- Baker, R. (2008). An order-statistics-based method for constructing multivariate distributions with fixed marginals. *J. Multivariate Anal.*, 99(10):2312–2327.
- Balkema, A. A. and De Haan, L. (1974). Residual life time at great age. *The Annals of probability*, pages 792–804.
- Basrak, B., Davis, R. A., and Mikosch, T. (2002). A characterization of multivariate regular variation. *The Annals of Applied Probability*, 12(3):908–920.
- Bäuerle, N. and Jaśkiewicz, A. (2015). Risk-sensitive dividend problems. *European J. Oper. Res.*, 242(1):161–171.
- Bäuerle, N. and Jaśkiewicz, A. (2017). Optimal dividend payout model with risk sensitive preferences. *Insurance Math. Econom.*, 73:82–93.

- Beirlant, J., Bardoutsos, A., de Wet, T., and Gijbels, I. (2016). Bias reduced tail estimation for censored pareto type distributions. *Statistics & Probability Letters*, 109:78–88.
- Beirlant, J., Boniphace, E., and Dierckx, G. (2011). Generalized sum plots. *REVSTAT-Statistical Journal*, 9(2):181–198.
- Beirlant, J., Dierckx, G., Guillou, A., and Fils-Villetard, A. (2007). Bias reduced tail estimation for censored pareto type distributions. *Extremes*, 10:151–174.
- Beirlant, J., Dierckx, G., Guillou, A., and Stărică, C. (2002). On exponential representations of log-spacings of extreme order statistics. *Extremes*, 5(2):157–180.
- Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. (2004). *Statistics of Extremes: Theory and Applications*. Wiley.
- Beirlant, J., Maribe, G., and Verster, A. (2018). Penalized bias reduction in extreme value estimation for censored pareto-type data, and long-tailed insurance applications. *Insurance: Mathematics and Economics*, 78:114–122.
- Beirlant, J., Vynckier, P., and Teugels, J. L. (1996). Tail index estimation, pareto quantile plots regression diagnostics. *Journal of the American Statistical Association*, 91(436):1659–1667.
- Beirlant, J., Worms, J., and Worms, R. (2019). Estimation of the extreme value index in a censorship framework: Asymptotic and finite sample behavior. *Journal of Statistical Planning and Inference*.
- Bhattacharya, S., Kallitsis, M., and Stoev, S. (2017). Trimming the Hill estimator: robustness, optimality and adaptivity. *arXiv preprint arXiv:1705.03088*.
- Bladt, M., Albrecher, H., and Beirlant, J. (2019). Trimming and threshold selection in extremes. *arXiv Preprint arXiv:1903.07942*.
- Bladt, M., Albrecher, H., and Beirlant, J. (2020). Combined tail estimation using censored data and expert information. *Scandinavian Actuarial Journal*, to appear.
- Bladt, M. and Nielsen, B. F. (2010). On the construction of bivariate exponential distributions with an arbitrary correlation coefficient. *Stochastic Models*, 26(2):295–308.
- Bladt, M. and Nielsen, B. F. (2017). *Matrix-Exponential Distributions in Applied Probability*. Springer, Berlin.
- Bladt, M., Nielsen, B. F., and Samorodnitsky, G. (2015). Calculation of ruin probabilities for a dense class of heavy tailed distributions. *Scandinavian Actuarial Journal*, pages 573–591.
- Bladt, M. and Rojas-Nandayapa, L. (2018). Fitting phase-type scale mixtures to heavy-tailed data and distributions. *Extremes*, 21(2):285–313.

- Blöschl, G., Hall, J., Parajka, J., Perdigão, R. A., Merz, B., Arheimer, B., Aronica, G. T., Bilibashi, A., Bonacci, O., Borga, M., et al. (2017). Changing climate shifts timing of european floods. *Science*, 357(6351):588–590.
- Bochner, S. (2005). *Harmonic analysis and the theory of probability*. Dover Publications.
- Bogaerts, K., Komarek, A., and Lesaffre, E. (2018). *Survival analysis with interval-censored data*. Chapman & Hall, Boca Raton.
