FACULTEIT ECONOMIE EN
BEDRIJFSWETENSCHAPPEN

KU LEUVEN

Data analytics for insurance loss modeling, telematics pricing and claims reserving

Proefschrift Voorgedragen tot
het Behalen van de Graad van
Doctor in de Toegepaste
Economische Wetenschappen

door

Roel VERBELEN

NUMMER 563 2017

Committee

Advisor:	
Prof. Dr. Gerda Claeskens	*KU Leuven*
Co-Advisor:	
Prof. Dr. Katrien Antonio	*KU Leuven and University of Amsterdam*
Chair:	
Prof. Dr. Robert Boute	*KU Leuven and Vlerick Business School*
Members:	
Prof. Dr. Jan Beirlant	*KU Leuven and University of the Free State*
Prof. Dr. Jan Dhaene	*KU Leuven*
Prof. Dr. Edward W. Frees	*University of Wisconsin–Madison*
Prof. Dr. Montserrat Guillén	*University of Barcelona*

Acknowledgments

"Pursuing a PhD? No thanks, not for me..."

That I needed quite some convincing is something that Katrien will surely confirm. In fact, she already started about 6 years ago when she guided me to the best possible educational path after my bachelor studies in mathematics at Ghent University. I started wondering about a future career in actuarial science and reached out to Katrien requesting help on the next step to take. I was blown away by her well-founded and comprehensive reply – that she wrote while being on vacation. She invited me for a personal meeting at her office in Leuven and convinced me to switch to KU Leuven and start a master program in statistics. Already back then, she suggested the possibility of afterwards starting a PhD. As supervisor of my master thesis, she continued to advocate a PhD and together with Gerda eventually persuaded me to embark on an interdisciplinary research project, combining actuarial science and statistics. Not a moment has gone by that I regretted this decision.

First and foremost, I would like to express my profound gratitude to Gerda and Katrien. During all these years, I have had the privilege to work closely with them and have them as mentors. Thank you for convincing me to take up this challenge and supporting me along the way. You were always there for me when I needed guidance, answered promptly to any of my questions and read article drafts with the utmost attention to detail. I have been impressed and inspired by your level of dedication to the university and to research and how you both manage to combine it with your family life.

Further, I want to thank my doctoral committee members from KU Leuven (Prof. Jan Beirlant and Prof. Jan Dhaene), my external jury members (Prof. Edward W. Frees and Prof. Montserrat Guillén) and chairman Prof. Robert Boute. Thank you for the thorough reading of my thesis and for all insightful comments

during the doctoral seminars and preliminary defense. Your constructive suggestions helped to improve the quality of this thesis. In particular, thank you Prof. Montserrat Guillén for travelling to Leuven to be present at my public defense.

I would also like to thank my co-authors of several research papers. Thanks to Prof. Andrei Badescu, Prof. Sheldon Lin and Lan Gong from the University of Toronto to introduce me to the class of mixtures of Erlang distributions which formed the starting point of my research in the field of loss modeling. I'm grateful for the collaboration with Tom Reynkens and Prof. Jan Beirlant which lead to a related follow-up paper. Thank you Jonas Crèvecoeur for the joint brainstorms and your feedback on the reserving project. I also very much enjoyed working with Maxime Clijsters and Roel Henckaerts on a research topic in the context of insurance pricing. I am proud to have co-supervised your master theses which were both rewarded with the IA|BE thesis prize and additionally with the Johan de Witt thesis prize for Maxime.

Data have been indispensable for the success of my project in actuarial statistics. In this regards, I could benefit from the professional network of Katrien in the actuarial community and the academic network of Gerda in the statistical community. Special thanks go out to Jonas Onkelinx for going to great lengths to share the telematics data with us and for his valuable support in handling this challenging data set. These telematics data created a strategic advantage and were of crucial importance for our research on usage-based insurance pricing. I also wish to thank Hans Laevens for permission to use the interesting mastitis data in the multivariate mixture of Erlangs chapter.

I am very grateful to the government agency for Innovation by Science and Technology (IWT), now called Flanders Innovation & Entrepreneurship (VLAIO), for financing this doctorate and to the Flemish Supercomputer Centre VSC for computational support involving simulation studies and analyzing big data sets.

Thanks to KU Leuven for providing me with the facilities to carry out this doctoral research. I have enjoyed being part of the stimulating, interactive and international research environment that the Faculty of Economics and Business has to offer. Since this Faculty hosts both a research group in statistics, as well as in actuarial science, this has been the perfect place to conduct my PhD research.

I am very thankful for having been surrounded by so many nice colleagues from ORSTAT, AFI, LSTAT in Heverlee, down the hall... It has been a pleasure to be part of the statistics group at ORSTAT as well as of the insurance group at AFI. I have been fortunate to get close with many of you and you formed a

big part in my life over the recent years. Thanks for all those social activities we shared together: the barbecues at Thomas's, the Turkish dinners at Deniz and Mehmet's, Peter's wedding, the Romania trip for the Pircalabelu wedding, the volleyball games, the soccer with the guys from MSI, the kayaking in the Ardennes, organizing the EDC quiz, the nights out at Oude Markt... I have enjoyed them all very much! In particular, I would like to thank Lore for having been such an incredible office mate. Thanks for all those laughs as well as sincere discussions on both life and research. I am lucky to have had an office mate who grew into a close friend. When Lore left for Boston, I was lucky once more to have Daumantas join the office. Thank you for all those little jokes and creating such a pleasant working atmosphere. Special thanks go out to Ines for all the advice, support and practical help over the years and for compensating my forgetfulness from time to time.

I want to thank Prof. Meelis Käärik for the enriching research stay at the University of Tartu in the final months of my doctorate. My office mate Annika has made me feel very welcome, aitäh!

Pursuing a PhD gave me the opportunity to present my work at international conferences, to attend scientific workshops and to meet many academics from all over the world. Above all, I was fortunate to have met Liivika in this way. Our impressive series of foreign encounters has turned into more than I could ever dream of. Thank you for all the wonderful journeys and experiences together. As a graduating PhD student yourself you understood as no one else how intensive these doctoral years have been for me and you managed to calm me down whenever I was stressed. I felt strengthened having you by my side and your support while writing this dissertation was invaluable. I'm grateful and proud to have you as a partner in life and I am excited to discover what lies ahead in our future together.

Last but definitely not least I want to thank my family. Words can do no justice to what you mean to me. You have my deepest and sincerest gratitude for your encouragement and unconditional support in every step of my life.

Roel Verbelen Leuven, June 2017.

Table of Contents

Chapter 1

Introduction

Today's society generates data more rapidly than ever before, creating many opportunities as well as challenges for statisticians. Many industries become increasingly dependent on high-quality data, and the demand for sound statistical analysis of these data is rising accordingly.

In the insurance sector, data have always played a major role. When selling a contract to a client, the insurance company is liable for the claims arising from this contract and will hold capital aside to meet these future liabilities. As such, the insurance premium has to be paid before the real costs are known. This is referred to as the *inversion of the production cycle*. It implies that the activities of *pricing* and *reserving* are strongly interconnected in actuarial practice. On the one hand, pricing actuaries have to determine a fair price for the insurance products they want to sell. Setting the premium levels charged to the insureds is done in a data driven way where statistical models are essential. Risk-based pricing is crucial in a competitive and well-functioning insurance market. On the other hand, an insurance company must safeguard its solvency and reserve capital to fulfill outstanding liabilities. Reserving actuaries thus must predict, with maximum accuracy, the total amount needed to pay claims that the insurer has legally committed himself to cover for. These reserves form the main item on the liability side of the balance sheet of the insurance company and therefore have an important economic impact.

The ambition of this research is the development of new, accurate predictive models for the actuarial work field. Non-life (e.g. motor, fire, liability), life and health insurers are constantly confronted with the challenges created by rapidly increasing computer facilities for data collection, storage and analysis. However,

using their state-of-the-art methodologies the insurance business will not be able to formulate an adequate response to these challenges, and interactions with the disciplines of statistics and big data analytics are necessary. Moreover, the increased focus on internal risk management and the changing supervisory guidelines motivate the relevance of improved tools for actuarial predictive modeling. In particular, the European Solvency II Directive[1] imposes new solvency requirements to enhance policyholder protection. With the recent introduction of these new regulatory guidelines, the measurement of future cash flows and their uncertainty becomes more important. At the same time, actuarial predictive models have to comply with existing and pending regulations. The Gender Directive[2] has prohibited the use of gender as a risk factor in insurance pricing and antidiscrimination laws may progress in the near future further limiting the contractual freedom of insurance companies.

The overall objective in this work is to improve actuarial practices for pricing and reserving by using sound and flexible statistical methods shaped for the actuarial data at hand. The tools we develop should lead to a better understanding of actuarial risks and an improved risk management. This thesis focuses on three related research avenues in the domain of non-life insurance: (1) flexible univariate and multivariate loss modeling in the presence of censoring and truncation, (2) car insurance pricing using telematics data and (3) micro-level claims reserving.

1.1 Innovations in loss modeling

Modeling claim losses – also called claim sizes or severities – is crucial when pricing insurance products, determining capital requirements, or managing risks within financial institutions. Various basic continuous distributions, such as the gamma or lognormal, have been employed to model nonnegative losses. However, these parametric distributions are not always appropriate for actuarial data, which may be multimodal or heavy-tailed. Furthermore, when constructing collective risk models or combining actuarial risks from multiple lines of business, these severity distributions do not lead to an analytical form for the corresponding aggregate loss distribution. While numerical or simulation algorithms are available, it is nevertheless convenient to utilize analytical techniques when possible. Of course, there is always a tradeoff between mathematical simplicity on the one hand and realistic modeling on the other. Ideally, loss models require on the one hand the

[1] See `http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32009L0138`.
[2] See `http://europa.eu/rapid/press-release_IP-12-1430_en.htm`.

flexibility of nonparametric approaches to describe the claims and on the other hand the feasibility to analytically quantify the risk.

In actuarial literature, the use of *mixtures of Erlang distributions* with a common scale parameter has been suggested to model insurance losses. An Erlang distribution is in fact a gamma distribution with an integer shape parameter and can be decomposed as the sum of independent, exponentially distributed random variables with the same mean (equal to the inverse of the scale parameter). A mixture of such distributions with a common scale parameter can be considered as a compound distribution of a random sum of exponential random variables with the same mean. The resulting class of distributions enjoys a wide variety of analytic properties because it can exploit the mathematical tractability of the exponential distribution. Many quantities of interest in connection with the aggregation of claims and stop-loss analysis are easily computable under the mixture of Erlangs assumption. At the same time, mixtures of Erlangs are extremely versatile in terms of possible shapes of the probability density function and are capable of multimodality as well as a wide range of degrees of skewness in the right tail, often the region of particular interest for risk management purposes. In fact, this class of distributions is dense in the space of positive continuous distributions. As such, any continuous distribution can be approximated to an arbitrary degree of accuracy by a mixture of Erlang distribution.

In Chapter 2, we discuss how to estimate mixtures of Erlangs using censored and truncated data. Parameter estimation is of course of utmost importance when we want to apply these mixtures of Erlangs in real-life applications. Our work is further inspired by the omnipresence of censoring and truncation in an actuarial context. Insurance contracts often do not provide full coverage of a loss. A policy modification such as a (franchise) deductible of €500 causes the insurer to only pay for the claim if it exceeds €500. Such kind of deductible is also used in excess-of-loss reinsurance treaties when an insurance company on his turn buys protection for a certain loss layer. As a consequence, only payments that exceed this threshold will be recorded by the reinsurer and can be used to estimate the loss distribution. The insurance losses are said to be left truncated at that threshold. Policy limits on the other hand define the maximum amount of coverage provided by the insurer. This policy modification has as effect that the observed insurance losses are right censored, meaning that the exact value of the loss, in case it exceeds this limit, is not recorded. Right censoring also arises when claims are not yet fully settled. For such unsettled claims only the payment to date is known whereas the final total payment will be at least as much.

In Chapter 3, we extend the estimation procedure under censoring and truncation to *multivariate mixtures of Erlang distributions*. This multivariate distribution generalizes the univariate mixture of Erlang distributions while preserving its flexibility and analytical tractability. When modeling multivariate insurance losses or dependent risks from different portfolios or lines of business, the inherent shape versatility of multivariate mixtures of Erlangs allows one to adequately capture both the marginals and the dependence structure. Moreover, its desirable analytical properties are particularly convenient in a wide variety of insurance related modeling situations.

1.2 Innovations in car insurance pricing through telematics technology

Telematics technology – the integrated use of telecommunication and informatics – may fundamentally change the car insurance industry by allowing insurers to base their prices on the real driving behavior instead of on traditional policyholder characteristics and historical claims information. The use of this technology in insured vehicles enables to transmit and receive information that allows an insurance company to better assess the accident risk of drivers and adjust the premiums accordingly through usage-based insurance. A small black box device is installed in the insured's car containing a GPS system, electronics that capture hundreds of sensor inputs, a SIM card and some computer software. It records the driving behavior directly and shares this information with the insurer.

On February 23, 2013 The Economist[3] reported *"Underwriters have traditionally used crude demographic data such as age, location and sex to separate the testosterone-fueled boy racers from their often tamer female counterparts. Now technology is giving insurers the chance to see just how skilled a driver really is. By monitoring their customers' motoring habits, underwriters can increasingly distinguish between drivers who are safe on the road from those who merely seem safe on paper. Many think that 'telematics insurance' will become the industry norm."*

This industry (r)evolution creates multiple opportunities from a business as well as statistical modeling perspective. With telematics insurance focus will be on how much time a car spends on the road (*pay-as-you-drive*) or on driver ability (*pay-how-you-drive*), as an alternative for the current practice where observable

[3] How's my driving? (2013, February 23) *The Economist.* `http://econ.st/Yd5x3C`

risk information (like age or gender) is used as a proxy for unobservable characteristics (like distance driven or driving style). The upswing of telematics data may also replace (in the near future) rating variables which are currently being banned from actuarial pricing practice by recent court decisions (such as the gender ban).

The availability of such data collected while driving creates a wide, but unexplored territory for statisticians. Usage-based insurance forces pricing actuaries to change their current practice and to develop innovative statistical tools to customize premiums based on the actual driving behavior. Analytic contributions on this topic in scientific research are scarce, probably because the collection of this type of data is immature and brand new.

In Chapter 4, we explore the vast potential of telematics insurance from a statistical point of view by analyzing a unique Belgian portfolio. Driving behavior data are collected in between 2010 and 2014 for young drivers who signed up for a telematics product. Since 2010, the Belgian insurance company offers young drivers a premium discount in exchange for a black box to be installed in their car. This telematics device collects data on when, where and how long the car is being used. The aim of our contribution is to develop the statistical methodology to incorporate this telematics information in statistical rating models, where we focus on predicting the number of claims, in order to adequately set premium levels based on individual policyholder's driving habits. We propose new tools and techniques that actuaries can use to improve their current pricing practices and to design new products that are better aligned with the potential these new technologies offer.

1.3 Innovations in claims reserving

To be able to fulfill future liabilities insurance companies will hold sufficient capital reserves. Loss reserving deals with the prediction of the remaining development of reported, open claims (*reported but not settled reserve*) and unreported claims (*incurred but not reported reserve*). Accurate, reliable and robust reserving methods for a wide range of products and lines of business are key factors in the stability and solvability of insurance companies. The industry-wide standard is the chain-ladder technique, which works on data aggregated in a run-off triangle. A run-off triangle summarizes the information registered during the lifetime of individual claims by aggregating loss payments over two dimensions, namely the year of occurrence of the claim and the period since claim event during which the payment took place.

Nowadays, insurance companies keep track of detailed information for each individual claim. Rich data sources record, for example, the occurrence date, the reporting delay, the date and amount of each loss payment, and the settlement date. The existing methods for claims reserving are designed for aggregated data, but through this data compression many useful information is lost. With the advent of Solvency II, insurers are required to not only provide a best estimate of their future liabilities, but also to have a better grasp of their uncertainty. Current techniques for loss reserving will have to be improved, adjusted or extended to meet the requirements of the new regulations.

In Chapter 5, we leave the track of aggregated data and focus on the underlying, more granular data. Stochastic loss reserving methods designed at the individual claim level are referred to as *micro-level reserving techniques*. The overall goal is to increase the predictive power of loss reserving methods and improve risk measurement by using the information stored in the insurer's data base system, instead of ignoring it. We focus on modeling the claims arrival and reporting delay using a micro-level approach. Due to time delays between the occurrence of the insured event and the notification of the claim to the insurer, not all of the claims that occurred in the past have been observed when the reserve needs to be calculated. We present a flexible regression framework to model and jointly estimate the occurrence and reporting of claims. This new technique models the claim arrival process on a daily basis in order to predict the number of incurred but not reported claim counts.

The various chapters in this thesis can be found in

(i) Verbelen, R., Gong, L., Antonio, K., Badescu, A., and Lin, X. S. (2015). Fitting mixtures of Erlangs to censored and truncated data using the EM algorithm. *ASTIN Bulletin*, 45(3):729-758.

(ii) Verbelen, R., Antonio, K., and Claeskens, G. (2016). Multivariate mixtures of Erlangs for density estimation under censoring. *Lifetime Data Analysis*, 22(3):429-455.

(iii) Verbelen, R., Antonio, K., and Claeskens, G. (2016). Unraveling the predictive power of telematics data in car insurance pricing. *FEB Research Report* KBI 1624.

(iv) Verbelen, R., Antonio, K., Claeskens, G. and Crèvecoeur, J. (2017). Predicting daily IBNR claim counts using a regression approach for the occurrence of claims and their reporting delay. *Working paper*.

The author also contributed to the following original publications

(i) Reynkens, T., Verbelen, R., Beirlant, J. and Antonio, K. (2016). Modeling censored losses using splicing: a global fit strategy with mixed Erlang and extreme value distributions, *arXiv:1608.01566*.

(ii) Henckaerts, R., Antonio, K., Clijsters, M. and Verbelen, R. (2017). A data driven binning strategy for the construction of risk classes, *Working paper*.

Chapter 2

Fitting mixtures of Erlangs to censored and truncated data using the EM algorithm

Abstract

We discuss how to fit mixtures of Erlangs to censored and truncated data by iteratively using the EM algorithm. Mixtures of Erlangs form a very versatile, yet analytically tractable, class of distributions making them suitable for loss modeling purposes. The effectiveness of the proposed algorithm is demonstrated on simulated data as well as real data sets.

This chapter is based on Verbelen, R., Gong, L., Antonio, K., Badescu, A., and Lin, X. S. (2015). Fitting mixtures of Erlangs to censored and truncated data using the EM algorithm. *ASTIN Bulletin*, 45(3):729-758

2.1 Introduction

The class of mixtures of Erlang distributions with a common scale parameter is very flexible in terms of the possible shapes of its members. Tijms (1994, p. 163) shows that mixtures of Erlangs are dense in the space of positive distributions in the sense that there always exists a series of mixtures of Erlangs that weakly converges, i.e. converges in distribution, to any positive distribution. As such, any continuous distribution can be approximated by a mixture of Erlang distributions

to any accuracy. Furthermore, via direct manipulation of the Laplace transform, a wide variety of distributions whose membership in this class is not immediately obvious can be written as a mixture of Erlangs. The class of mixtures of Erlangs with a common scale is also closed under mixture, convolution and compounding. At the same time, it is possible to work analytically with this class leading to explicit expressions for e.g. the Laplace transform, the hazard rate, a Tail-Value-at-Risk (TVAR) and stop-loss moments. A quantile or a Value-at-Risk (VaR) can be obtained by numerically inverting the cumulative distribution function. Klugman et al. (2013), Willmot and Lin (2011) and Willmot and Woo (2007) give an overview of these analytical and computational properties of mixtures of Erlangs.

Mixtures of Erlang distributions have received most attention in the field of actuarial science. Modeling data on claim sizes is crucial when pricing insurance products. Actuarial models help insurance companies to assess the risk associated with the portfolio, to set the level of premiums (Frees and Valdez, 2008) and reserves (Antonio and Plat, 2014), to determine optimal reinsurance levels (Beirlant et al., 2004) or to determine capital requirements for solvency purposes (Bolancé et al., 2012). Insurance data are often modeled using a parametric distribution such as a gamma, lognormal or Pareto distribution. The usual way to proceed in loss modeling, pricing and reserving is to calibrate the data using several of these parametric distributions and then select, among these, the most appropriate model based on a model selection tool (Klugman and Rioux, 2006). These classes of distributions may however not always be flexible enough in terms of the possible shapes of their members in order to obtain a satisfying fit (e.g. in the presence of multimodal data) and resulting models become intractable when aggregating risks in an insurance portfolio or arising from multiple lines of losses. Ideally, it would be useful to have a single approach to fitting loss models (Klugman and Rioux, 2006) with on the one hand the flexibility of nonparametric density estimation techniques to describe the insurance losses and on the other hand the feasibility to analytically quantify the risk. This is exactly what the class of mixtures of Erlangs has to offer. In particular, using these distributions in aggregate loss models leads to an analytical form of the corresponding aggregate loss distribution, which avoids the need for simulations to evaluate the model.

Mixture models are often used to reflect the heterogeneity in a population consisting of multiple groups or clusters (McLachlan and Peel, 2001). In some applications, these clusters can be physically identified and used to interpret the fitted distributions. This is however not the approach we follow; the components

in the mixture will not be identified with existing groups. Mixtures of Erlangs are discussed here for their great flexibility in modeling data and should be regarded as a semiparametric density estimation technique. The densities in the mixture are parametrically specified as Erlangs, whereas the associated weights form the nonparametric part. The number of Erlangs in the mixture with non-zero weights can be viewed as a smoothing parameter. Mixtures of Erlangs have much of the flexibility of nonparametric approaches and furthermore allow for tractable results.

The expectation-maximization (EM) algorithm, first introduced by Dempster et al. (1977), is an iterative method used to compute maximum likelihood (ML) estimates when the data can be viewed as being incomplete and direct maximization of the incomplete data likelihood is either not desirable or not possible (McLachlan and Krishnan, 2008). The EM algorithm is particularly useful in estimating the parameters of a finite mixture. The clue is to view data from a mixture as being incomplete since the associated component-label vectors are not available (McLachlan and Peel, 2001).

Lee and Lin (2010) iteratively use the EM algorithm (Dempster et al., 1977) for finite mixtures to estimate the parameters of a mixture of Erlang distributions with a common scale parameter. For a specified fixed set of shapes, the E- and M-step can be solved analytically without using any optimization method. This makes the EM algorithm for mixtures of Erlangs a pure iterative algorithm which is therefore simple, effective and easy to implement. The initialization is based on Tijm's proof of the denseness property of mixtures of Erlangs (Tijms, 1994, p. 163) which ensures good starting values and fast convergence. Since the number of Erlangs in the mixture and the corresponding shape parameters are pre-fixed and hence not estimated, Lee and Lin (2010) propose an adjustment procedure to identify the 'optimal' number of Erlang distributions and the 'optimal' shape parameters of these distributions in the mixture. The authors illustrate the flexibility of mixtures of Erlangs by generating data from parametric models (such as the uniform, lognormal, and generalized Pareto distribution) and by approximating the underlying distribution of this sample using a mixture of Erlangs. They further demonstrate the usefulness of mixtures of Erlangs in the context of quantitative risk management for the insurance business. However, modeling censored and/or truncated losses is not covered by the approach in Lee and Lin (2010).

In many practical problems data are censored and/or truncated, for example, due to the way how the data is collected or measured or by the design of the

experiment. Censoring entails that you only know in which interval an observation of a variable lies without knowing the exact value while truncation implies that you only observe values that lie within a given range. Interest however is in the underlying distribution of the uncensored and untruncated data instead of the observed censored and/or truncated data. Hence the censoring and truncation has to be accounted for in the analysis.

Survival analysis is the most common application in which data are often censored and truncated. A typical example is a medical study in which one follows patients over a period of time. In case the event of interest has not yet occurred before the end of the study, the patient drops out of the study or dies from another cause, independent of the cause of interest, the event time is right censored. In case the event of interest is known to have occurred between two dates, but the precise date is not known, the event time is interval censored. In actuarial science, insurance losses are often censored and truncated due to policy modifications such as deductibles (left truncation) and policy limits (right censoring). Left truncation is also present in life insurance where members of pension schemes and holders of insurance contracts only enter a portfolio at a certain adult age. Censored and truncated data occur in the context of claim reserving as well (Antonio and Plat, 2014). Indeed, the reserving actuary wants to predict the future development of claims when setting aside reserves at the present moment and has to deal with claims being reported but not yet settled (RBNS) and claims being incurred but not yet reported (IBNR). In operational risk, data are left truncated as they are only recorded in case they exceed a certain threshold. Badescu et al. (2015) use the EM algorithm to fit the correlated frequencies of such left truncated operational loss data using an Erlang-based multivariate mixed Poisson distribution.

Motivated by the large number of areas where censored and truncated data are encountered, the objective in this chapter is to develop an extension of the iterative EM algorithm of Lee and Lin (2010) for fitting mixtures of Erlangs with common scale parameter to censored and truncated data. The traditional way of dealing with (grouped and) truncated data using the EM algorithm involves treating the unknown number of truncated observations as a random variable and including it into the complete data vector (Dempster et al., 1977; McLachlan and Krishnan, 2008, p. 66; McLachlan and Peel, 2001, p. 257; McLachlan and Jones, 1988). We do not follow this approach and rather only include the uncensored observations and the component-label vectors in the complete data vector as is also done in Lee and Scott (2012). The fitting procedure is applicable to a wide range of applications. We demonstrate its use in actuarial science and econometrics.

In the following, we briefly introduce mixtures of Erlangs with a common scale parameter in Section 2.2. The adjusted EM algorithm, able to deal with censored and truncated data, is presented in Section 2.3. The procedures used to initialize the parameters, to adjust the shapes of the Erlangs in the mixture and to choose the number of components are discussed in Section 2.4. Examples follow in Section 2.5 and Section 2.6 concludes.

2.2 Mixtures of Erlangs with a common scale parameter

The Erlang distribution is a positive continuous distribution with density function

$$f(x; r, \theta) = \frac{x^{r-1} e^{-x/\theta}}{\theta^r (r-1)!} \qquad \text{for } x > 0\,, \tag{2.1}$$

where r, a positive integer, is the shape parameter and $\theta > 0$ the scale parameter (the inverse $\lambda = 1/\theta$ is called the rate parameter). The cumulative distribution function is obtained by integrating (2.1) by parts r times

$$F(x; r, \theta) = \int_0^x \frac{z^{r-1} e^{-z/\theta}}{\theta^r (r-1)!} dz = 1 - \sum_{n=0}^{r-1} e^{-x/\theta} \frac{(x/\theta)^n}{n!}\,. \tag{2.2}$$

Following Lee and Lin (2010) we consider mixtures of M Erlang distributions with common scale parameter $\theta > 0$ and having density

$$f(x; \boldsymbol{\alpha}, \boldsymbol{r}, \theta) = \sum_{j=1}^{M} \alpha_j \frac{x^{r_j-1} e^{-x/\theta}}{\theta^{r_j} (r_j-1)!} = \sum_{j=1}^{M} \alpha_j f(x; r_j, \theta) \qquad \text{for } x > 0\,, \tag{2.3}$$

where the positive integers $\boldsymbol{r} = (r_1, \ldots, r_M)$ with $r_1 < \ldots < r_M$ are the shape parameters of the Erlang distributions and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M)$ with $\alpha_j > 0$ and $\sum_{j=1}^{M} \alpha_j = 1$ are the weights used in the mixture. Similarly, the cumulative distribution function can be written as a weighted sum of terms (2.2) or (2.22).

Tijms (1994, p. 163) shows that the class of mixtures of Erlang distributions with a common scale parameter is dense in the space of distributions on $\mathbb{R}^+$. The formulation of the Theorem is given in Appendix 2.7. Lee and Lin (2010) give an alternative proof using characteristic functions.

2.3 The EM algorithm for censored and truncated data

Lee and Lin (2010) formulate the EM algorithm customized for fitting mixtures of Erlangs with a common scale parameter to complete data. Here, we construct an adjusted EM algorithm which is able to deal with censored and truncated data. We represent a censored sample truncated to the range $[t^l, t^u]$ by $\mathcal{X} = \{(l_i, u_i) | \, i = 1, \ldots, n\}$, where t^l and t^u represent the lower and upper truncation points, l_i and u_i the lower and upper censoring points and $t^l \leqslant l_i \leqslant u_i \leqslant t^u$ for $i = 1, \ldots, n$. $t^l = 0$ and $t^u = \infty$ mean no truncation from below and above, respectively. The censoring status is determined as follows:

Uncensored:	$t^l \leqslant l_i = u_i =: x_i \leqslant t^u$
Left Censored:	$t^l = l_i < u_i < t^u$
Right Censored:	$t^l < l_i < u_i = t^u$
Interval Censored:	$t^l < l_i < u_i < t^u$

For example, when the truncation interval equals $[t^l, t^u] = [0, 10]$, an uncensored observation at 1 gets denoted by $(l_i, u_i) = (1, 1)$, an observation left censored at 2 by $(l_i, u_i) = (0, 2)$, an observation right censored at 3 by $(l_i, u_i) = (3, 10)$ and an observation censored between 4 and 5 by $(l_i, u_i) = (4, 5)$. Thus, l_i and u_i should be seen as the lower and upper endpoints of the interval that contains observation i.

The parameter vector to be estimated is $\boldsymbol{\Theta} = (\boldsymbol{\alpha}, \theta)$. The number of Erlangs M in the mixture and the corresponding positive integer shapes $\boldsymbol{r}$ are fixed. The value of M is, in most applications, however unknown and has to be inferred from the available data, along with the shape parameters, see Section 2.4. The portion of the likelihood containing the unknown parameter vector $\boldsymbol{\Theta}$ is given by

$$\mathcal{L}(\boldsymbol{\Theta}; \mathcal{X}) = \prod_{i \in U} \frac{f(x_i; \boldsymbol{\Theta})}{F(t^u; \boldsymbol{\Theta}) - F(t^l; \boldsymbol{\Theta})} \prod_{i \in C} \frac{F(u_i; \boldsymbol{\Theta}) - F(l_i; \boldsymbol{\Theta})}{F(t^u; \boldsymbol{\Theta}) - F(t^l; \boldsymbol{\Theta})}$$

where U is the subset of observations in $\{1, \ldots, n\}$ which are uncensored and C is the subset of left, right and interval censored observations. In case there is no truncation, i.e. $[t^l, t^u] = [0, \infty]$, the contribution of a left censored observation to the likelihood equals $F(u_i; \boldsymbol{\Theta})$ since $l_i = 0$, of a right censored observation $1 - F(l_i; \boldsymbol{\Theta})$ with $u_i = \infty$, and of an interval censored observation $F(u_i; \boldsymbol{\Theta}) - F(l_i; \boldsymbol{\Theta})$.

The corresponding log-likelihood is

$$l(\boldsymbol{\Theta};\mathcal{X}) = \sum_{i\in U} \ln\left(\sum_{j=1}^{M} \alpha_j f(x_i;r_j,\theta)\right) + \sum_{i\in C} \ln\left(\sum_{j=1}^{M} \alpha_j \left(F(u_i;r_j,\theta) - F(l_i;r_j,\theta)\right)\right) - n\ln\left(\sum_{j=1}^{M} \alpha_j \left(F(t^u;r_j,\theta) - F(t^l;r_j,\theta)\right)\right), \tag{2.4}$$

which is difficult to optimize numerically.

2.3.1 Truncated mixture of Erlangs

The probability density function evaluated at an uncensored observation x_i after truncation (t^l, t^u) is given by

$$\begin{aligned} f(x_i;t^l,t^u,\boldsymbol{\Theta}) &= \frac{f(x_i;\boldsymbol{\Theta})}{F(t^u;\boldsymbol{\Theta}) - F(t^l;\boldsymbol{\Theta})} \\ &= \sum_{j=1}^{M} \alpha_j \cdot \frac{f(x_i;r_j,\theta)}{F(t^u;\boldsymbol{\Theta}) - F(t^l;\boldsymbol{\Theta})} \\ &= \sum_{j=1}^{M} \alpha_j \cdot \frac{F(t^u;r_j,\theta) - F(t^l;r_j,\theta)}{F(t^u;\boldsymbol{\Theta}) - F(t^l;\boldsymbol{\Theta})} \cdot \frac{f(x_i;r_j,\theta)}{F(t^u;r_j,\theta) - F(t^l;r_j,\theta)} \\ &= \sum_{j=1}^{M} \beta_j f(x_i;t^l,t^u,r_j,\theta), \end{aligned} \tag{2.5}$$

for $t^l \leqslant x_i \leqslant t^u$ and zero otherwise. This is again a mixture with mixing weights β_j and component density functions given by, respectively,

$$\beta_j = \alpha_j \cdot \frac{F(t^u;r_j,\theta) - F(t^l;r_j,\theta)}{F(t^u;\boldsymbol{\Theta}) - F(t^l;\boldsymbol{\Theta})} \tag{2.6}$$

and

$$f(x_i;t^l,t^u,r_j,\theta) = \frac{f(x_i;r_j,\theta)}{F(t^u;r_j,\theta) - F(t^l;r_j,\theta)}. \tag{2.7}$$

The component density functions $f(x_i;t^l,t^u,r_j,\theta)$ are truncated versions of the original component density functions $f(x_i;r_j,\theta)$. The weights β_j are obtained by reweighting the original weights α_j by means of the probabilities of the corresponding component to lie in the truncation interval.

2.3.2 Construction of the complete data vector

The EM algorithm provides a computationally easy way to fit this finite mixture to the censored and truncated data. The main clue is to regard the censored sample $\mathcal{X}$ as being incomplete since the uncensored observations $\boldsymbol{x} = (x_1, \ldots, x_n)$ and their associated component-indicator vectors $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ with

$$z_{ij} = \begin{cases} 1 & \text{if observation } x_i \text{ comes from the mixture component (2.7)} \\ & \text{corresponding to the shape parameter } r_j \\ 0 & \text{otherwise} \end{cases} \tag{2.8}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, M$, are not available. The component-label vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ are distributed according to a multinomial distribution consisting of one draw on M categories with probabilities $\beta_1, \ldots, \beta_M$ where

$$P(\boldsymbol{Z}_i = \boldsymbol{z}_i) = \beta_1^{z_{i1}} \ldots, \beta_M^{z_{iM}}$$

for $i = 1, \ldots, n$ with z_{ij} equal to 0 or 1 and $\sum_{j=1}^{M} z_{ij} = 1$. We write

$$\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n \overset{\text{i.i.d.}}{\sim} \text{Mult}_M(1, \boldsymbol{\beta}) \,.$$

Hence, the latent variables $\boldsymbol{Z}_i$ reveal which component density generated observation x_i. Whereas the unconditional truncated probability density function is given by (2.5), the conditional truncated probability density function of X_i given $Z_{ij} = 1$ is given by (2.7).

The complete data vector, $\mathcal{Y} = (x_1, \ldots, x_n, \boldsymbol{z}) = \{(x_i, \boldsymbol{z_i}) | i = 1 \ldots n\}$, contains all uncensored observations x_i and their corresponding mixing component vector $\boldsymbol{z}_i$. The log-likelihood of the complete sample $\mathcal{Y}$ then becomes

$$l(\boldsymbol{\Theta}; \mathcal{Y}) = \sum_{i=1}^{n} \sum_{j=1}^{M} z_{ij} \ln \left(\beta_j f(x_i; t^l, t^u, r_j, \theta) \right) \,, \tag{2.9}$$

which has a simpler form than the incomplete log-likelihood (2.4) as it does not contain logarithms of sums. The EM algorithm deals with the censored and truncated data from the mixture of Erlangs with common scale in the following steps.

2.3.3 Initial step

An initial guess for $\boldsymbol{\Theta}$ is needed to start the algorithm. The closer the starting value is to the true maximum likelihood estimator, the faster the algorithm will converge. Parameter initialization is often the sore point of an EM implementation and the study of good initial estimates is often not feasible and disregarded.

For mixtures of Erlangs however, the denseness property (see Tijms (1994, p. 163) and Appendix 2.7) provides an excellent way of coming up with good initial estimates. In the initial step, we deal with the censoring and truncation in a crude manner. We switch to an initializing data set, denoted by $\boldsymbol{d}$, in which we treat the left and right censored data points as being observed, i.e. we use u_i and l_i, respectively, and we replace the interval censored data points with the midpoint, i.e. we use $(l_i + u_i)/2$. Based on this initial data, we initialize the parameters θ and $\boldsymbol{\alpha}$ as:

$$\theta^{(0)} = \frac{\max(\boldsymbol{d})}{r_M} \quad \text{and} \quad \alpha_j^{(0)} = \frac{\sum_{i=1}^{n} I\left(r_{j-1}\theta^{(0)} < d_i \leqslant r_j\theta^{(0)}\right)}{n}, \tag{2.10}$$

for $j = 1, \ldots, M$, with $r_0 = 0$ for notational convenience. Inspired by Tijms's formulation of the denseness property, the initial scale $\theta^{(0)}$ is chosen such that $\theta^{(0)}r_M$ equals the maximum data point and the initial weights α_j for $j = 1, 2, \ldots, M$ are set to be the relative frequency of data points in the interval $(r_{j-1}\theta^{(0)}, r_j\theta^{(0)}]$. The truncation is only taken into account to transform the initial values for $\boldsymbol{\alpha}$ into the initial values for $\boldsymbol{\beta}$ via (2.6).

2.3.4 E-step

In the kth iteration of the E-step, we take the conditional expectation of the complete log-likelihood (2.9) given the incomplete data $\mathcal{X}$ and using the current estimate $\boldsymbol{\Theta}^{(k-1)}$ for $\boldsymbol{\Theta}$ with

$$\begin{aligned} Q(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(k-1)}) &= E(l(\boldsymbol{\Theta}; \mathcal{Y}) \mid \mathcal{X}; \boldsymbol{\Theta}^{(k-1)}) \\ &= E\left[\sum_{i \in U}\sum_{j=1}^{M} Z_{ij} \ln\left(\beta_j f(x_i; t^l, t^u, r_j, \theta)\right) \middle| \mathcal{X}; \boldsymbol{\Theta}^{(k-1)}\right] \\ &\quad + E\left[\sum_{i \in C}\sum_{j=1}^{M} Z_{ij} \ln\left(\beta_j f(X_i; t^l, t^u, r_j, \theta)\right) \middle| \mathcal{X}; \boldsymbol{\Theta}^{(k-1)}\right] \\ &= Q_u(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(k-1)}) + Q_c(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(k-1)}), \end{aligned} \tag{2.11}$$

where $Q_u(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(k-1)})$ and $Q_c(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(k-1)})$ are the conditional expectations of the uncensored and censored part of the complete log-likelihood, respectively.

Uncensored case. The truncation does not complicate the computation of the expectation for the uncensored data as

$$\begin{aligned} Q_u(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(k-1)}) &= E\left[\sum_{i \in U} \sum_{j=1}^{M} Z_{ij} \ln\left(\beta_j f(x_i; t^l, t^u, r_j, \theta)\right) \middle| \mathcal{X}; \boldsymbol{\Theta}^{(k-1)} \right] \\ &= \sum_{i \in U} \sum_{j=1}^{M} E\left[Z_{ij} | \, \mathcal{X}; \boldsymbol{\Theta}^{(k-1)} \right] \ln\left(\beta_j f(x_i; t^l, t^u, r_j, \theta)\right) \\ &= \sum_{i \in U} \sum_{j=1}^{M} {}^{u}z_{ij}^{(k)} \ln\left(\beta_j f(x_i; t^l, t^u, r_j, \theta)\right) \\ &= \sum_{i \in U} \sum_{j=1}^{M} {}^{u}z_{ij}^{(k)} \Big[\ln(\beta_j) + (r_j - 1)\ln(x_i) - \frac{x_i}{\theta} - r_j \ln(\theta) \\ &\quad - \ln((r_j - 1)!) - \ln\left(F(t^u; r_j, \theta) - F(t^l; r_j, \theta)\right) \Big] , \end{aligned} \tag{2.12}$$

with, for $i \in U$ and $j = 1, \ldots, M$,

$$\begin{aligned} {}^{u}z_{ij}^{(k)} &= P(Z_{ij} = 1 \mid x_i, t^l, t^u; \boldsymbol{\Theta}^{(k-1)}) \\ &= \frac{\beta_j^{(k-1)} f(x_i; t^l, t^u, r_j, \theta^{(k-1)})}{\sum_{m=1}^{M} \beta_m^{(k-1)} f(x_i; t^l, t^u, r_m, \theta^{(k-1)})} \\ &\overset{(2.7)}{=} \frac{\beta_j^{(k-1)} \frac{f(x_i; r_j, \theta^{(k-1)})}{F(t^u; r_j, \theta^{(k-1)}) - F(t^l; r_j, \theta^{(k-1)})}}{\sum_{m=1}^{M} \beta_m^{(k-1)} \frac{f(x_i; r_m, \theta^{(k-1)})}{F(t^u; r_m, \theta^{(k-1)}) - F(t^l; r_m, \theta^{(k-1)})}} \\ &\overset{(2.6)}{=} \frac{\alpha_j^{(k-1)} f(x_i; r_j, \theta^{(k-1)})}{\sum_{m=1}^{M} \alpha_m^{(k-1)} f(x_i; r_m, \theta^{(k-1)})} , \end{aligned} \tag{2.13}$$

where we plugged in definitions (2.6) and (2.7) of the weights and components of the truncated mixture in the last two equations in order to express this probability in terms of the original mixing weights and mixing components. The E-step for the uncensored part only requires the computation of the posterior probabilities ${}^{u}z_{ij}^{(k)}$ that observation i belongs to the jth component in the mixture, which remains the same in the truncated case and in the untruncated case.

Censored case. Denote by ${}^{c}z_{ij}^{(k)}$ the posterior probability that observation i belongs to the jth component in the mixture for a censored data point. Then

$$\begin{aligned}
&Q_c(\boldsymbol{\Theta};\boldsymbol{\Theta}^{(k-1)}) \\
&= E\left[\sum_{i\in C}\sum_{j=1}^{M} Z_{ij}\ln\left(\beta_j f(X_i;t^l,t^u,r_j,\theta)\right)\middle|\,\mathcal{X};\boldsymbol{\Theta}^{(k-1)}\right] \\
&= \sum_{i\in C} E\left[\sum_{j=1}^{M} Z_{ij}\ln\left(\beta_j f(X_i;t^l,t^u,r_j,\theta)\right)\middle|\,l_i,u_i,t^l,t^u;\boldsymbol{\Theta}^{(k-1)}\right] \\
&= \sum_{i\in C}\sum_{j=1}^{M} {}^{c}z_{ij}^{(k)} E\left[\ln\left(\beta_j f(X_i;t^l,t^u,r_j,\theta)\right)\middle|\,Z_{ij}=1,l_i,u_i,t^l,t^u;\theta^{(k-1)}\right] \\
&= \sum_{i\in C}\sum_{j=1}^{M} {}^{c}z_{ij}^{(k)}\left[\ln(\beta_j)+(r_j-1)E\left(\ln(X_i)\middle|Z_{ij}=1,l_i,u_i,t^l,t^u;\theta^{(k-1)}\right)\right. \\
&\quad -\frac{1}{\theta}E\left(X_i\middle|Z_{ij}=1,l_i,u_i,t^l,t^u;\theta^{(k-1)}\right)-r_j\ln(\theta)-\ln((r_j-1)!) \\
&\quad \left.-\ln\left(F(t^u;r_j,\theta)-F(t^l;r_j,\theta)\right)\right] \qquad (2.14)
\end{aligned}$$

where we used the tower rule in the third equality. Again using Bayes' rule, we can compute these posterior probabilities, for $i\in C$ and $j=1,\ldots,M$, as

$$\begin{aligned}
{}^{c}z_{ij}^{(k)} &= P(Z_{ij}=1\mid l_i,u_i,t^l,t^u;\boldsymbol{\Theta}^{(k-1)}) \\
&= \frac{\beta_j^{(k-1)}\left(F(u_i;t^l,t^u,r_j,\theta^{(k-1)})-F(l_i;t^l,t^u,r_j,\theta^{(k-1)})\right)}{\sum_{j=1}^{M}\beta_j^{(k-1)}\left(F(u_i;t^l,t^u,r_j,\theta^{(k-1)})-F(l_i;t^l,t^u,r_j,\theta^{(k-1)})\right)} \\
&= \frac{\beta_j^{(k-1)}\frac{F(u_i;r_j,\theta^{(k-1)})-F(l_i;r_j,\theta^{(k-1)})}{F(t^u;r_j,\theta^{(k-1)})-F(t^l;r_j,\theta^{(k-1)})}}{\sum_{j=1}^{M}\beta_j^{(k-1)}\frac{F(u_i;r_j,\theta^{(k-1)})-F(l_i;r_j,\theta^{(k-1)})}{F(t^u;r_j,\theta^{(k-1)})-F(t^l;r_j,\theta^{(k-1)})}} \\
&\overset{(2.6)}{=} \frac{\alpha_j^{(k-1)}\left(F(u_i;r_j,\theta^{(k-1)})-F(l_i;r_j,\theta^{(k-1)})\right)}{\sum_{j=1}^{M}\alpha_j^{(k-1)}\left(F(u_i;r_j,\theta^{(k-1)})-F(l_i;r_j,\theta^{(k-1)})\right)}\,. \qquad (2.15)
\end{aligned}$$

The expression for the posterior probability in the censored case has the same form as in the uncensored case (2.13), but with the densities replaced by the probabilities in between the upper and lower censoring points. The terms in (2.14) for $Q_c(\boldsymbol{\Theta};\boldsymbol{\Theta}^{(k-1)})$ containing $E(\ln(X_i)|Z_{ij}=1,l_i,u_i,t^l,t^u;\theta^{(k-1)})$ will not play a role in the EM algorithm as they do not depend on the unknown parameter vector $\boldsymbol{\Theta}$. The E-step requires the computation of the expected value of X_i conditional on the censoring times and the mixing component $\boldsymbol{Z}_i$ for the current value $\boldsymbol{\Theta}^{(k-1)}$

of $\boldsymbol{\Theta}$:

$$\begin{aligned}
&E\left(X_i \,\middle|\, Z_{ij}=1, l_i, u_i, t^l, t^u; \theta^{(k-1)}\right) \\
&\quad= \int_{l_i}^{u_i} x \frac{f(x; r_j, \theta^{(k-1)})}{F(u_i; r_j, \theta^{(k-1)}) - F(l_i; r_j, \theta^{(k-1)})} dx \\
&\quad= \frac{r_j \theta^{(k-1)}}{F(u_i; r_j, \theta^{(k-1)}) - F(l_i; r_j, \theta^{(k-1)})} \int_{l_i}^{u_i} \frac{x^{r_j} e^{-x/\theta^{(k-1)}}}{\left(\theta^{(k-1)}\right)^{r_j+1} r_j!} dx \\
&\quad= \frac{r_j \theta^{(k-1)} \left(F(u_i; r_j+1, \theta^{(k-1)}) - F(l_i; r_j+1, \theta^{(k-1)})\right)}{F(u_i; r_j, \theta^{(k-1)}) - F(l_i; r_j, \theta^{(k-1)})},
\end{aligned}$$

for $i \in C$ and $j = 1, \ldots, M$, which has a closed-form expression.

2.3.5 M-step

In the M-step, we maximize the expected value (2.11) of the complete data log-likelihood obtained in the E-step with respect to the parameter vector $\boldsymbol{\Theta}$ over all $(\boldsymbol{\beta}, \theta)$ with $\beta_j > 0$, $\sum_{j=1}^{M} \beta_j = 1$ and $\theta > 0$. The expressions for $Q_u(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(k-1)})$ and $Q_c(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(k-1)})$ are given in (2.12) and (2.14), respectively. The maximization over the mixing weights $\boldsymbol{\beta}$, requires the maximization of

$$\sum_{i \in U} \sum_{j=1}^{M} {}^{u}z_{ij}^{(k)} \ln(\beta_j) + \sum_{i \in C} \sum_{j=1}^{M} {}^{c}z_{ij}^{(k)} \ln(\beta_j).$$

We implement the restriction $\sum_{j=1}^{M} \beta_j = 1$ by setting $\beta_M = 1 - \sum_{j=1}^{M-1} \beta_j$. Setting the partial derivatives at $\boldsymbol{\beta}^{(k)}$ equal to zero implies that the optimizer satisfies

$$\beta_j^{(k)} = \frac{\sum_{i \in U} {}^{u}z_{ij}^{(k)} + \sum_{i \in C} {}^{c}z_{ij}^{(k)}}{\sum_{i \in U} {}^{u}z_{iM}^{(k)} + \sum_{i \in C} {}^{c}z_{iM}^{(k)}} \beta_M^{(k)} \qquad \text{for } j = 1, \ldots, M-1.$$

By the sum constraint we have

$$\beta_M^{(k)} = \frac{\sum_{i \in U} {}^{u}z_{iM}^{(k)} + \sum_{i \in C} {}^{c}z_{iM}^{(k)}}{n},$$

and the same form also follows for $j = 1, \ldots, M-1$:

$$\beta_j^{(k)} = \frac{\sum_{i \in U} {}^{u}z_{ij}^{(k)} + \sum_{i \in C} {}^{c}z_{ij}^{(k)}}{n} \qquad \text{for } j = 1, \ldots, M. \tag{2.16}$$

The new estimate for the prior probability β_j in the truncated mixture is the average of the posterior probabilities of belonging to the jth component in the mixture. The optimizer indeed corresponds to a maximum since the matrix of second order partial derivatives is negative definite matrix with a compound symmetry structure.

In order to maximize $Q(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(k-1)})$ with respect to θ, we set the first order partial derivatives equal to zero (see Appendix 2.8). This leads to the following M-step equation for θ:

$$\theta^{(k)} = \frac{\left(\sum_{i \in U} x_i + \sum_{i \in C} E\left(X_i \left| l_i, u_i, t^l, t^u; \theta^{(k-1)} \right.\right)\right)/n - T^{(k)}}{\sum_{j=1}^{M} \beta_j^{(k)} r_j}, \tag{2.17}$$

with

$$T^{(k)} = \sum_{j=1}^{M} \beta_j^{(k)} \frac{\left(t^l\right)^{r_j} e^{-t^l/\theta} - \left(t^u\right)^{r_j} e^{-t^u/\theta}}{\theta^{r_j - 1}(r_j - 1)!\left(F(t^u; r_j, \theta) - F(t^l; r_j, \theta)\right)}\Bigg|_{\theta = \theta^{(k)}}.$$

As in the uncensored case, the new estimate $\theta^{(k)}$ in (2.17) for the common scale parameter θ again has the interpretation of the sample mean divided by the average shape parameter in the mixture, but in the formula for the sample mean, we now take the expected value of the censored data points given the censoring times and subtract a correction term $T^{(k)}$ due to the truncation. However, $T^{(k)}$ in (2.17) depends on $\theta^{(k)}$ and has a complicated form. Therefore, it is not possible to find an analytical solution and we resort to a Newton-type algorithm to solve (2.17) numerically using the previous value $\theta^{(k-1)}$ as starting value.

The E- and M-steps are iterated until $l(\boldsymbol{\Theta}^{(k)}; \mathcal{X}) - l(\boldsymbol{\Theta}^{(k-1)}; \mathcal{X})$ is sufficiently small. The maximum likelihood estimator of the original mixing weights α_j for $j = 1, \ldots, M$ can be retrieved by inverting expression (2.6). This is most easily done by first computing

$$\widetilde{\alpha}_j = \frac{\widehat{\beta}_j}{F(t^u; r_j, \widehat{\theta}) - F(t^l; r_j, \widehat{\theta})} \qquad \text{for } j = 1, \ldots, M,$$

where $\widehat{\beta}_j$ and $\widehat{\theta}$ denote the values in the final EM step, and then normalizing the weights such that they sum to 1.

2.4 Choice of the shape parameters and of the number of Erlangs in the mixture

2.4.1 Initialization

We start by making an initial choice for the number of Erlangs M in the mixture and set the shapes equal to $r_j = j$ for $j = 1, 2, \ldots, M$. Extending Lee and Lin (2010), we introduce a spread factor s by which we multiply the shapes in order to get a wider spread at the initial step, i.e. $r_j = sj$ for $j = 1, 2, \ldots, M$.

The initialization of θ and $\boldsymbol{\alpha}$ is based on the denseness of mixtures of Erlangs (see (Tijms, 1994, p. 163) and Appendix 2.7), as explained in Section 2.3.3. Each weight α_j gets initialized as the relative frequency of data points in the interval corresponding to the shape parameter r_j. In case this interval does not contain any data points for some j, the initial weight corresponding to the Erlang in the mixture with shape r_j will be zero and consequently the weight α_j will remain zero at each subsequent iteration. This is clear from the updating scheme (2.16) in the M-step and the expressions (2.13) and (2.15) of the posterior probabilities in the E-step. The shapes r_j with initial weight α_j equal to zero are therefore removed from the mixture at the initial step.

Numerical experiments show that the iterative scheme performs well and results in fast convergence using the above choice of initial estimates for θ and $\boldsymbol{\alpha}$.

2.4.2 Adjusting the shapes

Since the initial shape parameters are pre-fixed and hence not estimated, the fitted mixture might be sub-optimal. Adjustment of the shape parameters is necessary. Ideally, for a given number of Erlangs M, we want to choose optimal values for the shapes. The choice of the shapes for a given M however is an optimization problem over $\mathbb{N}^M$ which is impossible to solve. We have to resort to a practical procedure which explores the parameter space efficiently in order to obtain a satisfying choice for the shapes.

After applying the EM algorithm a first time to obtain the maximum likelihood estimates corresponding to the initial choice of the shape parameters, we perform stepwise variations of the shapes, each time refitting the scale and the weights using the EM algorithm, and compare the log-likelihoods of the results. We hereby follow the procedure proposed by Lee and Lin (2010):

(i) Run the algorithm starting from the shapes $\{r_1, \ldots, r_{M-1}, r_M + 1\}$ with initial scale θ and weights $\{\beta_1, \ldots, \beta_{M-1}, \beta_M\}$ equal to the final estimates of the previous execution of the EM algorithm. Repeat this step for as long as the log-likelihood improves, each time replacing the old set of parameters by the new ones. This procedure is then applied on the $(M-1)$th shape and so forth until all the shapes are treated.

(ii) Run the algorithm starting from the shapes $\{r_1 - 1, r_2, \ldots, r_M\}$ with initial scale θ and weights $\{\beta_1, \beta_2, \ldots, \beta_M\}$ the final estimates of the previous execution of the EM algorithm. Repeat this step for as long as the log-likelihood improves, each time replacing the old set of parameters by the new ones. This procedure is then applied on the 2nd shape and so forth until all the shapes are treated.

(iii) Repeat the loops described in the previous steps until the log-likelihood can no longer be increased.

Using this algorithm we eventually reach a local maximum of the log-likelihood, by which we mean that the fit can no longer be improved by either increasing or decreasing any of the r_j.

2.4.3 Reducing the number of Erlangs

Too many Erlangs in the mixture will result in an issue of overfitting, which is always a problem in statistical modeling. A decision rule such as Akaike's information criterion (AIC, Akaike, 1974) or Schwartz's Bayesian information criterion (BIC, Schwarz, 1978) helps to decide on the value of M. Models with smaller AIC and BIC values are preferred. Any other information criterion (IC) or objective function could be optimized depending on the purpose for which the model is used.

The problem of testing for the number of components is of both theoretical and practical importance and has attracted considerable attention of many studies over the years and still is a major contemporary issue in a mixture modeling context where the underlying population can be conceptualized as being composed of a finite number of subpopulations. Since mixtures of Erlangs are employed here as a semi-parametric density estimation technique and not as model-based clustering, the commonly used criteria of AIC and BIC are adequate for choosing the number of components (McLachlan and Peel, 2001).

We use a backward stepwise search. As mixtures of Erlangs are dense in the space of positive continuous distributions, we start from a close-fitting mixture of

M Erlangs resulting from the shape adjustment procedure described in Section 2.4.2 and compute the value of the IC. We next reduce the number of Erlangs M in the mixture by deleting the mixture component of which the shape r_j has smallest weight β_j, refit the scale and weights using the EM algorithm and readjust the shapes using the same shape adjustment procedure. If the resulting fit with $M-1$ Erlangs attains a lower value of the IC, the new parameter values replace the old ones. We continue reducing the number of Erlangs in the mixture until the value of the IC does no longer decrease by deleting an additional mixture component.

A backward selection has the advantage of providing initial values close to the maximum likelihood estimates of the new set of shapes which greatly reduces the run time (Lee and Lin (2010)). In contrast, by using a forward stepwise procedure it is not clear which additional shape parameter to use and how the parameters from the previous run can be used to provide useful information on parameter initialization.

As a guideline, we recommend to start from an initial choice for the number of Erlangs M and a spread s resulting in a close-fitting or even overfitting of the data.

2.4.4 Compare the resulting fit using different initializing parameters

Since the log-likelihood has multiple local maxima, the value of the initializing parameters M and s can influence the result. Therefore, it is wise to compare the final fits, after the shape adjustment procedure and reduction of the number of Erlangs using an IC, starting from different choices for the initial number of Erlangs M and/or the spread factor s in the initial step. Tuning of such initializing parameters is common in different numerical algorithms and fitting strategies as well (Hastie et al., 2009). Specifically for the case of mixture of Erlangs, many values for the tuning parameters M and s can lead to a satisfying resulting fit, while using a different mixture of Erlangs representation. This is illustrated in the first data example (Section 2.5.1, Table 2.1). In order not to limit the flexibility of the fitting procedure, we do not prefix the value of M and s up front and do not propose any stringent rule. The examples in Section 2.5 show how a small search for these values is often sufficient to obtain satisfactory results. The freedom of doing an even wider search is left as an option to the user.

2.5 Examples

The usefulness of the proposed fitting procedure is demonstrated using several examples. A first example involves simulated data from a bimodal distrubution which we censor and truncate allowing us to compare the original density and the entire uncensored and untruncated sample to the fitted mixture of Erlangs. The second example illustrates the use of mixtures of Erlangs to represent right-censored unemployment durations. In the third example, we illustrate the use of mixtures of Erlangs in actuarial science in the context of loss modeling. We fit a mixture of Erlang distribution to truncated claim size data and demonstrate how the fitted mixture can be used to analytically price reinsurance contracts. In the final example, we generate data from a generalized Pareto distribution to explore limitations in modeling heavy-tailed distributions.

2.5.1 Simulated censored and truncated bimodal data

We generate a random sample of 5000 observations from the bimodal mixture of gamma distributions with density function given by

$$f_u(x) = 0.4f(x; r = 5, \theta = 0.5) + 0.6f(x; r = 10, \theta = 1)\,. \tag{2.18}$$

Next we truncate the data by rejecting all observations beneath the 5% sample quantile or above the 95% sample quantile. The remaining 4500 data points are subsequently being right censored by generating 4500 observations from another mixture of gamma distributions with density function

$$f_{rc}(x) = pf(x; r = 5, \theta = 2/3) + (1 - p)f(x; r = 9, \theta = 1.25)\,, \tag{2.19}$$

with $p = 0.4$. The resulting data set is composed of 2595 uncensored and 1905 right censored data points, and is used to calibrate the Erlang mixture, keeping the lower and upper truncation into account.

Using the automatic search from Section 2.4.4 we start from $M = 10$ Erlangs in the mixture and let the spread factor s used in the initial step range from 1 to 10. AIC is used to decide upon the number of Erlangs to use in the mixture as explained in Section 2.4.3. The right censored data points are treated as being observed at the initialization in (2.10). The different values of the initializing spread all lead to a different final Erlang mixture, which are reported in Table 2.1. This illustrates the importance of varying the initial spread. Based on the

AIC and BIC values (and plots of the fits not shown here), the different models all represent the data quite well.

Table 2.1: *Demonstration of initialization and fitting procedure on the data generated from (18). Starting point is a mixture of 10 Erlangs. The initial spread factor s ranges from 1 to 10. The superscripts in the last two columns represent the preference order according to that information criterium.*

s	$\boldsymbol{r}$	$\boldsymbol{\alpha}$	θ	AIC	BIC
1	3; 12	0.46; 0.54	0.83	13961.09^{5}	13993.15^{1}
2	4; 14; 18	0.44; 0.34; 0.22	0.63	13956.31^{2}	14001.19^{3}
3	6; 15; 23; 31	0.39; 0.12; 0.35; 0.15	0.41	13959.51^{3}	14017.22^{4}
4	5; 15; 21	0.42; 0.20; 0.38	0.51	13955.61^{1}	14000.50^{2}
5	9; 15; 29; 43; 58	0.23; 0.17; 0.14; 0.31; 0.15	0.22	13961.03^{4}	14031.56^{5}
6	8; 14; 29; 43; 59	0.21; 0.20; 0.15; 0.31; 0.13	0.22	13962.63^{6}	14033.16^{6}
7	14; 23; 34; 45; 58; 74; 96	0.20; 0.17; 0.05; 0.07; 0.14; 0.24; 0.13	0.13	13970.25^{10}	14066.42^{10}
8	10; 16; 24; 40; 55; 69; 89	0.12; 0.18; 0.11; 0.10; 0.16; 0.21; 0.12	0.15	13966.94^{8}	14063.11^{8}
9	11; 18; 28; 46; 63; 79; 101	0.11; 0.19; 0.11; 0.10; 0.17; 0.21; 0.11	0.13	13969.23^{9}	14065.41^{9}
10	13; 21; 32; 50; 67; 84; 107	0.14; 0.18; 0.09; 0.10; 0.17; 0.21; 0.11	0.12	13966.63^{7}	14062.81^{7}

The lowest AIC value was reached using spread factor $s = 4$ with a corresponding mixture of 3 Erlangs. The parameter estimates of this final model are given in Table 2.2.

Table 2.2: *Parameter estimates of the mixture of 3 Erlangs fitted to the censored and truncated data with underlying density (2.18).*

r_j	α_j	θ
5	0.4206869	0.5081993
15	0.2018598	
21	0.3774533	

In order to verify the goodness-of-fit, we might consider analytical tests such as the Kolmogorov-Smirnov test. However, the form of the test statistic and the corresponding distribution is not at all obvious in a censored and truncated setting. For the case of power-law distributions, Clauset et al. (2009) used Kolmogorov-Smirnov tests to evaluate whether the hypothesized distribution adequately describes the tail. Dufour and Maag (1978) modify the form of the test statistic to allow for truncated and censored data. Guilbaud (1988) derive an exact Kolmogorov-Smirnov test for left-truncated and/or right-censored data. In an actuarial context, Chernobai et al. (2014) discuss goodness-of-fit tests for left-truncated loss samples. We mainly focus on graphical goodness-of fit evaluation

in this chapter.

A graphical comparison of the fitted distribution and the originally generated data can be found in Figure 2.1. We compare the fitted mixture of Erlangs density to the true density (2.18) and a histogram of all 5000 generated data points before truncation and censoring in the left plot in Figure 2.1. The right plot in Figure 2.1 compares the truncated mixture of Erlangs density to the true truncated density and a histogram of the 4500 data points after truncation and before censoring. The fitted mixture of Erlangs density shows to be a very close approximation of the true density. Varying the spread from 1 to 10 in the initial mixture of 10 Erlangs is sufficient to obtain a satisfactory result, so there is no need to increase the number of Erlangs in the initial step.

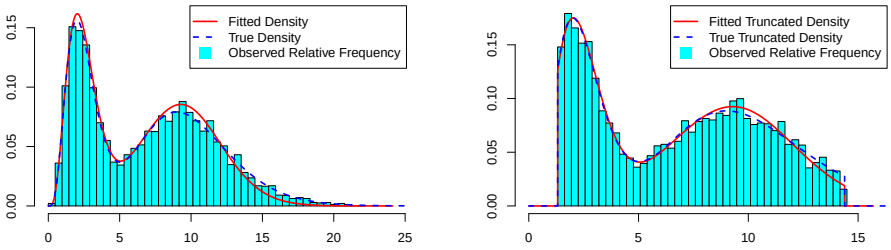

Figure 2.1: *Graphical comparison of the density of the fitted mixture of 3 Erlangs, the true underlying density (2.18) and the histogram of the generated data before censoring and truncation (left) and of the truncated density of the fitted mixture of 3 Erlangs, the true truncated density and the histogram of the generated data after truncated and before censoring (right).*

In actuarial practice, loss data can sometimes be of multimodal nature due to the fact that the property and casualty losses often come from multiple sources. Clearly, using standard parametric distributions will result in unsatisfactory approximations as they are incapable of reflecting the multimodal characteristic. Moreover, applying straightforward estimation techniques may lead to non-convergence issues due to the censoring and truncation. On the contrary, convergence is guaranteed in the presented EM algorithm for mixtures of Erlangs and captures the bimodality of the data very flexibly.

Next, we investigate the sensitivity with respect to the level of censoring in the data. To that end, we fix the data generated from (2.18), truncate them at the 5% and 95% sample quantile and vary the value of the mixing weight p in the

density (2.19) of the right censoring distribution from 0 to 1 by 0.1. Let $f(x)$ and $F(x)$ denote the true density and distribution function and $\widehat{f}(x)$ and $\widehat{F}(x)$ the estimated mixture of Erlangs density and distribution function. We measure the performance of both the underlying and the truncated mixture of Erlangs density estimator in approximating the underlying and the truncated true density by calculating the L^1 and L^2 norms:

$$\begin{aligned}
\mathrm{L}^1 &= \int_0^\infty \left|\widehat{f}(x) - f(x)\right| \mathrm{d}x \\
\mathrm{L}_t^1 &= \int_{t^l}^{t^u} \left|\frac{\widehat{f}(x)}{\widehat{F}(t^u) - \widehat{F}(t^l)} - \frac{f(x)}{F(t^u) - F(t^l)}\right| \mathrm{d}x \\
\mathrm{L}^2 &= \left(\int_0^\infty \left(\widehat{f}(x) - f(x)\right)^2 \mathrm{d}x\right)^{1/2} \\
\mathrm{L}_t^2 &= \left(\int_{t^l}^{t^u} \left(\frac{\widehat{f}(x)}{\widehat{F}(t^u) - \widehat{F}(t^l)} - \frac{f(x)}{F(t^u) - F(t^l)}\right)^2 \mathrm{d}x\right)^{1/2}.
\end{aligned}$$

For each value of p in the right censoring distribution (2.19), we generate 100 censoring samples of size 4500 and each time fit an Erlang mixture to the right censored data set using the automatic search starting from $M = 10$ Erlangs in the mixture and letting the initial spread s vary from 1 to 10. The averages of the performance measures over the 100 best-fitting resulting mixtures are shown in Table 2.3. The L^1 and L^2 norms over the truncation interval deteriorate when increasing the censoring level, but remain quite low. This reveals that the performance of the estimator remains excellent when the level of censoring increases, except at the highest level where the estimated Erlang mixture is still bimodal but the second mode and the tail of the true density are underestimated. The L^1 and L^2 norms over the entire positive real line do not run as parallel with the censoring level as the truncated versions. Note in this context the limitations of accurately estimating the density outside of the truncation interval, since no data has been observed in that region. One should hence not rely on probability statements made using the fitted Erlang mixture outside of the data range.

Table 2.3: *Results of the sensitivity analysis with respect to the level of censoring. For each value of p in the right censoring distribution (2.19), we generate 100 censoring samples and report the average censoring level and average performance measures of the best-fitting mixtures of Erlang distributions.*

p	censoring %	L^1	L^2	L_t^1	L_t^2
0.0	0.2172	0.0862	0.0227	0.0266	0.0097
0.1	0.2695	0.0594	0.0170	0.0280	0.0099
0.2	0.3224	0.0740	0.0197	0.0278	0.0099
0.3	0.3753	0.0864	0.0226	0.0309	0.0109
0.4	0.4289	0.1438	0.0343	0.0329	0.0114
0.5	0.4806	0.1129	0.0277	0.0367	0.0126
0.6	0.5330	0.0905	0.0235	0.0412	0.0140
0.7	0.5844	0.1527	0.0349	0.0465	0.0157
0.8	0.6383	0.1597	0.0377	0.0594	0.0199
0.9	0.6903	0.1787	0.0416	0.0705	0.0236
1.0	0.7426	0.5156	0.1199	0.2276	0.0997

2.5.2 Unemployment duration

We examine the economic data from the January Current Population Survey's Displaced Workers Supplements (DWS) for the years 1986, 1988, 1990, and 1992 which was first analyzed in McCall (1996). A thorough discussion of this data set is available in Cameron and Trivedi (2005). The variable under consideration is unemployment duration (`spell`) or more accurately joblessness duration, measured in two-week intervals. All other covariates in the data set are ignored in the analysis. Following Cameron and Trivedi (2005), a spell is considered complete if the person is re-employed at a full-time job (`CENSOR1` = 1) and right-censored otherwise (`CENSOR1` = 0). This results in 1073 uncensored data points and 2270 right censored data points.

The parameter estimates of the Erlang mixture, obtained by using the automatic search procedure starting from $M = 10$ Erlangs in the mixture with spread factor s in the initial step ranging from 1 to 10, are given in Table 2.4. AIC is again used to decide upon the number of Erlangs in the mixture and the right censored data points are treated as being observed at initialization. The lowest AIC value was obtained with a mixture of 8 Erlangs. This optimal choice of shapes was reached using spread factor $s = 10$.

Table 2.4: *Parameter estimates of the mixture of 8 Erlangs fitted to the right-censored unemployment data.*

r_j	α_j	θ
8	0.10563305	0.1477264
17	0.09443584	
33	0.08578746	
50	0.09099055	
73	0.04273362	
99	0.14814091	
135	0.07546787	
199	0.35681069	

The Kaplan-Meier estimator (Kaplan and Meier (1958)), also known as the product limit estimator, is the standard non-parametric estimator of the survival function in case of right censored data. The resulting survival curve is a step function with jumps at the observed event times of which the size not only depends on the number of events observed at each event time, but also on the pattern of the censored observations prior to that event time. In order to graphically evaluate the fit, we compare the Kaplan-Meier survival curve, along with 95% confidence bounds, to the survival function of the estimated Erlang mixture in Figure 2.2. Marks are added on the Kaplan-Meier estimate to indicate censoring times. The fitted survival function provides a smooth fit of the data, closely resembling the non-parametric estimate.

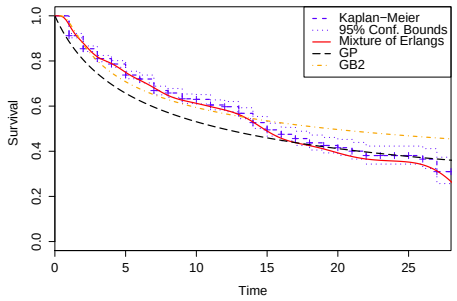

Figure 2.2: *Graphical comparison of the survival function of the fitted mixture of 8 Erlangs and the Kaplan-Meier estimator with 95% confidence bounds for the right-censored unemployment data.*

As an illustration, we also compare our approach to two commonly used parametric models, the generalized Pareto distribution (GP) and the generalized beta

distribution of the second kind (GB2). In Figure 2.2, we see how mixtures of Erlangs offer much more flexibility and lead to a more appropriate fit for these data at the cost of requiring more parameters. However, AIC and BIC strongly prefer the mixture of Erlangs approach, see Table 2.5.

Table 2.5: *Comparison of information criteria for the different models fitted to the right-censored unemployment data.*

Model	AIC	BIC
Mixtures of Erlangs	8066.281	8170.230
Generalized Pareto (GP)	8733.718	8745.947
Generalized beta 2 (GB2)	8280.168	8304.627

2.5.3 Secura Re, Belgian insurance data

The Secura Re data set discussed in Beirlant et al. (2004) contains 371 automobile claims from 1988 until 2001 gathered from several European insurance companies. The data are uncensored, but left truncated at 1 200 000 since a claim is only reported to the reinsurer if the claim size is at least as large as 1 200 000 euro. The sizes of the claims are corrected among others for inflation. Based on these observations, the reinsurer wants to calibrate a model in order to price reinsurance contracts.

The search procedure using AIC prefers a mixture of only two Erlangs with shapes 5 and 16. The parameter estimates of this best-fitting mixture are shown in Table 2.6. In Figure 2.3 (left) we compare the histogram of the truncated data to the fitted truncated density. Figure 2.3 (right) illustrates that the truncated survival function of the mixture of two Erlangs perfectly coincides with the Kaplan-Meier estimate.

Table 2.6: *Parameter estimates of the mixture of 2 Erlangs fitted to the left-truncated claim sizes in the Secura Re data set.*

r_j	α_j	θ
5	0.97103229	360 096.1
16	0.02896771	

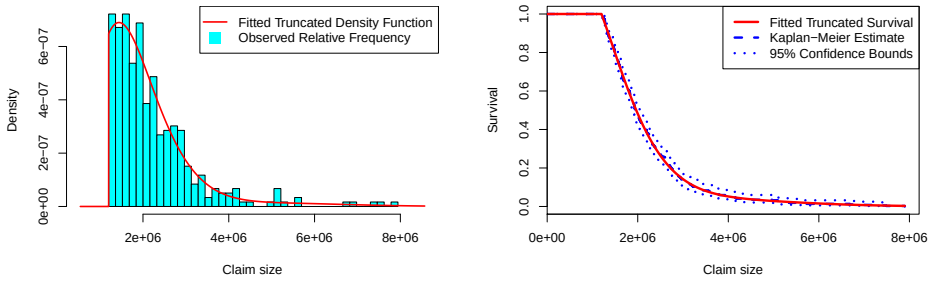

Figure 2.3: *Graphical comparison of the truncated density of the fitted mixture of 2 Erlangs and the histogram of the left-truncated claim sizes (left) and of the truncated survival function and the Kaplan-Meier estimator with 95% confidence bounds (right) for the Secura Re data set.*

In Figure 2.4, we validate the fit in the tail by plotting the QQ-plot on the left and the log-log plot of the empirical truncated survival function (black dots) and the truncated survival function of the best-fitting Erlang mixture (red line) on the right. Both figures show how the mixture of only two Erlangs achieves a adequate approximation in the tail.

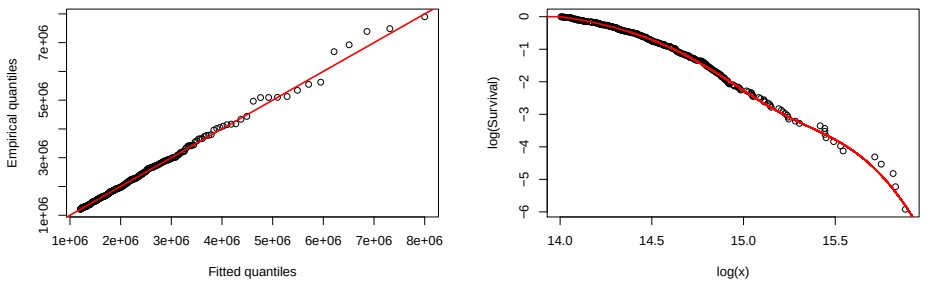

Figure 2.4: *QQ-plot of the empirical quantiles and the quantiles of the fitted mixture of 2 Erlangs with identity line (left) and log-log plot of the empirical truncated survival function and the truncated survival function of the fitted Erlang mixture (right) for the Secura Re data set.*

Following Beirlant et al. (2004, p. 188), we use the calibrated Erlang mixture to price an excess-of-loss (XL) reinsurance contract, where the reinsurer pays for the claim amount in excess of a given limit. The net premium $\Pi(R)$ of such a contract with retention level $R > 1\,200\,000$ is given by

$$\Pi(R) = E((X - R)_+ \mid X > 1\,200\,000)$$

where X denotes the claim size and $(\cdot)_+ = \max(\cdot, 0)$. In case X follows a mixture of M Erlang distributions, where we assume without loss of generality $r_i = i$ for $i = 1, \ldots, M$, the net premium is

$$\begin{aligned}\Pi(R) &= \frac{\theta e^{-R/\theta}}{1 - F(1\,200\,000; \boldsymbol{\alpha}, \boldsymbol{r}, \theta)} \sum_{n=0}^{M-1} \left(\sum_{k=n}^{M-1} A_k \right) \frac{(R/\theta)^n}{n!} \\ &= \frac{\theta^2}{1 - F(1\,200\,000; \boldsymbol{\alpha}, \boldsymbol{r}, \theta)} \sum_{n=1}^{M} \left(\sum_{k=n-1}^{M-1} A_k \right) f(R; n, \theta)\,, \end{aligned} \tag{2.20}$$

with $A_k = \sum_{j=k+1}^{M} \alpha_j$ for $k = 0, \ldots, M-1$. The derivation of this property can be reconstructed using Willmot and Woo (2007) or Klugman et al. (2013, p. 21). In Table 2.7, we compare the non-parametric, Hill and Generalized Pareto (GP) based estimates of $\Pi(R)$ for the Secura Re data set from Table 6.1 in Beirlant et al. (2004, p. 191) to the estimates obtained using formula (2.20). The maximum claim size observed in the data set equals 7 898 639 which is the only data point on which the non-parametric estimate of the net premium with retention level $R = 7\,500\,000$ is based. The non-parametric estimate corresponding to retention level $R = 10\,000\,000$ is hence zero. The fitted Erlang mixture allows us to estimate the net premium using intrinsically all data points, but postulates a lighter tail compared to the Pareto-type alternatives since Erlang mixtures have an asymptotically exponential tail (Neuts (1981, p. 62)). Both the estimates based on the extreme value methodology and those based on the Erlang mixture keep pace with the non-parametric ones, but at the high-end of the sample range, the estimators differ strongly, as implied by the different tail behavior of the three approaches. The reinsurance actuary should carefully investigate the right tail behavior of the data in order to choose his approach.

Besides modeling the tail of the claim size distribution above a certain threshold, Beirlant et al. (2004, p. 198) also estimate a global statistical model to describe the whole range of all possible claim outcomes for the Secura Re data set. This is needed when trying to estimate $\Pi(R)$ for values of R smaller than the

Table 2.7: *Non-parametric, Hill, GP and Mixture of Erlangs-based estimates for* $\Pi(R)$.

R	Non-Parametric	Hill	GP	Mixture of Erlangs
3 000 000	161 728.1	163 367.4	166 619.6	163 987.7
3 500 000	108 837.2	108 227.2	111 610.4	110 118.5
4 000 000	74 696.3	75 581.4	79 219.0	77 747.6
4 500 000	53 312.3	55 065.8	58 714.1	55 746.3
5 000 000	35 888.0	41 481.6	45 001.6	39 451.6
7 500 000	1074.5	13 944.5	16 393.3	4018.6
10 000 000	0.0	6434.0	8087.8	159.6

threshold above which the extreme value distribution is fit. Based on the mean excess function, the authors propose the use of a mixture of an exponential and a Pareto distribution (Exp-Par). Instead of having to use this body-tail approach (a form a splicing, see Klugman et al. (2012)) explicitly, the implemented shape adjustment and reduction techniques when fitting the Erlang mixture have guided us to a mixture with two components of which the first one represents the body of the distribution and the second represents the tail. The fitting procedure for Erlang mixtures is able to make this choice implicitly in a data driven way, leading to a close representation of the data. In Table 2.8 we compare the estimated net premiums from Table 6.2 in Beirlant et al. (2004, p. 198) obtained using the Exp-Par model to the non-parametric and mixture of Erlangs estimates. The estimates based on the fitted Erlang mixture follow the non-parametric ones more closely than those obtained using the Exp-Par model.

Table 2.8: *Non-parametric, Exp-Par and Mixture of Erlangs-based estimates for* $\Pi(R)$.

R	Non-Parametric	Exp-Par	Mixture of Erlangs
1 250 000	981 238.0	944 217.8	981 483.1
1 500 000	760 637.6	734 371.6	760 912.9
1 750 000	583 403.6	571 314.1	582 920.1
2 000 000	445 329.8	444 275.5	444 466.6
2 250 000	340 853.2	344 965.2	339 821.4
2 500 000	263 052.7	267 000.7	262 314.6

Note that when $R = 1\,200\,000$, the net premium equals the mean excess loss $E(X - R \mid X > R)$, which is called the mean residual lifetime in survival context. (Klugman et al., 2013, p. 20) show that the distribution of the excess loss or

residual lifetime is again a mixture of M Erlangs with the same scale θ and different weights which we can compute analytically:

$$\alpha_j^* = \frac{\sum_{n=0}^{M-j} \alpha_{n+j} f(R; n+1, \theta)}{\sum_{n=0}^{M-1} A_n f(R; n+1, \theta)} \qquad \text{for } j = 1, \ldots, M\,.$$

2.5.4 Simulated generalized Pareto data

When modeling claim sizes, the insurer or reinsurer is often confronted with heavy tailed distributions. To safeguard the company against extreme losses that might jeopardize their solvency, an accurate description of the upper tail of the claim size distribution is of utmost importance. In order to explore the limits of Erlang mixtures in approximating heavy-tailed distribution using the presented method, we consider the generalized Pareto distribution with density

$$f_X(x; \mu, \sigma, \xi) = \frac{1}{\sigma}\left(1 + \frac{\xi(x-\mu)}{\sigma}\right)^{\left(-\frac{1}{\xi}-1\right)} \qquad \text{for } x \geqslant \mu\,. \tag{2.21}$$

with location $\mu > 0$, scale $\sigma > 0$ and shape $\xi > 0$. The generalized Pareto family is known for its tail thickness and is used for insurance branches with a high probability of large claims, such as liability insurance. The shape parameter coincides with the extreme value index (EVI) and determines the heaviness of the tail (Beirlant et al., 2004). The higher the value of the EVI, the heavier the tail. The variance is finite for $\xi < 1/2$ and the mean is finite for $\xi < 1$. In general is the kth moment finite for $\xi < 1/k$. When modeling the Secura Re data of the previous example using Pareto-type modeling, Beirlant et al. (2004) estimate the corresponding EVI around 0.3. Using the presented method, we were able to obtain a very good approximation in the tail with a mixture of Erlangs. We now want to illustrate what happens when the EVI further increases, by generating 1000 observations from a generalized Pareto distribution with location $\mu = 10$, scale $\sigma = 2$ and shape $\xi = 1$. In this extreme setting, the EVI equals 1 and none of the moments exist. Location $\mu = 10$ implies that the distribution is left truncated at 10.

In order to obtain a decent approximation of this sample, the initial values of the number of Erlangs M and the spread s become even more important. Due to the fact that the data is very skew and heavy-tailed, the maximum in the data set is extremely high, i.e. $\max(\boldsymbol{x}) = 10\,636.49$, and many of the initial shape parameters in the mixture will get a corresponding weight equal to zero. To ensure

that we start our calibration procedure with sufficient non-zero shape parameters, we decided – after some exploratory choices for M and s – to try all combinations of spread s between 1 and 10 and initial number of Erlangs $M = \left\lceil \frac{\max(\boldsymbol{x})}{i} \right\rceil$ for $i = 1, \ldots, 10$, leading to initial mixtures with 30 to 85 non-zero weight Erlang components. The best-fitting Erlang mixture according to AIC was obtained starting from $M = \left\lceil \frac{\max(\boldsymbol{x})}{7} \right\rceil = 1520$ and $s = 4$, corresponding to a mixture of 34 non-zero weight Erlang components at the initial step. The parameter estimates of the final mixture of 16 Erlangs, after the shape adjustment procedure and the reduction of the number of Erlangs based on AIC, are given in Table 2.9.

Table 2.9: *Parameter estimates of the mixture of 16 Erlangs fitted to the simulated generalized Pareto data.*

r_j	α_j	θ
2	0.9973387302	1.334924
13	0.0016914393	
20	0.0002066144	
28	0.0003513364	
47	0.0001826860	
74	0.0000809294	
120	0.0000458669	
163	0.0000079065	
211	0.0000286491	
286	0.0000073181	
488	0.0000073471	
613	0.0000219147	
3338	0.0000073155	
4472	0.0000073155	
6307	0.0000073155	
7964	0.0000073155	

The underlying untruncated mixture contains 16 components and is dominated by an Erlang distribution with shape 2, modeling the main bulk of the data, whereas the approximation of the tail requires a combination of 15 Erlangs with shapes ranging from 13 to 7964. A graphical comparison of the fitted Erlang mixture and the underlying true distribution up to the 95% sample quantile is shown in Figure 2.5. The QQ-plot in Figure 2.6 (left) shows that this mixture does a great job in fitting the sample in the tail. However, the log-log plot of the empirical truncated survival function and the truncated survival function of the best-fitting Erlang mixture in Figure 2.6 (right) reveals that this approximation is obtained by letting separate Erlang components with a very small weight coincide

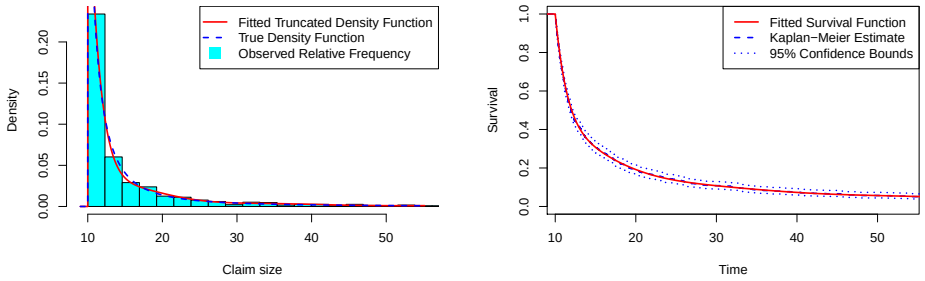

Figure 2.5: *Graphical comparison of the truncated density of the fitted mixture of 16 Erlangs and the histogram (left) and of the truncated survival function and the Kaplan-Meier estimator with 95% confidence bounds (right) for the simulated generalized Pareto data up to the 95% empirical quantile.*

with the largest data points that lie very far apart. Moreover, all moments of a finite mixture of Erlangs are finite whereas the expected value of the underlying distribution is infinite. We thus conclude that in this extreme setting with EVI equal to 1, the fitted finite mixture of Erlang distributions follows the observed data set closely, but is not able to extrapolate the heaviness in the tail in comparison to the extreme value methodology based on the Fisher-Tippett-Gnedenko theorem.

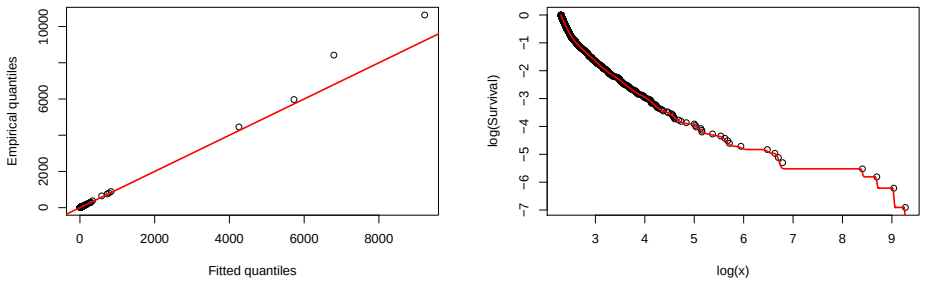

Figure 2.6: *QQ-plot of the empirical quantiles and the quantiles of the fitted mixture of 16 Erlangs with identity line (left) and log-log plot of the empirical truncated survival function and the truncated survival function of the fitted Erlang mixture (right) for the simulated generalized Pareto data.*

2.6 Discussion

We extend the Lee and Lin (2010) EM algorithm for fitting mixtures of Erlangs with a common scale parameter to censored and truncated data. The EM algorithm able to deal with censored and truncated data remains a simple iterative algorithm. The initialization of the parameters can be done in a similar way as in Lee and Lin (2010) based on the denseness property (Tijms, 1994, p. 163) and provides close starting values making the algorithm converge fast. The shape adjustment procedure explores the parameter space in a clever way such that, when adjusting and reducing the shapes, the previous estimates for the scale and the weights provide a very close approximation to the maximum likelihood estimates corresponding to the new set of shapes, which greatly reduces the run time. Extending Lee and Lin (2010), we suggest the use of a spread factor to achieve a wider spread for the shapes at the initial step. We recommend comparing the resulting fits starting from different initial values obtained by varying the spread factor and changing the initial number of Erlangs.

We implement the fitting procedure in `R` and show how mixtures of Erlangs can be used to adequately represent any univariate distribution in a wide variety of applications where data is allowed to be censored and truncated. We focus on the domain of actuarial science, where claim severity data is often censored and truncated due to policy modifications such as deductibles and policy limits. The use of mixtures of Erlangs offers on the one hand the flexibility of nonparametric density estimation techniques to describe the insurance losses and on the other hand the feasibility to analytically quantify the risk. The examples on several simulated and real data sets illustrate the effectiveness of our proposed algorithm and demonstrate the approximation strength of mixtures of Erlangs.

Future research may explore incorporating regressor variables in the mixture of Erlangs with common scale and introducing the flexibility of this approach in a regression context. We detected some limitations of mixtures of Erlangs in approximating heavy-tailed distributions and suggest combining our methodology with the extreme value methodology using a body-tail approach (Lee et al., 2012; Pigeon and Denuit, 2011). Adjusting the EM algorithm tailored to the class of multivariate mixtures of Erlangs, introduced by Lee and Lin (2012), to the case of censored and truncated data is another appealing extension.

2.7 Appendix A: Denseness

Theorem 2.7.1. *(Tijms, 1994, p. 163) The class of mixtures of Erlang distributions with a common scale parameter is dense in the space of distributions on* $\mathbb{R}^+$*. More specifically, let* $G(x)$ *be the cumulative distribution function of a positive random variable. Define the following cumulative distribution function of a mixture of Erlang distributions with a common scale parameter* $\theta > 0$*,*

$$F(x;\theta) = \sum_{j=1}^{\infty} \alpha_j(\theta) F(x;j,\theta)\,,$$

where $F(x;j,\theta)$ *denotes the cumulative distribution function of an Erlang distribution with shape* j *and scale* θ*,*

$$F(x;j,\theta) = 1 - \sum_{n=0}^{j-1} e^{-x/\theta} \frac{(x/\theta)^n}{n!}\,,$$

and the mixing weights are given by

$$\alpha_j(\theta) = G(j\theta) - G((j-1)\theta) \qquad \text{for } j = 1, 2, \ldots\,.$$

Then

$$\lim_{\theta \to 0} F(x;\theta) = G(x)\,,$$

for each point x *at which* $G(\cdot)$ *is continuous.*

2.8 Appendix B: Partial derivative of Q

We first introduce the lower incomplete gamma function

$$\gamma(s,x) = \int_0^x z^{s-1} e^{-z} dz\,,$$

by which we can write the cumulative distribution function of an Erlang distribution as

$$F(x;r,\theta) = \int_0^x \frac{z^{r-1} e^{-z/\theta}}{\theta^r (r-1)!} dz = \frac{1}{(r-1)!} \int_0^{x/\theta} u^{r-1} e^{-u} du = \frac{\gamma(r, x/\theta)}{(r-1)!}\,. \quad (2.22)$$

In order to maximize $Q(\boldsymbol{\Theta};\boldsymbol{\Theta}^{(k-1)})$ with respect to θ, we set the first order partial derivative at $\theta^{(k)}$ equal to zero

$$
\begin{aligned}
&\left.\frac{\partial Q(\boldsymbol{\Theta};\boldsymbol{\Theta}^{(k-1)})}{\partial\theta}\right|_{\theta=\theta^{(k)}} \\
&= \sum_{i\in U}\sum_{j=1}^{M} {}^{u}z_{ij}^{(k)}\left(\frac{x_i}{\theta^2}-\frac{r_j}{\theta}-\frac{\frac{\partial}{\partial\theta}\left[F(t^u;r_j,\theta)-F(t^l;r_j,\theta)\right]}{F(t^u;r_j,\theta)-F(t^l;r_j,\theta)}\right) \\
&\quad+\sum_{i\in C}\sum_{j=1}^{M} {}^{c}z_{ij}^{(k)}\left(\frac{E\left(X_i\left|Z_{ij}=1,l_i,u_i,t^l,t^u;\theta^{(k-1)}\right.\right)}{\theta^2}-\frac{r_j}{\theta}\right. \\
&\quad\left.-\frac{\frac{\partial}{\partial\theta}\left[F(t^u;r_j,\theta)-F(t^l;r_j,\theta)\right]}{F(t^u;r_j,\theta)-F(t^l;r_j,\theta)}\right)\Bigg|_{\theta=\theta^{(k)}} \\
&\overset{(2.22)}{=} \frac{1}{\theta^2}\sum_{i\in U}\left(\sum_{j=1}^{M} {}^{u}z_{ij}^{(k)}\right)x_i \\
&\quad+\frac{1}{\theta^2}\sum_{i\in C}\left(\sum_{j=1}^{M} {}^{c}z_{ij}^{(k)}E\left(X_i\left|Z_{ij}=1,l_i,u_i,t^l,t^u;\theta^{(k-1)}\right.\right)\right) \\
&\quad-\frac{n}{\theta}\sum_{j=1}^{M}\left(\frac{\sum_{i\in U}{}^{u}z_{ij}^{(k)}+\sum_{i\in C}{}^{c}z_{ij}^{(k)}}{n}\right)r_j \\
&\quad-\sum_{i\in U}\sum_{j=1}^{M} {}^{u}z_{ij}^{(k)}\frac{\frac{\partial}{\partial\theta}\left(\gamma(r_j,t^u/\theta)-\gamma(r_j,t^l/\theta)\right)}{(r_j-1)!\left(F(t^u;r_j,\theta)-F(t^l;r_j,\theta)\right)} \\
&\quad-\sum_{i\in C}\sum_{j=1}^{M} {}^{c}z_{ij}^{(k)}\frac{\frac{\partial}{\partial\theta}\left(\gamma(r_j,t^u/\theta)-\gamma(r_j,t^l/\theta)\right)}{(r_j-1)!\left(F(t^u;r_j,\theta)-F(t^l;r_j,\theta)\right)}\Bigg|_{\theta=\theta^{(k)}} \\
&\overset{(2.16)}{=} \frac{1}{\theta^2}\sum_{i\in U}x_i+\frac{1}{\theta^2}\sum_{i\in C}E\left(X_i\left|l_i,u_i,t^l,t^u;\theta^{(k-1)}\right.\right)-\frac{n}{\theta}\sum_{j=1}^{M}\beta_j^{(k)}r_j \\
&\quad-\sum_{i\in U}\sum_{j=1}^{M} {}^{u}z_{ij}^{(k)}\frac{t^l/\theta^2\left(t^l/\theta\right)^{r_j-1}e^{-t^l/\theta}-t^u/\theta^2\left(t^u/\theta\right)^{r_j-1}e^{-t^u/\theta}}{(r_j-1)!\left(F(t^u;r_j,\theta)-F(t^l;r_j,\theta)\right)} \\
&\quad-\sum_{i\in C}\sum_{j=1}^{M} {}^{c}z_{ij}^{(k)}\frac{t^l/\theta^2\left(t^l/\theta\right)^{r_j-1}e^{-t^l/\theta}-t^u/\theta^2\left(t^u/\theta\right)^{r_j-1}e^{-t^u/\theta}}{(r_j-1)!\left(F(t^u;r_j,\theta)-F(t^l;r_j,\theta)\right)}\Bigg|_{\theta=\theta^{(k)}} \\
&= \frac{1}{\theta^2}\sum_{i\in U}x_i+\frac{1}{\theta^2}\sum_{i\in C}E\left(X_i\left|l_i,u_i,t^l,t^u;\theta^{(k-1)}\right.\right)-\frac{n}{\theta}\sum_{j=1}^{M}\beta_j^{(k)}r_j
\end{aligned}
$$

$$-\frac{n}{\theta^2}\sum_{j=1}^{M}\beta_j^{(k)}\frac{\left(t^l\right)^{r_j}e^{-t^l/\theta}-\left(t^u\right)^{r_j}e^{-t^u/\theta}}{\theta^{r_j-1}(r_j-1)!\left(F(t^u;r_j,\theta)-F(t^l;r_j,\theta)\right)}\Bigg|_{\theta=\theta^{(k)}}=0\,,$$

where we used expression (2.22) of the cumulative distribution of an Erlang.