- Brodsky, E. and Darkhovsky, B. S. (1993). *Nonparametric methods in change point problems*. Springer.
- Chan, J. C. and Kroese, D. P. (2011). Rare-event probability estimation with conditional monte carlo. *Annals of Operations Research*, 189(1):43–61.
- Chen, J. and Gupta, A. K. (2011). *Parametric statistical change point analysis: with applications to genetics, medicine, and finance*. Springer.
- Chikrii, A. A. and Eidel'man, S. D. (2000). Generalized Mittag-Leffler matrix functions in game problems for evolutionary equations of fractional order. *Cybernetics and Systems Analysis*, 36(3):315–338.
- Clifton, D. A., Clifton, L., Hugueny, S., and Tarassenko, L. (2014). Extending the generalised pareto distribution for novelty detection in high-dimensional spaces. *Journal of signal processing systems*, 74(3):323–339.
- Clifton, D. A., Clifton, L., Hugueny, S., Wong, D., and Tarassenko, L. (2013). An extreme function theory for novelty detection. *IEEE Journal of Selected Topics in Signal Processing*, 7(1):28–37.
- Clifton, D. A., Hugueny, S., and Tarassenko, L. (2011). Novelty detection with multivariate extreme value statistics. *Journal of signal processing systems*, 65(3):371–389.
- Constantinescu, C. D., Ramirez, J. M., and Zhu, W. R. (in press). An application of fractional differential equations to risk theory. *Finance and Stochastics*.
- Csörgő, S., Deheuvels, P., and Mason, D. (1985). Kernel estimates of the tail index of a distribution. *Ann. Statist.*, 13(3):1050–1077.
- Csörgő, M. and Horváth, L. (1997). *Limit theorems in change-point analysis*. John Wiley & Sons.
- Czymzik, M., Brauer, A., Dulski, P., Plessen, B., Naumann, R., von Grafenstein, U., and Scheffler, R. (2013). Orbital and solar forcing of shifts in mid-to late holocene flood intensity from varved sediments of pre-alpine lake ammersee (southern germany). *Quaternary Science Reviews*, 61:96–110.

- Daley, D. J. and Vere-Jones, D. (2007). *An introduction to the theory of point processes: volume II: general theory and structure*. Springer Science & Business Media.
- Danielsson, J., de Haan, L., Peng, L., and de Vries, C. G. (2001). Using a bootstrap method to choose the sample fraction in tail index estimation. *Journal of Multivariate analysis*, 76(2):226–248.
- de Finetti, B. (1957). Su un'impostazione alternativa della teoria collettiva del rischio. *Transactions of the 15th Int. Congress of Actuaries*, 2:433–443.
- de Haan, L. and Ferreira, A. (2007). *Extreme value theory: an introduction*. Springer Science & Business Media.
- De Sousa, B. and Michailidis, G. (2004). A diagnostic plot for estimating the tail index of a distribution. *Journal of Computational and Graphical Statistics*, 13(4):974–995.
- Diggle, P. and Marron, J. S. (1988). Equivalence of smoothing parameter selectors in density and intensity estimation. *Journal of the American Statistical Association*, 83(403):793–800.
- Draisma, G., de Haan, L., Peng, L., and Pereira, T. T. (1999). A bootstrap-based method to achieve optimality in estimating the extreme-value index. *Extremes*, 2(4):367–404.
- Drees, H., de Haan, L., and Resnick, S. (2000). How to make a Hill plot. *The Annals of Statistics*, 28(1):254–274.
- Drees, H. and Kaufmann, E. (1998). Selecting the optimal sample fraction in univariate extreme value estimation. *Stochastic Processes and their Applications*, 75(2):149–172.
- Dybvig, P. H. (1995). Dusenberry's ratcheting of consumption: optimal dynamic consumption and investment given intolerance for any decline in standard of living. *The Review of Economic Studies*, 62(2):287–313.
- Einmahl, J. H., Fils-Villetard, A., Guillou, A., et al. (2008). Statistics of extremes under random censoring. *Bernoulli*, 14(1):207–227.
- Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997). *Modelling extremal events*, volume 33 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin. For insurance and finance.
- Embrechts, P., Klüppelberg, C., and Mikosch, T. (2013). *Modelling extremal events: for insurance and finance*, volume 33. Springer Science & Business Media.
- Eo, S.-H., Hong, S.-M., and Cho, H. (2014). Identification of outlying observations with quantile regression for censored data. *arXiv preprint arXiv:1404.7710*.

- Erdélyi, A., Magnus, W., Oberhettinger, F., and Tricomi, F. G. (1955). *Higher transcendental functions. Vol. III.* McGraw-Hill Book Company, Inc., New York-Toronto-London. Based, in part, on notes left by Harry Bateman.
- Falk, M., Padoan, S. A., and Wisheckel, F. (2019). Generalized Pareto copulas: A key to multivariate extremes. *J. Multivariate Anal.*, 174(104538):17 pp.
- Feller, W. (1971). *An introduction to probability theory and its applications. Vol. II.* John Wiley & Sons Inc., New York.
- Field, C. and Van Aalst, M. (2014). Climate change 2014: impacts, adaptation, and vulnerability.
- Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 24, pages 180–190. Cambridge University Press.
- Foss, S., Korshunov, D., and Zachary, S. (2013). *An introduction to heavy-tailed and subexponential distributions.* Springer Series in Operations Research and Financial Engineering. Springer, New York, second edition.
- Frances, F., Salas, J. D., and Boes, D. C. (1994). Flood frequency analysis with systematic and historical or paleoflood data based on the two-parameter general extreme value models. *Water Resources Research*, 30(6):1653–1664.
- Frees, E. W. (2009). *Regression modeling with actuarial and financial applications.* Cambridge University Press.
- Garrappa, R. and Popolizio, M. (2018). Computing the matrix Mittag-Leffler function with applications to fractional calculus. *Journal of Scientific Computing*, 77(1):129–153.
- Geiger, D. and Adekpedjou, A. (2019). On corrected phase-type approximations of the time value of ruin with heavy tails. *Statistics and Risk Modelling*, 36:57–75.
- Genest, C. and Rémillard, B. (2004). Test of independence and randomness based on the empirical copula process. *Test*, 13(2):335–369.
- Gerber, H. U. (1969). Entscheidungskriterien fuer den zusammengesetzten Poisson-Prozess. *Schweiz. Aktuarver. Mitt.*, (1):185–227.
- Gerber, H. U. and Shiu, E. (2006a). On optimal dividend strategies in the compound Poisson model. *North American Actuarial Journal*, 10(2):76–93.
- Gerber, H. U. and Shiu, E. S. (2006b). On optimal dividends: from reflection to refraction. *Journal of Computational and Applied Mathematics*, 186(1):4–22.
- Gerber, H. U. and Shiu, E. S. W. (1998). On the time value of ruin. *N. Am. Actuar. J.*, 2(1):48–78. With discussion and a reply by the authors.

- Ghamami, S. and Ross, S. M. (2012). Improving the asmussen–kroese-type simulation estimators. *Journal of Applied Probability*, 49(4):1188–1193.
- Gomes, M. I., de Haan, L., and Rodrigues, L. H. (2008). Tail index estimation for heavy-tailed models: accommodation of bias in weighted log-excesses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):31–52.
- Gomes, M. I. and Guillou, A. (2015). Extreme value theory and statistics of univariate extremes: a review. *International Statistical Review*, 83(2):263–292.
- Gomes, M. I. and Oliveira, O. (2001). The bootstrap methodology in statistics of extremes—choice of the optimal sample fraction. *Extremes*, 4(4):331–358.
- Gomes, M. I. and Pestana, D. (2007). A sturdy reduced-bias extreme quantile (var) estimator. *Journal of the American Statistical Association*, 102(477):280–292.
- Gorenflo, R., Kilbas, A. A., Mainardi, F., and Rogosin, S. V. (2014). *Mittag-Leffler functions, related topics and applications*. Springer Monographs in Mathematics. Springer.