Chapter 3

Multivariate mixtures of Erlangs for density estimation under censoring

Abstract

Multivariate mixtures of Erlang distributions form a versatile, yet analytically tractable, class of distributions making them suitable for multivariate density estimation. We present a flexible and effective fitting procedure for multivariate mixtures of Erlangs, which iteratively uses the EM algorithm, by introducing a computationally efficient initialization and adjustment strategy for the shape parameter vectors. We furthermore extend the EM algorithm for multivariate mixtures of Erlangs to be able to deal with randomly censored and fixed truncated data. The effectiveness of the proposed algorithm is demonstrated on simulated as well as real data sets.

This chapter is based on Verbelen, R., Antonio, K., and Claeskens, G. (2016). Multivariate mixtures of Erlangs for density estimation under censoring. *Lifetime Data Analysis*, 22(3):429-455.

3.1 Introduction

We present an estimation technique for fitting multivariate mixtures of Erlang distributions (MME). We suggest an efficient initialization method and adjust-

ment strategy for the values of the shape parameter vectors of an MME, which has been underexposed in the literature. The fitting procedure is also extended to take random censoring and fixed truncation into account. Data are censored in case you only observe an interval in which a data point is lying without knowing its exact value. Truncation entails that it is only possible to observe the data of which the values lie in a certain range. Censoring and/or truncation is often the case in applications such as loss modeling (finance and actuarial science), clinical experiments (survival/failure time analysis), veterinary studies (e.g. mastitis studies), and duration data (econometric studies).

The class of MME is introduced by Lee and Lin (2012). MME form a highly flexible class of distributions as they are dense in the space of positive continuous multivariate distributions in the sense of weak convergence, extending this property of the univariate class (Tijms, 1994). An overview of the analytical and distributional properties of mixtures of Erlangs can be found in Klugman et al. (2013), Willmot and Lin (2011) and Willmot and Woo (2007). Parameter estimation in the univariate case is treated in Lee and Lin (2010) and extended to be able to deal with randomly censored and fixed truncated data in Verbelen et al. (2015).

Mixtures of Erlangs have received most attention in the field of actuarial science. Cossette et al. (2013a) model the joint distribution of a portfolio of dependent risks using univariate mixtures of Erlangs as marginals along with the Farlie-Gumbel-Morgenstern (FGM) copula. Cossette et al. (2013b) and Mailhot (2012) study the bivariate lower and upper orthant Value-at-Risk and use MME as an illustration. Willmot and Woo (2015) study the analytical properties of the MME class. They motivate the use of MME in actuarial science and illustrate how their tractability leads to closed-form expressions.

The use of MME should be regarded as a multivariate density estimation technique, not as a type of model-based clustering. The MME model can be seen as semiparametric, since the mixture components have a specific parametric form, whereas the mixing weights can have a nonparametric nature, and is an interesting alternative to the use of copulas, which is the dominant choice to model multivariate data in a two stage procedure, separating the dependence structure from the marginal distributions (see e.g. Joe, 1997; Nelsen, 2006). In contrast, MME are able to model the multivariate data directly on the original scale. The MME model enjoys many desirable properties of a multivariate model as listed by Joe (1997, p. 84), see Lee and Lin (2012), with regard to interpretability, closure, flexibility and wide range of dependence, and closed-form representation,

often not satisfied for the commonly used copula structures. Lee and Lin (2012) demonstrate the flexibility of MME by fitting 12-dimensional data generated from a multivariate lognormal distribution and extremely dependent bivariate data with Spearman's rho very close to 1 or -1.

An extensive literature exists on mixtures of multivariate normals (see e.g. McLachlan and Peel, 2001). Lee and Scott (2012) discuss the estimation of multivariate Gaussian mixtures in case the data can be randomly censored and fixed truncated. Due to the limitations of Gaussian mixtures, such as the difficulty in modeling skewed data, non-Gaussian approaches have received an increasing interest over the last years. Important examples include mixtures of multivariate t-distributions (see e.g. Peel and McLachlan, 2000), mixtures of multivariate skew-normal distributions (see e.g. Lin, 2009), and mixtures of multivariate skew-t distributions (see e.g. Lee and McLachlan, 2014). All of these mixture models involve modeling real-valued multivariate random variables, whereas in this chapter we consider multivariate positive-valued random variables.

Lee and Lin (2012) show in Theorem 2.3 that a finite multivariate Erlang mixture is a multivariate phase-type distribution, a generalization of the class of univariate phase-type distributions introduced by Assaf et al. (1984). Parameter estimation for phase-type distributions in the bivariate case (Eisele, 2005; Zadeh and Bilodeau, 2013), as in the univariate case (Asmussen et al., 1996; Olsson, 1996), uses the expectation-maximization (EM) algorithm, first introduced by Dempster et al. (1977)

The EM algorithm forms the key to fit an MME to multivariate positive data. Taking censoring and truncation into account when calibrating data using copulas is cumbersome, especially in more than two dimensions, due to complicated forms of the likelihood (see e.g. Georges et al., 2001) which are hard to optimize numerically. This is, as we will show, not the case for the MME class due to the EM algorithm. As opposed to the traditional way of dealing with grouped and truncated data using the EM algorithm (Dempster et al., 1977; McLachlan and Krishnan, 2008, p. 66; McLachlan and Peel, 2001, p. 257; McLachlan and Jones, 1988), we follow the approach of Lee and Scott (2012), as was done in the univariate setting (Verbelen et al., 2015).

We demonstrate the effectiveness of our proposed algorithm and the practical use of MME on a simulated data set, the old faithful geyser data and a four-dimensional data set of interval and right censored udder quarter infection times, each time highlighting one of the analytical aspects of MME.

3.2 Multivariate Erlang mixtures with a common scale parameter

In this section, we briefly revise the definition of a multivariate mixture of Erlang distributions with a common scale parameter and the denseness property of this distributional class. These formulas are extended in Section 3.3.1 and 3.3.2 towards censoring and truncation.

The Erlang distribution is a positive continuous distribution with density function

$$f(x;r,\theta) = \frac{x^{r-1}e^{-x/\theta}}{\theta^r(r-1)!} \qquad \text{for } x>0\,, \tag{3.1}$$

where r, a positive integer, is the shape parameter and $\theta > 0$ the scale parameter (the inverse $\lambda = 1/\theta$ is called the rate parameter). The cumulative distribution function is obtained by integrating (3.1) by parts r times

$$F(x;r,\theta) = 1 - \sum_{n=0}^{r-1} e^{-x/\theta}\frac{(x/\theta)^n}{n!} = \frac{\gamma(r,x/\theta)}{(r-1)!}\,, \tag{3.2}$$

using the lower incomplete gamma function defined as $\gamma(s,x) = \int_0^x z^{s-1}e^{-z}dz$.

A univariate Erlang distribution is in fact a gamma distribution of which the shape parameter is a positive integer and can therefore be seen as the distribution of a sum of i.i.d. exponential random variables. Lee and Lin (2012) define a d-variate Erlang mixture as a mixture such that each mixture component is the joint distribution of d independent Erlang distributions with a common scale parameter $\theta > 0$. The dependence structure is captured by the combination of the positive integer shape parameters of the Erlangs in each dimension. We denote the positive integer shape parameters of the jointly independent Erlang distributions in a mixture component by the vector $\boldsymbol{r} = (r_1,\ldots,r_d)$ and the set of all shape vectors with non-zero weight by $\mathcal{R}$. The mixture weights are denoted by $\boldsymbol{\alpha} = \{\alpha_{\boldsymbol{r}}\,|\boldsymbol{r}\in\mathcal{R}\}$ and must satisfy $\alpha_{\boldsymbol{r}} \geqslant 0$ and $\sum_{\boldsymbol{r}\in\mathcal{R}}\alpha_{\boldsymbol{r}} = 1$. The density of a d-variate Erlang mixture evaluated in $\boldsymbol{x} = (x_1,\ldots,x_d)$ with $x_j > 0$ for $j = 1,\ldots,d$ can then be written as

$$\begin{aligned} f(\boldsymbol{x};\boldsymbol{\alpha},\boldsymbol{r},\theta) &= \sum_{\boldsymbol{r}\in\mathcal{R}}\alpha_{\boldsymbol{r}} f(\boldsymbol{x};\boldsymbol{r},\theta) = \sum_{\boldsymbol{r}\in\mathcal{R}}\alpha_{\boldsymbol{r}}\prod_{j=1}^{d} f(x_j;r_j,\theta) \\ &= \sum_{\boldsymbol{r}\in\mathcal{R}}\alpha_{\boldsymbol{r}}\prod_{j=1}^{d}\frac{x_j^{r_j-1}e^{-x_j/\theta}}{\theta^{r_j}(r_j-1)!}\,. \end{aligned} \tag{3.3}$$

The following property states that for any positive multivariate distribution there exists a sequence of multivariate Erlang distributions that weakly converges to the target distribution. The proof is given in the appendix of Lee and Lin (2012).

Property 1 (Lee and Lin 2012)**.** The class of multivariate Erlang mixtures of form (3.3) is dense in the space of positive continuous multivariate distributions in the sense of weak convergence. More specifically, let $g(\boldsymbol{x})$ be the density function of a d-variate positive random variable with cumulative distribution function $G(\boldsymbol{x})$. For any given $\theta > 0$, define the following d-variate Erlang mixture

$$f(\boldsymbol{x};\theta) = \sum_{r_1=1}^{\infty} \cdots \sum_{r_d=1}^{\infty} \alpha_{\boldsymbol{r}}(\theta) \prod_{j=1}^{d} f(x_j; r_j, \theta)\,, \tag{3.4}$$

with mixing weights

$$\alpha_{\boldsymbol{r}}(\theta) = \int_{(r_1-1)\theta}^{r_1\theta} \cdots \int_{(r_d-1)\theta}^{r_d\theta} g(\boldsymbol{x}) d\boldsymbol{x}\,. \tag{3.5}$$

Then $\lim_{\theta\to 0} F(\boldsymbol{x};\theta) = G(\boldsymbol{x})$ for each point $\boldsymbol{x}$ at which F is continuous.

In Property 1, for any given common scale $\theta > 0$, an infinite multivariate mixture of Erlangs in (3.4) is considered using combinations of shapes from 1 to infinity in each marginal dimension. The weights in (3.5) of the components in the mixture are defined by integrating the density over the corresponding d-dimensional rectangle of the d-dimensional grid formed by the shape parameters multiplied with the common scale. When the value of the common scale θ decreases, this grid becomes more refined and the sequence of Erlang mixtures converges to the underlying cumulative distribution function.

Next to its flexibility, Lee and Lin (2012) show that it is easy to work analytically with this class of distributions due to the independence structure of the Erlang distributions within each mixture component. This leads to explicit expressions of many distributional quantities such as the characteristic function, the joint moments and bivariate measures of association (Kendall's tau and Spearman's rho). The authors further reveal interesting closure properties, such as the fact that each p-variate marginal or conditional distribution with $p \leqslant d$ can again be written as a p-variate Erlang distribution. The same property holds for the distribution of the multivariate excess losses (actuarial science context) or multivariate residual lifetimes (survival analysis context). Furthermore, the dis-

tribution of the sum of the component random variables of an MME distributed random variable is a univariate Erlang mixture distribution.

Willmot and Woo (2015) consider an extension of the MME class, allowing different scale parameters in each dimension. However, in Proposition 1 they show how a multivariate mixture of Erlangs distribution with different scale parameters can be rewritten as a multivariate mixture of Erlangs distribution with a common scale parameter, which is smaller than all original scales. We thus concentrate on models with a common scale parameter.

3.3 Parameter estimation

The parameters of an MME to be estimated are the common scale parameter θ, the mixture weights $\boldsymbol{\alpha} = \{\alpha_{\boldsymbol{r}} \,|\boldsymbol{r} \in \mathcal{R}\}$ and the set of corresponding shape parameter vectors $\mathcal{R}$. Lee and Lin (2012) propose an EM algorithm in order to find the maximum likelihood estimators for $\boldsymbol{\Theta} = (\boldsymbol{\alpha}, \theta)$, given a fixed set of shape parameter vectors $\mathcal{R}$. Model selection for the number of mixture components and the corresponding values of the shape parameter vectors is based on an information criterion, similar to the univariate strategy of Lee and Lin (2010) and Verbelen et al. (2015).

The two main novelties we present in this chapter are (i) an extension of the EM algorithm to be able to deal with randomly censored and fixed truncated data and (ii) a computationally more efficient initialization and adjustment strategy for the shape parameter vectors in order to make the estimation procedure more flexible and effective. The improvements (i) and (ii) allow us to analyze realistic data with diverse forms of dependence in contrast to the simulated example in Lee and Lin (2012) with a simple structure.

First we discuss how we represent a censored and truncated sample and evaluate the expression of the likelihood. The form of the complete data log-likelihood is given next, followed by the adjusted EM algorithm and a discussion on some asymptotic properties. In Section 3.4, we present the initialization and selection of the shape parameter vectors.

3.3.1 Randomly censored and fixed truncated data

We represent a censored sample, truncated to the fixed range $[\boldsymbol{t}^l, \boldsymbol{t}^u]$, by $\mathcal{X} = \{(\boldsymbol{l}_i, \boldsymbol{u}_i) | \, i = 1, \ldots, n\}$. The lower and upper truncation points are $\boldsymbol{t}^l = (t_1^l, \ldots, t_d^l)$ and $\boldsymbol{t}^u = (t_1^u, \ldots, t_d^u)$, which are common to each observation $i = 1, \ldots, n$. The

lower and upper censoring points are $\boldsymbol{l}_i = (l_{i1}, \ldots, l_{id})$ and $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{id})$. It holds that $\boldsymbol{t}^l \leqslant \boldsymbol{l}_i \leqslant \boldsymbol{u}_i \leqslant \boldsymbol{t}^u$ for $i = 1, \ldots, n$. $t_j^l = 0$ and $t_j^u = \infty$ mean no truncation from below and above for the jth dimension, respectively. The censoring status for the jth dimension of observation i is determined as follows:

Uncensored:	$t_j^l \leqslant l_{ij} = u_{ij} =: x_{ij} \leqslant t_j^u$
Left Censored:	$t_j^l = l_{ij} < u_{ij} < t_j^u$
Right Censored:	$t_j^l < l_{ij} < u_{ij} = t_j^u$
Interval Censored:	$t_j^l < l_{ij} < u_{ij} < t_j^u$.

Thus, l_{ij} and u_{ij} should be interpreted as the lower and upper endpoints of the interval that contains the jth element of observation i. A missing value in dimension j for observation i can also be dealt with by setting $l_{ij} = t_j^l$ and $u_{ij} = t_j^u$, i.e. treating the missing value as a data point being interval censored between the lower and upper truncation points.

The likelihood of a censored and truncated sample of a multivariate Erlang distribution is given by

$$\mathcal{L}(\boldsymbol{\Theta}; \mathcal{X}) = \prod_{i=1}^{n} \frac{\sum_{\boldsymbol{r} \in \mathcal{R}} \alpha_{\boldsymbol{r}} \prod_{j=1}^{d} f(l_{ij}, u_{ij}; r_j, \theta)}{\mathbb{P}(\boldsymbol{t}^l \leqslant \boldsymbol{X}_i \leqslant \boldsymbol{t}^u; \boldsymbol{\Theta})}$$

with

$$f(l_{ij}, u_{ij}; r_j, \theta) = \begin{cases} f(x_{ij}; r_j, \theta) & \text{if } l_{ij} = u_{ij} = x_{ij} \\ F(u_{ij}; r_j, \theta) - F(l_{ij}; r_j, \theta) & \text{if } l_{ij} < u_{ij}\,, \end{cases}$$

and

$$\mathbb{P}(\boldsymbol{t}^l \leqslant \boldsymbol{X}_i \leqslant \boldsymbol{t}^u; \boldsymbol{\Theta}) = \sum_{\boldsymbol{r} \in \mathcal{R}} \alpha_{\boldsymbol{r}} \prod_{j=1}^{d} \left[F(t_j^u; r_j, \theta) - F(t_j^l; r_j, \theta) \right] .$$

The corresponding log-likelihood is

$$\begin{aligned} l(\boldsymbol{\Theta}; \mathcal{X}) = &\sum_{i=1}^{n} \ln \left(\sum_{\boldsymbol{r} \in \mathcal{R}} \alpha_{\boldsymbol{r}} \prod_{j=1}^{d} f(l_{ij}, u_{ij}; r_j, \theta) \right) \\ &- n \ln \left(\sum_{\boldsymbol{r} \in \mathcal{R}} \alpha_{\boldsymbol{r}} \prod_{j=1}^{d} \left[F(t_j^u; r_j, \theta) - F(t_j^l; r_j, \theta) \right] \right) . \end{aligned} \tag{3.6}$$

This expression is however not workable as it involves the logarithm of a sum

and cannot be used to easily find the maximum likelihood estimators for $\boldsymbol{\Theta}$ for a fixed set of positive integer shape parameters $\mathcal{R}$.

3.3.2 Construction of the complete data likelihood

For an uncensored observation $\boldsymbol{x}_i$, truncated to $[\boldsymbol{t}^l, \boldsymbol{t}^u]$, the probability density function can be rewritten as a mixture

$$\begin{aligned} f(\boldsymbol{x}_i; \boldsymbol{t}^l, \boldsymbol{t}^u, \boldsymbol{\Theta}) &= \frac{f(\boldsymbol{x}_i; \boldsymbol{\Theta})}{\mathbb{P}(\boldsymbol{t}^l \leqslant \boldsymbol{X}_i \leqslant \boldsymbol{t}^u; \boldsymbol{\Theta})} = \frac{\sum_{\boldsymbol{r} \in \mathcal{R}} \alpha_{\boldsymbol{r}} \prod_{j=1}^{d} f(x_{ij}; r_j, \theta)}{\mathbb{P}(\boldsymbol{t}^l \leqslant \boldsymbol{X}_i \leqslant \boldsymbol{t}^u; \boldsymbol{\Theta})} \\ &= \sum_{\boldsymbol{r} \in \mathcal{R}} \alpha_{\boldsymbol{r}} \cdot \frac{\mathbb{P}(\boldsymbol{t}^l \leqslant \boldsymbol{X}_i \leqslant \boldsymbol{t}^u; \boldsymbol{r}, \theta)}{\mathbb{P}(\boldsymbol{t}^l \leqslant \boldsymbol{X}_i \leqslant \boldsymbol{t}^u; \boldsymbol{\Theta})} \cdot \frac{\prod_{j=1}^{d} f(x_{ij}; r_j, \theta)}{\mathbb{P}(\boldsymbol{t}^l \leqslant \boldsymbol{X}_i \leqslant \boldsymbol{t}^u; \boldsymbol{r}, \theta)} \\ &= \sum_{\boldsymbol{r} \in \mathcal{R}} \beta_{\boldsymbol{r}} \cdot f(\boldsymbol{x}_i; \boldsymbol{t}^l, \boldsymbol{t}^u, \boldsymbol{r}, \theta)\,, \end{aligned}$$

for $\boldsymbol{t}^l \leqslant \boldsymbol{x}_i \leqslant \boldsymbol{t}^u$ and zero otherwise. The mixing weights $\beta_{\boldsymbol{r}}$ and component density functions are given by, respectively,

$$\begin{aligned} \beta_{\boldsymbol{r}} &= \alpha_{\boldsymbol{r}} \cdot \frac{\mathbb{P}(\boldsymbol{t}^l \leqslant \boldsymbol{X}_i \leqslant \boldsymbol{t}^u; \boldsymbol{r}, \theta)}{\mathbb{P}(\boldsymbol{t}^l \leqslant \boldsymbol{X}_i \leqslant \boldsymbol{t}^u; \boldsymbol{\Theta})} \\ &= \alpha_{\boldsymbol{r}} \cdot \frac{\prod_{j=1}^{d} \left[F(t_j^u; r_j, \theta) - F(t_j^l; r_j, \theta)\right]}{\sum_{\boldsymbol{m} \in \mathcal{R}} \alpha_{\boldsymbol{m}} \prod_{j=1}^{d} \left[F(t_j^u; m_j, \theta) - F(t_j^l; m_j, \theta)\right]} \end{aligned} \tag{3.7}$$

and

$$f(\boldsymbol{x}_i; \boldsymbol{t}^l, \boldsymbol{t}^u, \boldsymbol{r}, \theta) = \frac{\prod_{j=1}^{d} f(x_{ij}; r_j, \theta)}{\mathbb{P}(\boldsymbol{t}^l \leqslant \boldsymbol{X}_i \leqslant \boldsymbol{t}^u; \boldsymbol{r}, \theta)} = \prod_{j=1}^{d} \frac{f(x_{ij}; r_j, \theta)}{F(t_j^u; r_j, \theta) - F(t_j^l; r_j, \theta)}\,. \tag{3.8}$$

The weights $\beta_{\boldsymbol{r}}$ are re-weighted versions of the original weights $\alpha_{\boldsymbol{r}}$ by means of the probabilities of the corresponding mixture component to lie in the d-dimensional truncation interval. The component density functions $f(\boldsymbol{x}_i; \boldsymbol{t}^l, \boldsymbol{t}^u, \boldsymbol{r}, \theta)$ are truncated versions of the original component density functions $f(\boldsymbol{x}_i; \boldsymbol{r}, \theta)$.

The EM algorithm forms the solution to fit this finite mixture to the censored and truncated data. The idea is to regard the censored sample $\mathcal{X}$ as being incomplete since the uncensored observations $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})$ and their associated

component-indicators $\boldsymbol{z}_i = \{z_{i\boldsymbol{r}} \,|\boldsymbol{r} \in \mathcal{R}\}$ with

$$z_{i\boldsymbol{r}} = \begin{cases} 1 & \text{if observation } \boldsymbol{x}_i \text{ comes from the mixture component} \\ & \text{corresponding to the shape parameter vector } \boldsymbol{r} \\ 0 & \text{otherwise} \end{cases} \tag{3.8}$$

(3.9)

for $i = 1, \ldots, n$ and $\boldsymbol{r} \in \mathcal{R}$, are not available. The complete data vector, $\mathcal{Y} = \{(\boldsymbol{x}_i, \boldsymbol{z_i})|i = 1, \ldots, n\}$, contains all uncensored observations $\boldsymbol{x}_i$ and their corresponding mixing component indicator $\boldsymbol{z}_i$. The log-likelihood of the complete sample $\mathcal{Y}$ can then be written as

$$l(\boldsymbol{\Theta}; \mathcal{Y}) = \sum_{i=1}^{n} \sum_{\boldsymbol{r}\in\mathcal{R}} z_{i\boldsymbol{r}} \ln\left(\beta_{\boldsymbol{r}} f(\boldsymbol{x}_i; \boldsymbol{t}^l, \boldsymbol{t}^u, \boldsymbol{r}, \theta)\right) . \tag{3.10}$$

3.3.3 The EM algorithm for censored and truncated data

The EM algorithm finds the maximum likelihood estimators for $\boldsymbol{\Theta} = (\boldsymbol{\alpha}, \theta)$, given a fixed set $\mathcal{R}$ of positive integer shape parameter vectors, based on a (possibly) censored and truncated sample by iteratively repeating the following two steps.

E-step Conditional on the incomplete data $\mathcal{X}$ and using the current estimate $\boldsymbol{\Theta}^{(k-1)}$ for $\boldsymbol{\Theta}$, we compute the expectation of the complete log-likelihood (3.10) in the kth iteration of the E-step:

$$\begin{aligned}
&Q(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(k-1)}) \\
&\quad = \quad E(l(\boldsymbol{\Theta}; \mathcal{Y}) \mid \mathcal{X}; \boldsymbol{\Theta}^{(k-1)}) \\
&\quad = \quad \sum_{i=1}^{n} E\left[\sum_{\boldsymbol{r}\in\mathcal{R}} Z_{i\boldsymbol{r}} \ln\left(\beta_{\boldsymbol{r}} f(\boldsymbol{X}_i; \boldsymbol{t}^l, \boldsymbol{t}^u, \boldsymbol{r}, \theta)\right) \middle| \boldsymbol{l}_i, \boldsymbol{u}_i, \boldsymbol{t}^l, \boldsymbol{t}^u; \boldsymbol{\Theta}^{(k-1)}\right] \\
&\quad = \quad \sum_{i=1}^{n} \sum_{\boldsymbol{r}\in\mathcal{R}} z_{i\boldsymbol{r}}^{(k)} E\left[\ln\left(\beta_{\boldsymbol{r}} f(\boldsymbol{X}_i; \boldsymbol{t}^l, \boldsymbol{t}^u, \boldsymbol{r}, \theta)\right) \middle| Z_{i\boldsymbol{r}} = 1, \boldsymbol{l}_i, \boldsymbol{u}_i, \boldsymbol{t}^l, \boldsymbol{t}^u; \theta^{(k-1)}\right] \\
&\quad = \quad \sum_{i=1}^{n} \sum_{\boldsymbol{r}\in\mathcal{R}} z_{i\boldsymbol{r}}^{(k)} \left[\ln(\beta_{\boldsymbol{r}}) + \sum_{j=1}^{d} (r_j - 1) E\left(\ln(X_{ij}) \middle| Z_{i\boldsymbol{r}} = 1, l_{ij}, u_{ij}, t_j^l, t_j^u; \theta^{(k-1)}\right)\right. \\
&\qquad \left. - \frac{1}{\theta} \sum_{j=1}^{d} E\left(X_{ij} \middle| Z_{i\boldsymbol{r}} = 1, l_{ij}, u_{ij}, t_j^l, t_j^u; \theta^{(k-1)}\right) - \sum_{j=1}^{d} r_j \ln(\theta)\right.
\end{aligned}$$

$$-\sum_{j=1}^{d}\ln((r_j-1)!)-\sum_{j=1}^{d}\ln\left(F(t_j^u;r_j,\theta)-F(t_j^l;r_j,\theta)\right)\Bigg]\,. \tag{3.11}$$

In the fourth equality, we apply the law of total expectation and denote the posterior probability that observation i belongs to the mixture component corresponding to the shape parameters $\boldsymbol{r}$ as $z_{i\boldsymbol{r}}^{(k)}$. These posterior probabilities can be computed using Bayes' rule,

$$\begin{aligned} z_{i\boldsymbol{r}}^{(k)} &= P(Z_{i\boldsymbol{r}}=1 \mid \boldsymbol{l}_i,\boldsymbol{u}_i,\boldsymbol{t}^l,\boldsymbol{t}^u;\boldsymbol{\Theta}^{(k-1)}) \\ &= \frac{\beta_{\boldsymbol{r}}^{(k-1)}\prod_{j=1}^{d}\frac{f(l_{ij},u_{ij};r_j,\theta^{(k-1)})}{F(t_j^u;r_j,\theta^{(k-1)})-F(t_j^l;r_j,\theta^{(k-1)})}}{\sum_{\boldsymbol{m}\in\mathcal{R}}\beta_{\boldsymbol{m}}^{(k-1)}\prod_{j=1}^{d}\frac{f(l_{ij},u_{ij};m_j,\theta^{(k-1)})}{F(t_j^u;m_j,\theta^{(k-1)})-F(t_j^l;m_j,\theta^{(k-1)})}} \\ &= \frac{\alpha_{\boldsymbol{r}}^{(k-1)}\prod_{j=1}^{d}f(l_{ij},u_{ij};r_j,\theta^{(k-1)})}{\sum_{\boldsymbol{m}\in\mathcal{R}}\alpha_{\boldsymbol{m}}^{(k-1)}\prod_{j=1}^{d}f(l_{ij},u_{ij};m_j,\theta^{(k-1)})}\,. \end{aligned} \tag{3.12}$$

using (3.7), for $i=1,\ldots,n$ and $\boldsymbol{r}\in\mathcal{R}$.

Since the terms in (3.11) for $Q(\boldsymbol{\Theta};\boldsymbol{\Theta}^{(k-1)})$ containing $E(\ln(X_{ij}) \mid Z_{i\boldsymbol{r}}=1, l_{ij}, u_{ij}, t_j^l, t_j^u;\theta^{(k-1)})$ do not depend on the unknown parameter vector $\boldsymbol{\Theta}$, they will not play a role in the EM algorithm. In the E-step, we need to compute the expected value of X_{ij} conditional on the censoring and truncation points and the mixing component $Z_{i\boldsymbol{r}}$ for the current value $\boldsymbol{\Theta}^{(k-1)}$ of the parameter vector. For $i=1,\ldots,n$ and $\boldsymbol{r}\in\mathcal{R}$, we have

$$\begin{aligned} &E\left(X_{ij}\,\middle|\,Z_{i\boldsymbol{r}}=1,l_{ij},u_{ij},t_j^l,t_j^u;\theta^{(k-1)}\right) \\ &\quad=\int_{l_{ij}}^{u_{ij}}x\frac{f(x;r_j,\theta^{(k-1)})}{F(u_{ij};r_j,\theta^{(k-1)})-F(l_{ij};r_j,\theta^{(k-1)})}dx \\ &\quad=\frac{r_j\theta^{(k-1)}}{F(u_{ij};r_j,\theta^{(k-1)})-F(l_{ij};r_j,\theta^{(k-1)})}\int_{l_{ij}}^{u_{ij}}\frac{x^{r_j}e^{-x/\theta^{(k-1)}}}{\left(\theta^{(k-1)}\right)^{r_j+1}r_j!}dx \\ &\quad=\frac{r_j\theta^{(k-1)}\left(F(u_{ij};r_j+1,\theta^{(k-1)})-F(l_{ij};r_j+1,\theta^{(k-1)})\right)}{F(u_{ij};r_j,\theta^{(k-1)})-F(l_{ij};r_j,\theta^{(k-1)})}\,, \end{aligned} \tag{3.13}$$

in case $l_{ij}<u_{ij}$ and in case $l_{ij}=u_{ij}=x_{ij}$, the observation is uncensored and the expression is equal to x_{ij}.

M-step In the kth iteration of the M-step, we maximize the expected value (3.11) of the complete data log-likelihood obtained in the E-step with respect to the parameter vector $\boldsymbol{\Theta}$ over all $(\boldsymbol{\beta},\theta)$ with $\beta_{\boldsymbol{r}}\geqslant 0$, $\sum_{\boldsymbol{r}\in\mathcal{R}}\beta_{\boldsymbol{r}}=1$ and $\theta>0$. The

maximization with respect to the mixing weights $\boldsymbol{\beta}$, requires the maximization of

$$\sum_{i=1}^{n} \sum_{\boldsymbol{r} \in \mathcal{R}} z_{i\boldsymbol{r}}^{(k)} \ln(\beta_{\boldsymbol{r}}) \,,$$

which can be done analogously as in the univariate case, yielding

$$\beta_{\boldsymbol{r}}^{(k)} = n^{-1} \sum_{i=1}^{n} z_{i\boldsymbol{r}}^{(k)} \qquad \text{for } \boldsymbol{r} \in \mathcal{R} \,. \tag{3.14}$$

The average over the posterior probabilities of belonging to the jth component in the mixture forms the new estimator for the prior probability β_j in the truncated mixture.

We set the first order partial derivative with respect to θ equal to zero in order to maximize $Q(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(k-1)})$ over θ (see Appendix 3.7), leading to the following M-step equation:

$$\theta^{(k)} = \frac{n^{-1} \sum_{i=1}^{n} \sum_{\boldsymbol{r} \in \mathcal{R}} z_{i\boldsymbol{r}}^{(k)} \sum_{j=1}^{d} E\left(X_{ij} \left| Z_{i\boldsymbol{r}} = 1, l_{ij}, u_{ij}, t_j^l, t_j^u; \theta^{(k-1)} \right.\right) - T^{(k)}}{\sum_{\boldsymbol{r} \in \mathcal{R}} \beta_{\boldsymbol{r}}^{(k)} \sum_{j=1}^{d} r_j} \tag{3.15}$$

with

$$T^{(k)} = \sum_{\boldsymbol{r} \in \mathcal{R}} \beta_{\boldsymbol{r}}^{(k)} \sum_{j=1}^{d} \left. \frac{\left(t_j^l\right)^{r_j} e^{-t_j^l/\theta} - \left(t_j^u\right)^{r_j} e^{-t_j^u/\theta}}{\theta^{r_j - 1} (r_j - 1)! \left(F(t_j^u; r_j, \theta) - F(t_j^l; r_j, \theta)\right)} \right|_{\theta = \theta^{(k)}} .$$

Similar to the univariate case (Verbelen et al., 2015), the new estimator $\theta^{(k)}$ in (3.15) for the common scale parameter θ has the interpretation of the expected total mean divided by the weighted total shape parameter in the mixture minus a correction term $T^{(k)}$ due to the truncation. Since $T^{(k)}$ in (3.15) depends on $\theta^{(k)}$ and has a complicated form, it is not possible to find an analytical solution. Therefore, we use a Newton-type algorithm, with the previous value of θ, i.e. $\theta^{(k-1)}$, as starting value, to solve the equation.

We iterate the E- and M-step until the difference in log-likelihood $l(\boldsymbol{\Theta}^{(k)}; \mathcal{X}) - l(\boldsymbol{\Theta}^{(k-1)}; \mathcal{X})$ between two iterations becomes sufficiently small. By inverting expression (3.7), we retrieve the maximum likelihood estimator of the original mixing

weights $\alpha_{\boldsymbol{r}}^{(k)}$ for $\boldsymbol{r} \in \mathcal{R}$. We first compute

$$\widetilde{\alpha}_{\boldsymbol{r}} = \frac{\widehat{\beta_{\boldsymbol{r}}}}{\prod_{j=1}^{d} \left[F(t_j^u; r_j, \widehat{\theta}) - F(t_j^l; r_j, \widehat{\theta}) \right]} \qquad \text{for } \boldsymbol{r} \in \mathcal{R}\,, \tag{3.16}$$

where $\widehat{\beta_{\boldsymbol{r}}}$ and $\widehat{\theta}$ denote the values in the final EM step, and then normalize the weights such that they sum to 1.

Using the EM algorithm, the log-likelihood (3.6) increases with each iteration (McLachlan and Krishnan, 2008). The estimator for $\boldsymbol{\Theta} = (\boldsymbol{\alpha}, \theta)$ obtained from the EM algorithm has the same limit as the maximum likelihood estimator, whenever the starting value is adequately chosen. Hence, the maximum likelihood asymptotic theory in terms of consistency, asymptotic normality and asymptotic efficiency applies. Within the EM framework, the asymptotic covariance matrix of the maximum likelihood estimator can be assessed (McLachlan and Krishnan, 2008).

These asymptotic results can only be applied with respect to $\boldsymbol{\Theta}$, given a fixed shape set $\mathcal{R}$. However, the number of mixture components and the corresponding values of the shape parameter vectors also have to be estimated for which we discuss a strategy in the next section. The asymptotic results stated here do not take this form of model selection into account. In Section 3.5.3 we apply a bootstrap approach to obtain bootstrap confidence intervals for the value of Kendall's τ and Spearman's ρ.

3.4 Computational details

An efficient multivariate extension of the univariate EM estimation procedure for Erlang mixtures is not straightforward. Indeed, initialization of the parameter values and model selection are the main difficulties when estimating a multivariate Erlang mixture to a data sample and are crucial for its practical use in data analysis. We fill this gap and suggest an effective method to initialize the parameters of a multivariate Erlang mixture and a strategy to select the best set of shape parameter vectors using a model selection criterion.

3.4.1 Initialization and first run of the EM algorithm

Property 1 ensures that any positive continuous distribution can be approximated by an MME. The formulation of the property also shows how this approximation

can be achieved in case the density to be approximated is available. Therefore, it serves as a starting point on how to come up with initial values in case of a sample of observations. A priori, it is however not clear how to translate the property to a finite sample setting.

Initializing data In a finite sample setting, we do not have the underlying density function at our disposal and initialize the parameters making use of an initializing data matrix $\boldsymbol{y}$ of dimension $n \times d$ which contains x_{ij} if the jth element of observation i is uncensored, l_{ij} in the case of right censoring, u_{ij} in the case of left censoring, and $(l_{ij} + u_{ij})/2$ in case of interval censoring. Hence, we use popular simple imputation techniques (see e.g. Leung et al., 1997) to deal with the censoring in the initial step. If the jth element of observation i is missing or right censored at 0, we set y_{ij} equal to missing.

Shapes For any given initial common scale $\theta^{(0)}$, instead of using an infinite set of positive integer shape parameters in each dimension (cfr. Property 1), we restrict this to a maximum number M of shape parameters in each dimension. We select these shape parameters in a sensible way by using M quantiles ranging from the minimum to the maximum in each dimension in order to make a data-driven decision on the locations of the shape parameters. Denoting the p-percent quantile of the initializing data in dimension j by $Q(p; \boldsymbol{y}_j)$, and taking into account that the expected value of a univariate Erlang distribution with shape r and scale θ equals $r\theta$, the set of positive integer shapes in dimension j is chosen as

$$\{r_{1,j}, \ldots, r_{M_j,j}\} = \left\{ \left\lceil \frac{Q(p; \boldsymbol{y}_j)}{\theta^{(0)}} \right\rceil \middle| \, p = 0, \frac{1}{M-1}, \frac{2}{M-1}, \ldots, 1 \right\} . \tag{3.17}$$

where $\lceil \ \rceil$ denotes upwards rounding, due to the fact that the shapes have to be positive integers. Consequently, several shapes might coincide which results in $M_j \leqslant M$ shape parameters in dimension j. The initial shape set is then constructed as the Cartesian product of the d sets of positive integer shape parameters in each dimension:

$$\mathcal{R} = \{r_{1,1}, \ldots, r_{M_1,1}\} \times \cdots \times \{r_{1,d}, \ldots, r_{M_d,d}\} . \tag{3.18}$$

Weights The shape parameters in each dimension, multiplied with the common scale parameter $\theta^{(0)}$, form a grid that covers the sample range. As an empirical version of Property 1, the weights $\alpha_{\boldsymbol{r}}$, for each shape parameter vector

$\boldsymbol{r} = (r_{m_1,1}, \ldots, r_{m_d,d})$ in $\mathcal{R}$, with $1 \leqslant m_j \leqslant M_j$ for all $j = 1, \ldots, d$, are initialized by the relative frequency of data points in the d-dimensional rectangle $(r_{m_1-1,1}\theta^{(0)}, r_{m_1,1}\theta^{(0)}] \times \cdots \times (r_{m_d-1,d}\theta^{(0)}, r_{m_d,d}\theta^{(0)}]$ defined by the grid:

$$\alpha^{(0)}_{\boldsymbol{r}=(r_{m_1,1},\ldots,r_{m_d,d})} = n^{-1} \sum_{i=1}^{n} \prod_{j=1}^{d} I\left(r_{m_j-1,j}\theta^{(0)} < y_{ij} \leqslant r_{m_j,j}\theta^{(0)}\right), \tag{3.19}$$

with $r_{0,j} = 0$ for notational convenience and the indicator equal to $1/M_j$ in case y_{ij} is missing. If this hyperrectangle does not contain any data points, the initial weight corresponding to the multivariate Erlang in the mixture with that shape vector will be set equal to zero. Consequently, the weight will remain zero at each subsequent iteration of the EM algorithm (see formulas (3.12) and (3.14)). Therefore, these shape vectors can immediately be removed from the set $\mathcal{R}$. At initialization, the truncation is only taken into account to transform the initial values for $\boldsymbol{\alpha}$ into the initial values for $\boldsymbol{\beta}$ via (3.7).

The maximal number of shape vectors is limited to M^d at the initial step. However, due to the fact that $M_j \leqslant M$ and many shape parameter vectors will receive an initial weight equal to zero, the actual number of shape vectors at the initial step will be lower.

Common scale The initial value of the common scale θ is the most influential for the performance of the initial multivariate Erlang mixture, as is the case in the univariate setting (Verbelen et al., 2015). A value which is too large will result in a multivariate mixture which is too flat (*'underfit'*); a value which is too small will lead to a mixture which is too peaky (*'overfit'*). A priori, it is not evident how one can make an insightful decision on θ. Similar to Verbelen et al. (2015), we therefore introduce an additional tuning parameter: an integer spread factor s. We propose to initialize the common scale as

$$\theta^{(0)} = \frac{\min_j(\max_i(y_{ij}))}{s}. \tag{3.20}$$

Due to the use of marginal quantiles in (3.17) , the range of the shape parameters varies according to the sample ranges in each dimension $j = 1, \ldots, d$ with a maximum shape parameter equal to

$$r_{M_j,j} = \left\lceil \frac{\max_i(y_{ij})}{\theta^{(0)}} \right\rceil = \left\lceil \frac{\max_i(y_{ij})}{\min_j(\max_i(y_{ij}))} s \right\rceil. \tag{3.21}$$

Hence, the spread factor s determines the maximum shape parameter in the dimension with the smallest maximum. The fact that the common scale parameter is equal across all dimensions is compensated by the different choice of the shape parameters in each dimension based on marginal quantiles. This ensures that the initialization works well when the ranges in each dimension are different and also gives reasonable initial approximations in case the data are skewed.

Apply EM algorithm Given an initial choice for the set $\mathcal{R}$ of shape parameter vectors, the initial common scale estimate $\theta^{(0)}$ and the initial weights $\boldsymbol{\beta}^{(0)} = \{\beta_{\boldsymbol{r}}^{(0)} \,|\boldsymbol{r} \in \mathcal{R}\}$, we find the maximum likelihood estimators for $(\boldsymbol{\beta}, \theta)$ corresponding to this initial multivariate mixtures of Erlangs, denoted by MME_{init}, via the EM algorithm as explained in section 3.3.3. An overview of the initialization and the EM algorithm written in pseudo code is given in Algorithm 1.

Algorithm 1 EM algorithm for a multivariate Erlang mixture.

{**Initial step**}
Choose M and s
Compute: θ as in (3.20)
shape parameters in each dimension as in (3.17)
shape set $\mathcal{R}$ as in (3.18)
mixture weights $\boldsymbol{\alpha}$ as in (3.19)
$\mathcal{R} \leftarrow \{\boldsymbol{r} \in \mathcal{R} \,|\alpha_{\boldsymbol{r}} \neq 0\}$
Transform weights $\boldsymbol{\alpha}$ to $\boldsymbol{\beta}$ as in (3.7)
{**EM algorithm**}
while log-likelihood (3.6) improves **do**
  {**E-step**}
  Compute: posterior probabilities (3.12)
  conditional expectations (3.13)
  {**M-step**}
  Update: weights $\boldsymbol{\beta}$ as in (3.14)
  scale θ by numerically solving (3.15)
end while
Transform weights $\boldsymbol{\beta}$ to $\boldsymbol{\alpha}$ using (3.16)
return $\text{MME}_{init} = (\mathcal{R}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \theta)$

3.4.2 Reduction of the shape vectors

The initial shape set $\mathcal{R}$ might not be optimal. After application of the EM algorithm, we reduce the number of mixture components of the fitted multivariate Erlang mixture. We use a backward stepwise search based on an information cri-

Algorithm 2 Reduction of the shape vectors

input $\text{MME}_{init} = (\mathcal{R}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \theta)$
while BIC (3.22) improves **and** $|\mathcal{R}| > 1$ **do**
  $\mathcal{R}_{red} \leftarrow \{\boldsymbol{r} \in \mathcal{R} \,|\beta_{\boldsymbol{r}} \neq \min_{\boldsymbol{r}\in\mathcal{R}} \beta_{\boldsymbol{r}}\}$
  $(\boldsymbol{\beta}^{(0)}, \theta^{(0)})_{red} \leftarrow (\{\beta_{\boldsymbol{r}} / \sum_{\boldsymbol{r}\in\mathcal{R}_{red}} \beta_{\boldsymbol{r}} \,|\boldsymbol{r} \in \mathcal{R}_{red}\}, \theta)$
  Compute MLE for $(\boldsymbol{\beta}, \theta)_{red}$ using the EM algorithm with initial values $(\boldsymbol{\beta}^{(0)}, \theta^{(0)})_{red}$
  if BIC (3.22) improves **then**
    $\mathcal{R} \leftarrow \mathcal{R}_{red}$
    $(\boldsymbol{\beta}, \theta) \leftarrow (\boldsymbol{\beta}, \theta)_{red}$
  end if
end while
return $\text{MME}_{red} = (\mathcal{R}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \theta)$

terion. Information criteria, such as Akaike's information criterion (AIC, Akaike, 1974) and Schwartz's Bayesian information criterion (BIC, Schwarz, 1978), measure the quality of the model as a trade-off between the goodness-of-fit, via the log-likelihood, and the model complexity, via the number of parameters in the model. Models with a smaller value of the information criterion are preferred. Based on numerical experiments, we prefer the use of BIC over AIC since it has a stronger penalty term for the number of parameters in the model and hence leads to more parsimonious models. BIC is computed as

$$\text{BIC} = -2 \cdot l(\boldsymbol{\Theta}; \mathcal{X}) + \ln(n) \cdot |\mathcal{R}| \cdot (d+1)\,, \tag{3.22}$$

where $|\mathcal{R}|$ indicates the number of shape parameter vectors in the shape set $\mathcal{R}$.

We reduce the number of mixture components by removing all redundant shape vectors from the initial mixture based on BIC. In the backward selection strategy, depicted in pseudo code in Algorithm 2, we delete the shape parameter vector $\boldsymbol{r}$ from the set $\mathcal{R}$ for which the corresponding mixture component has the smallest weight $\beta_{\boldsymbol{r}}$. The remaining weights are standardized to sum to one. Along with the previous maximum likelihood estimate for the common scale, they serve as initial estimates to find the maximum likelihood estimators for $(\boldsymbol{\beta}, \theta)$ corresponding to the reduced set $\mathcal{R}_{red}$ of shape parameter vectors by again applying the EM algorithm. In case this maximum likelihood estimate achieves a lower BIC value, the reduced set $\mathcal{R}_{red}$ of shape parameters is accepted and we reduce the number of components further in the same manner. If not, we keep the previous set. This backward approach provides efficient initial parameter estimates

for the reduced set of shape parameter vectors and ensures a fast convergence of the EM algorithm.

3.4.3 Adjustment of the shape vectors

Algorithm 3 Adjustment of the shape combinations

input $\text{MME}_{red} = (\mathcal{R}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \theta)$
while log-likelihood (3.6) improves **do**
  for $j \in \{1, \ldots, d\}$ **do**
    for $\widetilde{\boldsymbol{r}} \in \mathcal{R}$ **do**
      repeat
        if $(\widetilde{r}_1, \ldots, \widetilde{r}_j + 1, \ldots, \widetilde{r}_d) \notin \mathcal{R}$ **then**
          $\mathcal{R}_{adj} \leftarrow \{\boldsymbol{r} \in \mathcal{R} \,|\boldsymbol{r} \neq \widetilde{\boldsymbol{r}}\} \cup \{(\widetilde{r}_1, \ldots, \widetilde{r}_j + 1, \ldots, \widetilde{r}_d)\}$
          Compute MLE for $(\boldsymbol{\beta}, \theta)_{adj}$ using the EM algorithm with initial values $(\boldsymbol{\beta}, \theta)$
          if log-likelihood (3.6) improves **then**
            $\mathcal{R} \leftarrow \mathcal{R}_{adj}$
            $(\boldsymbol{\beta}, \theta) \leftarrow (\boldsymbol{\beta}, \theta)_{adj}$
          end if
        end if
      until $(\widetilde{r}_1, \ldots, \widetilde{r}_j + 1, \ldots, \widetilde{r}_d) \in \mathcal{R}$ **or** log-likelihood (3.6) no longer improves
    end for
    for $\widetilde{\boldsymbol{r}} \in \mathcal{R}$ **do**
      repeat
        if $(\widetilde{r}_1, \ldots, \widetilde{r}_j - 1, \ldots, \widetilde{r}_d) \notin \mathcal{R}$ **and** $\widetilde{r}_j - 1 \geqslant 1$ **then**
          $\mathcal{R}_{adj} \leftarrow \{\boldsymbol{r} \in \mathcal{R} \,|\boldsymbol{r} \neq \widetilde{\boldsymbol{r}}\} \cup \{(\widetilde{r}_1, \ldots, \widetilde{r}_j - 1, \ldots, \widetilde{r}_d)\}$
          Compute MLE for $(\boldsymbol{\beta}, \theta)_{adj}$ using the EM algorithm with initial values $(\boldsymbol{\beta}, \theta)$
          if log-likelihood (3.6) improves **then**
            $\mathcal{R} \leftarrow \mathcal{R}_{adj}$
            $(\boldsymbol{\beta}, \theta) \leftarrow (\boldsymbol{\beta}, \theta)_{adj}$
          end if
        end if
      until $(\widetilde{r}_1, \ldots, \widetilde{r}_j - 1, \ldots, \widetilde{r}_d) \in \mathcal{R}$ **or** $\widetilde{r}_j - 1 = 0$ **or** log-likelihood (3.6) no longer improves
    end for
  end for
end while
return $\text{MME}_{adj} = (\mathcal{R}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \theta)$

In a next step we improve the shape parameter vectors of the remaining Erlang components in the mixture. Each time we adjust one of the components of a shape parameter vector by shifting its value by one (increase or decrease) and use the maximum likelihood estimates $(\widehat{\boldsymbol{\beta}}, \widehat{\theta})$ corresponding to the current shape parameter set $\mathcal{R}$ as initial values $(\boldsymbol{\beta}^{(0)}, \theta^{(0)})_{adj}$ of the mixture of Erlang distributions with slightly adjusted shape parameter vector set $\mathcal{R}_{adj}$. These initial values are close to the maximum likelihood estimates which guarantees fast convergence. In case the maximum likelihood estimate corresponding to the adjusted set $\mathcal{R}_{adj}$ achieves a lower log-likelihood value (3.6), the adjusted set $\mathcal{R}_{adj}$ is accepted and we continue adjusting the value of the shape parameter in the same direction. If not, we keep the previous set of shape parameter combinations.

The gradual adjustment strategy of the shape parameter combinations is described in detail in Algorithm 3. While the log-likelihood improves, we continue to consecutively increase or decrease the value of a component of a shape parameter vector if it leads to a better fit. The algorithm converges when no single addition or subtraction of the value of any of the components of any of the shape parameter vectors leads to an improvement in the log-likelihood.

After adjusting the shape parameters, we apply the reduction step in combination with the adjustment step. Based on BIC we further reduce the number of shape parameter vectors by deleting the shape vector with the smallest mixture weight and adjusting the values of the remaining ones. The outline of this adjustment and further reduction of the shape parameter vectors, which results in the final MME, is given in Algorithm 4.

Algorithm 4 Adjustment and further reduction of the shape vectors

input $\text{MME}_{adj} = (\mathcal{R}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \theta)$
while BIC (3.22) improves **and** $|\mathcal{R}| > 1$ **do**
  $\mathcal{R}_{red} \leftarrow \{\boldsymbol{r} \in \mathcal{R} \,|\beta_{\boldsymbol{r}} \neq \min_{\boldsymbol{r}\in\mathcal{R}} \beta_{\boldsymbol{r}}\}$
  $(\boldsymbol{\beta}^{(0)}, \theta^{(0)})_{red} \leftarrow (\{\beta_{\boldsymbol{r}} / \sum_{\boldsymbol{r}\in\mathcal{R}_{red}} \beta_{\boldsymbol{r}} \,|\boldsymbol{r} \in \mathcal{R}_{red}\}, \theta)$
  Compute MLE for $(\boldsymbol{\beta}, \theta)_{red}$ using the EM algorithm with initial values $(\boldsymbol{\beta}^{(0)}, \theta^{(0)})_{red}$
  Apply adjustment algorithm 3
  if BIC (3.22) improves **then**
    $\mathcal{R} \leftarrow \mathcal{R}_{adj}$
    $(\boldsymbol{\beta}, \theta) \leftarrow (\boldsymbol{\beta}, \theta)_{adj}$
  end if
end while
return $\text{MME}_{adj} = (\mathcal{R}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \theta)$

3.5 Examples

We demonstrate the proposed fitting procedure on three data sets, each time highlighting a different aspect of multivariate mixtures of Erlangs. In a first simulated two-dimensional example, we explicitly illustrate the different steps of the estimation procedure. Second, we model the waiting time between eruptions and the duration of the eruptions of the old faithful geyser data set. Based on the fitted two-dimensional MME, we immediately obtain the distribution of the sum of the waiting time and the duration, representing the total cycle time. In the third example, we use multivariate mixtures of Erlangs to model the udder infection times of dairy cows observed in a mastitis study, and use the fitted MME to analytically quantify the positive correlation between the udder infection times using the explicit expression of the bivariate measures of association Kendall's tau and Spearman's rho in the MME setting.

The resulting MME after applying the different steps in choosing the shape vectors depends heavily on the starting values. Therefore it is crucial to sufficiently explore the effect of changing the value of the tuning parameters M and s and compare the results of several different initial starting points for the shape set. In addition to the value of BIC, graphs aid the assessment of the fitted model.

3.5.1 Simulated data

As a first example, we generate 1000 uncensored and untruncated observations from a bivariate normal copula with correlation coefficient 0.75 and Erlang distributed margins with shape parameter equal to 2 and 10, respectively, and scale parameter equal to 3 and 20, resp. A scatterplot of this simulated data set is shown in Figure 3.1a. Due to the parameter choice, the ranges in each dimension are quite different.

We now apply the different steps of the estimation procedure on this data set and graphically illustrate the interpretations and effects of these steps. First we consider the initialization strategy for the shape set $\mathcal{R}$, the scale parameter θ and the mixture weights $\boldsymbol{\beta}$, based on the denseness property of MME in Property 1, as explained in Section 3.4.1. This strategy is controlled by two tuning parameters, a maximum number M of shape parameters in each dimension and a spread factor s. In this illustration, we use $M = 10$ and $s = 20$. For this choice, the scale θ is initialized as

$$\theta^{(0)} = \frac{\min_j(\max_i(x_{ij}))}{s} = \frac{27.32452}{20} = 1.366226\,.$$

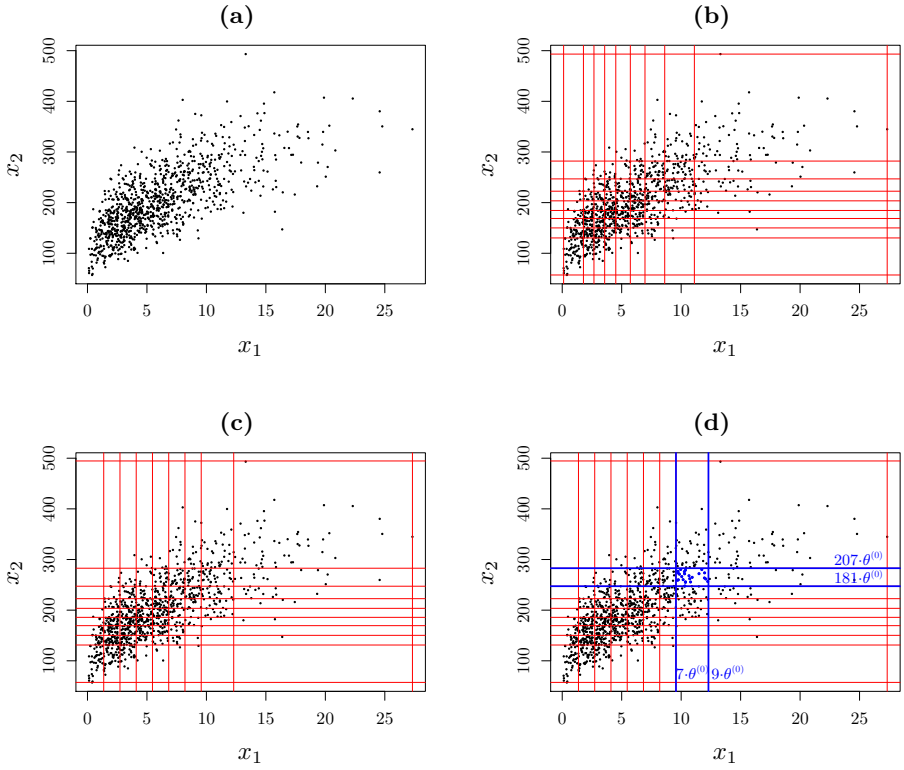

Figure 3.1: *Simulated example: (a) scatterplot, (b) marginal quantile grid, (c) grid formed by multiplying the shapes (3.17) by the common scale (3.20) and (d) initial weight* $\alpha^{(0)}_{\boldsymbol{r}=(9,207)} = 0.024$.

In order to make a data driven choice for the initial positions of the shape parameters, we compute M marginal quantiles in each dimension, which are depicted in Figure 3.1b and form a grid that covers the data range. These marginal quantiles are then divided by the initial scale $\theta^{(0)}$ and rounded upwards to initialize the shape parameters in each dimension:

$$\{r_{1,j}, \ldots, r_{M_j,j}\} = \left\{ \left\lceil \frac{Q(p; \boldsymbol{x}_j)}{\theta^{(0)}} \right\rceil \middle| p = 0, \frac{1}{9}, \frac{2}{9}, \ldots, 1 \right\} \quad \text{for } j = 1, 2\,.$$

The shape set $\mathcal{R}$ is constructed as the Cartesian product of the set of shape

parameters in each dimension:

$$\begin{aligned}\mathcal{R} &= \{r_{1,1}, \ldots, r_{M_1,1}\} \times \{r_{1,2}, \ldots, r_{M_2,2}\} \\ &= \{1, 2, 3, 4, 5, 6, 7, 9, 20\} \times \{42, 96, 110, 124, 136, 149, 163, 181, 207, 362\}\,.\end{aligned}$$

Due to the rounding, shape 2 appears twice in the first dimension and only 9 instead of 10 shapes remain in that dimension. Due to the choice of $\theta^{(0)}$, $s = 20$ is the maximal shape parameter in the first dimension, the dimension with the smallest maximum. The maximal shape in the second dimension is s times the ratio of the maximum in the second dimension and the lowest maximum, rounded upwards (see (3.21)). If we multiply this shape set $\mathcal{R}$ with the initial scale $\theta^{(0)}$, we obtain a grid that covers the entire sample range which is depicted in Figure 3.1c. This grid differs from the marginal quantile grid due to the rounding and is used to initialize the weights as the relative frequency of data points in the 2-dimensional rectangle corresponding to each shape vector:

$$\alpha^{(0)}_{\boldsymbol{r}=(r_{m_1,1}, r_{m_2,2})} = 0.001 \sum_{i=1}^{1000} \prod_{j=1}^{2} I\left(r_{m_j-1,j}\theta^{(0)} < y_{ij} \leqslant r_{m_j,j}\theta^{(0)}\right)\,.$$

For example, for the shape vector $\boldsymbol{r} = (r_{m_1,1}, r_{m_2,2}) = (9, 207)$, we consider the 2-dimensional rectangle $(r_{m_1-1,1}\theta^{(0)}, r_{m_1,1}\theta^{(0)}] \times (r_{m_2-1,2}\theta^{(0)}, r_{m_2,2}\theta^{(0)}] = (7 \cdot \theta^{(0)}, 9 \cdot \theta^{(0)}] \times (181 \cdot \theta^{(0)}, 207 \cdot \theta^{(0)}]$ shown in Figure 3.1d, leading to an initial weight of

$$\begin{aligned}\alpha^{(0)}_{\boldsymbol{r}=(9,207)} &= 0.001 \sum_{i=1}^{1000} I\left(7 \cdot \theta^{(0)} < y_{i1} \leqslant 9 \cdot \theta^{(0)}\right) I\left(181 \cdot \theta^{(0)} < y_{i2} \leqslant 207 \cdot \theta^{(0)}\right) \\ &= 0.024\,,\end{aligned}$$

since 24 of the 1000 observations lie in this rectangle. The resulting initial MME contains 71 shape vectors with a nonzero weight and already forms a reasonable approximation for the main portion of the data. In Figure 3.2a, we show the scatterplot of the data with an overlay of the density of the initial MME using a contour plot and heat map. In the margins, we plot the marginal histograms with an overlay of the true densities in blue and the fitted densities in red. In the second dimension, there is too much weight in the tail and too little near the origin. After applying the EM algorithm a first time with these initial estimates, we obtain the maximum likelihood estimates of the weights and scale corresponding to this choice of the shape set (Section 3.4). In Figure 3.2b, we observe that the fit is

better in the tail, but there is still too little weight in the second dimension near the origin, due to a bad positioning of the first shape in second dimension.

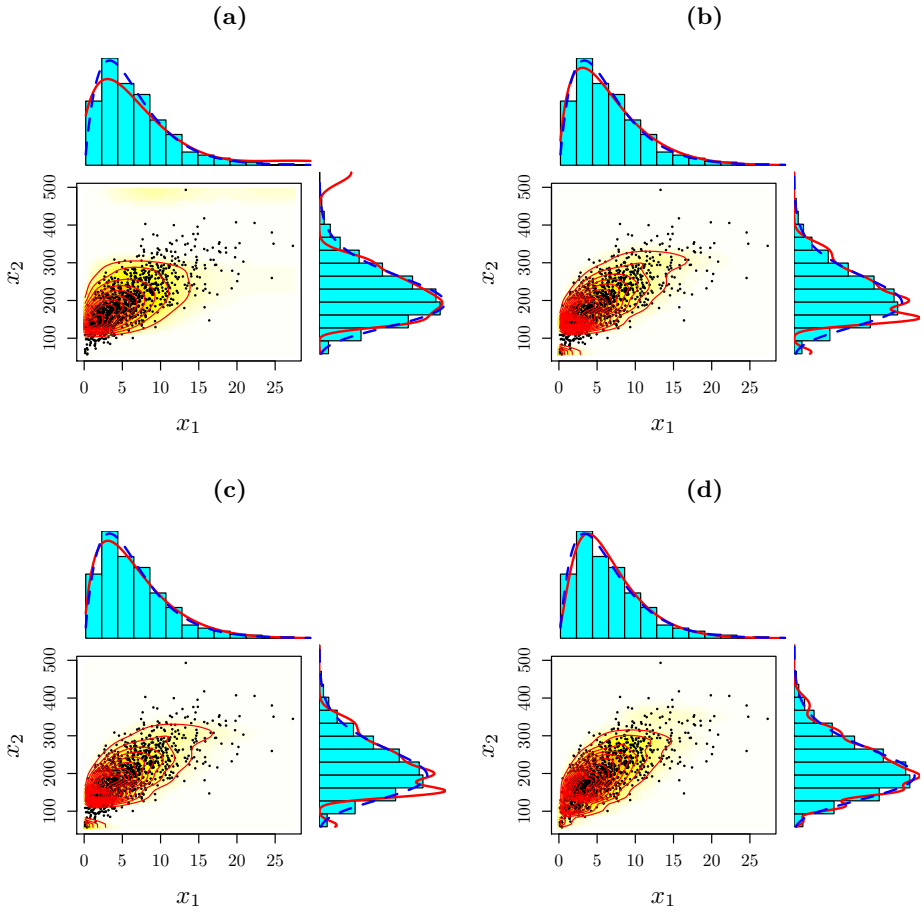

Figure 3.2: *Scatterplot of the simulated data with an overlay of the fitted density of the MME using a contour plot and heat map. In the margins, we plot the marginal histograms with an overlay of the true densities in blue and the fitted densities in red. In (a), we display the fit after initialization, in (b) after applying the EM algorithm a first time, in (c) after applying the reduction step and in (d) after applying the adjustment and further reduction step.*

Hence, the initial set of shape parameter vectors is not ideal and additional steps are required to improve the shape set. First, we reduce the number of mix-

ture components from 71 to 17 by subsequently removing the mixture component having the smallest weight if it is found to be redundant based on BIC (Section 3.4.2). The fit of this reduced mixture in Figure 3.2c nearly coincides with the one in Figure 3.2b. Second, we adjust the values of the shape parameter vectors and further reduce the number of mixture components based on BIC (Section 3.4.3) until we obtain a close-fitting MME with 11 shape parameter vectors (Figure 3.2d). The parameter estimates of this final MME are given in Table 3.1.

Table 3.1: *Parameter estimates of the MME with 11 mixture components fitted to the simulated data.*

$\boldsymbol{r}$	$\alpha_{\boldsymbol{r}}$	θ
(1, 56)	0.0124	1.2889
(2, 84)	0.0814	
(3, 112)	0.1773	
(3, 132)	0.1005	
(4, 143)	0.1568	
(4, 164)	0.0257	
(5, 164)	0.1320	
(6, 189)	0.1586	
(8, 223)	0.1097	
(11, 273)	0.0446	
(11, 382)	0.0010	

3.5.2 Old faithful geyser data

We consider the waiting time between eruptions and the duration of the eruption for the Old faithful geyser in Yellowstone National Park, Wyoming, USA. We use the version of Azzalini and Bowman (1990) which contains 299 observations. This data set is popular in the field of nonparametric density estimation (see e.g. Silverman, 1986; Härdle, 1991). We stress that we use MME as a multivariate density estimation technique, and not as a mixture modeling technique to identify subgroups in this data.

We fit a two-dimensional MME to the data using the fitting strategy explained in Section 3.4. We perform a grid search to identify good values for the tuning parameters M and s. We let s vary between 10 and 90 by 10 and between 100 and 1000 by 100 and set M equal to 5, 10 and 20. To illustrate the importance and effect of the tuning parameters, we report part of the results of the search grid, up to $s = 200$, in Figure 3.3 and Table 3.2. Values of s beyond 200 resulted in MME which were overfitting the data.

Table 3.2: *BIC values and number of mixture components when fitting an MME to the Old Faithful geyser data, starting from different values of the tuning parameters. The minimum BIC value is underlined and obtained for $M = 10$ and $s = 90$.*

	$M = 5$		$M = 10$		$M = 20$	
s	BIC	$\lvert\mathcal{R}\rvert$	BIC	$\lvert\mathcal{R}\rvert$	BIC	$\lvert\mathcal{R}\rvert$
10	3211.134	2	3211.134	2	3211.134	2
20	3133.564	5	3148.824	5	3148.824	5
30	3069.731	6	3069.731	6	3083.757	6
40	3056.588	8	3024.869	9	3051.427	6
50	3026.997	8	3011.941	12	3023.951	15
60	3011.567	8	3008.350	14	3040.962	16
70	3008.319	8	3008.350	14	3018.867	15
80	3015.743	8	3007.694	15	3039.017	17
90	3028.742	8	<u>2998.870</u>	15	3047.314	18
100	3029.431	8	3005.343	15	3023.761	17
200	3037.532	8	3026.490	23	3224.578	36

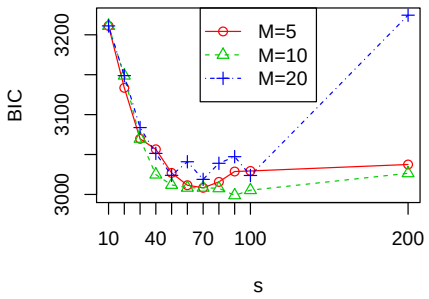

Figure 3.3: *BIC values when fitting an MME to the Old Faithful geyser data, starting from different values of the tuning parameters. The minimum BIC value is obtained for $M = 10$ and $s = 90$.*

Table 3.3: *Parameter estimates of the best-fitting MME with 15 mixture components fitted to the Old Faithful geyser data.*

$\boldsymbol{r}$	$\alpha_{\boldsymbol{r}}$	θ
(791, 79)	0.0061	0.0556
(893, 81)	0.1103	
(964, 79)	0.0798	
(1047, 77)	0.0795	
(1121, 83)	0.0378	
(1193, 79)	0.0402	
(1314, 74)	0.0893	
(1319, 37)	0.0387	
(1418, 73)	0.1284	
(1425, 36)	0.1380	
(1543, 73)	0.0633	
(1551, 36)	0.1249	
(1660, 72)	0.0142	
(1672, 34)	0.0462	
(1940, 36)	0.0033	

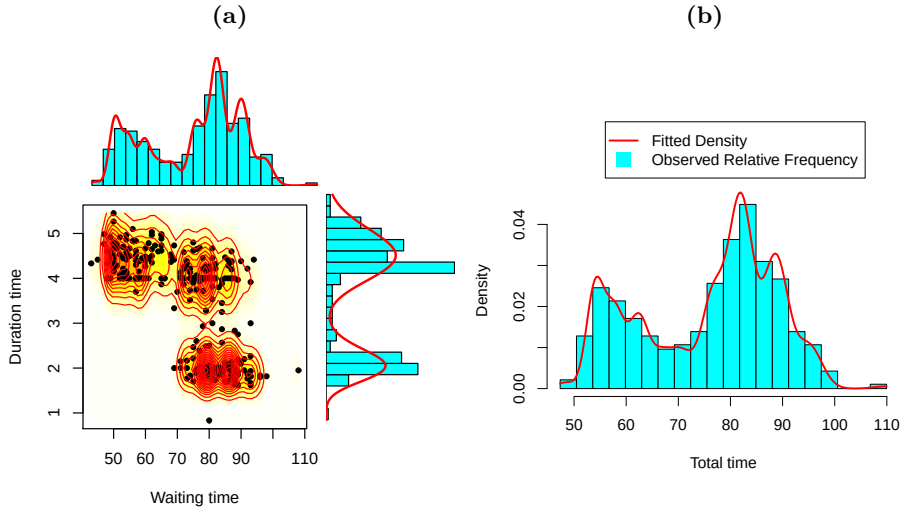

Figure 3.4: *Graphical evaluation of the best-fitting MME to the Old Faithful geyser data. In (a), we display the scatterplot of the data with an overlay of the fitted density using a contour plot and heat map. The margins show the marginal histograms with an overlay of the fitted densities in red. In (b), we compare the fitted density of the sum of the components and the histogram of the observed total cycle times.*

The resulting MME depends on the value of the tuning parameters. However, multiple MME can result in a satisfactory fit of the data. BIC indicates that the best-fitting MME is obtained for $M = 10$ and $s = 90$. The parameter estimates of this MME are reported in Table 3.3. Both the marginals as well as the dependence structure are adequately represented by this MME as is confirmed graphically in Figure 3.4a. Since the maximum of the waiting times is about 20 times as big as the maximum of the duration times whereas the scale parameter of the MME is the same across dimensions, the fitted marginal density is more capricious in the dimension of the waiting times and smoother in the dimension of the duration times.

We are interested in the distribution of the duration of the total cycle, i.e. the sum of the waiting time until the eruption and the duration of the eruption. Based on the fitted two-dimensional MME and due to the analytical properties of MME, we immediately obtain the distribution of this sum, which is a univariate mixture of Erlang distributions with the same scale, the sum of the shape parameters across the dimensions as shape parameters and the same corresponding weights

in (Lee and Lin, 2012, Theorem 5.1). Hence, the parameters of this univariate mixture of Erlang distributions are readily available from Table 3.3. Comparing the histogram of the observed total times to the fitted density in Figure 3.4b reveals a close approximation.

3.5.3 Mastitis study

Mastitis is economically one of the most important diseases in the dairy sector since it leads to reduced milk yield and milk quality. In this example, we consider infectious disease data from a mastitis study by Laevens et al. (1997). This data set has also been used in Goethals et al. (2009) and Ampe et al. (2012).

We focus on the infection times of individual cow udder quarters with a bacterium. As each udder quarter is separated from the three other quarters, one quarter might be infected while the other quarters remain infection-free. However, the dependence must be modeled since the data are hierarchical, with individual observations at the udder quarter level being correlated within the cow. Additionally, the infection times are not known exactly due to a period follow-up, which is often the case in observational studies since a daily checkup would not be feasible. Roughly each month, the udder quarters are sampled and the infection status is assessed, from the time of parturition, at which the cow was included in the cohort and assumed to be infection-free, until the end of the lactation period. This generates interval-censored data since for udder quarters that experience an event it is only known that the udder quarter got infected between the last visit at which it was infection-free and the first visit at which it was infected. Observations can also be right censored if no infection occurred before the end of the lactation period, which is roughly 300-350 days but different for every cow, before the end of the study or if the cow is lost to follow-up during the study, for example due to culling.

The data we consider contains information on 100 dairy cows on the time to infection of the four udder quarters by different types of bacteria. This data set is used in Goethals et al. (2009), who model the data using an extended shared gamma frailty model that is able to handle the interval censoring and clustering simultaneously. We treat the infection times at the udder quarter level of the cow as four-dimensional interval and right censored data of which we estimate the underlying density using MME. The udder quarters are denoted as RL (rear left), FL (front left), RR (rear right) and FR (front right).

In search for the best values of the tuning parameters in the MME estimation

procedure, we first fixed $M = 20$ and let s vary between 10 and 100 by 10 and between 100 and 1000 by 100. As the best final fit was obtained for $s = 10$, we varied M between 10 and 100 by 10 for s fixed at 10. The resulting fits did, however, not depend on M when s is as low as 10 since the starting values were identical. Varying s from 5 tot 15 for $M = 20$ confirmed that the best fit is obtained for $M = 20$ and $s = 10$. For this setting, the initial number of shape vectors was 73, which got reduced to 6 after the reduction step and to 4 after the adjustment step. The final parameter estimates of the best-fitting mixture are given in Table 3.4.

Table 3.4: *Parameter estimates of the best-fitting MME with four mixture components fitted to the mastitis data (infections by all bacteria).*

$\boldsymbol{r}$				$\alpha_{\boldsymbol{r}}$	θ
(2,	2,	2,	2)	0.4897	37.8621
(3,	5,	8,	4)	0.1331	
(7,	5,	2,	7)	0.2262	
(10,	14,	11,	8)	0.1510	

In order to graphically examine the goodness-of-fit of the fitted MME, we construct in Figure 3.5 a generalization of the scatterplot matrix. On the diagonal we compare the Turnbull nonparametric estimate of the survival curve for right and interval censored data (Turnbull, 1976), along with the log-transformed equal precision simultaneous confidence intervals (Nair, 1984), to the univariate marginal survival function of the fitted MME. On the off-diagonal, we construct bivariate scatterplots of interval and right censored data points, represented using the effective visualization of Li et al. (2015). Interval censored observation are depicted as segments or rectangles ranging from the lower to the upper censoring points and right censored observations are depicted as arrows starting from the lower censoring point and pointing to the censoring direction. On top, we display the contour plot and heat map representing the density of the bivarite marginal of the fitted MME. Based on this graph, we observe that in four dimensions, with 100 interval and right censored observations, we are able to fit an MME with four shape parameter vectors which appropriately captures the marginals as well as the dependence structure.

As a measure of the infectivity of the agent causing the disease, we are interested in the correlation between udder infection times. Due to the fact that the bivariate marginals again belong to the MME class and the analytical qual-

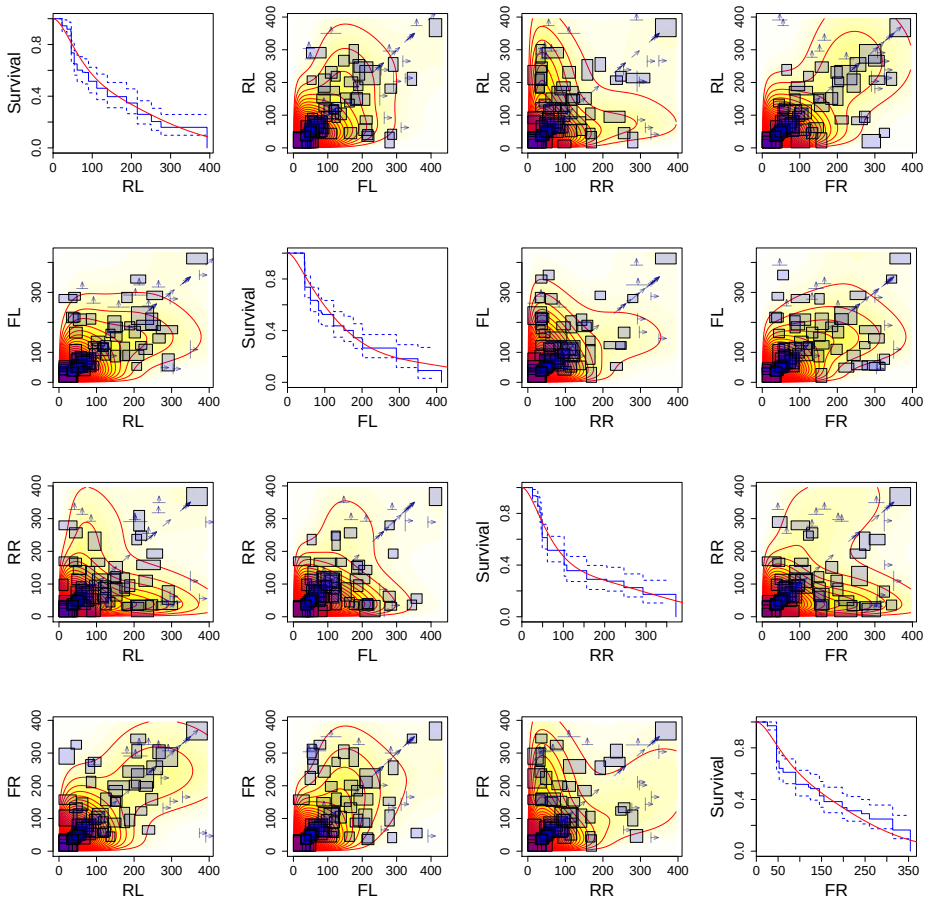

Figure 3.5: *Scatterplot matrix comparing the fitted four-dimensional MME to the observed interval and right censored observations of the mastitis data (infections by all bacteria). For more explanation, see Section 3.5.3*

ities of MME, we have closed-form expressions for Kendall's τ and Spearman's ρ (Lee and Lin, 2012, Theorem 3.2 and 3.3). Note that these do not depend on the common scale parameter. For the interval and right censored sample, we can hence estimate these measures based on the fitted MME to analytically quantify the positive correlation between each pair of udder quarter infection times (Table 3.5).

Inference is not straightforward due to the model selection as pointed out in

Section 3.3.3. In order to quantify the uncertainty and construct an approximate confidence interval for the bivariate measures of association, we resort to a bootstrapping procedure (Efron and Tibshirani, 1994). By sampling with replacement from the original four-dimensional data set of size 100, we generate 1000 bootstrap samples of the same size 100. For each of these bootstrap samples, we fit an MME where we set the tuning parameter M equal to 20 and let s vary between 5 and 25. We choose this fixed grid for each bootstrap sample since the optimal tuning parameters for the full sample were $M = 20$ and $s = 10$ and the starting values are not that sensitive with respect to M for low values of s. We thereby obtain 1000 estimates for each measure of association. The 5% and 95% quantiles of these estimates are used to construct a 90% bootstrap percentile confidence interval for each Kendall's τ and Spearman's ρ in Table 3.5.

Table 3.5: *Estimates and 90% bootstrap confidence intervals for the bivariate measures of association Kendall's τ and Spearman's ρ based on the fitted MME for the mastitis data (infections by all bacteria).*

		RL	FL	RR
FL	τ	0.4187		
		$(0.3329, 0.5515)$		
	ρ	0.6019		
		$(0.4727, 0.7439)$		
RR	τ	0.2018	0.3307	
		$(0.1693, 0.3989)$	$(0.2585, 0.4784)$	
	ρ	0.3004	0.4852	
		$(0.2423, 0.5616)$	$(0.3806, 0.6664)$	
FR	τ	0.4326	0.4105	0.2119
		$(0.3598, 0.5538)$	$(0.2701, 0.4883)$	$(0.1543, 0.3968)$
	ρ	0.6354	0.5994	0.3122
		$(0.5066, 0.7608)$	$(0.3875, 0.6794)$	$(0.2206, 0.5577)$

3.6 Discussion

MME form a highly flexible class of distributions which are at the same time mathematically tractable. From Property 1, we know that any positive continuous multivariate distribution can be approximated up to any accuracy by an infinite multivariate mixture of Erlang distributions. Our contribution presents a computationally efficient initialization and adjustment strategy for the shape parameter vectors, translating this theoretical aspect in a strong point in practice

as well. In the examples, we demonstrate how the fitting procedure is able to estimate an MME that adequately represents both the marginals and the dependence structure. By extending the EM algorithm, we are now also able to deal with left, interval or right censored and truncated data. MME therefore form a valuable multivariate density estimation technique to analyze realistic data, even in incomplete data settings, and to model the dependence directly in a low dimensional setting.

Their tractability allows to derive explicit expression of properties of interest. Willmot and Woo (2015) have paved the way for applying MME in insurance loss modeling, survival analysis and ruin theory. When modeling insurance losses or dependent risks from different portfolios or lines of business using MME, the aggregate and excess losses have again a univariate and multivariate mixture of Erlangs distribution. Stop-loss moments, several types of premiums, risk capital allocation based on the Tail-Value-at-Risk (TVaR) or covariance rule for regulatory risk capital requirements (see e.g. Dhaene et al., 2012) have analytical expressions. When modeling bivariate lifetimes and pricing joint-life and last-survivor insurance (see e.g. Frees et al., 1996) using MME, the distribution of the minimum and maximum is again a univariate mixture of Erlangs. Such kind of data are always left truncated and right censored. The extension of the fitting procedure for MME presented in this chapter, allows to take the right censoring into account. Left truncation can only be properly handled when the left truncation points are fixed for each observation. This is however not the case when pricing joint-life and last-survivor insurance since the ages at which policyholders enter a contract vary.

The reduction and adjustment steps of the shape parameters in the fitting procedure iteratively make use of the EM algorithm and can be time consuming. Further adjustment is needed to estimate parameters in high dimensional settings. As also acknowledged in the univariate case (Verbelen et al., 2015), the modeling of heavy-tailed distributions using MME is challenging since MME are not able to extrapolate the heaviness in the tail.

3.7 Appendix: Partial derivative of Q

In order to maximize $Q(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(k-1)})$ with respect to θ, we set the first order partial derivative at $\theta^{(k)}$ equal to zero. In the second equation, expression (3.2) of the cumulative distribution of an Erlang, while (3.14) is used to obtain the third equation.

$$
\begin{aligned}
&\left.\frac{\partial Q(\boldsymbol{\Theta};\boldsymbol{\Theta}^{(k-1)})}{\partial\theta}\right|_{\theta=\theta^{(k)}} \\
&\quad= \sum_{i=1}^{n}\sum_{\boldsymbol{r}\in\mathcal{R}} z_{i\boldsymbol{r}}^{(k)} \left(\frac{\sum_{j=1}^{d} E\left(X_{ij}\,\middle|Z_{i\boldsymbol{r}}=1, l_{ij}, u_{ij}, t_j^l, t_j^u; \theta^{(k-1)}\right)}{\theta^2} \right. \\
&\qquad \left. -\frac{\sum_{j=1}^{d} r_j}{\theta} - \sum_{j=1}^{d} \frac{\frac{\partial}{\partial\theta}\left[F(t_j^u; r_j, \theta) - F(t_j^l; r_j, \theta)\right]}{F(t_j^u; r_j, \theta) - F(t_j^l; r_j, \theta)} \right)\Bigg|_{\theta=\theta^{(k)}} \\
&\quad= \frac{1}{\theta^2}\sum_{i=1}^{n}\sum_{\boldsymbol{r}\in\mathcal{R}} z_{i\boldsymbol{r}}^{(k)} \sum_{j=1}^{d} E\left(X_{ij}\,\middle|Z_{i\boldsymbol{r}}=1, l_{ij}, u_{ij}, t_j^l, t_j^u; \theta^{(k-1)}\right) \\
&\qquad -\frac{n}{\theta}\sum_{\boldsymbol{r}\in\mathcal{R}}\left(\frac{\sum_{i=1}^{n} z_{i\boldsymbol{r}}^{(k)}}{n}\right)\sum_{j=1}^{d} r_j \\
&\qquad -\sum_{i=1}^{n}\sum_{\boldsymbol{r}\in\mathcal{R}} z_{i\boldsymbol{r}}^{(k)} \sum_{j=1}^{d} \frac{\frac{\partial}{\partial\theta}\left(\gamma(r_j, t_j^u/\theta) - \gamma(r_j, t_j^l/\theta)\right)}{(r_j-1)!\left(F(t_j^u; r_j, \theta) - F(t_j^l; r_j, \theta)\right)}\Bigg|_{\theta=\theta^{(k)}} \\
&\quad= \frac{1}{\theta^2}\sum_{i=1}^{n}\sum_{\boldsymbol{r}\in\mathcal{R}} z_{i\boldsymbol{r}}^{(k)} \sum_{j=1}^{d} E\left(X_{ij}\,\middle|Z_{i\boldsymbol{r}}=1, l_{ij}, u_{ij}, t_j^l, t_j^u; \theta^{(k-1)}\right) \\
&\qquad -\frac{n}{\theta}\sum_{\boldsymbol{r}\in\mathcal{R}}\beta_{\boldsymbol{r}}^{(k)}\sum_{j=1}^{d} r_j \\
&\qquad -\frac{n}{\theta^2}\sum_{\boldsymbol{r}\in\mathcal{R}}\beta_{\boldsymbol{r}}^{(k)} \sum_{j=1}^{d} \frac{\left(t_j^l\right)^{r_j} e^{-t_j^l/\theta} - \left(t_j^u\right)^{r_j} e^{-t_j^u/\theta}}{\theta^{r_j-1}(r_j-1)!\left(F(t_j^u; r_j, \theta) - F(t_j^l; r_j, \theta)\right)}\Bigg|_{\theta=\theta^{(k)}} = 0\,,
\end{aligned}
$$

where we used expression (3.2) of the cumulative distribution of an Erlang in the second equality and (3.14) in the third.

Chapter 4

Unraveling the predictive power of telematics data in car insurance pricing

Abstract

A data set from a Belgian telematics product aimed at young drivers is used to identify how car insurance premiums can be designed based on the telematics data collected by a black box installed in the vehicle. In traditional pricing models for car insurance, the premium depends on self-reported rating variables (e.g. age, postal code) which capture characteristics of the policy(holder) and the insured vehicle and are often only indirectly related to the accident risk. Using telematics technology enables tailor-made car insurance pricing based on the driving behavior of the policyholder. We develop a statistical modeling approach using generalized additive models and compositional predictors to quantify and interpret the effect of telematics variables on the expected claim frequency. We find that such variables increase the predictive power and render the use of gender as a discriminating rating variable redundant.

This chapter is based on Verbelen, R., Antonio, K., and Claeskens, G. (2016). Unraveling the predictive power of telematics data in car insurance pricing. *FEB Research Report* KBI 1624.

4.1 Introduction

For a unique Belgian portfolio of young drivers in the period between 2010 and 2014, telematics data on how many kilometers are driven, during which time slots and on which type of roads were collected using black box devices installed in the insureds' cars. The aim in this chapter is to incorporate this information in statistical rating models, where we focus on predicting the number of claims, in order to adequately set premium levels based on individual policyholder's driving habits.

Determining a fair and correct price for an insurance product (also called *ratemaking*, *pricing* or *tarification*) is crucial for both insureds and insurance companies. Pricing through risk classification or segmentation is the mechanism insurance companies use to compete and to reduce the price of insurance contracts. Insurance Europe, the European insurance and reinsurance federation, reports[1] a total motor premium income amounting to €124 billion in 2014. Car insurance is the most widely purchased non-life insurance product in Europe, accounting for 27.3% of non-life premiums. To avoid lapses in this competitive market many rating factors are used to classify risks and differentiate prices. Besides the fierce competition, high acquisition and retention costs, low customer engagement, no brand loyalty and a high cost of retention have put a huge pressure on the car insurance industry. Car insurance is traditionally priced based on self-reported information from the insured, most importantly: age, license age, postal code, engine power, use of the vehicle, and claims history. However, these observable risk factors are only proxy variables, not reflecting present patterns of driving habits and the driving style, and consequently tariff cells are still quite heterogeneous.

Telematics technology – the integrated use of telecommunication and informatics – may fundamentally change the car insurance industry. The use of this technology in insured vehicles enables to transmit and receive information that allows an insurance company to better quantify the accident risk of drivers and adjust the premiums accordingly through usage-based insurance (UBI). By monitoring their customers' motoring habits, underwriters can increasingly distinguish between drivers who are safe on the road from those who merely seem safe on paper.[2] Young drivers and drivers in other high risk groups, who are typically facing hefty insurance premiums, can be judged based on how they really drive. Regulation also plays a role as the use of indirect indicators of risk is being questioned

[1] `http://www.insuranceeurope.eu/european-motor-insurance-markets-addendum`
[2] How's my driving? (2013, February 23) *The Economist.* *http://econ.st/Yd5x3C*

by the European Court of Justice. In 2012, a European Union (EU) ruling came into force, banning price differentiation based on gender.[3] Through telematics, women may be able to confirm that they really are safer drivers.

The use of telematics risk factors potentially enables an improved method for determining the cost of insurance. Due to a more refined customer segmentation and greater monitoring of the driving behavior, UBI addresses the problems of adverse selection and moral hazard that arise from the information asymmetry between the insurer and the policyholders (Filipova-Neumann and Welzel, 2010). Closer aligning insurance policies to the actual risks increases actuarial fairness and reduces cross-subsidization compared to grouping the drivers into too general actuarial classes (Desyllas and Sako, 2013). In addition, some positive externalities are to be expected (Parry, 2005; Litman, 2015; Tselentis et al., 2016). Telematics insurance gives a high incentive to change the current driving pattern and stimulates more responsible driving. Users' feedback on driving behavior and gamification of UBI can further enhance the customer experience by making it more interactive, gratifying and even exciting (Toledo et al., 2008). Less and safer driving is encouraged, leading to improved road safety and reduced vehicle travel with less congestion, pollution, fuel consumption, road cost, and crashes (Greenberg, 2009).

Usage-based insurance includes *Pay-as-you-drive* (PAYD) and *pay-how-you-drive* (PHYD) schemes (Tselentis et al., 2016). PAYD focuses on the driving habits, e.g. the driven distance, the time of day, how long the insured has been driving, and the location. PHYD goes even further by also considering the driving style, e.g. the speed, harsh or smooth braking, aggressive acceleration or deceleration, cornering and parking skills. Furthermore, the telematics data collected can be enriched using other sources of data, for example road maps with corresponding speed limitations to infer road types and speeding violations.

Telematics insurance started as a niche market when the technology first surfaced more than 10 years ago. The high implementation costs and its complexity limited its success. Advances in technology and telecommunication have however reduced the cost substantially. Early adopters of UBI were seen primarily in the United States (US), Italy and the United Kingdom (UK) due to the higher premiums, particularly for young drivers, the highly competitive markets, and a higher incidence of fraudulent claims and vehicle theft. Monti's decree of 2012[4],

[3] `http://europa.eu/rapid/press-release_IP-11-1581_en.htm`

[4] Law Decree of 24 January 2012, n.1 "Urgent provisions for competition, infrastructure development and competitiveness" (the so-called "Cresci Italia"), converted by law 24 March 2012, n.27.

encouraging Italian insurers to provide a telematics option, has made Italy the most active country in Europe in telematics insurance, with the overall penetration level around 15% in June 2016.[5] Ptolemus further reports that at that moment insurance companies have launched 292 telematics programs or active trials worldwide (see Husnjak et al., 2015, for some examples of UBI solutions implemented worldwide). The number of UBI policies is over 7.9 million in the US, over 5 million in Italy and over 860 000 in the UK.[6] Moreover, on 28 April 2015 the European Parliament voted in favor of eCall regulation which forces all new cars in the EU from April 2018 onwards to be equipped with a telematics device that will automatically dial 112 in the event of an accident, providing precise location and impact data.[7] However, legislation also gives rise to legal concerns and challenges in the telematics insurance market. In particular, insurers have to comply with the aspects of data protection and privacy in the evolving legal environment.