- Guillou, A. and Hall, P. (2001). A diagnostic for selecting the threshold in extreme value analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):293–305.
- Hall, P. (1982). On some simple estimates of an exponent of regular variation. *J. Roy. Statist. Soc. Ser. B*, 44(1):37–42.
- Hall, P. (1990). Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. *Journal of multivariate analysis*, 32(2):177–203.
- Hall, P., Welsh, A., et al. (1985). Adaptive estimates of parameters of regular variation. *The Annals of Statistics*, 13(1):331–341.
- Harteringer, J. and Kortschak, D. (2009). On the efficiency of the asmussen–kroese-estimator and its application to stop-loss transforms. *Blätter der DGVMF*, 30(2):363.
- Haubold, H. J., Mathai, A. M., and Saxena, R. K. (2011). Mittag-Leffler functions and their applications. *Journal of Applied Mathematics*, 2011. Article ID 298628, 51 pages.
- He, Q.-M., Zhang, H., and Vera, J. C. (2012). On some properties of bivariate exponential distributions. *Stochastic Models*, 28(2):187–206.
- Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3:1163–1174.
- Ho, Z. W. O. and Dombry, C. (2019). Simple models for multivariate regular variation and the Hüsler-ReiPareto distribution. *J. Multivariate Anal.*, 173:525–550.

- Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures*. John Wiley & Sons, Inc., New York, NY, USA.
- Jeanblanc-Picqué, M. and Shiryaev, A. N. (1995). Optimization of the flow of dividends. *Uspekhi Mat. Nauk*, 50(2):25–46.
- Joe, H. and Li, H. (2011). Tail risk of multivariate regular variation. *Methodology and Computing in Applied Probability*, 13(4):671–693.
- Jones, A. F., Macklin, M. G., and Brewer, P. A. (2012). A geochemical record of flooding on the upper river severn, uk, during the last 3750 years. *Geomorphology*, 179:89–105.
- Jongman, B., Hochrainer-Stigler, S., Feyen, L., Aerts, J. C., Mechler, R., Botzen, W. W., Bouwer, L. M., Pflug, G., Rojas, R., and Ward, P. J. (2014). Increasing stress on disaster-risk finance due to large floods. *Nature Climate Change*, 4(4):264.
- Jose, K. K., Uma, P., Lekshmi, V. S., and Haubold, H. J. (2010). Generalized Mittag-Leffler distributions and processes for applications in astrophysics and time series modeling. In *Proceedings of the Third UN/ESA/NASA Workshop on the International Heliophysical Year 2007 and basic space science*, pages 79–92. Springer.
- Juneja, S. (2007). Estimating tail probabilities of heavy tailed distributions with asymptotically zero relative error. *Queueing Systems*, 57(2-3):115–127.
- Kallenberg, O. (2006). *Foundations of modern probability*. Springer Science & Business Media.
- Kämpf, L., Brauer, A., Swierczynski, T., Czymzik, M., Mueller, P., and Dulski, P. (2014). Processes of flood-triggered detrital layer deposition in the varved lake mondsee sediment record revealed by a dual calibration approach. *Journal of Quaternary Science*, 29(5):475–486.
- Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.
- Kiriliouk, A., Rootzén, H., Segers, J., and Wadsworth, J. L. (2019). Peaks over thresholds modeling with multivariate generalized Pareto distributions. *Technometrics*, 61(1):123–135.
- Klugman, S. A., Panjer, H. H., and Willmot, G. E. (2012). *Loss models: from data to decisions*. John Wiley & Sons.
- Kozubowski, T. J. (2001). Fractional moment estimation of Linnik and Mittag-Leffler parameters. *Math. Comput. Modelling*, 34(9-11):1023–1035. Stable non-Gaussian models in finance and econometrics.
- Kulkarni, V. G. (1989). A new class of Multivariate Phase type distributions. *Operations Research*, 37:151–158.

- Kyprianou, A. (2006). *Introductory lectures on fluctuations of Lévy processes with applications*. Springer Science & Business Media.