This potentially high dimensional telematics data, collected on the fly, forces pricing actuaries to change their current practice, both from a business as well as a statistical point of view. New statistical models have to be developed to adequately set premiums based on an individual policyholder's driving habits and style and the current literature on insurance rating does not adequately address this question. In this chapter, we take a first step in this direction. We use a Belgian telematics insurance data set with in total over 297 million kilometers driven. Based on how many kilometers the insured drives, on which kind of roads and during which moments in the day, we quantify the impact of individual driving habits on expected claim frequencies. Combined with a similar predictive model for claim severities, which is outside of the scope in this chapter, this allows for tailor-made car insurance pricing. We first discuss how a car insurance policy is traditionally priced and relate this to the literature investigating the impact of vehicle usage on the accident risk in Section 4.2. The data set is described in Section 4.3, along with the necessary preliminary data processing steps to combine the telematics information with the policy and claims information. By constructing predictive models for the claim frequency, we compare the performance of different sets of predictor variables (e.g. traditional vs. purely telematics) and unravel the relevance and impact of adding telematics insights. In particular, we

[5] `http://www.ptolemus.com/ubi-study/telematics-insurance-infographic/`
[6] Ptolemus Consulting Group (2016). Usage-based insurance (global study), free abstract.
[7] Regulation (EU) 2015/758 of the European Parliament and of the Council of 29 April 2015 concerning type-approval requirements for the deployment of the eCall in-vehicle system based on the 112 service and amending Directive 2007/46/EC.

contrast the use of time and distance as exposure-to-risk measures. The statistical methodology, including in particular the challenges when incorporating the divisions of the driven distance by road type and time slots as predictors in the model, is presented in Section 4.4. In Section 4.5, we present the results and, finally, in Section 4.6, we conclude.

4.2 Statistical background and related modeling literature

Insurance pricing is the calculation of a fair premium, given the policy(holder) characteristics, as well as information on claims reported in the past (if available). The pure premium represents the expected cost of the claims a policyholder will declare during the insured period. Pricing relies on regression techniques and requires a data set with policy(holder) information and corresponding claim frequencies and severities, where severity is the ultimate total impact of a claim.

A priori pricing refers to the statistical problem of pricing without incorporating the claim history of the policyholder, thus neither frequency nor severity of past claims is taken into account. The construction of an a priori tariff traditionally relies on a frequency-severity modeling framework in which the claim frequency and severity components are typically modeled separately using regression techniques (Frees, 2014). A policyholder's pure premium is obtained by multiplying the expected claim frequency and expected claim severity, given the observable risk factors. The current state-of-the-art (see Denuit et al., 2007; de Jong and Heller, 2008, for an overview) uses generalized linear models (GLMs; McCullagh and Nelder, 1989), with typically a Poisson GLM for the claim counts and a gamma GLM for the claim severities. Modeling the claim severities is difficult, since only those observations corresponding to policyholders who filed a claim can be used to estimate the claim severity model and due to the complexity of the phenomenon (Denuit and Charpentier, 2005). On the one hand, there is a long delay to assess the cost of bodily injury and other severe claims and on the other hand the cost of an accident is, for most part, beyond the control of the driver. In practice, covariates are much less informative to predict claim amounts than to predict frequency (Boucher and Charpentier, 2014).

A posteriori pricing refers to experience rating systems which penalize or reward policyholders based on (usually) the number of claims reported in the past. The idea is that, over time, insurers try to refine their a priori risk classification

and restore fairness using no-claim discounts and claim penalties. A bonus-malus system is a typical example (Lemaire, 1995). A bonus-malus scale consists a finite number of levels, each with its own relativity that is applied to the base premium. Transitional rules determine how a policy moves up or down based on the number of claims at fault. As such, on the basis of the insured's individual claim experience, the amount of the premium is adjusted each year with penalties in the form of premium surcharges (corresponding to higher bonus-malus levels) for one or more accidents in the current year and rewards in the form of premium discounts (corresponding to lower bonus-malus levels) for claim-free policyholders. From a statistical point of view a posteriori rating requires the analysis of multilevel data (Gelman and Hill, 2007).

In car insurance, the duration of the policy period during which coverage is provided, is referred to as the exposure-to-risk, the basic rating unit underlying the insurance premium. The expected number of claims is in practice modeled directly proportional to the exposure. The logic behind this is to make the premiums proportional to the length of coverage. As such, a premium related to an insured period of 6 months will be half of the one-year premium, for a given risk profile. From a theoretical point of view, this can also be motivated by the probabilistic framework of Poisson processes (Denuit et al., 2007). It is however suggested (see e.g. Butler, 1993) that every kilometer traveled by a vehicle transfers risk to its insurer and hence the number of driven kilometers (*car-kilometer*) should be adopted as the exposure unit instead of the policy duration (*car-year*). Statistical studies show how claim frequencies significantly increase with kilometers (Bordoff and Noel, 2008; Ferreira and Minikel, 2010; Litman, 2011; Boucher et al., 2013; Lemaire et al., 2016). Most of these studies show a relationship between claim frequencies and the number of driven kilometers which is less than proportional. They suggest that possibly high-kilometer drivers are more experienced, have newer and safer vehicles, or drive more on low-risk motorways rather than high-risk urban areas.

Data collected using telematics technology offers more insight in the driving habits. Instead of relying only on the self-reported annual number of driven kilometers, pay-as-you-drive insurance can also account for the type of road and the time of the day when an insured has been driving. A next step is to also take data on driving style into account, leading to a pay-how-you-drive insurance (Weiss and Smollik, 2012). Statistical analysis of these types of data has been the subject of limited academic scrutiny.

Ayuso et al. (2014, 2016a) study the traveled time and distance to the first

accident using Weibull regression models involving both policy and telematics predictors. Paefgen et al. (2014) investigate the relationship between the accident risk and driving habits using logistic regression models. Their case-control study design does not allow for inference on the probability of accident involvement. The difference in time exposure between the vehicles with accident involvement (6 months prior to the accident) and the control group (24 months) is however only used to obtain a per-month distance exposure, but is further neglected in the study. Traditional risk factors were not accounted for, since that information was not available, and the compositional nature of the constructed telematics predictor variables was ignored. In contrast, combining the new telematics variables with traditional policy(holder) information through a careful model and variable selection process as well as recognizing the compositional structure in the analysis are main focus points in our research, see Section 4.3.2.

4.3 Telematics insurance data

We consider data from a Belgian portfolio of drivers with motor third party liability (MTPL) insurance. MTPL insurance is the legally compulsory minimum insurance covering damage to third parties' health and property caused by an accident for which the driver of the vehicle is responsible. The special type of MTPL product we are considering, is specifically aiming for young drivers who are traditionally facing high insurance premiums. Insureds were offered a substantial discount on their premium if they agree to install a telematics black box device in their car. The telematics box collects statistics on the driving habits: how often one drives, how many kilometers, where and when. Information on the driving style (such as speeding, braking, accelerating, cornering or parking) is not registered. The telematics data have so far no effect on the (future) premium levels of the insureds and do not induce any restrictions on how much or where they can drive.

4.3.1 Data processing

The unstructured telematics data, collected by the telematics box installed in the vehicle, are first transmitted to the data provider who structures and aggregates these data each day and then reports them to the insurance company as a CSV file (Figure 4.1a). Only the structured, aggregated telematics information is available to us. Each daily file contains information on the daily driven distance (in

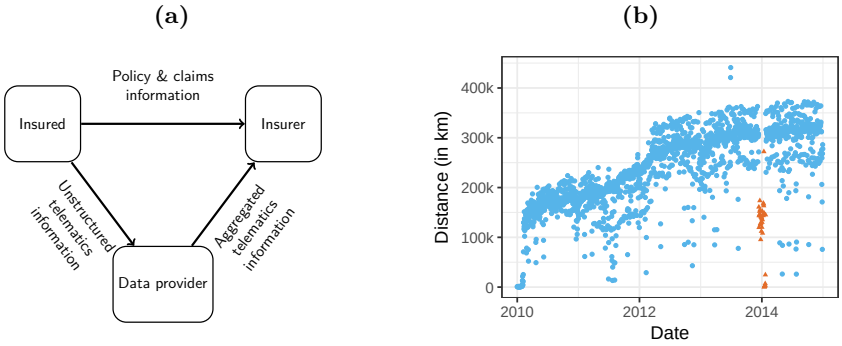

Figure 4.1: *(a) A schematic overview of the flow of information. (b) The number of registered kilometers on each day on an aggregate, portfolio level for the telematics data observed between January 1, 2010 and December 31, 2014. The outliers by the turn of the year 2014, corresponding to a technical malfunction, are indicated as triangles.*

meters) for each policyholder. This number of meters is split into 4 road types (*urban areas, other, motorways* and *abroad*) and 5 time slots (*6h-9h30, 9h30-16h, 16h-19h, 19h-22h* and *22h-6h*). The nature of the data does not allow for a classification of a driven meter by road type and time slot simultaneously. The number of trips, measured as key-on/key-off events, is also reported. This is a typical setup (see Paefgen et al., 2014). In this study, we analyze the telematics data collected between January 1, 2010 and December 31, 2014.

The telematics data are linked with the policy(holder) and claims information of the insurance company corresponding to the portfolio under consideration (see Table 4.1 for a complete list). Policy data, such as age, gender and characteristics of the car, are directly reported by the insured to the insurer at underwriting (see Figure 4.1a). They are updated over time which enables us to link the claims occurring at a specific moment in time to the correct policy information. Each observation of a policyholder in the policy data set refers to a policy period over which the MTPL insurance coverage holds and contains the most recent policy information. For most insureds, this coverage period is one year, however, it can be smaller for several reasons. If for instance the policyholder decides to add a comprehensive coverage, buys a new vehicle, or changes his residence during the term of the contract, the policy period will be restricted to that date and an additional observation line will be added for the subsequent period. A policy period can also be split when the coverage is suspended for a certain time.

Using the policy number and period we first merge the telematics information

Table 4.1: *Description of the variables contained in the data set arising from the different sources of information.*

Claims information	
`claims`	number of reported MTPL claims at fault during the policy period
Policy information	
`policy period`	duration in days of the policy period (minimal 30 days and at most one year)
`age`	age of the least experienced driver listed on the policy at the start of the policy period, measured as the number of years between the birth date and the start of the policy period
`experience`	experience of the least experienced driver listed on the policy, measured as the number of years between the date when the driver's permit was obtained and the start of the policy period
`gender`	gender of the least experienced driver listed on the policy (*male* or *female*)
`material damage cover`	indicator whether the insurance policy also covers material damage (*yes* or *no*)
`postal code`	Belgian postal code where the policyholder resides
`bonus-malus`	bonus-malus level of the policy, reflecting the past individual claims experience, between -4 and 22 with lower values indicating a better history
`age vehicle`	age of the vehicle, measured as the number of years between the date when the car was registered and the start of the policy period
`kwatt`	horsepower of the vehicle, measured in kilowatt
`fuel`	fuel type of the vehicle (*petrol* or *diesel*)
Telematics information	
`distance`	distance in meters driven during the policy period
`yearly distance`	distance in meters driven during the policy period, rescaled to a full year by dividing by duration in days of the policy period and multiplying by 365
`trips`	number of trips (*key-on, key-off*) during the policy period
`average distance`	distance in meters driven on average during one trip, obtained by dividing the distance by the number of trips
`road type`	division of the `distance` into 4 road types (*motorways, urban areas, abroad* and *other*)
`time slot`	division of the `distance` into 5 time slots (*22h-6h, 6h-9h30, 9h30-16h, 16h-19h* and *19h-22h*)
`week/weekend`	division of `distance` into *week* (Monday to Friday) and *weekend* (Saturday, Sunday)

on daily level with the policy data set. Next, we adjust the start and end date of the policy periods based on the first and last day at which telematics data are observed for each policy period of each insured. This ensures that the adjusted policy periods reflect time periods over which both the insurance coverage holds and telematics data are collected. Based on Figure 4.1b, where we plot the evolution of the driven distance on each day by all drivers of the portfolio, we suspect that technical deficiencies of the data provider can cause an underreporting of the number of meters driven on an aggregate level. The outliers indicated as triangles by the turn of the year 2014 could be linked to a serious technical failure preventing telematics information from being reported for a significant part of our portfolio. We dealt with this by removing this period of roughly one month from the policy periods of all insureds. In the remainder of the observation period between January 1, 2010 and December 31, 2014, clear causes of underreporting could not be identified and hence we did not take any other corrective action. However, this illustrates that data reliability forms a challenge for this new telematics technology. We further removed those observations with a policy duration of less than 30 days in order to avoid senseless observations of only a couple of days and retained only the complete observations with no missing policyholder information.

Next, we aggregate the telematics information by policyholder and period. This means that we sum the driven distance, their divisions into 4 road types and 5 time slots, and the number of trips made. Finally, we use the claims information to link the number of MTPL claims at fault that occurred between the start and end date of the adjusted policy periods for each policy record.

Over the time period of this study, we end up with a data set of 33 259 observations. Table 4.1 gives an overview of the available variables coming from the three data sources (claims, policy, and telematics). These observations correspond to 10 406 unique policyholders, who are followed over time, have jointly driven over 297 million kilometers during a combined insured policy period of 17 681 years and reported 1481 MTPL claims at fault. Hence, on average, there were 0.0838 claims per insured year or 0.0499 claims per 10 000 driven kilometers. For over 95% of the observations no claim occurred during the corresponding policy period, whereas for 52 observations two claims occurred and for a single observation even three during the same policy period.

4.3.2 Risk classification using policy and telematics information

The goal of this research is to build a rating model to express the number of claims as a function of the available covariates. Two sources of information are combined which are described in detail in Table 4.1. First, there is the self-reported policy information which contains all rating variables traditionally used in car insurance pricing. The second source of information is derived from the telematics data. The main objective is to discover the relevance and impact of adding the new telematics insights using flexible statistical modeling techniques in combination with appropriate model and variable selection tools. One of the key questions is whether the amount of risk transferred from the policyholder to the insurer is proportional to the duration of the policy period or the driven distance during that time. Telematics technology allows a shift to be made from time as exposure to distance as exposure. This would lead to a form of pay-as-you-drive insurance, where a driver pays for every kilometer driven. Histograms of both potential exposure variables are contrasted in Figure 4.2a and 4.2b.

In order to investigate the influence and explanatory power of the telematics variables in predicting the risk of an accident, we compare the performance of four sets of predictor variables used to model the number of claims, see Figure 4.2c. The *classic* set only contains policy information and uses time as exposure-to-risk. The *telematics* set only contains telematics information and uses the distance in meters as exposure-to-risk. The two other models, *time hybrid* and *meter-hybrid*, both contain policy and telematics information. Whereas the first one uses time as an exposure measure, the second one uses distance. These four predictor sets contrast on the one hand the use of traditional policy rating variables and telematics variables and on the other hand the use of policy duration as exposure and the use of distance as exposure in the assessment of the risk.

The main predictors based on the policy information besides the duration of the policy period include the age of the driver, the experience as measured using the driver's license age, the gender, characteristics of the car and the postal code where the policyholder lives. In the case of multiple insured drivers (around 18% of the observations), we select (in consultation with the insurer) the age, gender, experience and postal code belonging to the driver with the most recent permit and hence the lowest experience. This is in line with the strategy of the insurer who offers this type of insurance contract to young drivers. The bonus-malus level is a special kind of variable that reflects the past individual claims experience.

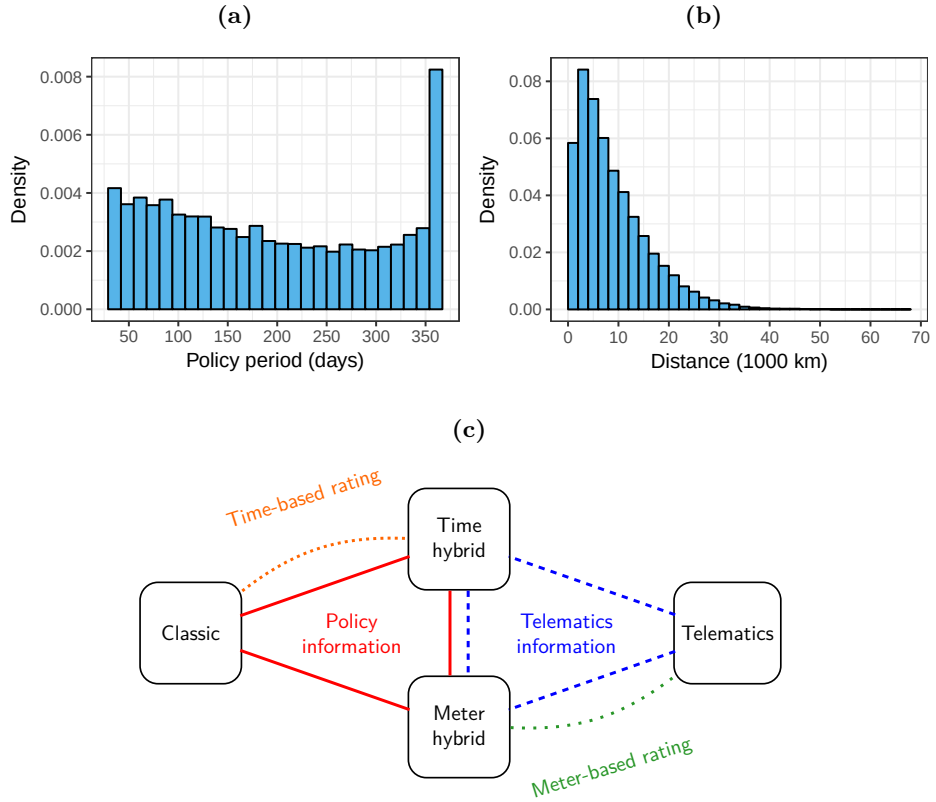

Figure 4.2: *Histogram of (a) the duration (in days) of the policy period (at most one year) and (b) the driven distance (in 1000 km) during the policy period. (c) A graphical representation of the similarities and differences between the four predictor sets.*

It is a function of the number of claims reported in previous years with values between -4 and 22 where lower levels indicate a better history. The insurer uses a slightly modified version of the former compulsory Belgian bonus-malus system, which all companies operating in Belgium have been obliged to use from 1992 to 2002, with minor refinements for the policyholders occupying the lowest levels in the scale. Despite the deregulation, many insurers in the Belgian market still apply the former mandatory system (Denuit et al., 2007). Even though, the bonus-malus scale level is not a covariate of the same type as the other a priori variables, we keep it in the analysis to have an idea of the information contained in this variable (as is also done in, for instance, Denuit and Lang, 2004). From a statistical point of view, it tries to structure dependencies between observations

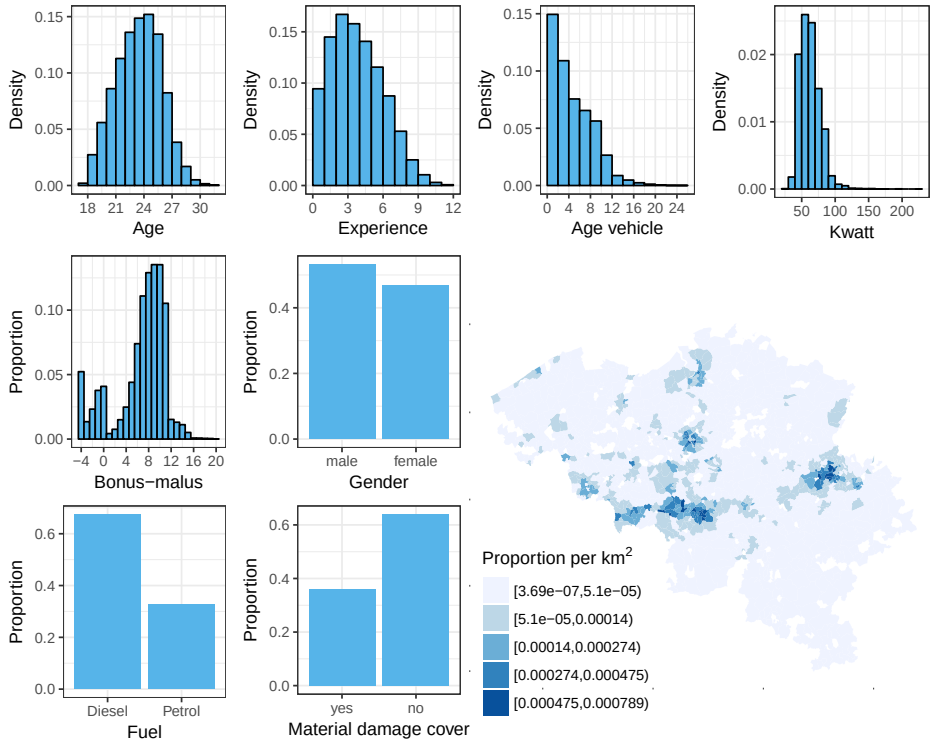

Figure 4.3: *Histograms and bar plots of the continuous and categorical policy variables contained in the data set. The map in the lower right depicts the geographical information by showing the proportion of insureds per squared kilometer living in each of the different postal codes in Belgium. The five class intervals have been created using k-means clustering.*

arising from the same policyholder. An overview of the policy predictor variables and their sample distributions is given in Figure 4.3.

In the telematics information set we use the driven distance during the policy period as a predictor, but we also create two additional telematics variables, the yearly and average distance driven, see Table 4.1. Histograms of these variables are shown in Figure 4.4. The divisions of the driven distance by time slot, road type and week/weekend are highly correlated with the total driven distance as they sum up to this amount. To distinguish the absolute information measured by the driven distance in a certain policy period from the compositional information of the distance split into different categories, we consider box plots of the relative

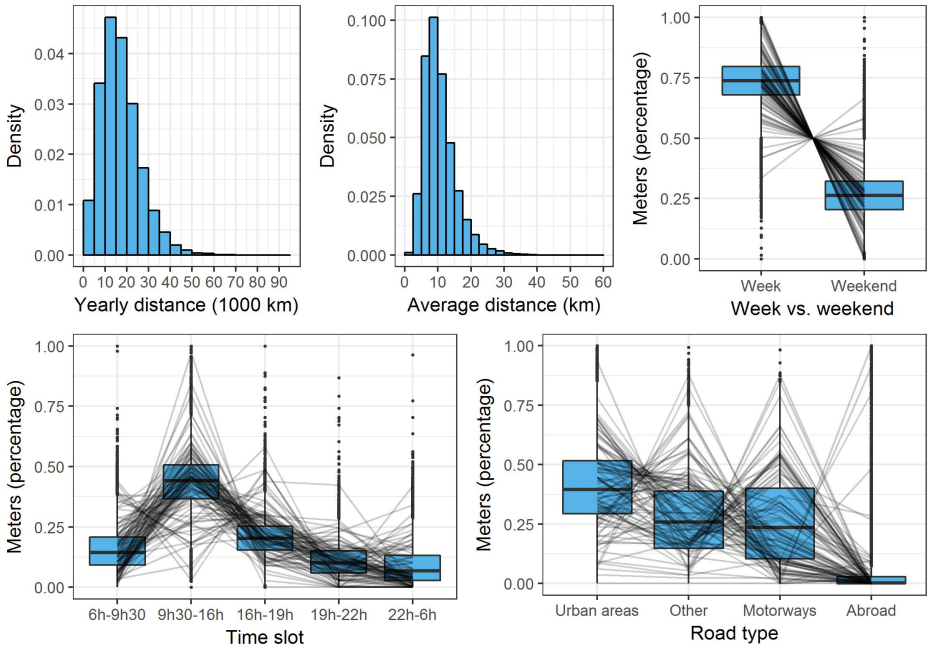

Figure 4.4: *Graphical illustration of the telematics variables contained in the data set. For the yearly and average distance, we construct histograms. For the division of the driven distance by road types, time slots and week/weekend, we construct box plots of the relative proportions. To highlight the dependencies intrinsic to the fact that the division in different categories sums to one, we plot profile lines for 100 randomly selected observations in the data set.*

proportions in Figure 4.4. These relative proportions sum to one for each observation in our data set. To stress this interconnectedness present in the different splits, we show the compositional profiles of a sample of 100 drivers on top of the marginal box plots. Another important point to stress is that not all components of a certain division of the distance are present for each observation. For instance, if an insured does not drive abroad during the policy period, the relative proportion of the driven distance abroad will be zero. The use of such compositional information as predictors in statistical modeling is another key issue in this research.

4.4 Model building and selection

We model the frequencies of claims by constructing Poisson and negative binomial (NB) regression models. We denote by N_{it} the number of claims for policyholder i in policy period t with $i = 1, \ldots, I$ and $t = 1, \ldots, T_i$. The model is denoted by $N_{it} \sim \text{Poisson}(\mu_{it})$ or $N_{it} \sim \text{NB}(\mu_{it}, \phi)$, where $\mu_{it} = \mathbb{E}(N_{it})$ represents the expected number of claims reported by policyholder i in policy period t and ϕ is the parameter of the NB distribution such that $\text{Var}(N_{it}) = \mu_{it} + \mu_{it}^2/\phi$, allowing for overdisperion. A log linear relationship between the mean and the predictor variables is specified by the log link function. This means that we set $\mu_{it} = \exp(\eta_{it})$ where η_{it} is a predictor function of the available explanatory factors. The probability mass functions for the Poisson and the NB models are, respectively, expressed as

$$\mathbb{P}(N_{it} = n_{it}) = \frac{\exp(-\mu_{it})\mu_{it}^{n_{it}}}{n_{it}!} \quad \text{and} \quad \mathbb{P}(N_{it} = n_{it}) = \frac{\Gamma(\phi + n_{it})}{n_{it}!\Gamma(\phi)} \frac{\phi^{\phi}\mu_{it}^{n_{it}}}{(\phi + \mu_{it})^{\phi + n_{it}}}\,.$$

For each of the predictor sets in Figure 4.2c we construct the best model using the allowed information based on AIC, see Section 4.4.3. Additionally, we identify the best models under the restriction that the risk is proportional to the time or meter exposure. This is accomplished by incorporating the logarithm of the exposure-to-risk, either duration of the policy period or total distance driven during the policy period, as an offset term in the predictor, i.e. a regression variable with a constant coefficient of 1 for each observation. In the most general case, the predictor has the form

$$\eta_{it} = \beta_0 + \text{offset} + \eta_{it}^{\text{cat}} + \eta_{it}^{\text{cont}} + \eta_{it}^{\text{spatial}} + \eta_{it}^{\text{re}} + \eta_{it}^{\text{comp}}\,, \tag{4.1}$$

where β_0 denotes the intercept, the categorical effects are bundled in η_{it}^{cat}, the term η_{it}^{cont} contains the effects of the continuous predictors, $\eta_{it}^{\text{spatial}}$ represents the geographical effect, η_{it}^{re} the policyholder-specific random effect and the term η_{it}^{comp} embodies the effects of the compositional predictors. Under the offset restriction, the continuous effect of the exposure-to-risk, either the duration of the policy period (time based rating) or the driven distance (meter based rating), gets replaced by the logarithm of the exposure-to-risk as an offset.

Zero inflated variants of these models could also be considered but is not done here for interpretability reasons since they are not able to capture the effect of a varying exposure-to-risk in a transparent and intuitive way.

4.4.1 Generalized additive models

The model framework we work with in this study is the one of generalized additive models (GAMs), introduced by Hastie and Tibshirani (1986). GAMs allow to incorporate continuous covariates in a more flexible way as compared to the traditional GLMs used in actuarial practice (see e.g. Klein et al., 2014). From an accuracy standpoint, GAMs are competitive with popular black box machine learning techniques (such as neural networks, random forests or support vector machines), but they have the important advantage of interpretability. In insurance pricing it is of crucial importance to have interpretable results in order to understand the premium structure and explain this to clients and regulators. Using a semiparametric additive structure, GAMs define nonparametric relationships between the response and the continuous predictors in the predictor in the following way

$$\eta_{it}^{\text{cat}} + \eta_{it}^{\text{cont}} = \boldsymbol{Z}_{it}\boldsymbol{\beta} + \sum_{j=1}^{J} f_j(x_{jit})\,,$$

where $\boldsymbol{Z}_{it}$ represents the row corresponding to policyholder i in policy period t of the model matrix of parametric terms for the categorical predictors with parameter vector $\boldsymbol{\beta}$ and f_j represents a smooth function of the jth continuous predictor variable. To estimate f_j, we choose cubic spline basis functions B_{jk}, such that f_j can be represented as $f_j(x) = \sum_{k=1}^{q} \gamma_{jk} B_{jk}(x)$. The knots are chosen using 10 quantiles of the unique x_j values. Cardinal basis functions parametrize the spline in terms of its values at the knots (Lancaster and Salkauskas, 1986). For identifiability, we impose constraints by centering each smooth component around zero, thus $\sum_{i=1}^{I}\sum_{t=1}^{T_i} f_j(x_{jit}) = 0$ for $j = 1, \ldots, J$. To avoid overfitting, the cubic splines are penalized by the integrated squared second derivative (Green and Silverman, 1994), which yields a measure for the overall curvature of the function. For each component, this penalty can be written as a quadratic function,

$$\int \left(f_j''(x)\right)^2 dx = \sum_{k=1}^{q}\sum_{l=1}^{q} \gamma_{jk}\gamma_{jl} \int B_{jk}''(x) B_{jl}''(x) dx = \boldsymbol{\gamma}_j^t \boldsymbol{S}_j \boldsymbol{\gamma}_j\,,$$

with $(\boldsymbol{S}_j)_{kl} = \int B_{jk}''(x) B_{jl}''(x) dx$. Given these penalty functions for each component, we define the penalized log-likelihood as

$$\ell(\boldsymbol{\psi}) - \frac{1}{2}\sum_{j=1}^{J} \lambda_j \boldsymbol{\gamma}_j^t \boldsymbol{S}_j \boldsymbol{\gamma}_j\,, \tag{4.2}$$

where $\ell(\boldsymbol{\psi})$ denotes the log-likelihood as a function of all model parameters $\boldsymbol{\psi} = (\boldsymbol{\beta}, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_J)^t$ and λ_j denotes the smoothness parameter that controls the tradeoff between goodness of fit and the degree of smoothness of component f_j for $j = 1, \ldots, J$. Different smoothing parameters for each component allow to penalize the smooth functions differently.

The model parameters $\boldsymbol{\psi}$ are estimated by maximizing (4.2) using penalized iteratively reweighted least squares (P–IRLS) (Wood, 2006). For the Poisson model, the smoothing parameters $\lambda_1, \ldots, \lambda_J$ are estimated using an unbiased risk estimator criterion (UBRE), which is a rescaled version of Akaike's information criterion (AIC; Akaike, 1974). For the negative binomial model, we estimate the smoothing parameters and the scale parameter ϕ using maximum likelihood (ML).

In addition to categorical and continuous covariates, the data set contains spatial information, namely the postal code where the policyholder resides. Insurance companies tend to use the geographical information of the insured's residence as a proxy for the traffic density and for other unobserved socio-demographic factors of the neighborhood. We model the spatial heterogeneity of claim frequencies by adding a spatial term $\eta_{it}^{\text{spatial}} = f_s(\texttt{lat}_{it}, \texttt{long}_{it})$ in the additive predictor η_{it}, using the latitude and longitude coordinates (in degrees) of the center of the postal code where the policyholder resides. We use second order smoothing splines on the sphere (Wahba, 1981) to model f_s. This allows us to quantify the effect of the geographic location while taking the regional closeness of the neighboring postal codes into account.

In our data set, many policyholders $i = 1, \ldots, I$ are observed over multiple policy periods $t = 1, \ldots, T_i$. This longitudinal aspect of the data can be modeled by including policyholder-specific random effects η_{it}^{re} in the predictor. The generalized additive model considered thus far is extended in this way by exploiting the link between penalized estimation and random effects (see e.g. Ruppert et al., 2003). We assess whether such random effects are needed to take the correlations between observations of the same policyholder into account using the approximate test for a zero random effect developed by Wood (2013).

4.4.2 Compositional data

The divisions of the total driven distance in the different categories – road types (4), time slots (5) and week/weekend (2), see Table 4.1 – are highly correlated with and sum up to the total driven distance. Incorporating these divisions in

a predictor also containing the total distance leads to a perfect multicollinearity problem. Furthermore, the corresponding model parameter estimators are not invariant to the ordering of the components: the statistical inference changes when permuting the components making interpretations misleading. The standard regression interpretation of a change in one of the components of the distance when the other components are held constant is not possible due to the sum constraint of adding up to the total distance.

The total distance in meters is used as a continuous predictor in the telematics models and its effect is modeled using a smooth function. Since the divisions of the distance only contribute additional relative information, we divide all components of each split by the total driven distance, see Figure 4.4. We obtain what is known as *compositional data* (Van den Boogaart and Tolosana-Delgado, 2013; Pawlowsky-Glahn et al., 2015). Such data are represented by real vectors with constant sum equal to one and positive components. The space of representations of compositions is called the simplex of D parts, denoted $\mathcal{S}^D$, defined by

$$\mathcal{S}^D = \left\{ \boldsymbol{x} = (x_1, \ldots, x_D)^t : x_i > 0, \sum_{i=1}^{D} x_i = 1 \right\}.$$

Only relative information is important, and multiplication of the vector of positive components by a positive constant does not change the ratios between the components. When data are considered compositional, classical statistics, that do not take the special geometry of the simplex into account, are not appropriate. Extending the current literature, we propose a new way of quantifying and interpreting the effect of the compositional explanatory variables on the outcome and propose an approach to deal with structural zeros.

The Aitchison geometry of the simplex

The vector space structure of the mathematical simplex was discovered by Aitchison (1986) who defined operations on compositional data leading to the Aitchison geometry of the simplex. Perturbation plays the role of addition on the simplex and is defined as a closed component-wise product $\boldsymbol{x} \oplus \boldsymbol{y} = \mathcal{C}(x_1 y_1, \ldots, x_D y_D)^t$, where the closing operation $\mathcal{C}$ ensures a total sum of one, i.e. the closure of $\boldsymbol{x}$ is $\mathcal{C}(\boldsymbol{x}) = \boldsymbol{x} / \sum_{i=1}^{D} x_i$. The product of a vector by a scalar is called powering and is defined as $\alpha \odot \boldsymbol{x} = \mathcal{C}(x_1^\alpha, \ldots, x_D^\alpha)^t$, for $\alpha \in \mathbb{R}$. The Aitchison inner product for

compositions is

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_a = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j} = \sum_{i=1}^{D} \ln(x_i) \ln(y_i) - \frac{1}{D} \left(\sum_{i=1}^{D} \ln(x_i) \right) \left(\sum_{j=1}^{D} \ln(y_j) \right)$$

and induces the following norm $\|\boldsymbol{x}\|_a = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle_a}$ and distance $\mathrm{d}_a(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} \ominus \boldsymbol{y}\|_a$, where $\ominus$ represents the opposite operation of $\oplus$, i.e. $\ominus \boldsymbol{y} = \oplus((-1) \odot \boldsymbol{y})$. The simplex along with these operations then forms a $(D-1)$-dimensional Euclidean vector space $(\mathcal{S}^D, \oplus, \odot, \langle \cdot, \cdot \rangle_a)$. Given this Euclidean structure, we can measure distances and angles, and define related geometrical concepts. Elementary statistical notions involving the metrics of the sample space can be adapted to the Euclidean structure of the simplex.

Egozcue et al. (2003) constructed orthonormal bases for this Euclidean space and deduced corresponding isometries between $\mathcal{S}^D$ and $\mathbb{R}^{D-1}$, called isometric logratio transformations (*ilr*). One possible ilr transformation maps a compositional data vector $\boldsymbol{x}$ in a $(D-1)$-dimensional real vector $\boldsymbol{z} = (z_1, z_2, \ldots, z_{D-1})^t$ with components

$$z_i = \mathrm{ilr}_i(\boldsymbol{x}) = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j}}, \quad i = 1, \ldots, D-1\,. \tag{4.3}$$

As the ilr transformation is isometric, all angles and distances are preserved. This means that, whenever compositions are transformed into coordinates, the metrics and operations in the Aitchison geometry of the simplex are translated into the ordinary Euclidean metrics and operations in real space. Let V be the $D \times (D-1)$ matrix with elements

$$V_{ij} = \frac{D-j}{\sqrt{(D-j+1)(D-j)}} \quad \text{for } i = j, \qquad \frac{-1}{\sqrt{(D-j+1)(D-j)}} \quad \text{for } i > j,$$

and 0 otherwise, for which it holds that $V^t V = I_{D-1}$ and $V V^t = I_D - (1/D)\mathbf{1}_D \mathbf{1}_D^t$, where I_D is the identity matrix of dimension D and $\mathbf{1}_D$ a D-vector of ones (Egozcue et al., 2011). Then we can rewrite this ilr transform and its inverse in matrix notation as

$$\boldsymbol{z} = \mathrm{ilr}(\boldsymbol{x}) = V^t \ln \boldsymbol{x}\,, \quad \text{and} \quad \boldsymbol{x} = \mathrm{ilr}^{-1}(\boldsymbol{z}) = \mathcal{C}(\exp(V\boldsymbol{z}))\,, \tag{4.4}$$

where the logarithmic and exponential function apply componentwise.

Even though the simplex $\mathcal{S}^D$ is a subset of the real space $\mathbb{R}^D$, Aitchison (1986) showed that the geometry is clearly different. Ignoring this aspect in a statistical context can lead to incompatible or incoherent results. The compositional nature of the data must not be ignored. The principle of working on coordinates in statistics (Mateu-Figueras et al., 2011) is to first express the compositional data with respect to an orthonormal basis of the underlying vector space with Euclidean structure. Next, to apply standard statistical techniques to the vectors of coordinates and, finally, to back-transform and describe the results in terms of the simplex. Final results do not depend on the chosen basis.

A new interpretation for compositional predictors

In our setting, it is key to incorporate the compositional data arising from the divisions of the distances into different categories as predictors in the claim count regression models. Hron et al. (2012) propose to first apply the isometric log ratio transform (4.3) to map the compositions in the D-part Aitchison simplex to a $(D-1)$ Euclidean space. Then, these terms are used as explanatory variables in a linear regression model. More generally, in any regression context involving a predictor, one can add a compositional predictor term η^{comp} using the ilr transformed variables, i.e.

$$\eta^{\text{comp}} = \beta_1 z_1 + \ldots + \beta_{D-1} z_{D-1}\,. \tag{4.5}$$

The fitted model does not depend on the choice of the orthonormal ilr basis since the coordinates of $\boldsymbol{x}$ with respect to different orthonormal bases are orthogonal transformations of each other. Using the ilr transformation the model parameters can be estimated without constraints and the ceteris paribus interpretation of altering one z_i without altering any other becomes possible. Only the first regression parameter, β_1, however has a comprehensible interpretation since z_1 explains relevant information about x_1. The remaining coefficients are not straightforward to interpret and hence Hron et al. (2012) suggest to permute the indices in formula (4.3) and construct D regression models, each time with a different component first for which we can interpret the corresponding coefficient. Having to refit the model multiple times is undesirable, especially in our case where we have more than one compositional predictor and each model fit is computationally intensive due to smooth continuous, spatial, and random effects. Hence, we develop a new strategy to include compositional predictors and interpret their effect.

By using the inverse ilr transform on the model coefficients, i.e. set $\boldsymbol{b} = \text{ilr}^{-1}(\boldsymbol{\beta})$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{D-1})^t$, we can rewrite the compositional predictor as

$$\eta^{\text{comp}} = \sum_{i=1}^{D-1} \beta_i z_i = \sum_{i=1}^{D-1} \text{ilr}_i(\boldsymbol{b})\text{ilr}_i(\boldsymbol{x}) = \langle \boldsymbol{b}, \boldsymbol{x} \rangle_a \,,$$

since the ilr transform preserves the inner product (Van den Boogaart and Tolosana-Delgado, 2013; Pawlowsky-Glahn et al., 2015). The composition $\boldsymbol{b} \in \mathcal{S}^D$ can be interpreted as the simplicial gradient of η^{comp} with respect to $\boldsymbol{x}$ (Barceló-Vidal et al., 2011) and is the compositional direction along which the predictor increases fastest. In particular, if we increase $\boldsymbol{x}$ to $\tilde{\boldsymbol{x}} = \boldsymbol{x} \oplus \frac{\boldsymbol{b}}{\|\boldsymbol{b}\|_a}$, then the predictor becomes

$$\tilde{\eta}^{\text{comp}} = \langle \boldsymbol{b}, \tilde{\boldsymbol{x}} \rangle_a = \langle \boldsymbol{b}, \boldsymbol{x} \oplus \frac{\boldsymbol{b}}{\|\boldsymbol{b}\|_a} \rangle_a = \langle \boldsymbol{b}, \boldsymbol{x} \rangle_a + \frac{1}{\|\boldsymbol{b}\|_a} \langle \boldsymbol{b}, \boldsymbol{b} \rangle_a = \eta^{\text{comp}} + \|\boldsymbol{b}\|_a \,.$$

When $D = 3$, the estimated regression model can be visualized as a surface on a ternary diagram (Van den Boogaart and Tolosana-Delgado, 2013). For $D > 3$, a graphical representation is not straightforward.

In order to overcome this shortcoming in interpretation and to develop a graphical representation for compositional explanatory variables, we propose to perturb the composition in the direction of each component. This offers a new interpretation for the effect of altering the composition on the predictor. For example, a relative ratio change of $\alpha > 1$ (increase) or $\alpha < 1$ (decrease) in the first component of $\boldsymbol{x}$ with constant ratios of the remaining components can be achieved by perturbing the composition $\boldsymbol{x}$ by $(\alpha, 1, \ldots, 1)^t$. This leads to a change of the predictor given by

$$\langle \boldsymbol{b}, (\alpha, 1, \ldots, 1)^t \rangle_a = \ln(b_1)\ln(\alpha) - \frac{1}{D}\left(\sum_{i=1}^{D} \ln(b_i)\right)\ln(\alpha) = \text{clr}_1(\boldsymbol{b})\ln(\alpha) \,, \quad (4.6)$$

which is independent of the original composition $\boldsymbol{x}$ and where

$$\text{clr}_i(\boldsymbol{b}) = \ln\left(\frac{b_i}{g_m(\boldsymbol{b})}\right), \qquad g_m(\boldsymbol{b}) = \left(\prod_{i=1}^{D} b_i\right)^{1/D}, \qquad i = 1, \ldots, D$$

denotes the centered log-ratio (clr) transform of $\boldsymbol{b}$ (Egozcue et al., 2011). The effect of a relative increase in any of the components can hence best be understood by considering the clr transform of $\boldsymbol{b}$, of which the elements sum to zero and indicate the positive or negative effect of each component on the predictor. A graphical representation of the effect of a compositional predictor

can be made by visualizing $\text{clr}(\boldsymbol{b})$ and comparing the elements to zero. Since $\boldsymbol{\beta} = \text{ilr}(\boldsymbol{b}) = V^t \ln(\boldsymbol{b}) = V^t \text{clr}(\boldsymbol{b})$ and $VV^t = I_D - (1/D)\mathbf{1}_D\mathbf{1}_D^t$, the clr transform of $\boldsymbol{b}$ can be written as $\text{clr}(\boldsymbol{b}) = V\boldsymbol{\beta}$. Confidence bounds can thus be constructed using the corresponding covariance matrix $V\widehat{\Sigma}V^t$ where $\widehat{\Sigma}$ is the estimated covariance matrix related to estimating $\boldsymbol{\beta}$. To interpret the effect on the level of the expected outcome in the Poisson and NB models, we can transform these confidence intervals using the exponential function. The exponentiated clr transform of $\boldsymbol{b}$ has to be compared to one and the effect of a relative ratio change of α in component $i = 1, \ldots, D$ is given by $\alpha^{\text{clr}_i(\boldsymbol{b})}$.

Dealing with structural zeros in compositional predictors

An additional difficulty when incorporating the compositional information as predictors in the analysis of the claim counts is the presence of proportions of a specific component that are exactly zero. In the division of the driven distance by road type, for instance, many insureds did not drive abroad during the observed policy period. Since compositional data are always analyzed by considering logratios of the components (see Section 4.4.2), a workaround is necessary.

In the compositional data literature, different types of zeros are being distinguished (Pawlowsky-Glahn et al., 2015). *Rounded zeros* occur when certain components may be unobserved because their true values are below the detection limit (cfr. geochemical studies). *Count zeros* refer to zero values due to the limited size of the sample in compositional data arising from count data. In our setting, the zero values are truly zero and are not due to imprecise or insufficient measurements. Such kind of zeros are called *structural zeros*. The structural zeros patterns in the data set are listed in Appendix 4.7. The presence of zeros is most prominent for splitting distance by road types as 40% of the drivers did not go abroad. Zeros are most often dealt with using replacement strategies (see e.g. Martín-Fernández et al., 2011, for an overview), which do not make sense for structural zeros. A general methodology is still to be developed (see e.g. Aitchison and Kay, 2003; Bacon Shone, 2003). In particular, there does not exist a method that deals with compositional data with structural zeros as predictor in regression models. Applying the ilr transform to the compositional data $\boldsymbol{x}$ and using the transformed $\boldsymbol{z}$ as explanatory variables in the predictor as discussed in Section 4.4.2 is no longer possible.

We propose to treat the structural zero patterns of the compositional predictors as different subgroups within the data and model the effect conditional on the zero pattern. In the most general situation, $2^D - 1$ possible zero patterns can

occur when dealing with compositional data with D components (a structural zero for every component being excluded). We introduce indicator variables for each zero pattern and use these in the compositional predictor term η^{comp} of the regression model to specify the effect on the outcome separately for each zero pattern. More specifically, we define the variables

$$d_{(i_1,\ldots,i_k)} = \begin{cases} 1 & \text{if components } i_1,\ldots,i_k \text{ of } \boldsymbol{x} \text{ are nonzero and all other are zero}\,, \\ 0 & \text{otherwise} \end{cases}$$

for all $k = 1,\ldots,D$ and $1 \leqslant i_1 < \ldots < i_k \leqslant D$. Conditional on the zero pattern $(i_1,\ldots,i_k)$ of the compositional data vector $\boldsymbol{x}$, the contribution to the predictor is given by the Aitchison inner product $\langle \boldsymbol{b}_{(i_1,\ldots,i_k)}, \boldsymbol{x}_{(i_1,\ldots,i_k)} \rangle_a$ of the subcomposition $\boldsymbol{x}_{(i_1,\ldots,i_k)}$ existing of the nonzero components of $\boldsymbol{x}$ and a subcompositional simplicial gradient $\boldsymbol{b}_{(i_1,\ldots,i_k)}$, which is different for each zero pattern. In case of only one nonzero component, the contribution is given by a simple categorical effect $b_{(i)}$. Note that the subscript $(i_1,\ldots,i_k)$ has a different interpretation for the dummy variable, simplicial gradient and compositional data vector. The proposed compositional predictor reads

$$\eta^{\text{comp}} = \sum_{i=1}^{D} d_{(i)} b_{(i)} + \sum_{k=2}^{D} \sum_{1 \leqslant i_1 < \ldots < i_k \leqslant D} d_{(i_1,\ldots,i_k)} \langle \boldsymbol{b}_{(i_1,\ldots,i_k)}, \boldsymbol{x}_{(i_1,\ldots,i_k)} \rangle_a \,.$$

Zero pattern specific intercepts can be added in the second term if deemed necessary.

4.4.3 Model selection and assessment

Using the same form as Akaike's information criterion, AIC for a GAM is defined as

$$\text{AIC} = -2 \cdot \widehat{\ell} + 2 \cdot \text{EDF} \tag{4.7}$$

where $\widehat{\ell}$ is the log-likelihood, evaluated at the estimated model parameters obtained using penalized likelihood maximization, and the effective degrees of freedom (EDF) is used instead of the actual number of model parameters. The EDF is defined as the trace of the hat or smoothing matrix in the corresponding working linear model at the last P-IRLS iteration (Hastie and Tibshirani, 1990). As such, (4.7) measures the quality of the model as a trade-off between the goodness-of-fit and the model complexity.

For each of the four predictor sets, see Figure 4.2c, variables are selected by AIC using an exhaustive search over all the possible combinations of variables given in Table 4.1. We limit ourselves to additive regression models (i.e. no interactions) such that an exhaustive search is still feasible and the marginal impact of a single variable can be easily assessed, interpreted and visualized. Even though the 2011 EU ruling prohibits a distinction between men and women in car insurance pricing, we allow gender to be selected as a categorical predictor in the model. For the division of the number of meters in different categories, 10 structural zero patterns occur for the road types, 20 for the time slots, and 3 for week/weekend. However, based on their relative frequencies, we only allow an additional compositional predictor for the distinction by road type in the case that a car did not drive abroad, which occurs for 40% of the observations. All remaining zero patterns are bundled into one residual group and their effect is modeled using a categorical effect b_0, see Table 4.8 of Appendix 4.7. The most comprehensive compositional predictor term we allow to be selected in the hybrid and telematics models is

$$\begin{aligned}\eta_{it}^{\text{comp}} = {} & d_{(1111)}^{\text{road}} \langle \boldsymbol{b}_{(1111)}^{\text{road}}, \boldsymbol{x}_{(1111)} \rangle_a + d_{(1110)}^{\text{road}} \langle \boldsymbol{b}_{(1110)}^{\text{road}}, \boldsymbol{x}_{(1110)} \rangle_a \\ & + (1 - d_{(1111)}^{\text{road}} - d_{(1110)}^{\text{road}}) b_0^{\text{road}} + d_{(11111)}^{\text{time}} \langle \boldsymbol{b}_{(11111)}^{\text{time}}, \boldsymbol{x}_{(11111)} \rangle_a \\ & + (1 - d_{(11111)}^{\text{time}}) b_0^{\text{time}} + d_{(11)}^{\text{week}} \langle \boldsymbol{b}_{(11)}^{\text{week}}, \boldsymbol{x}_{(11)} \rangle_a + (1 - d_{(11)}^{\text{week}}) b_0^{\text{week}} .\end{aligned}$$

In total, 165 888 model specifications are estimated under both the Poisson and the negative binomial framework.

Predictive performance of these models is assessed using *proper scoring rules* for count data, see Table 4.2 (Czado et al., 2009). Scoring rules assess the quality of probabilistic forecasts through a numerical score $s(P, n)$ based on the predictive distribution P and the observed count n. Lower scores indicate a better quality of the forecast. A scoring rule is proper (Gneiting and Raftery, 2007) if $s(Q, Q) \leqslant s(P, Q)$ for all P and Q with $s(P, Q)$ the expected value of $s(P, \cdot)$ under Q. In general, we define by $p_k = \mathbb{P}(N = k)$ and $P_k = \mathbb{P}(N \leqslant k)$ the probability mass function and cumulative probability function of the predictive distribution P for count variable N. The probability mass at the observed count n is denoted as p_n. The mean and standard deviation of P are written as μ_P and σ_P, respectively, and we set $\|p\| = \sum_{k=0}^{\infty} p_k^2$.

We compare the predictive performance of the best models according to AIC under the four predictor sets, with or without offset in the predictor (4.1), and using a Poisson or negative binomial distribution. We apply the proper scoring

Table 4.2: *Proper scoring rules for count data.*

Score	Formula
logarithmic	$\text{logs}(P, n) = -\log p_n$
quadratic	$\text{qs}(P, n) = -2p_n + \|\|p\|\|$
spherical	$\text{sphs}(P, n) = -\frac{p_n}{\|\|p\|\|}$
ranked probability	$\text{rps}(P, n) = \sum_{k=0}^{\infty}\{P_k - \mathbf{1}(n \leqslant k)\}^2$
Dawid-Sebastiani	$\text{dss}(P, n) = \left(\frac{n-\mu_P}{\sigma_P}\right)^2 + 2\log\sigma_P$
squared error	$\text{ses}(P, n) = (n - \mu_P)^2$

rules to the predictive count distributions of the observed claim counts. We adopt a K-fold cross-validation approach (Hastie et al., 2009) with $K = 10$ and apply the same partition to assess each model specification. Let $\kappa_{it} \in 1, 2, \ldots K$ be the part of the data to which the observed claim count n_{it} of policyholder i in policy period t is allocated by the randomization. Denote by $\widehat{P}_{it}^{-\kappa_{it}}$ the predictive count distribution for observation n_{it} estimated without the κ_{it}th part of the data. The K-fold cross-validation score $\text{CV}(s)$ is then given by

$$\text{CV}(s) = \frac{1}{\sum_{i=1}^{I} T_i} \sum_{i=1}^{I} \sum_{t=1}^{T_i} s(\widehat{P}_{it}^{-\kappa_{it}}, n_{it}) \,,$$

where s is any of the aforementioned proper scoring rules and smaller values of $\text{CV}(s)$ indicate better forecasts.

4.5 Results

4.5.1 Model selection

All computations are performed with `R` 3.2.5 (R Core Team, 2016) and, in particular, the `R` package `mgcv` version 1.8-11 (Wood, 2011) is used for the parameter estimation in the GAMs. The variables selected for each of the predictor sets were identical for the Poisson and NB models, see Table 4.3. The functional forms of the selected best models are given in Appendix 4.8. The offset versions of the classic and time-hybrid model replace the term $f_1(\texttt{time}_{it})$ by $\ln(\texttt{time}_{it})$, without any regression coefficient in front. This causes the expected number of reported MTPL claims, $\mu_{it} = \mathbb{E}(N_{it}) = \exp(\eta_{it})$, to be proportional to the duration of the policy period. In the offset versions of the meter-hybrid and telematics model, the flexible term related to `distance` gets replaced by an offset $\ln(\texttt{distance}_{it})$,

Table 4.3: *Variables contained in the best Poisson model for each of the predictor sets. The second column of each predictor set refers to the model with the offset restriction for either time or meter. The best NB models were identical to the best Poisson models.*

	Predictor	Classic		Time-hybrid		Meter-hybrid		Telematics	
Policy	Time	×	offset	×	offset				
	Age								
	Experience	×	×	×	×	×	×		
	Sex	×	×						
	Material	×	×	×	×	×	×		
	Postal code	×	×	×	×	×	×		
	Bonus-malus	×	×	×	×	×	×		
	Age vehicle	×	×	×	×	×	×		
	Kwatt			×	×	×	×		
	Fuel	×	×	×		×			
Telematics	Distance					×	offset	×	offset
	Yearly distance			×	×				
	Average distance			×	×	×	×	×	×
	Road type 1111			×	×	×	×	×	×
	Road type 1110			×	×	×	×	×	
	Time slot			×	×	×	×	×	×
	Week/weekend			×	×	×	×	×	×

imposing the risk to be proportional to the distance. Both hybrid models drop the `fuel` term in the best offset variants and the telematics model drops `road type 1110`.

The models which are allowed to use the policyholder information prefer the use of `experience`, measured as the years since obtaining the driver's license, instead of `age` to segment the risk in young drivers. `Gender` is only selected as an important covariate in the classic models, not in any of the hybrid models, indicating that the telematics information renders the use of gender as a rating variable redundant. The newly introduced telematics predictors are selected in both the hybrid and the telematics models and hence contribute to the quality of these models.

The second best models, with only a slightly higher AIC value, show that adding `kwatt` to the classic model gives a comparable model fit and the same holds for adding `road type 1110` to the telematics model with offset restriction. Furthermore, `fuel` and `kwatt` can easily be left out of the hybrid models without deteriorating the fit.

For each of these best model formulations, we added a policyholder-specific random effect in the predictor (4.1) to account for possible dependence from observing policyholders over multiple policy periods. However, none of the added random effects were deemed necessary at the 5% significance level using the approximate test of Wood (2013).

4.5.2 Model assessment

Table 4.4 reports AIC and all 6 proper scoring rules obtained using 10-fold cross validation for each predictor set under the Poisson model specification. These performance tools unanimously indicate that the time-hybrid model without offset scores best. The meter-hybrid model is a close second. Their respective versions with an offset restriction and the telematics model without offset conclude the top five according to all criteria except the Dawid–Sebastiani score. This demonstrates the significant impact of telematics constructed variables on the predictive power of the model. In addition, the telematics model without offset outperforms the classic models across all assessment criteria. Hence, using only telematics predictors is considered to be better than the use of the traditional rating variables.

Across all predictor sets, the use of an offset for the exposure-to-risk, either time or meter, is too restrictive for these data. From a statistical point of view, the time or meter rating unit cannot be considered to be directly proportional to the risk. However, from a business point of view, it is convenient to consider a proportional approach due to its simplicity and explainability.

Similar results are obtained under the negative binomial model specification. The rankings according to AIC are the same as in Table 4.4. The AIC values for each predictor set under the NB model specification compared to their Poisson counterpart were slightly higher for the classic and hybrid models and slightly lower for the telematics models indicating that only the telematics predictor sets benefit from the additional parameter to capture overdispersion. The model assessment using proper scoring rules led to the same conclusions as before.

Beside an exhaustive search among additive terms, we have explored the use of interactions among categorical, among continuous, between categorical and continuous, and between categorical and compositional predictors. Slight marginal improvements in AIC could only be achieved in the classic model by further refining the effects of `experience`, `age vehicle` and `material` by `gender` without changing the rankings in Table 4.4 of the best models.

Table 4.4: *Model assessment of the best models according to AIC for each of the four predictor sets under the Poisson model specification. The second row of each predictor set refers to the model with the offset restriction for either time or meter. For each model we list the effective degrees of freedom (EDF), Akaike information criterion (AIC) and 6 cross-validated proper scoring rules: logarithmic (logs), quadratic (qs), spherical (sphs), ranked probability (rps), Dawid-Sebastiani (dss), and squared error scores (ses). For AIC and the proper scoring rules, the first column represents the value and the second column the rank.*

Predictor set	Offset	EDF	AIC		logs		qs		sphs		rps		dss		ses	
			value	rank	value	rank	value	rank	value	rank	value	rank	value	rank	value	rank
Classic	no	32.15	11 896	6	0.1790	6	−0.918 58	6	−0.958 22	6	0.042 24	6	−2.206	5	0.045 35	6
	yes	27.27	11 995	8	0.1804	8	−0.918 39	8	−0.958 16	8	0.042 34	8	−2.130	8	0.045 46	8
Time-hybrid	no	39.66	11 727	1	0.1764	1	−0.919 10	1	−0.958 37	1	0.041 95	1	−2.275	1	0.045 01	1
	yes	36.22	11 811	3	0.1777	3	−0.918 90	3	−0.958 31	3	0.042 06	3	−2.212	4	0.045 14	3
Meter-hybrid	no	41.47	11 736	2	0.1766	2	−0.919 08	2	−0.958 36	2	0.041 96	2	−2.266	2	0.045 02	2
	yes	36.23	11 856	5	0.1784	5	−0.918 80	4	−0.958 27	4	0.042 12	4	−2.158	6	0.045 22	4
Telematics	no	20.58	11 855	4	0.1782	4	−0.918 73	5	−0.958 26	5	0.042 15	5	−2.231	3	0.045 24	5
	yes	14.38	11 976	7	0.1800	7	−0.918 47	7	−0.958 18	7	0.042 30	7	−2.134	7	0.045 46	7

4.5.3 Visualization and discussion

The effects of each predictor variable in the best time-hybrid model without offset restriction are graphically displayed in Figure 4.5 for the policy model terms and Figure 4.6 for the telematics model terms. By exponentially transforming the additive effects, we show the multiplicative effects on the expected number of claims for each categorical parametric, continuous smooth or geographical term in the fitted model. For the categorical predictors we quantify the uncertainty of those estimates by constructing individual 95% confidence intervals based on the large sample normality of the model parameter estimators. Bayesian 95% confidence pointwise intervals are used for the smooth components of the GAM and include the uncertainty about the intercept (Marra and Wood, 2012). For the compositional data predictors, we visualize the exponentiated clr transform of the corresponding model parameters with 95% confidence intervals along with a reference line at one (see Section 4.4.2). Similar graphs for the other three predictor sets, see Figure 4.2c, are shown in Appendix 4.9 and the relative importance of these predictors is quantified and visualized in Appendix 4.10. In the remainder of this section, we discuss the insights and interpretations for both the policy and telematics variables in each of these models.

Policy variables The rating unit `policy period` in the classic and time-hybrid models always has a monotone increasing estimated effect. The longer a policyholder is insured, the higher the premium amount, ceteris paribus. Using the fact that the level of the nonlinear smooth components are not uniquely identifiable (see Section 4.4.1), we vertically translated the estimated smooth term to pass the point (365, 0) on the predictor scale (and hence (365, 1) on the response scale) for ease of interpretation.

The smooth effect of `experience` embodies the higher risk posed by younger, less experienced drivers. The increased risk is more outspoken in the first two years for the hybrid models as compared to the classic model.

In the classic model, the significant effect of `gender` indicates that women are 16% less risky drivers than men. However, when telematics predictors are taken into account in the hybrid models, the categorical variable `gender` is no longer selected as predictor. Neither did any interaction term between gender and a categorical, a continuous or a compositional predictor improve AIC. The perceived difference between women and men can hence be explained through differences in driving habits. In particular, female drivers in the portfolio drive significantly fewer kilometers on a yearly basis compared to men (15 409 vs 18 570

on average, with a p-value smaller than 0.001 using a two sample t-test). Similar findings are reported in Ayuso et al. (2016a,b). In light of the EU rules on gender-neutral pricing in insurance, this shows how moving towards car insurance rating based on individual driving habits and style can resolve possible discrimination of basing the premium on proxies such as gender.

The smooth effects of `bonus-malus` in the classic and hybrid models are non-linear and somewhat counterintuitive. Given the lack of a lengthy claim history of the young drivers of this portfolio, the BM level of the insureds are not yet fully developed and stabilized. The majority of the drivers has a bonus-malus (BM) level between 4 and 12 for which the effect on the claim frequency is increasing. For the highest BM levels however, the effect is declining, albeit with a high uncertainty due to a lack of observations in this region. Furthermore, the effect does not decrease for the lowest BM levels. This can be explained by an improper use of the BM scale as marketing tool to attract new customers. By lowering the initial value of the BM scale, the insurer can reduce the premium a potential new policyholder has to pay.

When it comes to characteristics of the car, insureds driving older vehicles have an estimated higher risk of accidents. The smooth effect of `age vehicle` is estimated as a straight line on the predictor scale in the classic and hybrid models. The effect of `kwatt` in the hybrid models also reduced to a straight line on the predictor scale. When the insured vehicle has more horsepower, the estimated expected claims number is lower, although this effect is of lesser importance for the model fit as indicated earlier. The categorical model term `fuel` shows that vehicles using petrol have an estimated lower risk for accidents compared to diesel. This difference is however smaller and no longer statistically significant in the hybrid models compared to the classic model.

In both the classic and hybrid models, the policies without `material damage cover` have a 20% lower estimated expected number of claims. This may be explained by the reluctance of some insureds without additional material damage coverage to report small accidents. Due to bonus-malus mechanisms being independent of the claim amount, filing a claim leads to premium surcharges which may be more disadvantageous for policyholders than for them to defray the third party. This phenomenon is known as the hunger for bonus (Denuit et al., 2007). Insureds with an additional material damage cover are less inclined to do so since their own, first party costs are also covered making it more worthwhile to report a claim at fault. Including telematics variables in the model does not affect this discrepancy.

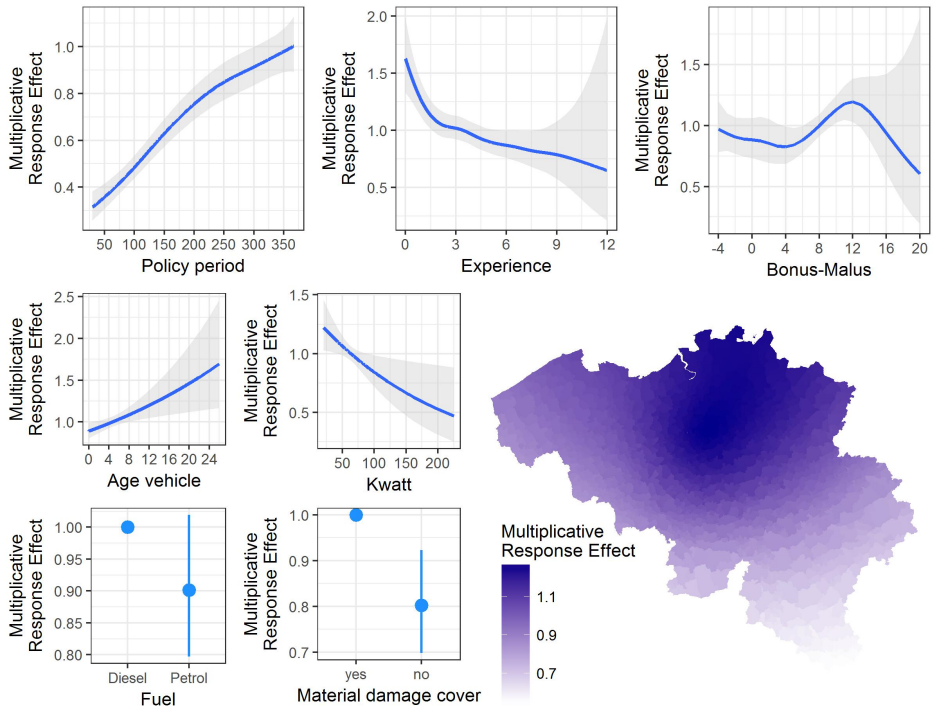

Figure 4.5: *Multiplicative response effects of the policy model terms of the time-hybrid model.*

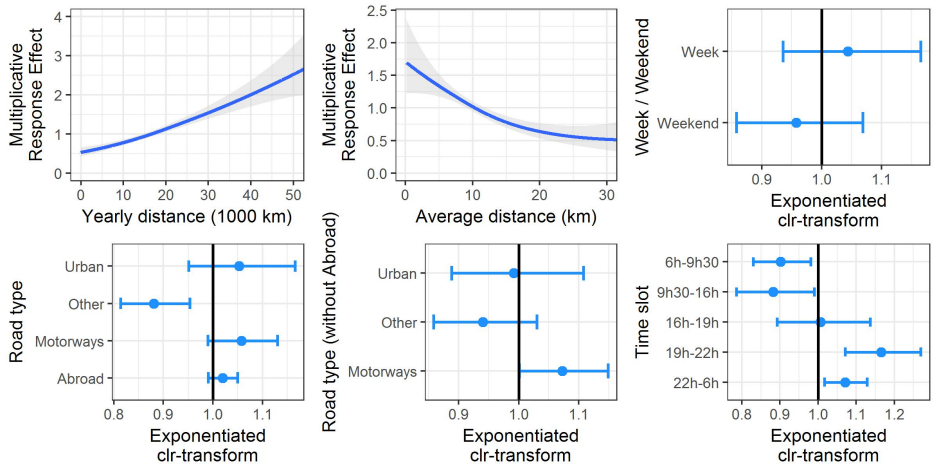

Figure 4.6: *Multiplicative response effects of the telematics model terms of the time-hybrid model.*

The geographical effect (`postal code`), plotted on top of a map of Belgium for the classic and hybrid models, captures the remaining spatial heterogeneity based on the postal code where the policyholder resides. For the classic model, the graph shows higher claim frequencies for urban areas like Brussels in the middle, Antwerp in the north and Liège in the east and lower claim frequencies in the more sparsely populated regions in the south. The geographic variation however decreases strongly in the hybrid models due to the inclusion of telematics predictors not taken into account in the classic model. The EDF corresponding to the spatial smooth reduced from 15.8 in the classic model to 6.4 in both hybrid models. This is satisfactory as it means, instead of overrelying on geographical proxies, the hybrid models are basing the insurance premium on actual differences in driving habits (such as the proportion driven on urban roads) which is more closely related to the accident risk.

Telematics variables In the meter-hybrid and telematics models, `distance` is used as the rating unit. Similar to the time effect in the classic and time-hybrid model, the effect of the risk exposure is estimated as a monotone increasing function. The accident risk however does not vanish for insureds who hardly drive any kilometers during the observation period.

The `yearly distance` is used in the time-hybrid model, which uses time as exposure, to differentiate between drivers who travel many versus few kilometers on a yearly basis. In this way, the driven distance is rescaled on a yearly basis (see Section 4.3.2) and used as an additional risk factor having a weaker effect on the claim frequency compared to the meter-hybrid and telematics models where distance is used a rating unit. In both hybrid models and the telematics model, the estimated `average distance` effect shows lower claim frequencies for insureds who on average drive long distances.

The exponentiated clr transforms of the model coefficients related to the compositional `road type` predictor in the telematics model show how insureds who drive relatively more on urban roads have higher claim frequencies and insureds who drive relative more on the road type 'other' have lower claim frequencies. The same interpretation holds for insureds who do not drive abroad during the policy period. In the hybrid models, these effects are headed in the same direction with the exception that motorways is perceived as riskier. The elevated accident risk for insureds driving more on urban roads is in line with Paefgen et al. (2014), where the driven distance is divided over 'highway', 'urban' and 'extra-urban' road types. The authors however neglect the compositional nature of this predic-

tor in the analysis and do not incorporate any of the classical policy risk factors in the logistic regression model. In Ayuso et al. (2014), the percentage of urban driving is considered an important variable to predict either the time or the distance to the first accident, although percentages driven on different road types are not considered. Using either a quadratic effect or a categorical effect (urban driving $> 25\%$) in Weibull regression models shows how increased percentages of urban driving reduce both the expected time or distance to the first accident.

The compositional `time slot` predictor in the hybrid and telematics models indicates that policyholders who drive relatively more in the morning have lower claim frequencies and policyholders who drive relatively more in the evening and during the night have higher claim frequencies. In Paefgen et al. (2014), the accident risk is considered to be lower during the daytime (between 5 and 18h) compared to the evening (between 18h and 21h), based on the estimated coefficients of linear model terms of the log transformed percentages of the driven distance in these time slots. Ayuso et al. (2014) reports how a higher percentage of driving at night reduces the expected time to a first accident, where the effect is modeled linearly, with no further distinction in time slots.

Driving more in the week than in the weekend increases the probability of having a claim. An increased accident risk in case of more driving in the week is also found in Paefgen et al. (2014), though they define weekend from Friday to Sunday. The compositional effect of `week/weekend` is retained in both hybrid models as well as the telematics model according to AIC even though it is not statistically significant. This is due to a highly significant and positive estimated categorical effect b_0^{week} for the 73 observations with structural zeros belonging to the rest group, see Table 4.8 of Appendix 4.7. These drivers have jointly driven 58 000 kilometers during a combined insured policy period of 16.5 years and reported the remarkably high number of 5 claims.

4.6 Conclusion

Telematics insurance offers new opportunities for insurers to differentiate drivers based on their driving habits and style. By aggregating the telematics data on the level of the policy period by policyholder and combining it with traditional policy(holder) rating variables, we construct predictive models for the frequency of MTPL claims at fault. Generalized additive models with a Poisson or negative binomial response are used to model the effects of predictors in a smooth, yet interpretive way. The divisions of the driven distance into 4 road types and 5

time slots forms a challenge from a methodological point of view that has not been addressed in the literature. We demonstrate how to include this information as compositional predictors in the regression and formulate a new way of how to interpret their effect on the average claim frequency.

Our research reveals the significant impact of the use of telematics data through an exhaustive model selection and an assessment of the predictive performance. The time-hybrid is the best model according to AIC and all proper scoring rules, closely followed by the meter-hybrid model. The model using only telematics variables is ranked higher than the best classic model using only traditional policy information.

The compositional predictors show that a further classification of the driven distance based on the location and the time is relevant. Our contribution indicates that driving more on urban roads, in the evening or at night and during the week contributes to a riskier driving pattern. The best hybrid models highlight that certain popular pricing factors (gender, fuel, postcode) are indeed proxies for the driving habits and part of their predictive power is taken over by the distance driven and the splits into different categories. Hence, we demonstrate using careful statistical modeling how the use of telematics variables is an answer to the European regulation on insurance pricing practices that bans the use of gender as a rating factor.

In the case of multiple insured drivers, it is unclear which characteristics (such as age, experience and gender) the insurer must use to determine the premium. We proceed, in consultation with the Belgian insurer providing the data, by identifying the driver with the lowest experience as the main driver and use his policyholder information as predictors in the regression for tarification purposes. In practice, when a parent adds a child as a driver in the policy, a premium surcharge is often avoided to prevent the policyholder from lapsing. By shifting towards pricing based on telematics information as we do in this research, this tarification issue becomes less of a problem because the premium will be usage-based.

Pricing using telematics data can be seen as falling in between *a priori* and *a posteriori* pricing. The driving habits and style are no traditional a priori variables since they cannot be determined before the policyholder starts to drive. Insurers now reason that available UBI products are only purchased by drivers who consider themselves to be either safe or low-kilometer drivers. This potential form of positive selection, which could not be quantified based on the studied portfolio alone, validates an upfront discount on the traditional insurance premium. Based on the telematics data collected over time, insurers can set up a

discount structure to adapt the premium in an a posteriori way. The discount structure can depend on the actual driven distance, with a further personalized differentiation based on the riskiness of the profile as perceived from the driving habits of the insured. The insights provided in this chapter reveal which elements can be adopted in such a structure, for instance, by making kilometers driven on urban roads or in the evening or at night more expensive.

In conclusion, telematics technology provides means to insurers to better align premiums with risk. Pay-as-you-drive insurance is a first step in which the number of driven kilometers, the type of road and the time of day are combined with the traditional self-reported information such as policyholder and car characteristics to calculate insurance premiums. A next step is pay-how-you-drive insurance, where on top of these driving habits also the driving style is considered to assess how risky someone drives by monitoring for instance speed infringements, harsh braking, excessive acceleration, and cornering style. The ideas and statistical framework presented can be extended to incorporate such additional pay-how-you-drive predictors if they are available.

4.7 Appendix A: Structural zero patterns of the compositional telematics predictors

We give an overview of the structural zero patterns for the division of the number of meters in road types (Table 4.5), time slots (Table 4.6) and week/weekend (Table 4.7). The pattern is represented in the first column by a code indicating which components are zero (0) or non-zero (1). For each structural zero pattern, we tabulate their absolute and relative frequency and the compositional mean of the nonzero components, which for M observations $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iD})^t$ and $i = 1, \ldots, M$ is defined as

$$\overline{\boldsymbol{x}} = \frac{1}{M} \odot \bigoplus_{i=1}^{M} \boldsymbol{x}_i = \mathcal{C}\left(\left(\prod_{i=1}^{M} x_{i1} \right)^{1/M}, \ldots, \left(\prod_{i=1}^{M} x_{iD} \right)^{1/M} \right)^t \tag{4.8}$$

resulting in the closed componentwise geometric mean. Following the principle of working on coordinates, we can alternatively write the compositional mean as

$$\overline{\boldsymbol{x}} = \mathrm{ilr}^{-1}\left(\frac{1}{M} \sum_{i=1}^{M} \mathrm{ilr}(\boldsymbol{x}_i) \right),$$

where we first transform the compositional data from $\mathcal{S}^D$ to $\mathbb{R}^{D-1}$ using the ilr transformation, then compute the mean in $\mathbb{R}^{D-1}$ and finally apply the inverse ilr transformation to obtain the compositional mean in $\mathcal{S}^D$.

Table 4.5: *Structural zero patterns for the division of meters in road types.*

Road type	Number	Percent	Urban	Other	Motorways	Abroad
1111	18821	0.5659	0.4421	0.2822	0.2516	0.0241
1110	13540	0.4071	0.5079	0.2782	0.2139	–
1100	481	0.0145	0.5923	0.4077	–	–
1101	258	0.0078	0.4960	0.4648	–	0.0392
0001	131	0.0039	–	–	–	1
1010	7	0.0002	0.9075	–	0.0925	–
1001	7	0.0002	0.0034	–	–	0.9966
1000	6	0.0002	1	–	–	–
0101	5	0.0001	–	0.0002	–	0.9998
0111	3	0.0001	–	0.0130	0.0833	0.9038

Table 4.6: *Structural zero patterns for the division of meters in time slots.*

Time slot	Number	Percent	6h-9h30	9h30-16h	16h-19h	19h-22h	22h-6h
11111	31886	0.9587	0.1472	0.4699	0.2159	0.1010	0.0661
11110	991	0.0298	0.2000	0.5090	0.2323	0.0587	–
11101	130	0.0039	0.2060	0.5953	0.1296	–	0.0691
11100	110	0.0033	0.2134	0.6238	0.1628	–	–
01111	47	0.0014	–	0.5398	0.1983	0.1339	0.1280
01110	23	0.0007	–	0.5850	0.2793	0.1357	–
01100	22	0.0007	–	0.7912	0.2088	–	–
11000	16	0.0005	0.1459	0.8541	–	–	–
11001	10	0.0003	0.0697	0.8000	–	–	0.1304
01000	7	0.0002	–	1	–	–	–
01001	3	0.0001	–	0.6803	–	–	0.3197
01010	2	0.0001	–	0.3054	–	0.6946	–
10000	2	0.0001	1	–	–	–	–
01101	2	0.0001	–	0.6698	0.1744	–	0.1558
10001	2	0.0001	0.1271	–	–	–	0.8729
11011	2	0.0001	0.0653	0.5536	–	0.2762	0.1049
00100	1	0.0000	–	–	1	–	–
00110	1	0.0000	–	–	0.8200	0.1800	–
10010	1	0.0000	0.9787	–	–	0.0213	–
10110	1	0.0000	0.2451	–	0.2935	0.4614	–

Table 4.7: *Structural zero patterns for the division of meters in week and weekend.*

Week/weekend	Number	Percent	Week	Weekend
11	33186	0.9978	0.7490	0.2510
10	72	0.0022	1	–
01	1	0.0000	–	1

In this chapter, infrequently observed patterns are bundled into a residual group when incorporating the compositional variables as predictors in the claim count models leading to the distinguished structural zero patterns of Table 4.8.

Table 4.8: *Structural zero patterns for the division of the number of meters in road types, time slots and week/weekend as recognized in the claim count models.*

Road type	Number	Percent	Urban	Other	Motorways	Abroad
1111	18821	0.5659	0.4421	0.2822	0.2516	0.0241
1110	13540	0.4071	0.5079	0.2782	0.2139	–
0	898	0.0270	–	–	–	–

Time slot	Number	Percent	6h-9h30	9h30-16h	16h-19h	19h-22h	22h-6h
11111	31886	0.9587	0.1472	0.4699	0.2159	0.1010	0.0661
0	1373	0.0413	–	–	–	–	–

Week/weekend	Number	Percent	Week	Weekend
11	33186	0.9978	0.7490	0.2510
0	73	0.0022	–	–

4.8 Appendix B: Functional forms of the selected best models

The functional form of the predictor in the preferred classic model is

$$\begin{aligned}\eta_{it}^{\text{classic}} = \beta_0 &+ \beta_1 \texttt{gender}_{it} + \beta_2 \texttt{material}_{it} + \beta_3 \texttt{fuel}_{it} + f_1(\texttt{time}_{it}) \\ &+ f_2(\texttt{experience}_{it}) + f_3(\texttt{bonus-malus}_{it}) + f_4(\texttt{age vehicle}_{it}) \\ &+ f_s(\texttt{lat}_{it}, \texttt{long}_{it})\,.\end{aligned}$$

The predictor in the best time-hybrid model can be written as

$$\begin{aligned}\eta_{it}^{\text{time-hybrid}} &= \beta_0 + \beta_1 \texttt{material}_{it} + \beta_2 \texttt{fuel}_{it} + f_1(\texttt{time})_{it} + f_2(\texttt{experience}_{it}) \\ &\quad + f_3(\texttt{bonus-malus}_{it}) + f_4(\texttt{age vehicle}_{it}) + f_s(\texttt{lat}_{it}, \texttt{long}_{it}) \\ &\quad + f_5(\texttt{yearly distance}_{it}) + f_6(\texttt{average distance}_{it}) \\ &\quad + d^{\text{road}}_{(1111)} \langle \boldsymbol{b}^{\text{road}}_{(1111)}, \boldsymbol{x}_{(1111)} \rangle_a + d^{\text{road}}_{(1110)} \langle \boldsymbol{b}^{\text{road}}_{(1110)}, \boldsymbol{x}_{(1110)} \rangle_a \\ &\quad + (1 - d^{\text{road}}_{(1111)} - d^{\text{road}}_{(1110)}) b_0^{\text{road}} + d^{\text{time}}_{(11111)} \langle \boldsymbol{b}^{\text{time}}_{(11111)}, \boldsymbol{x}_{(11111)} \rangle_a \\ &\quad + (1 - d^{\text{time}}_{(11111)}) b_0^{\text{time}} + d^{\text{week}}_{(11)} \langle \boldsymbol{b}^{\text{week}}_{(11)}, \boldsymbol{x}_{(11)} \rangle_a + (1 - d^{\text{week}}_{(11)}) b_0^{\text{week}},\end{aligned}$$

and for the preferred meter-hybrid model we have

$$\begin{aligned}\eta_{it}^{\text{meter-hybrid}} &= \beta_0 + \beta_1 \texttt{material}_{it} + \beta_2 \texttt{fuel}_{it} + f_1(\texttt{experience}_{it}) \\ &\quad + f_2(\texttt{bonus-malus}_{it}) + f_3(\texttt{age vehicle}_{it}) + f_s(\texttt{lat}_{it}, \texttt{long}_{it}) \\ &\quad + f_4(\texttt{distance}_{it}) + f_5(\texttt{average distance}_{it}) \\ &\quad + d^{\text{road}}_{(1111)} \langle \boldsymbol{b}^{\text{road}}_{(1111)}, \boldsymbol{x}_{(1111)} \rangle_a + d^{\text{road}}_{(1110)} \langle \boldsymbol{b}^{\text{road}}_{(1110)}, \boldsymbol{x}_{(1110)} \rangle_a \\ &\quad + (1 - d^{\text{road}}_{(1111)} - d^{\text{road}}_{(1110)}) b_0^{\text{road}} + d^{\text{time}}_{(11111)} \langle \boldsymbol{b}^{\text{time}}_{(11111)}, \boldsymbol{x}_{(11111)} \rangle_a \\ &\quad + (1 - d^{\text{time}}_{(11111)}) b_0^{\text{time}} + d^{\text{week}}_{(11)} \langle \boldsymbol{b}^{\text{week}}_{(11)}, \boldsymbol{x}_{(11)} \rangle_a + (1 - d^{\text{week}}_{(11)}) b_0^{\text{week}}.\end{aligned}$$

Finally, the predictor in the best telematics model is given by

$$\begin{aligned}\eta_{it}^{\text{telematics}} &= \beta_0 + f_1(\texttt{distance})_{it} + f_2(\texttt{average distance}_{it}) \\ &\quad + d^{\text{road}}_{(1111)} \langle \boldsymbol{b}^{\text{road}}_{(1111)}, \boldsymbol{x}_{(1111)} \rangle_a + d^{\text{road}}_{(1110)} \langle \boldsymbol{b}^{\text{road}}_{(1110)}, \boldsymbol{x}_{(1110)} \rangle_a \\ &\quad + (1 - d^{\text{road}}_{(1111)} - d^{\text{road}}_{(1110)}) b_0^{\text{road}} + d^{\text{time}}_{(11111)} \langle \boldsymbol{b}^{\text{time}}_{(11111)}, \boldsymbol{x}_{(11111)} \rangle_a \\ &\quad + (1 - d^{\text{time}}_{(11111)}) b_0^{\text{time}} + d^{\text{week}}_{(11)} \langle \boldsymbol{b}^{\text{week}}_{(11)}, \boldsymbol{x}_{(11)} \rangle_a + (1 - d^{\text{week}}_{(11)}) b_0^{\text{week}}.\end{aligned}$$

4.9 Appendix C: Graphical model displays

The effects of each predictor variable in the best classic model (resp. telematics model) without offset restriction are graphically displayed in Figure 4.7 (resp Figure 4.8). Similarly, the effects of each predictor variable in the best meter-hybrid model without offset restriction are graphically displayed in Figure 4.9 for the policy model terms and Figure 4.10 for the telematics model terms.

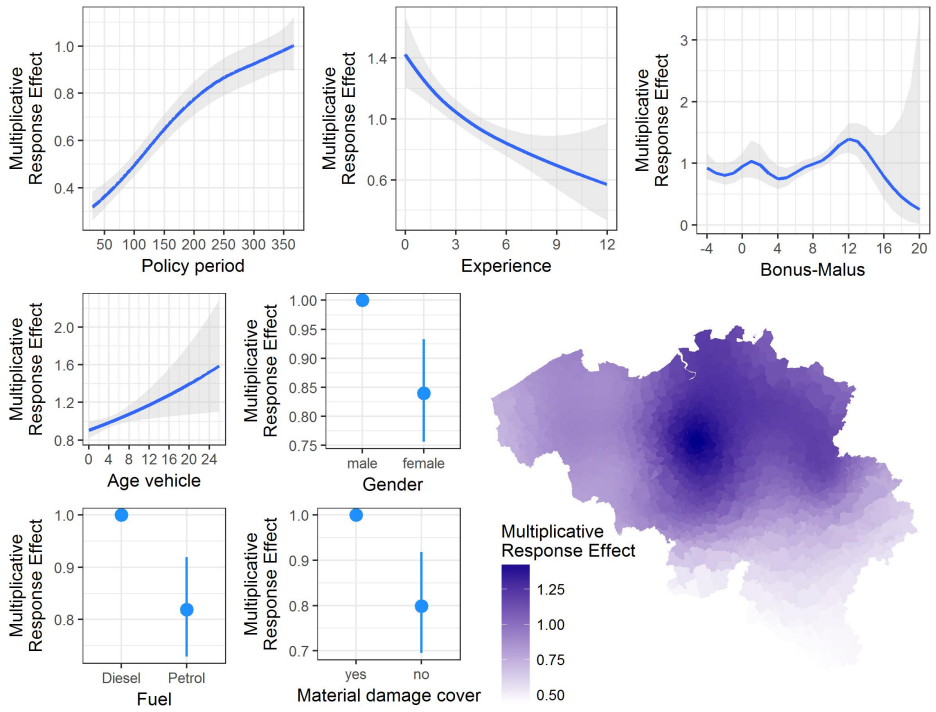

Figure 4.7: *Multiplicative response effects of the model terms of the classic model.*

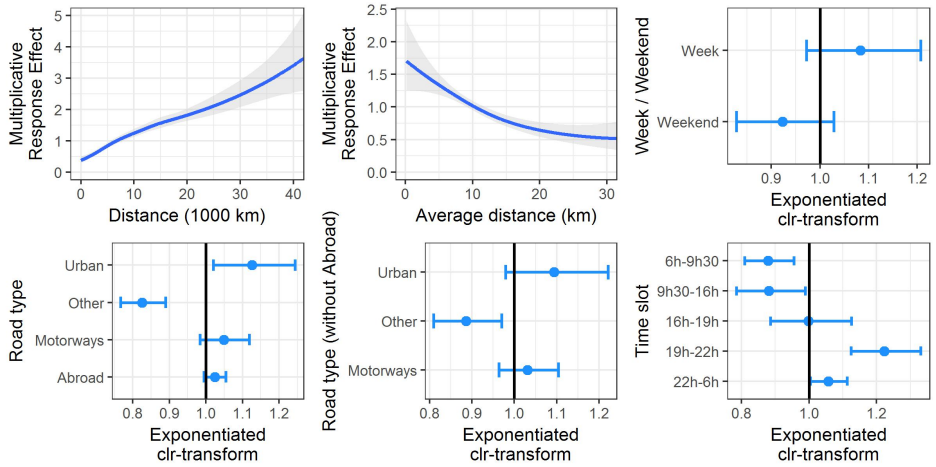

Figure 4.8: *Multiplicative response effects of the model terms of the telematics model.*

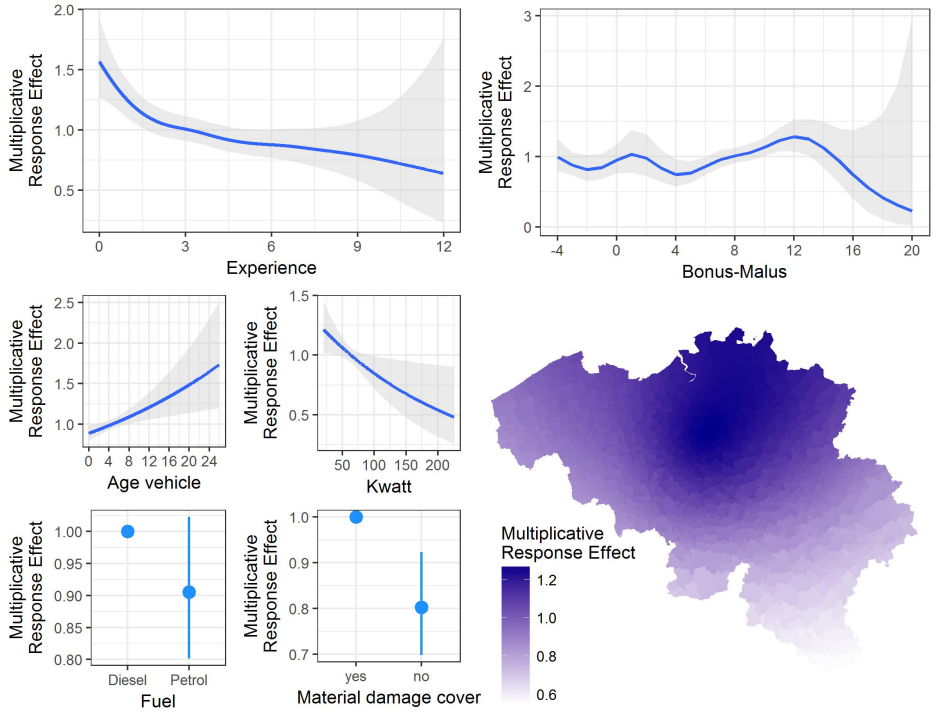

Figure 4.9: *Multiplicative response effects of the policy model terms of the meter-hybrid model.*

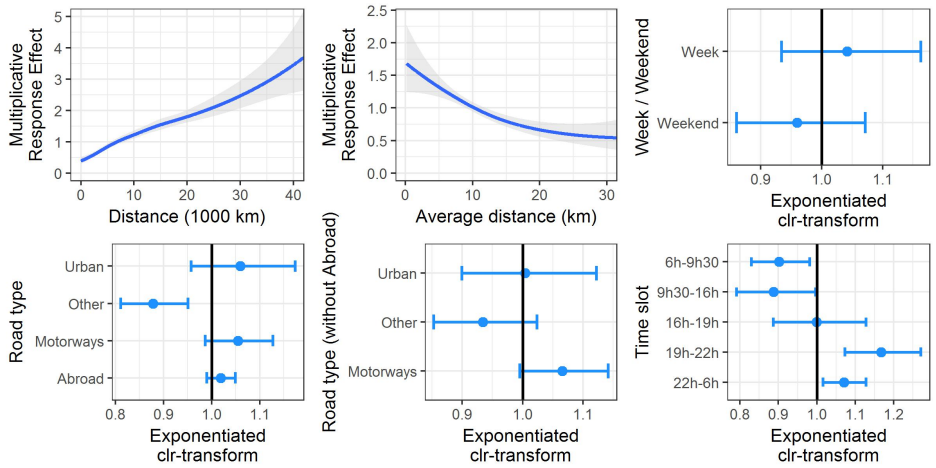

Figure 4.10: *Multiplicative response effects of the telematics model terms of the meter-hybrid model.*

4.10 Appendix D: Relative importance

To assess the relative importance of these variables in the model, we construct histograms of the multiplicative effects by predictor for each observation in the data set. This is done for the classic model in Figure 4.7, for the telematics model in Figure 4.8, for the time-hybrid model in Figures 4.13 and 4.14 and for the meter-hybrid model in Figures 4.9 and 4.10. For the hybrid models, we constructed separate graphs for the model terms derived from the policy and telematics information. For categorical predictors this reduces to a bar plot of the categorical effects and for the continuous and geographical predictors to a histogram of the exponentiated smooth effects. For a compositional predictor, such as time slot, we plot a histogram of the exponential of the term $\langle \hat{\boldsymbol{b}}^{\text{time}}_{(11111)}, \boldsymbol{x}_{(11111)} \rangle_a$ for all observations with pattern 11111. With the division in road types, we consider simultaneously the terms related to patterns 1111 and 1110. To rank the influence of the different policy and telematics variables on the claim frequency, we use the standard deviations over all observations of the effects on the predictor scale, see Table 4.9. Under the offset restriction, the logarithm of time or meter is used as an explanatory variable in the predictor without any regression coefficient in front and we report its standard deviation.