- Kyprianou, A. E. (2014). *Fluctuations of Lévy processes with applications*. Universitext. Springer Heidelberg.
- Kyprianou, A. E. and Loeffen, R. (2010). Refracted Lévy processes. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 46(1):24–44.
- Lauterbach, S., Brauer, A., Andersen, N., Danielopol, D. L., Dulski, P., Hüls, M., Milecka, K., Namiotko, T., Obremaska, M., Von Grafenstein, U., et al. (2011). Environmental responses to lateglacial climatic fluctuations recorded in the sediments of pre-alpine lake mondsee (northeastern alps). *Journal of Quaternary Science*, 26(3):253–267.
- Lee, D., Li, W. K., and Wong, T. S. T. (2012). Modeling insurance claims via a mixture exponential model combined with peaks-over-threshold approach. *Insurance: Mathematics and Economics*, 51(3):538–550.
- Leung, K.-M., Elashoff, R. M., and Afifi, A. A. (1997). Censoring issues in survival analysis. *Annual review of public health*, 18(1):83–104.
- Lim, S. C. and Teo, L. P. (2010a). Analytic and asymptotic properties of multivariate generalized Linnik's probability densities. *J. Fourier Anal. Appl.*, 16(5):715–747.
- Lim, S. C. and Teo, L. P. (2010b). Analytic and asymptotic properties of multivariate generalized Linnik's probability densities. *J. Fourier Anal. Appl.*, 16(5):715–747.
- Lin, X. S. and Pavlova, K. P. (2006). The compound Poisson risk model with a threshold dividend strategy. *Insurance Math. Econom.*, 38(1):57–80.
- Loeffen, R. (2008). On optimality of the barrier strategy in de Finetti's dividend problem for spectrally negative Lévy processes. *Ann. Appl. Probab.*, 18(5):1669–1680.
- Luca, S., Clifton, D. A., and Vanrumste, B. (2016). One-class classification of point patterns of extremes. *The Journal of Machine Learning Research*, 17(1):6581–6601.
- Luca, S., Karsmakers, P., and Vanrumste, B. (2014). Anomaly detection using the poisson process limit for extremes. In *Data Mining (ICDM), 2014 IEEE International Conference on*, pages 370–379. IEEE.
- Luca, S. E., Pimentel, M. A., Watkinson, P. J., and Clifton, D. A. (2018). Point process models for novelty detection on spatial point patterns and their extremes. *Computational Statistics & Data Analysis*, 125:86–103.
- Mack, T. et al. (1994). Which stochastic model is underlying the chain ladder method. *Insurance: mathematics and economics*, 15(2-3):133–138.

- Mai, J.-F. and Scherer, M. (2017). *Simulating copulas*, volume 6 of *Series in Quantitative Finance*. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ.
- Markou, M. and Singh, S. (2003a). Novelty detection: a review—part 1: statistical approaches. *Signal processing*, 83(12):2481–2497.
- Markou, M. and Singh, S. (2003b). Novelty detection: a review—part 2:: neural network based approaches. *Signal processing*, 83(12):2499–2521.
- Matychyn, I. and Onyshchenko, V. (2018). Matrix Mittag-Leffler function in fractional systems and its computation. *Bulletin of the Polish Academy of Sciences: Technical Sciences*, 66(4):495 – 500.
- McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*, volume 37. CRC press.
- McNeil, A. J., Frey, R., and Embrechts, P. (2015). *Quantitative risk management*. Princeton Series in Finance. Princeton University Press, Princeton, NJ, revised edition.
- Merz, B., Dung, N. V., Apel, H., Gerlitz, L., Schröter, K., Steirou., E., and Vorogushyna, S. (2018). Spatial coherence of flood-rich and flood-poor periods across germany. *Journal of Hydrology*, 559:813–826.
- Merz, B., Nguyen, V. D., and Vorogushyn, S. (2016). Temporal clustering of floods in Germany: Do flood-rich and flood-poor periods exist? *Journal of Hydrology*, 541:824–838.
- Merz, R. and Blöschl, G. (2003). A process typology of regional floods. *Water Resources Research*, 39(12):1340.
- Mikosch, T. (1999). *Regular variation, subexponentiality and their applications in probability theory*. Eurandom Report 99013, Eindhoven University of Technology.
- Mikosch, T. (2006). Copulas: tales and facts. *Extremes*, 9(1):3–20.
- Mikosch, T. (2009). *Non-life insurance mathematics: an introduction with the Poisson process*. Springer Science & Business Media.
- Mittag-Leffler, M. G. (1904). Sopra la funzione $E\alpha(x)$. *Rend. Accad. Lincei*, 13(5):3–5.
- Mudelsee, M. (2014). *Climate time series analysis*. Springer.
- Mudelsee, M., Börngen, M., Tetzlaff, G., and Grünewald, U. (2003). No upward trends in the occurrence of extreme floods in central europe. *Nature*, 425(6954):166.
- Ndao, P., Diop, A., and Dupuy, J.-F. (2014). Nonparametric estimation of the conditional tail index and extreme quantiles under random censoring. *Computational Statistics & Data Analysis*, 79:63–79.

- Nešlehová, J., Embrechts, P., and Chavez-Demoulin, V. (2006). Infinite mean models and the lda for operational risk. *Journal of Operational Risk*, 1(1):3–25.
- Neuts, M. (1975). Probability distributions of phase type. In *Liber Amicorum Professor Emeritus H. Florin*, pages 173–206. Department of Mathematics, University of Louvain, Belgium.
- Neuts, M. (1979). A versatile Markovian point process. *Journal of Applied Probability*, 16(4):764–779.
- Neuts, M. F. (1994). *Matrix-Geometric Solutions in Stochastic Models*. Dover Publications Inc. Corrected reprint of the 1981 original.
- Nguyen, Q. H. and Robert, C. Y. (2014). New efficient estimators in rare event simulation with heavy tails. *Journal of Computational and Applied Mathematics*, 261:39–47.
- Payraastre, O., Gaume, E., and Andrieu, H. (2011). Usefulness of historical information for flood frequency analyses: Developments based on a case study. *Water Resources Research*, 47(8).
- Petrow, T. and Merz, B. (2009). Trends in flood magnitude, frequency and seasonality in germany in the period 1951–2002. *Journal of Hydrology*, 371:129–141.
- Pickands III, J. et al. (1975). Statistical inference using extreme order statistics. *the Annals of Statistics*, 3(1):119–131.
- Pigeon, M. and Denuit, M. (2011). Composite lognormal-Pareto model with random threshold. *Scand. Actuar. J.*, pages 177–192.
- Pillai, R. N. (1990). On Mittag-Leffler functions and related distributions. *Ann. Inst. Statist. Math.*, 42(1):157–161.
- Pimentel, M. A., Clifton, D. A., Clifton, L., and Tarassenko, L. (2014). A review of novelty detection. *Signal Processing*, 99:215–249.
- Prettenthaler, F., Albrecher, H., Asadi, P., and Köberl, J. (2017). On flood risk pooling in europe. *Natural Hazards*, 88(1):1–20.
- Prettenthaler, F., Kortschak, D., Hochrainer-Stigler, S., Mechler, R., Urban, H., and Steininger, K. W. (2015). Catastrophe management: riverine flooding. In *Economic evaluation of climate change impacts*, pages 349–366. Springer.
- Resnick, S. (2002). Hidden regular variation, second order regular variation and asymptotic independence. *Extremes*, 5(4):303–336.
- Rolski, T., Schmidli, H., Schmidt, V., and Teugels, J. (1999). *Stochastic Processes for Insurance and Finance*. Wiley Series in Probability and Statistics. John Wiley & Sons Ltd.

- Sabatier, P., Wilhelm, B., Gentile, F. F., Moiroux, F., Poulenard, J., Develle, A.-L., Bichet, A., Chen, W., Pignol, C., Reyss, J.-L., et al. (2017). 6-kyr record of flood frequency and intensity in the western mediterranean Alps – interplay of solar and temperature forcing. *Quaternary Science Reviews*, 170:121–135.