Table 4.9: *Standard deviations of the effects on the predictor scale in the best Poisson model for each of the predictor sets. The second column of each predictor set refers to the model with the offset restriction for either time or meter.*

	Predictor	Classic		Time-hybrid		Meter-hybrid		Telematics	
Policy	Time	0.36	0.69	0.37	0.69				
	Age								
	Experience	0.18	0.14	0.16	0.11	0.15	0.12		
	Gender	0.09	0.09						
	Material	0.11	0.11	0.11	0.10	0.11	0.10		
	Postal code	0.21	0.20	0.14	0.14	0.14	0.16		
	Bonus-malus	0.16	0.18	0.11	0.15	0.14	0.15		
	Age vehicle	0.08	0.10	0.09	0.10	0.10	0.11		
	Kwatt			0.07	0.06	0.07	0.08		
	Fuel	0.09	0.09	0.05		0.05			
Telematics	Distance					0.44	0.95	0.45	0.95
	Yearly distance			0.30	0.36				
	Average distance			0.23	0.25	0.21	0.32	0.23	0.34
	Road type			0.13	0.14	0.12	0.15	0.16	0.18
	Time slot			0.20	0.20	0.20	0.18	0.23	0.22
	Week/weekend			0.03	0.03	0.03	0.04	0.05	0.05

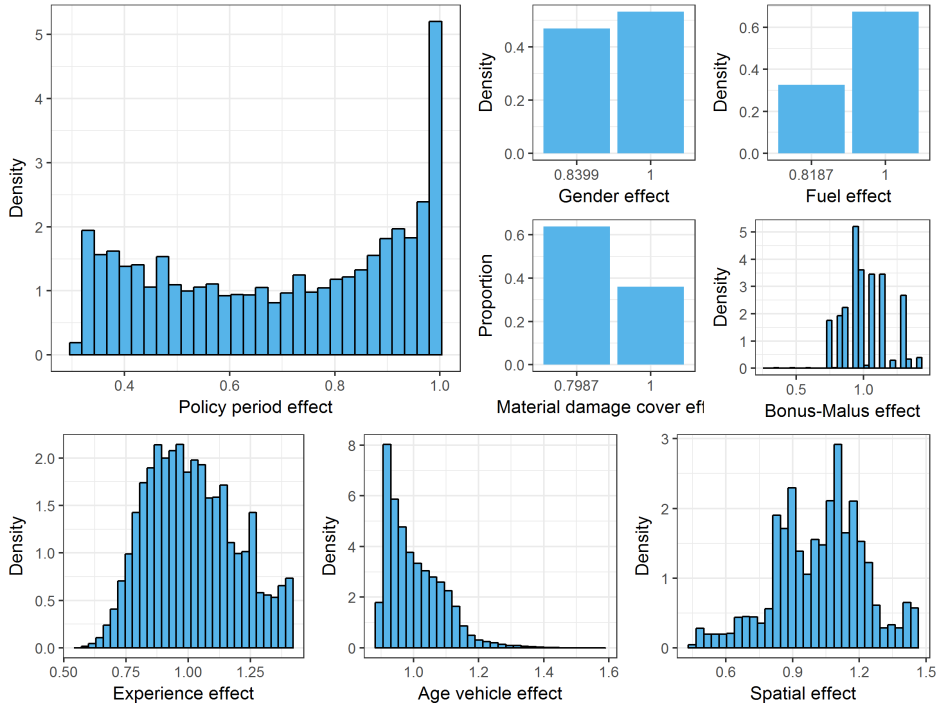

Figure 4.11: *Relative frequencies of the multiplicative response effects of the model terms of the classic model.*

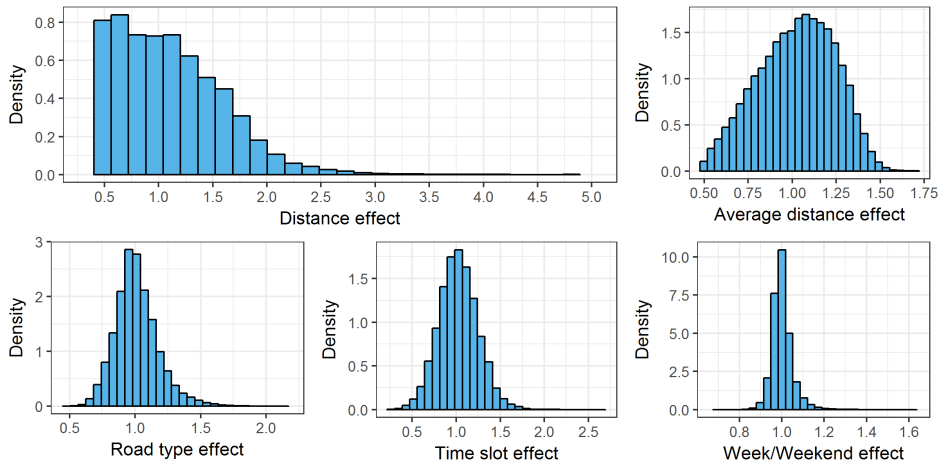

Figure 4.12: *Relative frequencies of the multiplicative response effects of the model terms of the telematics model.*

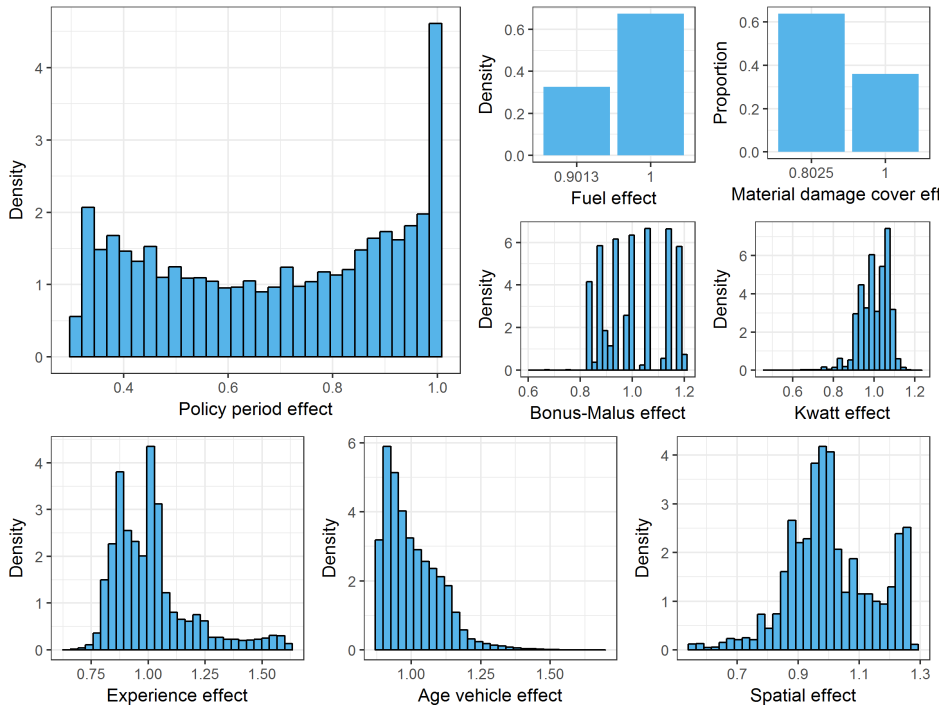

Figure 4.13: *Relative frequencies of the multiplicative response effects of the policy model terms of the time-hybrid model.*

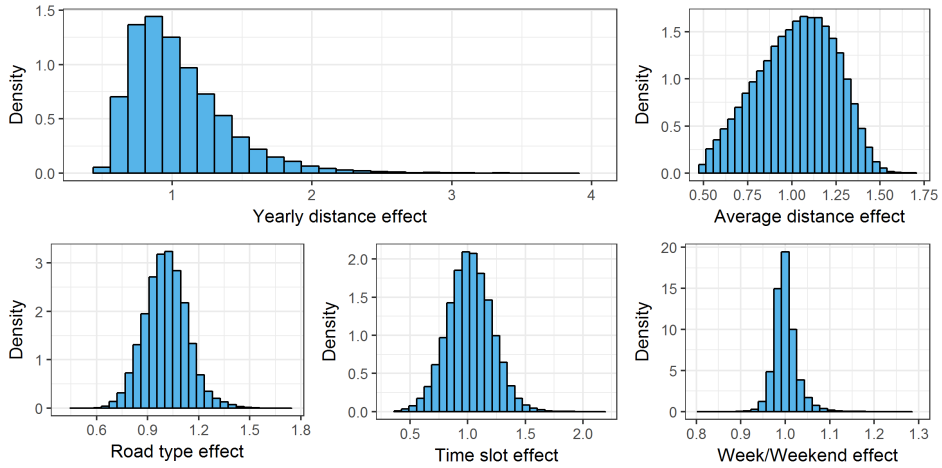

Figure 4.14: *Relative frequencies of the multiplicative response effects of the telematics model terms of the time-hybrid model.*

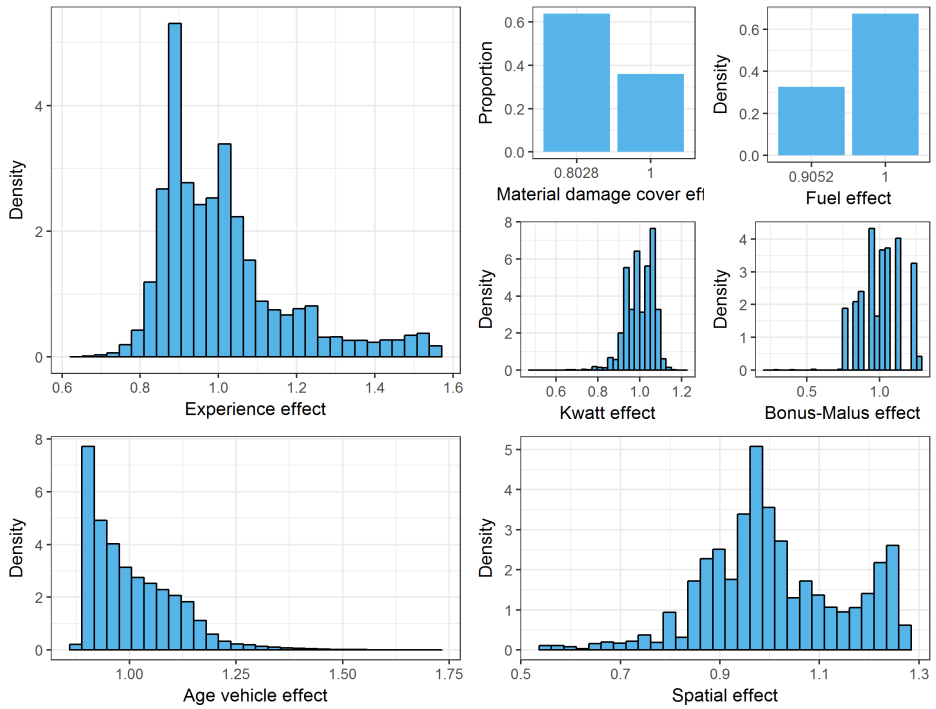

Figure 4.15: *Relative frequencies of the multiplicative response effects of the policy model terms of the meter-hybrid model.*

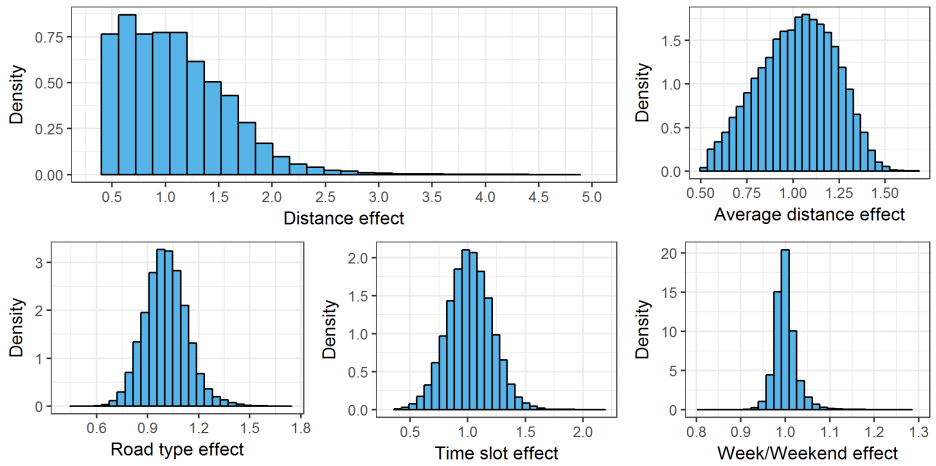

Figure 4.16: *Relative frequencies of the multiplicative response effects of the telematics model terms of the meter-hybrid model.*

Chapter 5

Predicting daily IBNR claim counts using a regression approach for the occurrence of claims and their reporting delay

Abstract

Insurance companies need to hold capital to be able to fulfill future liabilities with respect to the policies they write. Due to the delay in the reporting of claims, not all of the claims that occurred in the past have been observed yet. The accurate estimation of the number of incurred but not reported claims forms an essential part of claims reserving. We present a flexible framework to model and jointly estimate the occurrence and reporting of claims. A regression approach is used to capture the seasonal effects of the month, day of the week and day of the month of the occurrence date and to incorporate the proportional effect of exposure on claim occurrences. Parameter estimates are obtained using the EM algorithm by regarding the daily run-off triangle of claims as an incomplete data problem. The resulting method is elegant, easy to understand and implement, and provides refined forecasts on a daily level. The proposed methodology is applied to a European general liability portfolio. Initial insight into the data set motivates us to model the reporting delays in weeks combined with day-specific reporting probabilities. The performance of our model is evaluated based on out-of-sample data.

This chapter is based on Verbelen, R., Antonio, K., Claeskens, G. and Crèvecoeur, J. (2017). Predicting daily IBNR claim counts using a regression approach for the occurrence of claims and their reporting delay. *Working paper*.

5.1 Introduction

Insurance companies need to hold sufficient reserves in order to be able to fulfill future liabilities with respect to outstanding claims. These reserves are a key factor on the liability side of the balance sheet of the insurance company. Accurate, reliable and stable reserving methods for a wide range of products and lines of business are crucial to safeguard solvency, stability and profitability. With the introduction of new regulatory guidelines for the European insurance business in the form of Solvency II, the insurance industry has regained interest in using more elaborate methodology to model future cash flows and meet regulators' increasing requirements. Insurance companies are strongly encouraged to supplement ad hoc, deterministic methods with fully stochastic approaches, aiming at accurately reflecting the riskiness in the portfolio under consideration.

The development of a single claim is visualized in the time line of Figure 5.1. A claim occurs at a certain occurrence date s_{occ}, consequently it is declared to the insurer at reporting date s_{rep} and one or several payments follow (at times s_1, s_2 and s_3) until the closure of the claim at settlement date s_{set}. Claims are not reported instantaneously to the insurer, but always after a certain reporting delay. This delay reflects the time gap between the occurrence of the claim and the reporting to the insurance company, which can for instance be due to the fact that the policyholder did not immediately notify his agent or only noticed the claim after a while. After notification, claims are also not settled immediately because it usually takes time to evaluate the whole size of the claim. Experts have to ascertain the loss or damage and the insured and the insurance company must come to an agreement. The settlement delay is sometimes further extended due to additional investigations or disputes which have to be settled in court. Intermediate payments of justified claim benefits are paid along the way which can lead to a sequence of multiple cash flows before final settlement.

When an insurance company closes its books, it needs to predict future cash flows of claims that have occurred in the past and are only settled in the future in order to set aside adequate premium reserves (see e.g. Wüthrich and Merz, 2008). This assessment of the outstanding loss liabilities of past claims is referred

to as *claims reserving*. At the present moment, when the reserve is calculated, say at date τ, a claim which has already occurred ($s_{\text{occ}} \leqslant \tau$) but has not yet been reported ($s_{\text{rep}} > \tau$) is called a *Incurred But Not Reported (IBNR)* claim. Between occurrence of the accident and notification to the insurance company, the insurer is unaware of the claim's existence but liable for the claim amount. A claim which has already been reported ($s_{\text{rep}} \leqslant \tau$) but has not yet been settled ($s_{\text{set}} > \tau$) is referred to as an *Reported But Not Settled (RBNS)* claim. Often, a distinction is made between the IBNR reserve and the RBNS reserve.

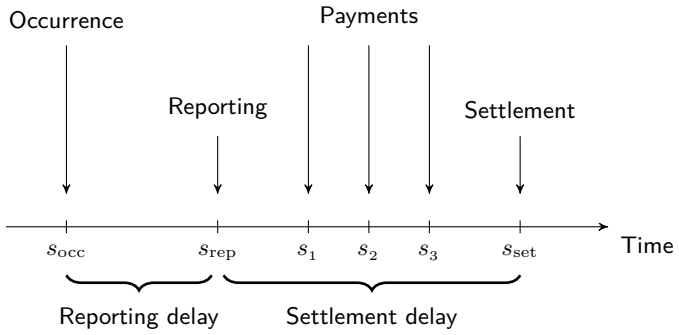

Figure 5.1: *Time line representing the development of a single claim.*

In this chapter, we analyze and model the arrival of claims together with the reporting delays. As such, we focus on the first part of the development of a claim in Figure 5.1, from occurrence until reporting, and not on the settlement delay and the claim payments. The goal is to obtain an accurate estimation of the number of IBNR claims based on the history of reported claims. This is an essential component to obtain a reliable estimate for the IBNR reserve.

Most existing methods for estimating the number of IBNR claims are designed for aggregated data, conveniently summarized in a so-called run-off triangle. A *run-off triangle* summarizes the reported claims by aggregating claim counts into an incomplete two-dimensional contingency table, representing the period of occurrence of the claim and the reporting period (where both periods are most often expressed in years). The industry-wide standard to estimate the future claim counts in the lower triangle is the chain-ladder model (Mack, 1993) and its related extensions. For an overview of this type of methods, see Taylor (2000); England and Verrall (2002); Wüthrich and Merz (2008).

Nowadays, insurance companies keep track of more detailed information, including the occurrence date and the reporting date of each individual claim. In

the so-called *macro-level* reserving techniques, such as the chain-ladder method, the available data is not fully used. Over the recent years, there has been increasing interest in *micro-level* reserving techniques, which make use of the insurance data on a more granular level. We briefly discuss a number of recent contributions from the actuarial science literature which use a micro-level approach to predict the number of IBNR claims.

Martínez Miranda et al. (2013) extend the traditional chain-ladder framework for the claim count data to a continuous chain-ladder setting. They reformulate the classical actuarial technique of chain-ladder as a histogram type of estimator and replacing this histogram by a two-dimensional kernel density estimator with support on the triangle. By assuming a multiplicative kernel, the local linear density estimate can be extrapolated to the whole square which provides a forecast for the IBNR claims in the lower triangle. The model can be applied to data recorded in continuous time, although it is illustrated in the paper on data aggregated on a monthly level.

Verrall and Wüthrich (2016) construct an inhomogeneous marked Poisson process to explicitly model the claims arrival process and reporting delay in continuous time based on individual claims data. The intensity of the Poisson process incorporates a weekly period piece-wise constant pattern and a monthly seasonal parameter. A spliced distribution with three layers (small, middle and large) is used for the reporting delay. Due to the delay in the reporting of claims, the marked Poisson process is thinned which complicates direct maximum likelihood estimation.

Badescu et al. (2016b,a) and Avanzi et al. (2016) propose to model the claim arrival process along with its reporting delays as a marked Cox process to allow for overdispersion and serial dependency. A Cox process, or doubly stochastic Poisson process, extends a Poisson process by modeling the intensity as a non-negative stochastic process.

Badescu et al. (2016b) use a weekly piecewise constant stochastic process generated by a hidden Markov model (HMM) with state-dependent Erlang distributions. The discrete process of the number of observed claims during each week then follows a Pascal-HMM with scale parameters depending on the exposure and the reporting delay distribution. Instead of joint estimation of all parameters, a two-stage method is applied. In a first stage, the reporting delay distribution is estimated using a mixture of Erlangs. Observable reporting delays are however right-truncated at different thresholds, which the fitting algorithm of Verbelen et al. (2015) is not able to handle. This is dealt with by extracting information

from the whole data set instead of only the training part, which is not possible in practice. In a second stage, the parameters of the Pascal-HMM are estimated using an *Expectation-Maximization (EM) algorithm* by plugging in time-varying scale parameters based on the fitted reporting delay distribution.

Avanzi et al. (2016) use a continuous time shot noise process to model the claim occurrence process, allowing for varying exposure and reporting delays. For parameter estimation, claim counts are no longer regarded in continuous time but discretized by week and, when calculating the reporting delay probabilities, it is assumed that the arrival time is the middle of the week. Joint estimation of all parameters relies on a complex Monte Carlo Expectation-Maximization (MCEM) algorithm with a Reversible Jump Markov Chain Monte Carlo (RJMCMC) filter since the likelihood of the Cox process unconditional on the shot noise process involves a high dimensional integral, which is not computationally efficient to calculate.

Beyond the field of actuarial science, similar statistical problems are also encountered in the research fields of biostatistics and epidemiology (see e.g. Harris, 1990; Lawless, 1994; Pagano et al., 1994; Midthune et al., 2005). For instance, when estimating the incidence of a disease, it is necessary to account for delays in the reporting of cases. Moreover, statistical surveillance systems for the timely detection of outbreaks of infectious disease have to properly adjust for these reporting delays in order to take timely preventive action (see e.g. Noufaily et al., 2015, 2016).

In our work, we present a new technique to estimate the number of events subject to a reporting delay by explicitly modeling both the occurrence process of the events and the reporting delay distribution using flexible regression approaches. We specifically focus on the case of IBNR claims in insurance, but our work can also be applied in the fields mentioned earlier. In practice, insurance companies register the occurrence date and the corresponding reporting date for each observed claim in their administrative systems, rather than the exact occurrence times or reporting times. We recognize this natural time unit of one day by constructing our models on this level instead of considering continuous time models (such as the Poisson or Cox processes) or aggregated versions by week, month or year (such as the traditional chain-ladder method).

A regression approach allows us to incorporate seasonal effects in both the occurrences of claims and the reporting delays. These effects can be caused by various time factors, such as the day of week, the day of month, or the month of the occurrence date, as well as relevant external information (if available), such

as economic conditions or expert-knowledge indicators which might impact the number of claims occurring or their corresponding reporting delays. The expected number of claims can also be made proportional to the *exposure* of the portfolio, which reflects the risk the insurer is taking on and is most often measured using the number of policyholders, the sum of the premiums, or the total sum insured.

We develop a novel estimation framework which allows for a joint estimation of both the occurrence and the reporting delay model parameters. The key is to treat the complexity of observing only reported claims due to reporting delays as a missing data problem and to use the EM algorithm to simplify the estimation significantly. Our estimation approach can be used more broadly and can be applied, for instance, to the setting of Badescu et al. (2016a) or Verrall and Wüthrich (2016). Its main advantage is that it avoids the use of ad hoc methods or two-step approaches to adjust for the reporting delay.

5.2 Data and first insights

We demonstrate our methodology using the data from Antonio and Plat (2014) on a portfolio of general liability insurance policies for private individuals from a European insurance company. This data set has also been studied in Pigeon et al. (2013), Pigeon et al. (2014), Godecharle and Antonio (2015) and Antonio et al. (2016). Detailed claims information is available from January 1997 until August 2009. This includes the occurrence date of a claim and the time between occurrence and notification to the insurance company. Claims are also categorized into bodily injury or material damage claims, although we will not make a distinction between both.

As a measure for the exposure to risk, the main driver underlying the occurrences of claims, we use the number of policies. This is available by month from January 2000 onwards. Exposure is expressed as earned exposure, i.e. the exposure units actually exposed to risk during the period. This means that a policy covered during the whole month of January will contribute 31/365th to the exposure of that month, 10/365th if it is only covered during 10 days, and so on. Earned exposure is not available on a daily level so instead we transform the monthly exposure to daily exposure by diving by the number of days in each month. Figure 5.2a shows the resulting exposure per day which is an increasing stepwise function, indicating an increase in the portfolio size over time. Since exposure information is only available from January 2000 onwards and to enable out-of-sample prediction, we restrict our analysis to claims that have occurred

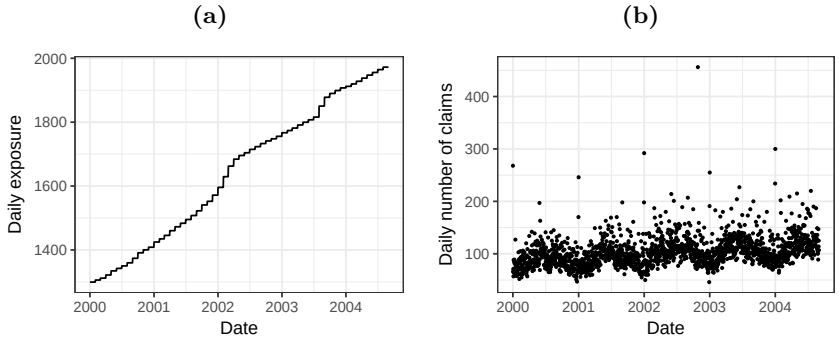

Figure 5.2: *From January 1, 2000 until August 31, 2004, we plot (a) the earned exposure per day and (b) the number of claims occurring on that day based on the full data set until August 2009.*

between January 1, 2000 and August 31, 2004.

We assume that at the end of this time window, on August 31, 2004, which will be referred to as the evaluation date, the insurance company has to set capital aside to cover for future payments related to the reported claims as well as to IBNR claims. This requires an estimate of the total number of IBNR claim counts, as well as their timing of reporting, as a building block to model such future cash flows. Based on the full data set until August 2009, we know that during this time frame 176 919 claims have occurred which are plotted by their occurrence date in Figure 5.2b. The graph shows a clear seasonal pattern, but also contains many days with an outlying high number of occurred claims. A large amount of these dates correspond to the 1st or the 15th of the month. The highest claim counts have occurred on October 27, 2002, due to a major storm, and on January 1st of each year. However, since claims are not immediately notified to the insurer, only 174 867 of these have been reported by the evaluation date, as depicted in blue the daily run-off triangle in Figure 5.3. The remaining 2052 are referred to as IBNR claims, i.e. claims which have occurred between January 2000 and August 2004 but have only been reported after the moment of evaluation (and before August 31, 2009). These are graphically illustrated in red in Figure 5.3.

An accurate estimation of IBNR claim counts requires an understanding and modeling of the reporting delay distribution. Due to the reporting delay, only a portion of the occurred claims is observed. If a claim occurs on a certain date t and the evaluation date is τ, then the claim is only observed if the reporting delay is smaller than or equal to $\tau - t$ days. In statistical terminology, we call the total number of claims occurring on day t right censored and the reporting delay

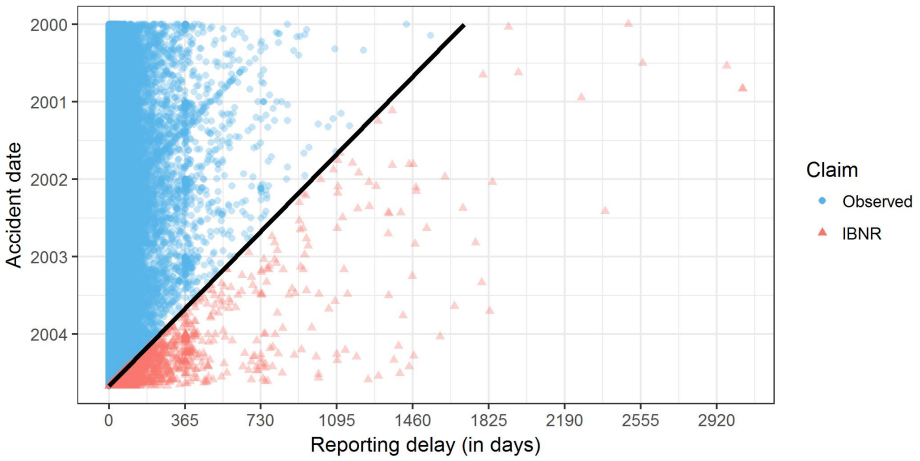

Figure 5.3: *Daily run-off triangle of claims with occurrence dates between January 1, 2000 and August 31, 2004. The black line indicates the evaluation date, August 31, 2004. Only the claims in the upper triangle depicted as blue dots are observed at the evaluation date. The remaining claims in the lower triangle depicted as red triangles are the IBNR claims based on the full data set until August 2009 and have to be predicted.*

distribution of the observed claims occurring on day t right-truncated at $\tau - t$ (see e.g. Klein and Moeschberger, 2003). Special care has to be taken in the analysis of these data, see Section 5.3.3. In the remainder of this section, we analyze the empirical distribution of the reporting delay. For now, in this exploratory analysis only, we circumvent the issue of right-truncation by extracting the reporting delays corresponding to claims that occurred between January 2000 and August 2004 and have been reported before August 2009, hence 5 years after the evaluation date, from the full data set.

In Figure 5.4, we show a bar plot of the empirical probability mass function of the reporting delay, limited to the first 4 weeks, for claims which occurred on a Monday in graph (a), a Thursday in graph (b) and a Saturday in graph (c). Note that a reporting delay of zero corresponds with a reporting on the day of occurrence. These graphs reveal two important features of the reporting delay distribution: a weekly declining pattern and a daily pattern within each week which depends on the day of the week of the occurrence date of the claim. First, we notice how the majority of claims is reported in the first couple of weeks after occurrence and the reporting delay probabilities decrease from one week to the

next. Second, from the second week onwards, the reporting delay probabilities are decreasing from the first working day of the reporting week (wday1) to the last working day (wday5) and are close to zero during the weekend. The ordering of these 7 days within the reporting week depends on the day of the week of the occurrence date of the claim as shown in Table 5.1. Only a small portion of claims are being reported on Saturdays and nearly none on Sundays. In fact, in the entire observed portion of the data, only 3 claims have been reported on Sunday.

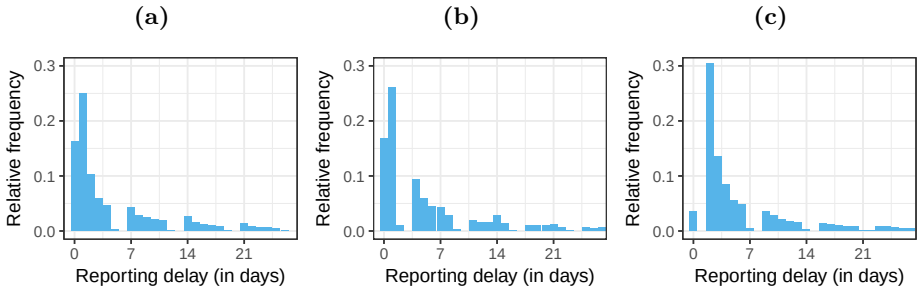

Figure 5.4: *Bar plot of the empirical reporting delay distribution in the first 4 weeks for claims that occurred on (a) Monday, (b) Thursday and (c) Saturday between January 2000 and August 2004 and have been reported before August 2009.*

In order to capture these phenomena, we model in the next section the reporting delay probabilities in weeks separately from the day probabilities. Reporting delay in weeks refers to the number of weeks that elapses between occurrence and reporting of the claim. A reporting delay in weeks equal to zero hereby implies that the claim is reported within the first week after its occurrence. The empirical reporting delay distribution in weeks can be well represented by a negative binomial distribution as is shown in Figures 5.5a and 5.5b. The reporting day probabilities model on which day a claim is reported within a given reporting week. The empirical day probabilities during the first reporting week are visualized in Figure 5.6a grouped by the day of the week of the occurrence date. From Monday to Thursday, the day probability in the first reporting week is highest on wday2, corresponding to one day after the claim occurred. For a Friday, the probability to report on the same day (wday1) is about as high as the probability to report on Monday after the weekend (wday2). For Saturday and Sunday, the day probability in the first reporting week is highest on wday1, corresponding to

Monday after the weekend. Since, on average, over 60% of the claims are reported in the first week (see Figure 5.5), the day probabilities during the first reporting week will be modeled separately for each day of the week of the occurrence date. From the second reporting week onwards, these probabilities behave very similarly within each reporting week, which is why we combine them in Figure 5.6b. The day probabilities also become comparable for each occurrence day of the week.

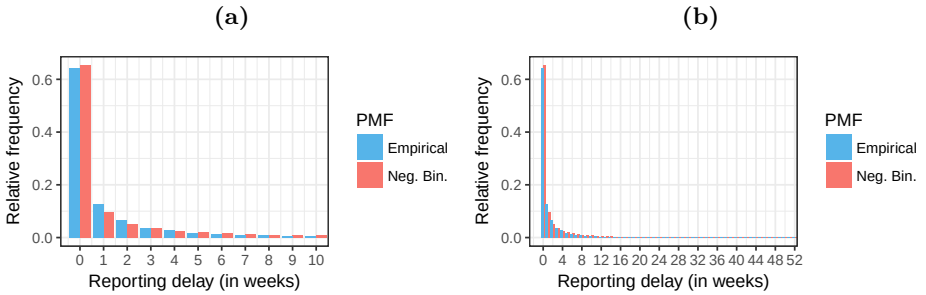

Figure 5.5: *Bar plot of the empirical reporting delay distribution in weeks and its negative binomial fit for the first 11 weeks in (a) and for the first year in (b) based on claims that occurred between January 2000 and August 2004 and have been reported before August 2009.*

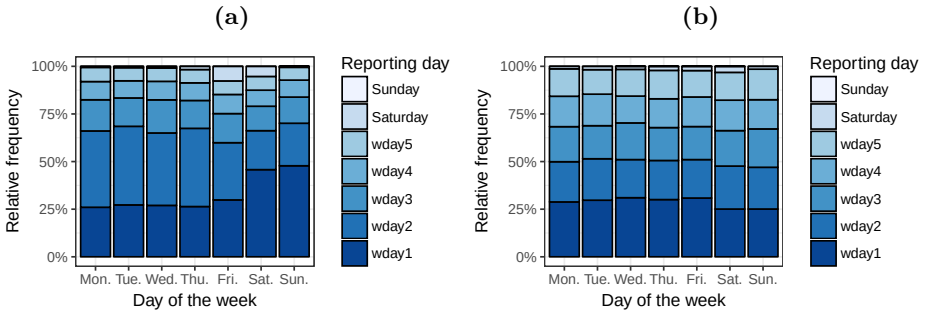

Figure 5.6: *Stacked bar plots of the empirical reporting delay day probabilities within a reporting week according to the day of the week of the occurrence date. Based on claims that occurred between January 2000 and August 2004 and have been reported before August 2009, we show the empirical day probabilities during the first reporting week in (a) and from the second reporting week onwards in (b). The ordering of the working days in a reporting week according to the day of the week of the occurrence date is clarified in Table 5.1.*

Table 5.1: *Ordering of the working days in the week (wday) by the day of the week (dow) of the occurrence date. wday3, for example, denotes the third working day of the reporting week, which is Wednesday when the claim occurred on Monday and a Monday when the claim occurred on Thursday, and so on.*

	wday						
dow	wday1	wday2	wday3	wday4	wday5	Saturday	Sunday
Monday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Tuesday	Tuesday	Wednesday	Thursday	Friday	Monday	Saturday	Sunday
Wednesday	Wednesday	Thursday	Friday	Monday	Tuesday	Saturday	Sunday
Thursday	Thursday	Friday	Monday	Tuesday	Wednesday	Saturday	Sunday
Friday	Friday	Monday	Tuesday	Wednesday	Thursday	Saturday	Sunday
Saturday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday

5.3 The statistical model

5.3.1 Daily claim count data

We model insurance claim counts on a daily level and denote the total number of claims which occurred on day t by N_t, where the integer t indicates the occurrence date and ranges from 1 to τ. The number of these claims which have been reported to the insurer after d days are denoted as N_{td} such that

$$N_t = \sum_{d=0}^{\infty} N_{td} \,.$$

Due to this reporting delay, only part of these claims have been reported to the insurer before or at the moment of evaluation, τ. Namely only those claims which have a reporting delay smaller than or equal to $\tau - t$. We denote the observed number of claims which occurred on day t by

$$N_t^{\mathrm{R}} = \sum_{d=0}^{\tau-t} N_{td} \,.$$

Only the observed claims $\boldsymbol{N}^{\mathrm{R}} = \{N_{td} \mid 1 \leqslant t \leqslant \tau, d \geqslant 0, t+d \leqslant \tau\}$ can be used on the moment of evaluation when the outstanding claims liabilities have to be calculated. These can be represented in a daily run-off triangle as shown in Table 5.2. The occurrence date is indicated in the rows and the reporting delay in the columns. Claim counts on the diagonal for which $t+d$ is constant are all reported

on the same calendar day $t+d$, with τ being the last calendar day observed. The objective is to predict the IBNR claim counts $\boldsymbol{N}^{\text{IBNR}} = \{N_{td} \mid 1 \leqslant t \leqslant \tau, d > 0, t+d > \tau\}$ in the lower part of the daily claims triangle in Table 5.2. We denote the total IBNR claim counts per day by

$$N_t^{\text{IBNR}} = \sum_{d=\tau-t+1}^{\infty} N_{td}\,,$$

and the total IBNR claim count over all occurrence days by

$$N^{\text{IBNR}} = \sum_{t=1}^{\tau} N_t^{\text{IBNR}} = \sum_{t=1}^{\tau} \sum_{d=\tau-t+1}^{\infty} N_{td}\,.$$

Table 5.2: *Run-off triangle with daily claim counts. Only the claim counts in the upper triangle are observed, whereas the claim counts in the lower triangle have to be predicted.*

Occurrence day	Reporting delay (in days)				
	0	$\cdots$	$\tau - t$	$\cdots$	$\tau - 1$
1	N_{10}	$\cdots$	$N_{1,\tau-t}$	$\cdots$	$N_{1,\tau-1}$
$\vdots$					
t	N_{t0}	$\cdots$	$N_{t,\tau-t}$		
$\vdots$				IBNR	
τ	$N_{\tau 0}$				

5.3.2 Model assumptions

The statistical analysis of the daily claim counts using our proposed model is based on the following two assumptions:

(A1) The daily total claim counts N_t for $t = 1, \ldots, \tau$ are independently Poisson distributed with intensity $\lambda_t = e_t \exp(\boldsymbol{x}_t'\boldsymbol{\alpha})$, where e_t is the exposure, $\boldsymbol{x}_t$ is the vector of covariate information corresponding to occurrence day t and $\boldsymbol{\alpha}$ is a parameter vector.

(A2) Conditional on N_t, the claim counts N_{td} for $d = 0, 1, 2, \ldots$, are multinomially distributed with probabilities p_{td}. These reporting delay probabilities are

structured as a product of week probabilities and day probabilities:

$$p_{td} = \begin{cases} p^{\mathrm{W}}_{t0} \cdot p^{1}_{td} & \text{for } d < 7 \\ p^{\mathrm{W}}_{t\left\lfloor \frac{d}{7} \right\rfloor} \cdot p^{2}_{td} & \text{otherwise.} \end{cases}$$

The reporting delay week probabilities,

$$p^{\mathrm{W}}_{tw} = \frac{\Gamma(\phi + w)}{w!\Gamma(\phi)} \frac{\phi^{\phi} \mu_t^{w}}{(\phi + \mu_t)^{\phi + w}} \quad \text{for } w = 0, 1, 2, \dots \,, \tag{5.1}$$

are modeled using the probability mass function of a negative binomial distribution with expected value $\mu_t = \exp(\boldsymbol{z}_t' \boldsymbol{\beta})$ and variance $\mu_t + \mu_t^2/\phi$, where ϕ is the dispersion parameter and $\boldsymbol{z}_t$ is the covariate vector corresponding to occurrence day t. The reporting delay day probabilities in the first week can be written in a symbolical way as

$$p^{1}_{td} = \mathbf{P}^{1}(\mathsf{dow}(t), \mathsf{wday}(t, t + d)) \,, \tag{5.2}$$

where $\mathsf{dow}(t)$ denotes the day of the week of occurrence date t and $\mathsf{wday}(t, t+d)$ denotes the working day of the week of the reporting date $t + d$, given that the corresponding occurrence date is t. $\mathbf{P}^1$ is a 7×7-matrix which has rows and columns as in Table 5.1 and contains the day probabilities related to the first week. Each element in $\mathbf{P}^1$ is between 0 and 1 and all row sums equal 1. Similarly, the reporting delay day probabilities from the second week onwards are given by

$$p^{2}_{td} = \mathbf{P}^{2}(\mathsf{wday}(t, t + d)) \,, \tag{5.3}$$

where $\mathbf{P}^2$ is a 1×7-matrix which has columns as in Table 5.1 and elements between 0 and 1 that sum up to 1.

Allowing for covariates in the model for the occurrences of claims as well as the model for reporting delays in weeks allows us to build flexible models. The expected number of claim occurrences can be made proportional to the exposure or depend on several measures of exposure. Evolutions over time or seasonal trends can be captured to improve forecast predictions. Fluctuations in both claim counts and their reporting delays by month, day of the month or day of the week of the occurrence date can be explicitly modeled. Additionally, an insurer can also model relationships with external covariates which might influence the

arrival process of claims. Potential effects might be plausible for economic circumstances, business cycles and weather conditions. On top of that, the day-specific particularities in the reporting delay we noticed in Section 5.1 and displayed in Figure 5.6 are captured using designated day probabilities.

5.3.3 Parameter estimation using the EM algorithm

We bundle the parameters to be estimated in $\boldsymbol{\Theta} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \phi, \mathbf{P}^1, \mathbf{P}^2\}$. Based on the assumptions in Section 5.3.2, the daily claim counts N_{td} are independently Poisson distributed with intensities $\lambda_t p_{td}$ for $t = 1, \ldots, \tau$ and $d = 0, 1, 2, \ldots$. This can be seen by writing the joint probability of $N_{td} = n_{td}$ for $d = 0, 1, 2, \ldots$ as the product of the probability of their sum N_t being equal to $n_t = \sum_{d=0}^{\infty} n_{td}$ and the conditional multinomial probability:

$$\begin{aligned}
&\mathbb{P}(N_{t0} = n_{t0}, N_{t1} = n_{t1}, N_{t2} = n_{t2}, \ldots) \\
&\quad = \mathbb{P}(N_t = n_t) \cdot \mathbb{P}(N_{t0} = n_{t0}, N_{t1} = n_{t1}, N_{t2} = n_{t2}, \ldots \mid N_t = n_t) \\
&\quad = \frac{\exp(-\lambda_t)\lambda_t^{n_t}}{n_t!} \cdot \frac{n_t!}{\prod_{d=0}^{\infty} n_{td}!} \cdot \prod_{d=0}^{\infty} (p_{td})^{n_{td}} \\
&\quad = \prod_{d=0}^{\infty} \frac{\exp(-\lambda_t p_{td})(\lambda_t p_{td})^{n_{td}}}{n_{td}!},
\end{aligned}$$

which factorizes into Poisson probabilities. This property is sometimes referred to as the thinning property of Poisson random variables. In particular, the observed claim count N_t^{R} on day t is Poisson distributed with intensity $\lambda_t p_t^{\mathrm{R}}$ where $p_t^{\mathrm{R}} = \sum_{d=0}^{\tau-t} p_{td}$ and the IBNR claim count N_t^{IBNR} is Poisson distributed with intensity $\lambda_t p_t^{\mathrm{IBNR}}$ where $p_t^{\mathrm{IBNR}} = 1 - p_t^{\mathrm{R}}$. Conditional on N_t^{R}, the observed daily claim counts $\{N_{td} \mid d = 0, 1, \ldots, \tau - t\}$ are multinomially distributed with parameters N_t^{R} and $\{p_{td}/p_t^{\mathrm{R}} \mid d = 0, 1, \ldots, \tau - t\}$, since we have to account for the right-truncation of the reporting delay.

The likelihood of the observed upper run-off triangle for our chosen model can then be written as the product of a Poisson likelihood and a multinomial likelihood,

$$\mathcal{L}(\boldsymbol{\Theta}; \boldsymbol{N}^{\mathrm{R}}) = \prod_{t=1}^{\tau} \frac{\exp(-\lambda_t p_t^{\mathrm{R}}) \left(\lambda_t p_t^{\mathrm{R}}\right)^{N_t^{\mathrm{R}}}}{N_t^{\mathrm{R}}!} \frac{N_t^{\mathrm{R}}!}{\prod_{d=0}^{\tau-t} N_{td}!} \prod_{d=0}^{\tau-t} \left(\frac{p_{td}}{p_t^{\mathrm{R}}}\right)^{N_{td}}. \tag{5.4}$$

Equivalently, the likelihood can also be constructed by treating the daily claim

counts as right censored, since the number of IBNR claims is unknown,

$$\mathcal{L}(\boldsymbol{\Theta}; \boldsymbol{N}^{\mathrm{R}}) = \prod_{t=1}^{\tau} \sum_{n=N_t^{\mathrm{R}}}^{\infty} \frac{\exp(-\lambda_t)\lambda_t^n}{n!} \cdot \frac{n!}{\prod_{d=0}^{\tau-t} N_{td}! \cdot (n - N_t^{\mathrm{R}})!} \cdot \prod_{d=0}^{\tau-t} (p_{td})^{N_{td}} \cdot \left(p_t^{\mathrm{IBNR}}\right)^{n-N_t^{\mathrm{R}}} .$$

Indeed, this expression reduces to (5.4) by rewriting the sum over n using the Taylor expansion for the exponential function. The corresponding log-likelihood equals

$$\log \mathcal{L}(\boldsymbol{\Theta}; \boldsymbol{N}^{\mathrm{R}}) = -\sum_{t=1}^{\tau} \lambda_t p_t^{\mathrm{R}} + \sum_{t=1}^{\tau} N_t^{\mathrm{R}} \log(\lambda_t) + \sum_{t=1}^{\tau} \sum_{d=0}^{\tau-t} N_{td} \log(p_{td}) - \sum_{t=1}^{\tau} \sum_{d=0}^{\tau-t} \log(N_{td}!) . \tag{5.5}$$

Note that, due to the right truncation of the reporting delay (or, the right censoring of the claim counts), the log-likelihood (5.5) contains terms which depend on the parameters of both the Poisson model for claim occurrences and the reporting delay distribution. This complicates direct maximum likelihood estimation as it prevents separate optimization with respect to each of these parameter blocks. Optimization using a standard numerical method such as Newton-Raphson is still feasible, but we cannot rely on statistical software packages and we need to derive the analytical expressions of the gradient and Hessian of the log-likelihood (5.5). To simplify computations, shortcuts have been used to estimate parameters in related works, such as plug-in estimates for the weekly periodic occurrence pattern in Verrall and Wüthrich (2016) or a two-stage method in which the reporting delay distribution is estimated first and then plugged in to estimate the parameters related to the occurrence process (Antonio and Plat, 2014; Badescu et al., 2016a).

Instead, we choose to treat the truncation as a missing data problem and employ the EM algorithm (Dempster et al., 1977; McLachlan and Krishnan, 2008) to simplify maximum likelihood parameter estimation. Consider the complete version of the data $\boldsymbol{N} = \boldsymbol{N}^{\mathrm{R}} \cup \boldsymbol{N}^{\mathrm{IBNR}} = \{N_{td} \mid 1 \leqslant t \leqslant \tau, d \geqslant 0\}$ which augments the observed daily claim counts from the upper part of the run-off triangle in Table 5.2 with the unknown values of the future claim counts in the lower triangle. Given

the complete data $\boldsymbol{N}$, we can construct the complete log-likelihood function

$$\begin{aligned}\log \mathcal{L}_c(\boldsymbol{\Theta}; \boldsymbol{N}) = &- \sum_{t=1}^{\tau} \lambda_t + \sum_{t=1}^{\tau} N_t \log(\lambda_t) \\ &+ \sum_{t=1}^{\tau} \sum_{d=0}^{\infty} N_{td} \log(p_{td}) - \sum_{t=1}^{\tau} \sum_{d=0}^{\infty} \log(N_{td}!)\,. \end{aligned} \tag{5.6}$$

which allows for a separate estimation of the parameters of the claim occurrence model (appearing in λ_t) and those of the reporting delay distribution (appearing in p_{td}). The observed data log-likelihood (5.5) can be maximized by iteratively maximizing the complete data log-likelihood (5.6) using the EM algorithm. However, as we do not fully observe the complete data, the complete log-likelihood is a random variable. Therefore, it is not possible to directly optimize (5.6). Yet, the EM algorithm exploits the simpler form of the complete log-likelihood by iterating between an E-step or expectation step and M-step or maximization step. Applied to our setting, the IBNR claim counts in the lower triangle of Table 5.2 are replaced by their expected values in the E-step and the log-likelihood of the augmented data is maximized in the M-step. The M-step will still require numerical optimization, but the parameters with respect to the claim occurrence model can be estimated separately from the reporting delay parameters and standard software routines can be utilized in the absence of truncation. We discuss these steps in more detail.