- Schmidli, H. (2008). *Stochastic Control in Insurance*. Springer, New York.
- Schmocker-Fackel, P. and Naef, F. (2010). More frequent flooding? changes in flood frequency in switzerland since 1850. *Journal of Hydrology*, 381(1-2):1–8.
- Shreve, S. E., Lehoczky, J. P., and Gaver, D. P. (1984). Optimal consumption for general diffusions with absorbing and reflecting barriers. *SIAM J. Control Optim.*, 22(1):55–75.
- Swierczynski, T., Brauer, A., Lauterbach, S., Martín-Puertas, C., Dulski, P., von Grafenstein, U., and Rohr, C. (2012). A 1600 yr seasonally resolved record of decadal-scale flood variability from the austrian pre-alps. *Geology*, 40(11):1047–1050.
- Swierczynski, T., Ionita, M., and Pino González, D. (2017). Using archives of past floods to estimate future flood hazards. *EOS transactions*, 98.
- Swierczynski, T., Lauterbach, S., Dulski, P., Delgado, J., Merz, B., and Brauer, A. (2013). Mid-to late holocene flood frequency changes in the northeastern alps as recorded in varved sediments of lake mondsee (Upper Austria). *Quaternary Science Reviews*, 80:78–90.
- Teugels, J. L. (1975). The class of subexponential distributions. *The Annals of Probability*, 3(6):1000–1011.
- Tzougas, G., Vrontos, S., and Frangos, N. (2014). Optimal bonus-malus systems using finite mixture models. *ASTIN Bulletin*, 44(2):417–444.
- Vatamidou, E., Adan, I. J. B. F., Vlasiou, M., and Zwart, B. (2013). Corrected phase-type approximations of heavy-tailed risk models using perturbation analysis. *Insurance: Mathematics and Economics*, 53(2):366–378.
- Wallemacq, P. (2018). Annual disaster statistical review 2017, CRED. <http://www.cred.be/emdat>.
- Wan, P. and Davis, R. A. (2019). Threshold selection for multivariate heavy-tailed data. *Extremes*, 22(1):131–166.
- Weissman, I. (1978). Estimation of parameters and large quantiles based on the k largest observations. *Journal of the American Statistical Association*, 73(364):812–815.
- Wilhelm, B., Ballesteros Cánovas, J. A., Macdonald, N., Toonen, W. H., Baker, V., Barriendos, M., Benito, G., Brauer, A., Corella, J. P., Denniston, R., et al. (2018a). Interpreting historical, botanical, and geological evidence to aid preparations for future floods. *Wiley Interdisciplinary Reviews: Water*, page e1318.

- Wilhelm, B., Canovas, J. A. B., Aznar, J. P. C., Kämpf, L., Swierczynski, T., Stoffel, M., Støren, E., and Toonen, W. (2018b). Recent advances in paleoflood hydrology: From new archives to data compilation and analysis. *Water Security*, 3:1–8.
- Wirth, S. B., Glur, L., Gilli, A., and Anselmetti, F. S. (2013). Holocene flood frequency across the central alps–solar forcing and evidence for variations in north atlantic atmospheric circulation. *Quaternary Science Reviews*, 80:112–128.
- Wolfe, S. J. (1975). On moments of probability distribution functions. In *Fractional Calculus and Its Applications*, pages 306–316. Springer.
- Worms, J. and Worms, R. (2014). New estimators of the extreme value index under random right censoring, for heavy-tailed distributions. *Extremes*, 17(2):337–358.
- Worms, J. and Worms, R. (2016). A lynden-bell integral estimator for extremes of randomly truncated data. *Statistics & Probability Letters*, 109:106–117.
- Worms, J. and Worms, R. (2018). Extreme value statistics for censored data with heavy tails under competing risks. *Metrika*, 81(7):849–889.
- Wüthrich, M. V. and Merz, M. (2008). *Stochastic claims reserving methods in insurance*, volume 435. John Wiley & Sons.