E-step In the kth iteration of the E-step, we take the conditional expectation of the complete log-likelihood (5.5) given the incomplete data $\boldsymbol{N}^{\mathrm{R}}$ and using the current estimate $\boldsymbol{\Theta}^{(k-1)}$ of the parameter vector $\boldsymbol{\Theta}$:

$$Q(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(k-1)}) = \mathbb{E}\left(\log \mathcal{L}_c(\boldsymbol{\Theta}; \boldsymbol{N}) \mid \boldsymbol{N}^{\mathrm{R}}; \boldsymbol{\Theta}^{(k-1)}\right)\,. \tag{5.7}$$

This requires us to compute the expected values of the future claim counts

$$N_{td}^{(k)} = \mathbb{E}\left[N_{td} \mid \boldsymbol{N}^{\mathrm{R}}; \boldsymbol{\Theta}^{(k-1)}\right] = \begin{cases} N_{td} & \text{if } d \leqslant \tau - t \\ \lambda_t^{(k-1)} p_{td}^{(k-1)} & \text{otherwise,} \end{cases} \tag{5.8}$$

for $t = 1, \ldots, \tau$ and the total daily claim counts $N_t^{(k)} = \sum_{d=0}^{\tau-t} N_{td} + \sum_{d=\tau-t+1}^{\infty} N_{td}^{(k)}$. The terms in (5.7) containing $E\left(\log(N_{td}!) \mid \boldsymbol{N}^{\mathrm{R}}; \boldsymbol{\Theta}^{(k-1)}\right)$ do not play a role in the EM algorithm as they do not depend on the unknown parameter vector $\boldsymbol{\Theta}$.

M-step In the kth iteration of the M-step, we maximize the expected value (5.7) of the complete data log-likelihood obtained in the E-step with respect to the parameter vector $\boldsymbol{\Theta}$. In order to optimize (5.7) with respect to $\boldsymbol{\alpha}$ as defined in model assumption (A1), we have to maximize the terms related to the claim occurrence model,

$$-\sum_{t=1}^{\tau}\lambda_t+\sum_{t=1}^{\tau}N_t^{(k)}\log(\lambda_t)=-\sum_{t=1}^{\tau}e_t\exp(\boldsymbol{x}_t'\boldsymbol{\alpha})+\sum_{t=1}^{\tau}N_t^{(k)}(\log(e_t)+\boldsymbol{x}_t'\boldsymbol{\alpha})\,, \quad (5.9)$$

which is a weighted Poisson log-likelihood with an offset term (related to the exposure). The parameter values optimizing (5.9) are denoted by $\boldsymbol{\alpha}^{(k)}$. Based on model assumption (A2), updating the estimates for the parameters of the reporting delay distribution requires the maximization of

$$\sum_{t=1}^{\tau}\sum_{d=0}^{\infty}N_{td}^{(k)}\log(p_{td})=\sum_{t=1}^{\tau}\left(\sum_{d=0}^{\infty}N_{td}^{(k)}\log\left(p^{\mathrm{W}}_{t\left\lfloor\frac{d}{7}\right\rfloor}\right)+\sum_{d=0}^{6}N_{td}^{(k)}\log(p_{td}^{1})\right.$$
$$\left.+\sum_{d=0}^{6}N_{td}^{(k)}\log(p_{td}^{1})+\sum_{d=7}^{\infty}N_{td}^{(k)}\log(p_{td}^{2})\right)\,.$$

From a numerical point of view, we truncate the infinite sums over the reporting delay d at $d=\tau-1$, which corresponds to completing the run-off triangle in Table 5.2 without further extending it. Numerical experiments have shown that this choice is sufficiently high in our setting as the subsequent terms are negligible. The new parameter estimates $\boldsymbol{\beta}^{(k)}$ and $\phi^{(k)}$ of the negative binomial distribution for the reporting delay in weeks are found by optimizing the weighted negative binomial log-likelihood,

$$\sum_{t=1}^{\tau}\sum_{d=0}^{\infty}N_{td}^{(k)}\log\left(p^{\mathrm{W}}_{t\left\lfloor\frac{d}{7}\right\rfloor}\right)=\sum_{t=1}^{\tau}\sum_{w=0}^{\infty}\left(\sum_{d=0}^{6}N_{t,7w+d}^{(k)}\right)\log\left(p^{\mathrm{W}}_{tw}\right)\,, \quad (5.10)$$

where p^{W}_{tw} is given in (5.1) with $\mu_t=\exp(\boldsymbol{z}_t'\boldsymbol{\beta})$. Both for optimizing (5.9) and (5.10), standard software packages can be used. Optimizing (5.7) with respect to the day probabilities (5.2) and (5.3) and under the restriction that the sums of the rows (the sums over the working days of the week) equal 1 leads to

$$\left(\mathbf{P}^1(u,v)\right)^{(k)}=\frac{\sum_{t=1,\ldots,\tau;\ \mathsf{dow}(t)=u}\ \sum_{d=0,\ldots,6;\ \mathsf{wday}(t,t+d)=v}N_{td}^{(k)}}{\sum_{t=1,\ldots,\tau;\ \mathsf{dow}(t)=u}\sum_{d=0}^{6}N_{td}^{(k)}}\,,$$

$u=\mathsf{Monday},\ldots,\mathsf{Sunday}$
$v=\mathsf{wday1},\ldots,\mathsf{Sunday}$

and

$$\left(\mathbf{P}^2(v)\right)^{(k)} = \frac{\sum_{t=1}^{\tau} \sum_{\substack{d=7,\ldots,\infty \\ \mathsf{wday}(t,t+d)=v}} N_{td}^{(k)}}{\sum_{t=1}^{\tau} \sum_{d=7}^{\infty} N_{td}^{(k)}}, \qquad v = \mathsf{wday1}, \ldots, \mathsf{Sunday}.$$

Initial step The first E-step of the EM algorithm with $k = 1$ requires a starting value $\boldsymbol{\Theta}^{(0)}$ for the parameter set. Our strategy is to first apply the chain-ladder method on the daily claim counts to obtain initial predictions $N_{td}^{(0)}$ of the future claim counts in the lower triangle of Table 5.2. Then, we initialize $\boldsymbol{\Theta}$ by applying an initial M-step based on these initial claim count estimates. More specifically, we define the cumulative claim counts as

$$C_{td} = \sum_{j=0}^{d} N_{tj} \quad \text{for } t = 1, \ldots, \tau, \text{and } d = 0, \ldots, \tau - 1,$$

and estimate the development factors of the chain-ladder technique on a daily level as

$$\widehat{f}_d = \frac{\sum_{t=1}^{\tau-d} C_{td}}{\sum_{t=1}^{\tau-d} C_{t,d-1}} \quad \text{for } d = 1, \ldots, \tau - 1.$$

The chain-ladder technique applies these development factors to the latest cumulative claim count in each row to produce forecasts of future cumulative claim counts:

$$\widehat{C}_{t,d} = C_{t,\tau-t}\widehat{f}_{\tau-t+1} \ldots \widehat{f}_d \quad \text{for } t = 2, \ldots, \tau, \text{and } d = \tau - t + 1, \ldots, \tau - 1.$$

We use these chain-ladder estimates for the daily cumulative claim counts to initialize the expected incremental claim counts as

$$N_{td}^{(0)} = \begin{cases} N_{td} & \text{if } d \leqslant \tau - t \\ \widehat{C}_{t,\tau-t+1} - C_{t,\tau-t} & \text{if } d = \tau - t + 1 \\ \widehat{C}_{td} - \widehat{C}_{t,d-1} & \text{otherwise,} \end{cases}$$

for $t = 1, \ldots, \tau$ and apply an M-step, as outlined above with $k = 0$, to find decent starting values $\boldsymbol{\Theta}^{(0)}$.

Convergence The log-likelihood (5.5) increases with each EM iteration (McLachlan and Krishnan, 2008). Given proper starting values, the sequence $\boldsymbol{\Theta}^{(k)}$ converges to the maximum likelihood estimate (MLE) of $\boldsymbol{\Theta}$ corresponding to the

(incomplete data) log-likelihood $\log \mathcal{L}(\boldsymbol{\Theta}; \boldsymbol{N}^{\mathrm{R}})$ in (5.5). The stopping criterion we apply is based on the relative change in the log-likelihood. Namely, we iterate until the absolute value of

$$\frac{\log \mathcal{L}(\boldsymbol{\Theta}^{(k)}; \boldsymbol{N}^{\mathrm{R}}) - \log \mathcal{L}(\boldsymbol{\Theta}^{(k-1)}; \boldsymbol{N}^{\mathrm{R}})}{0.1 + \log \mathcal{L}(\boldsymbol{\Theta}^{(k)}; \boldsymbol{N}^{\mathrm{R}})}$$

becomes sufficiently small. The parameter vector estimate upon convergence is denoted by

$$\widehat{\boldsymbol{\Theta}} = \{\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}, \widehat{\phi}, \widehat{\mathbf{P}^1}, \widehat{\mathbf{P}^2}\}\,.$$

5.3.4 Asymptotic variance-covariance matrix

The estimator for Θ obtained from the EM algorithm has the same limit as the MLE, whenever the starting value is adequately chosen. Hence, the maximum likelihood asymptotic theory in terms of consistency, asymptotic normality and asymptotic efficiency applies. In particular, if we denote the (incomplete data) score statistic as

$$\boldsymbol{S}(\boldsymbol{\Theta}; \boldsymbol{N}^{\mathrm{R}}) = \frac{\partial}{\partial \boldsymbol{\Theta}} \log \mathcal{L}(\boldsymbol{\Theta}; \boldsymbol{N}^{\mathrm{R}})\,,$$

and the (incomplete data) observed information matrix

$$\boldsymbol{I}(\boldsymbol{\Theta}; \boldsymbol{N}^{\mathrm{R}}) = -\frac{\partial^2}{\partial \boldsymbol{\Theta} \partial \boldsymbol{\Theta}'} \log \mathcal{L}(\boldsymbol{\Theta}; \boldsymbol{N}^{\mathrm{R}})\,,$$

then the asymptotic variance-covariance matrix of the MLE $\widehat{\boldsymbol{\Theta}}$ is equal to the inverse of the (incomplete data) expected (Fisher) information matrix $\boldsymbol{\mathcal{I}}(\boldsymbol{\Theta})$ given by

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\Theta}) = \mathbb{E}\left[\boldsymbol{I}(\boldsymbol{\Theta}; \boldsymbol{N}^{\mathrm{R}}) \mid \boldsymbol{\Theta}\right]\,. \tag{5.11}$$

The asymptotic variance-covariance matrix can be approximated by $\boldsymbol{\mathcal{I}}^{-1}(\widehat{\boldsymbol{\Theta}})$. It is also common practice to estimate this matrix using the inverse of the observed information matrix evaluated at $\boldsymbol{\Theta} = \widehat{\boldsymbol{\Theta}}$, i.e. $\boldsymbol{I}^{-1}(\widehat{\boldsymbol{\Theta}}; \boldsymbol{N}^{\mathrm{R}})$. This matrix is produced as a by-product when applying Newton-Raphson's method.

When the parameters are estimated using the EM algorithm, the observed information matrix is however not directly accessible. Moreover, the main reason why the EM algorithm is chosen over Newton-Raphson's method is because it avoids the computation of the first- and second-order partial derivatives of the incomplete data log-likelihood. In case an estimate of the covariance matrix of the MLE is required, Louis (1982) showed how the observed information matrix

can be expressed in terms of the gradient and curvature of the complete data log-likelihood function. For this purpose, we introduce the complete data score statistic

$$\boldsymbol{S}_c(\boldsymbol{\Theta}; \boldsymbol{N}) = \frac{\partial}{\partial \boldsymbol{\Theta}} \log \mathcal{L}_c(\boldsymbol{\Theta}; \boldsymbol{N}) \,,$$

and we let

$$\boldsymbol{I}_c(\boldsymbol{\Theta}; \boldsymbol{N}) = -\frac{\partial^2}{\partial \boldsymbol{\Theta} \partial \boldsymbol{\Theta}'} \log \mathcal{L}_c(\boldsymbol{\Theta}; \boldsymbol{N}) \,,$$

with its conditional expectation given $\boldsymbol{N}^{\mathrm{R}}$ denoted by

$$\boldsymbol{\mathcal{I}}_c(\boldsymbol{\Theta}; \boldsymbol{N}^{\mathrm{R}}) = \mathbb{E}\left[\boldsymbol{I}_c(\boldsymbol{\Theta}; \boldsymbol{N}) \mid \boldsymbol{N}^{\mathrm{R}}; \boldsymbol{\Theta} \right] . \tag{5.12}$$

The complete data expected information matrix is then given by

$$\boldsymbol{\mathcal{I}}_c(\boldsymbol{\Theta}) = \mathbb{E}\left[\boldsymbol{I}_c(\boldsymbol{\Theta}; \boldsymbol{N}) \mid \boldsymbol{\Theta} \right] .$$

The missing information principle writes the observed information (5.11) as the (conditional expected) complete information (5.12) minus the missing information,

$$\boldsymbol{I}(\boldsymbol{\Theta}; \boldsymbol{N}^{\mathrm{R}}) = \boldsymbol{\mathcal{I}}_c(\boldsymbol{\Theta}; \boldsymbol{N}^{\mathrm{R}}) - \boldsymbol{\mathcal{I}}_m(\boldsymbol{\Theta}; \boldsymbol{N}^{\mathrm{R}}) \tag{5.13}$$

where

$$\boldsymbol{\mathcal{I}}_m(\boldsymbol{\Theta}; \boldsymbol{N}^{\mathrm{R}}) = -\mathbb{E}\left[\frac{\partial^2}{\partial \boldsymbol{\Theta} \partial \boldsymbol{\Theta}'} \log \frac{\mathcal{L}_c(\boldsymbol{\Theta}; \boldsymbol{N})}{\mathcal{L}(\boldsymbol{\Theta}; \boldsymbol{N}^{\mathrm{R}})} \,\middle|\, \boldsymbol{N}^{\mathrm{R}}; \boldsymbol{\Theta} \right]$$

denotes the missing information matrix. Louis (1982) derived that the missing information matrix can be computed as

$$\begin{aligned} \boldsymbol{\mathcal{I}}_m(\boldsymbol{\Theta}; \boldsymbol{N}^{\mathrm{R}}) &= \mathbb{E}\left[\boldsymbol{S}_c(\boldsymbol{\Theta}; \boldsymbol{N}) \boldsymbol{S}_c'(\boldsymbol{\Theta}; \boldsymbol{N}) \mid \boldsymbol{N}^{\mathrm{R}}; \boldsymbol{\Theta} \right] - \boldsymbol{S}(\boldsymbol{\Theta}; \boldsymbol{N}^{\mathrm{R}}) \boldsymbol{S}'(\boldsymbol{\Theta}; \boldsymbol{N}^{\mathrm{R}}) \\ &= \mathrm{Cov}\left[\boldsymbol{S}_c(\boldsymbol{\Theta}; \boldsymbol{N}) \mid \boldsymbol{N}^{\mathrm{R}}; \boldsymbol{\Theta} \right] . \end{aligned} \tag{5.14}$$

As such, the observed information matrix in (5.13) can be expressed in terms of conditional moments of the first- and second-order partial derivatives of the complete data log-likelihood function, which is more amenable to analytical calculations than the incomplete data analog. By averaging both sides of (5.13) over the distribution of $\boldsymbol{N}^{\mathrm{R}}$, we get an expression for the expected information matrix

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\Theta}) = \boldsymbol{\mathcal{I}}_c(\boldsymbol{\Theta}) - \mathbb{E}\left[\boldsymbol{\mathcal{I}}_m(\boldsymbol{\Theta}; \boldsymbol{N}^{\mathrm{R}}) \mid \boldsymbol{\Theta} \right] . \tag{5.15}$$

In our framework, we are mainly interested in the parameter uncertainty con-

cerning the regressional parameters $\boldsymbol{\alpha}$ of the Poisson occurrence model and the $\boldsymbol{\beta}$ of the negative binomial reporting delay distribution in weeks. We assess both covariance matrices separately in Appendix 5.6.

5.3.5 Prediction of IBNR claim counts

Using the estimated parameter vector $\widehat{\boldsymbol{\Theta}}$ of the reserving model, we can predict the daily IBNR claim counts in the lower triangle of Table 5.2. Point estimates for all $N_{td} \in \boldsymbol{N}^{\mathrm{IBNR}}$ can be obtained using the expected values $\widehat{N}_{td} = \mathbb{E}\left[N_{td} \mid \boldsymbol{N}^{\mathrm{R}}; \widehat{\boldsymbol{\Theta}}\right] = \widehat{\lambda}_t \, \widehat{p}_{td}$. Similarly, the total IBNR claim counts per day are estimated by $\widehat{N}_t^{\mathrm{IBNR}} = \sum_{d=\tau-t+1}^{\infty} \widehat{N}_{td} = \widehat{\lambda}_t \, \widehat{p}_t^{\mathrm{IBNR}}$ and the total IBNR claim count over all occurrence days by $\widehat{N}^{\mathrm{IBNR}} = \sum_{t=1}^{\tau} \widehat{N}_t^{\mathrm{IBNR}}$. Moreover, under the model assumptions of Section 5.3.2, the future daily claim counts N_{td} are independently Poisson distributed and we thus have that

$$N_{td} \sim \mathrm{Poisson}(\lambda_t \, p_{td}) \,, \qquad N_t^{\mathrm{IBNR}} \sim \mathrm{Poisson}(\lambda_t \, p_t^{\mathrm{IBNR}}) \,,$$

$$\text{and} \qquad N^{\mathrm{IBNR}} \sim \mathrm{Poisson}\left(\sum_{t=1}^{\tau} \lambda_t \, p_t^{\mathrm{IBNR}}\right) .$$

This allows us to construct prediction intervals and to make probabilistic statements concerning the claim count component of the IBNR reserve by replacing the intensities by their maximum likelihood estimates.

5.4 Results

We apply our model outlined in Section 5.3 to the data set of general liability insurance policies discussed in Section 5.2. To illustrate the regressional approach of our methodology we use the month, the day of the week and the day of the month of the occurrence date as regressors in both the Poisson model for claim occurrences and the negative binomial model for the reporting delay in weeks. These categorical variables are incorporated into the covariate vectors $\boldsymbol{x}_t$ and $\boldsymbol{z}_t$ using dummy coding with the first level as reference category and by including an intercept term. Earned exposure (see Figure 5.2a) is used as the offset e_t in the Poisson occurrence model. We fit the model using the observed data up to the evaluation date, August 31, 2004. The remaining out-of-sample data until August 31, 2009 will be used to evaluate the model predictions.

5.4.1 Parameter estimates

The parameters $\boldsymbol{\Theta} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \phi, \mathbf{P}^1, \mathbf{P}^2\}$ are estimated using the EM algorithm of Section 5.3.3. The maximum likelihood estimates of the day probabilities $\mathbf{P}^1$ within the first reporting week are reported in Table 5.3 and those of the day probabilities $\mathbf{P}^2$ from the second reporting week onwards are given in Table 5.4. Recall that the ordering of the working days in a reporting week depends on the occurrence day of the week (dow), see Table 5.1. As motivated by Figure 5.6, the day probabilities in the first week have separate estimates by dow, whereas no distinction is made from the second week onwards. The estimated probabilities $\widehat{\mathbf{P}^1}$ are very close to the empirical values from Figure 5.6a and the estimated probabilities $\widehat{\mathbf{P}^2}$ are very close to the empirical values from Figure 5.6b, averaged over dow.

The effects related to the categorical predictors month, day of the month and day of the week of the occurrence date are visualized in Figure 5.7 for the Poisson regression model of the claim occurrences and in Figure 5.8 for the negative binomial regression model of the reporting delay in weeks. The corresponding maximum likelihood estimates of the parameter vectors $\boldsymbol{\alpha}$ (resp. $\boldsymbol{\beta}$), except for

Table 5.3: *Maximum likelihood estimates of the day probabilities* $\mathbf{P}^1$ *within the first reporting week. Separate reporting day probabilities are estimated for each day of the week (dow) of the occurrence date, as shown in the rows.*

	wday						
dow	wday1	wday2	wday3	wday4	wday5	Saturday	Sunday
Monday	0.2600	0.4006	0.1638	0.0957	0.0744	0.0055	0.0000
Tuesday	0.2722	0.4131	0.1486	0.0900	0.0689	0.0072	0.0000
Wednesday	0.2699	0.3802	0.1739	0.0972	0.0700	0.0088	0.0000
Thursday	0.2639	0.4106	0.1464	0.0925	0.0695	0.0170	0.0000
Friday	0.2985	0.3003	0.1527	0.1006	0.0712	0.0767	0.0000
Saturday	0.4575	0.2045	0.1284	0.0843	0.0722	0.0531	0.0000
Sunday	0.4778	0.2232	0.1375	0.0890	0.0673	0.0051	0.0001

Table 5.4: *Maximum likelihood estimates of the day probabilities* $\mathbf{P}^2$ *from the second reporting week onwards.*

wday1	wday2	wday3	wday4	wday5	Saturday	Sunday
0.2886	0.2117	0.1829	0.1542	0.1429	0.0196	0.0000

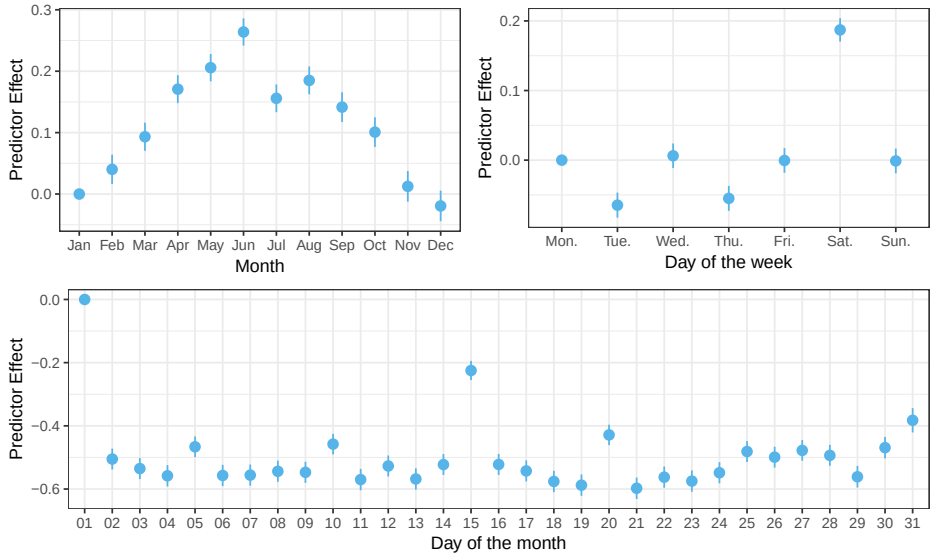

Figure 5.7: *Maximum likelihood estimates for the parameters in* $\boldsymbol{\alpha}$ *corresponding to the categorical effects of the month, the day of the week and the day of the month of the occurrence date in the Poisson claim occurrence model. 95% confidence intervals are constructed using the inverse of the expected information matrix* $\boldsymbol{\mathcal{I}}^{\boldsymbol{\alpha},\boldsymbol{\alpha}}(\widehat{\boldsymbol{\Theta}})$ *derived in Appendix 5.6.*

the intercept term, are plotted along with 95% confidence intervals based on the inverse of the expected information matrix $\boldsymbol{\mathcal{I}}^{\boldsymbol{\alpha},\boldsymbol{\alpha}}(\widehat{\boldsymbol{\Theta}})$ (resp. $\boldsymbol{\mathcal{I}}^{\boldsymbol{\beta},\boldsymbol{\beta}}(\widehat{\boldsymbol{\Theta}})$) as derived in Appendix 5.6. For completeness, we also report that in the Poisson model the intercept is estimated as -2.4044 (S.E. 0.0145) and in the negative binomial model the intercept is estimated as 1.8316 (S.E. 0.0323) and the dispersion parameter ϕ as 0.1775.

The month predictor reveals a seasonal pattern in which the number of claims rises in the middle of the year and falls around the year end. Most claims occur in June and least in December with an estimated difference in expected value of 32%. The reporting delay in weeks on the other hand has the highest expected values in winter months and the lowest in autumn months.

Modeling the seasonal variations in the occurrence process with respect to the day of the week shows an increase in the expectation of the number of claims on Saturdays and a slight decrease on Tuesdays and Thursdays. The reporting delays only vary mildly by day of the week with the highest estimates on Thursdays and Fridays and the lowest on Sundays.

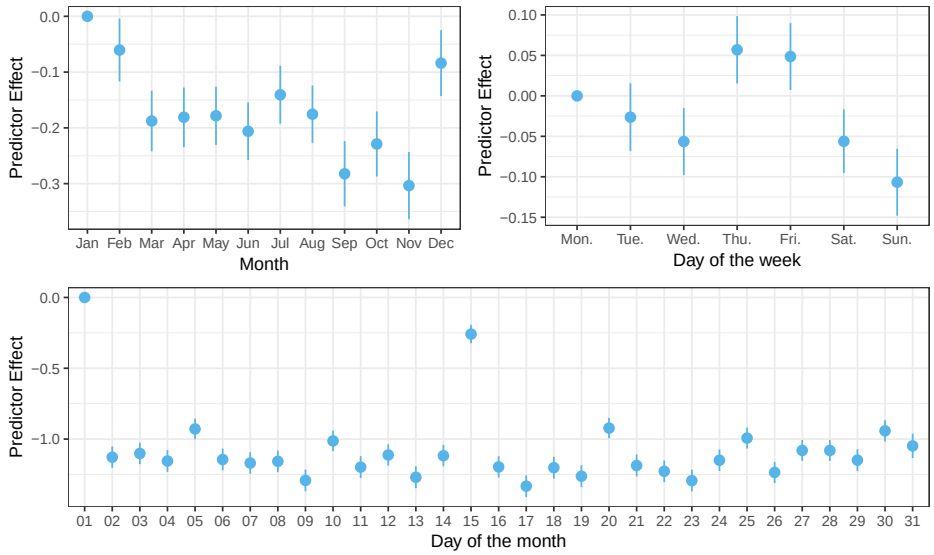

Figure 5.8: *Maximum likelihood estimates for the parameters in $\boldsymbol{\beta}$ corresponding to the categorical effects of the month, the day of the week and the day of the month of the occurrence date in the negative binomial reporting delay model. 95% confidence intervals are constructed using the inverse of the expected information matrix $\boldsymbol{\mathcal{I}}^{\boldsymbol{\beta},\boldsymbol{\beta}}(\widehat{\boldsymbol{\Theta}})$ derived in Appendix 5.6.*

The categorical effect of the day of the month shows a remarkable pattern which is similar in both the claim occurrence model and the reporting delay model. On the 1st and 15th, the number of claims as well as the reporting delays have significantly higher expected values. A similar effect, but of a lower degree, is also present for the 5th, 10th, 20th, and 30th or 31st day of each month. This pattern can most likely be explained by rounding errors of the occurrence date when insureds have to report a claim which took place several weeks or months ago. As the policyholder can no longer precisely remember the actual occurrence date, he simply reports the first day or the middle of the month in which the claim occurred or, to a lesser extent, replaces the month day by a value which is a multiple of 5. Many of the outlying observations from Figure 5.2b correspond to these values for the occurrence day of the month. Since this misreporting of dates is more likely to occur for claims which are only reported after a longer time period, we simultaneously see an increase in the expected reporting delay for claims occurring on these rounded month days.

5.4.2 Prediction of IBNR claim counts

Besides providing insight in the claim occurrence process, the main goal of our model is to estimate the number of IBNR claims. In our setting, the IBNR claims are those with occurrence date in between January 2000 and August 2004 that have only been reported after the evaluation date, August 31, 2004. We know there are 2052 such IBNR claims present in the full data set until August 2009 of which we know the corresponding occurrence date and reporting delay, see the lower triangle in Figure 5.3. These data are used to assess the out-of-sample predictive performance.

Based on the fitted model where τ corresponds to August 31, 2004, the total number of IBNR claims is estimated as $\widehat{N}^{\text{IBNR}} = 2066.03$, which is very close to the actual count. Moreover, the distributional assumptions of our model can be used to provide a 95% prediction interval given by $[1977, 2156]$, see Section 5.3.5. Furthermore, since the model is defined on a daily level, the total IBNR prediction can be divided into daily forecasts by occurrence date and by reporting date. This allows insurers to get a refined projection of the expected number of IBNR claims according to their occurrence time points and their future reporting times. To illustrate this strong point of our model, we predict the IBNR claim counts by occurrence dates in Figure 5.9 and by reporting dates in Figure 5.10.

In Figure 5.9a we plot point estimates and 95% prediction intervals for N_t^{IBNR} with t corresponding to occurrences dates in between July 1, 2004, and August 31, 2004, i.e. the last two months from our training period. The predictions follow the same trend as the actual IBNR claim counts derived from the full data set until August 2009. In particular, we notice for instance how IBNR claims are elevated on the first day and middle of each month, in line with our earlier findings. In Figure 5.9b (resp. Figure 5.9c) we group the occurrence dates by weeks (resp. months) prior to the evaluation date and show the IBNR claim count predictions corresponding to the past 26 weeks (resp. 12 months). We notice how, also over longer time spans, the predictions by occurrence week or month follow the pattern observed in the actual IBNR counts.

In Figure 5.10 we disperse the total predicted IBNR claim count according to the date on which the IBNR claims will be reported to the insurer. It means we now focus on estimating $\sum_{t=1}^{\tau} N_{t,\rho-t}$ for $\rho = \tau+1, \tau+2, \ldots$, i.e. the number of IBNR claims reported on day 1, 2, ... of the out-of-sample period. This forms an appealing way to use our model in practice as it gives the insurer a refined view on the reporting times of the IBNR claims. The predictions on a daily level in Figure 5.10a range from September 1, 2004, until November 7, 2004 and are again

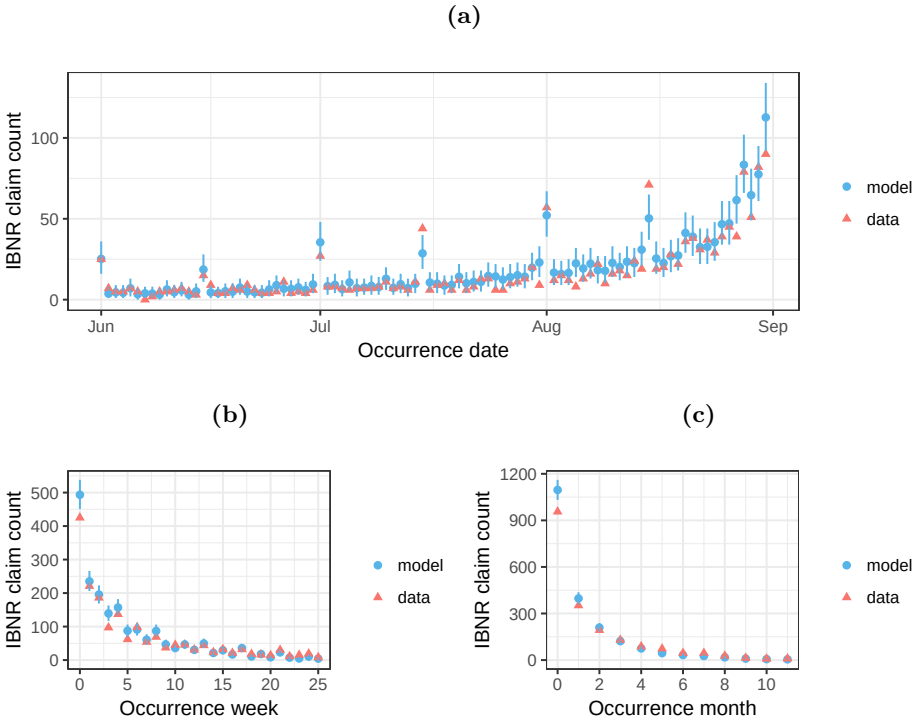

Figure 5.9: *Predictions of the IBNR claim counts by occurrence date. Prediction intervals are constructed at the 95% confidence level. The actual IBNR claim counts are derived based on the full data set until August 2009. In (a), we show predictions by day for occurrence dates in between July 1, 2004, and August 31, 2004. In (b), we group the occurrence dates by weeks (7 days) prior to the evaluation date and show the predictions corresponding to the past 26 weeks. In (c), we group the occurrence dates by months (30 days) from the evaluation date and show the predictions corresponding to the past 12 months.*

accompanied by 95% prediction intervals. When compared to the out-of-sample actual values, the forecasts clearly capture the downward trend in the reporting of IBNR claims and the nearly absence of reporting in weekends. This is primarily the case due to the day probabilities in our model which reflect the day-specific aspects of the reporting delay. Similar as before, in Figure 5.10b (resp. Figure 5.10c) we group the reporting dates by weeks (resp. months) after the evaluation date and show the IBNR claim count predictions corresponding to the next 26 weeks (resp. 12 months).

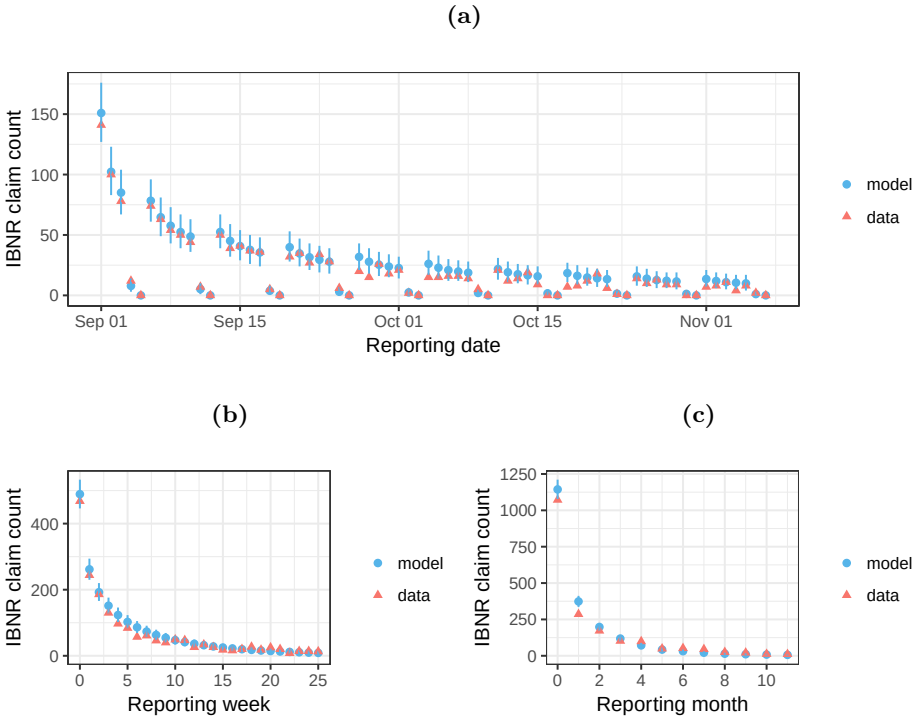

Figure 5.10: *Predictions of the IBNR claim counts by reporting date. Prediction intervals are constructed at the 95% confidence level. The actual IBNR claim counts are derived based on the full data set until August 2009. In (a), we show predictions by day for reporting dates in between September 1, 2004, and November 7, 2004. In (b), we group the reporting dates by weeks (7 days) after the evaluation date and show the predictions corresponding to the next 26 weeks. In (c), we group the reporting dates by months (30 days) after the evaluation date and show the predictions corresponding to the next 12 months.*

5.4.3 Prediction of total IBNR claim counts over time

Finally, we evaluate how the model performs in estimating the total IBNR claim count when it is refitted to a different subset of the data. In order to do so, we adjust the evaluation date τ, which was chosen to be August 31, 2004, to any date in between September 1, 2003, and August 31, 2004. For each such τ, we refit the model based on the observed data by that date, $\boldsymbol{N}^{\mathrm{R}} = \{N_{td} \mid$

$1 \leqslant t \leqslant \tau, d \geqslant 0, t+d \leqslant \tau\}$, and produce an estimate of the total IBNR claim count $N^{\text{IBNR}} = \sum_{t=1}^{\tau} N_t^{\text{IBNR}}$ corresponding to claims that occurred before τ. Figure 5.11 contrasts these predictions along with 95% prediction intervals to the actual total IBNR claim counts at each evaluation date based on the full data set. Although the model estimates follow the seasonal pattern also observed in the actual IBNR claim counts, the predictions are often too high. One possible reason for this is that the measure we used for the exposure to risk, namely the earned exposure, is too crude. If a more refined exposure unit would be available, such as the sum of the net earned premiums, these predictions might improve. Another reason is that we assume the seasonal monthly pattern to be the same over the different years. For the data at hand, it seems that this assumption might be too simplistic and that the effect of the month on the occurrences of claims and on the reporting delays changes over the years. As a result, the estimates in Figures 5.7 and 5.8 are averaged values. However, including effects of the calendar year or interactions of months with the calendar years in both the occurrence model and the reporting delay model is even more harmful for the predictive performance and leads to a more severe underestimation of the total number of IBNR claims (results not shown). This is due to the amount of missing information in the last calendar year, see Figure 5.3. As a consequence, the extra parameters related to the last occurrence year are used to further maximize the likelihood of the observed claims in the upper triangle of Table 5.2 but lead to bad extrapolations for the lower triangle. Imposing restrictions and allowing the calendar year to be only used in either the occurrence model is a better strategy to extrapolate the past observed patterns, but still cannot provide on-target predictions over the entire range of evaluation dates of Figure 5.11 (results not shown). This shows how the claim arrival process is an intrinsically hard process to model.

Possible reasons why the occurrence process of claims might change over time include changes in product design and conditions, changes in the business environment, changes in legislation, and changes in the registration of reported claims. If any of this is the case and corresponding expert-knowledge is available on how it impacts the claim arrival process, then the model could be appropriately adjusted. The regressors used in the Poisson distribution for the daily total claim counts and the negative binomial distribution of the reporting delay in weeks could be easily extended based on external covariate information of which the insurer believes it affects the claim occurrence process.

A final remark related to Figure 5.11 is the increase in the total IBNR claim count around the end of the year. Claim counts are indeed higher on New Year's

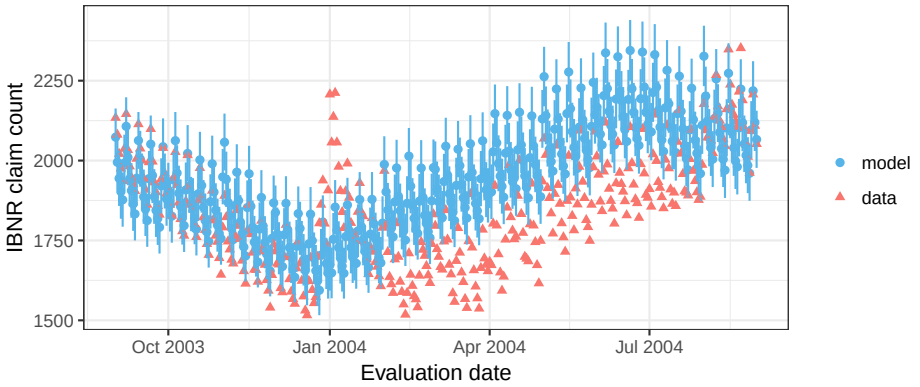

Figure 5.11: *Predictions of the total IBNR claim counts for varying evaluation dates τ in between September 1, 2003, and August 31, 2004. Prediction intervals are constructed at the 95% confidence level. The actual total IBNR claim counts are derived based on the full data set until August 2009.*

Eve and New Year as can be seen from Figure 5.2b. The model can incorporate this aspect using designated dummy indicators in the regression models, but this would not completely make the IBNR predictions at the end of the year in line with the actual values (results not shown). This is due to the fact that the insurance company is closed around the holidays, preventing any claims from being reported at that time. Tackling this issue would require us to adjust the day-specific probabilities to take the absence of reporting on holidays into account, which is not straightforward to do under the model assumptions of Section 5.3.2. For this reason, it is advisable not to estimate the number of IBNR claims exactly on the first or last day of the year.

5.5 Conclusions and outlook

We propose a new technique to model the claim arrival process on a daily basis in order to estimate the number of IBNR claim counts. The method uses regression models for count data for the occurrence of claims and their corresponding reporting delays. The main idea introduced in this work is to treat the right truncation of the reporting delays as a type of missing data. Applying the EM algorithm strongly simplifies maximum likelihood estimation as it allows for the use of standard statistical software to fit the regression models. We investigate

the performance of our micro-level IBNR reserving method in a case study with a European portfolio of general liability insurance policies for private individuals. The presented model provides a better understanding of the claim arrival process and can be used to predict IBNR claims on a daily level.

We indicate some possible directions for future research. First of all, we would like to stress that the provided estimation framework involving the EM algorithm can be applied to different models in this context. This provides a more desirable alternative over the ad hoc methods or two-step approaches used earlier in actuarial literature. The essence of the estimation procedure described in Section 5.3.3 would remain the same.

A direct extension of the model presented in this chapter would be to introduce a multinomial logistic regression model for the day probabilities within a reporting week shown in Figure 5.6. Incorporating covariate information would allow us to model possible evolutions of these reporting day probabilities over time. If say, for instance, in more recent years claims are also being reported on Sundays through online reporting the day-specific probabilities would be able to adapt. This would be easily implemented because the EM algorithm relies on complete data computations which enables using a statistical software package to fit the multinomial logit model.

It would also be interesting to explore different distributional assumptions for the daily total claim counts and the reporting delay distribution in weeks. The reporting delay can be easily altered within the given framework to, for instance, a zero-inflated or hurdle distribution or a more heavy-tailed distribution. Relaxing the Poisson assumption for the daily total claim counts is also feasible but might complicate the E-step in which we now relied on the thinning property of Poisson distributions. The EM framework is however compatible with latent underlying processes affecting the occurrence of claims such as hidden Markov models or shot noise process (see e.g. Badescu et al., 2016a; Avanzi et al., 2016). Another promising approach would be to investigate how time series models for counts (see Jung and Tremayne, 2011, for an overview) could be introduced in this setting.

5.6 Appendix: Derivation of the asymptotic variance-covariance matrix

Covariance matrix with respect to $\boldsymbol{\alpha}$ The asymptotic covariance matrix of the MLE $\widehat{\boldsymbol{\alpha}}$ can be estimated by the inverse of the submatrix of the observed

information matrix related to $\boldsymbol{\alpha}$, evaluated at $\boldsymbol{\Theta} = \widehat{\boldsymbol{\Theta}}$. Using relationship (5.13) between the incomplete data, complete data and missing information matrices, we have that

$$\boldsymbol{I}^{\boldsymbol{\alpha},\boldsymbol{\alpha}}(\boldsymbol{\Theta}; \boldsymbol{N}^{\mathrm{R}}) = -\frac{\partial^2}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}'} \log \mathcal{L}(\boldsymbol{\Theta}; \boldsymbol{N}^{\mathrm{R}}) = \boldsymbol{\mathcal{I}}_c^{\boldsymbol{\alpha},\boldsymbol{\alpha}}(\boldsymbol{\Theta}; \boldsymbol{N}^{\mathrm{R}}) - \boldsymbol{\mathcal{I}}_m^{\boldsymbol{\alpha},\boldsymbol{\alpha}}(\boldsymbol{\Theta}; \boldsymbol{N}^{\mathrm{R}}) \,.$$

The subvector of the complete data score statistic related to $\boldsymbol{\alpha}$ is equal to

$$\begin{aligned}
\boldsymbol{S}_c^{\boldsymbol{\alpha}}(\boldsymbol{\Theta}; \boldsymbol{N}) &= \frac{\partial}{\partial \boldsymbol{\alpha}} \log \mathcal{L}_c(\boldsymbol{\Theta}; \boldsymbol{N}) \\
&= \frac{\partial}{\partial \boldsymbol{\alpha}} \left[-\sum_{t=1}^{\tau} e_t \exp(\boldsymbol{x}_t' \boldsymbol{\alpha}) + \sum_{t=1}^{\tau} N_t (\log(e_t) + \boldsymbol{x}_t' \boldsymbol{\alpha}) \right] \\
&= -\sum_{t=1}^{\tau} e_t \exp(\boldsymbol{x}_t' \boldsymbol{\alpha}) \boldsymbol{x}_t + \sum_{t=1}^{\tau} N_t \boldsymbol{x}_t \,.
\end{aligned}$$

The missing information matrix with respect to $\boldsymbol{\alpha}$ can be derived using (5.14) as

$$\begin{aligned}
\boldsymbol{\mathcal{I}}_m^{\boldsymbol{\alpha},\boldsymbol{\alpha}}(\boldsymbol{\Theta}; \boldsymbol{N}^{\mathrm{R}}) &= \mathrm{Cov}\left[\boldsymbol{S}_c^{\boldsymbol{\alpha}}(\boldsymbol{\Theta}; \boldsymbol{N}) \mid \boldsymbol{N}^{\mathrm{R}}; \boldsymbol{\Theta} \right] \\
&= \mathrm{Cov}\left[-\sum_{t=1}^{\tau} e_t \exp(\boldsymbol{x}_t' \boldsymbol{\alpha}) \boldsymbol{x}_t + \sum_{t=1}^{\tau} N_t \boldsymbol{x}_t \,\middle|\, \boldsymbol{N}^{\mathrm{R}}; \boldsymbol{\Theta} \right] \\
&= \mathrm{Cov}\left[\sum_{t=1}^{\tau} N_t \boldsymbol{x}_t \,\middle|\, \boldsymbol{N}^{\mathrm{R}}; \boldsymbol{\Theta} \right] \\
&= \mathrm{Cov}\left[\sum_{t=1}^{\tau} \left(\sum_{d=\tau-t+1}^{\infty} N_{td} \right) \boldsymbol{x}_t \,\middle|\, \boldsymbol{N}^{\mathrm{R}}; \boldsymbol{\Theta} \right] \\
&= \sum_{t=1}^{\tau} \lambda_t p_t^{\mathrm{IBNR}} \boldsymbol{x}_t \boldsymbol{x}_t' \,, \qquad (5.16)
\end{aligned}$$

where we use the assumption that the daily total claim counts are independently Poisson distributed. Furthermore, we compute

$$\begin{aligned}
\boldsymbol{I}_c^{\boldsymbol{\alpha},\boldsymbol{\alpha}}(\boldsymbol{\Theta}; \boldsymbol{N}) &= -\frac{\partial^2}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}'} \log \mathcal{L}(\boldsymbol{\Theta}; \boldsymbol{N}^{\mathrm{R}}) \\
&= -\frac{\partial}{\partial \boldsymbol{\alpha}} \boldsymbol{S}_c^{\boldsymbol{\alpha}}(\boldsymbol{\Theta}; \boldsymbol{N}) \\
&= \sum_{t=1}^{\tau} e_t \exp(\boldsymbol{x}_t' \boldsymbol{\alpha}) \boldsymbol{x}_t \boldsymbol{x}_t' \,, \qquad (5.17)
\end{aligned}$$

which does not depend on $\boldsymbol{N}$ such that

$$\boldsymbol{\mathcal{I}}_c^{\boldsymbol{\alpha},\boldsymbol{\alpha}}(\boldsymbol{\Theta};\boldsymbol{N}^{\mathrm{R}}) = \mathbb{E}\left[\boldsymbol{I}_c^{\boldsymbol{\alpha},\boldsymbol{\alpha}}(\boldsymbol{\Theta};\boldsymbol{N}) \mid \boldsymbol{N}^{\mathrm{R}};\boldsymbol{\Theta}\right] = \boldsymbol{I}_c^{\boldsymbol{\alpha},\boldsymbol{\alpha}}(\boldsymbol{\Theta};\boldsymbol{N}),$$

and

$$\boldsymbol{\mathcal{I}}_c^{\boldsymbol{\alpha},\boldsymbol{\alpha}}(\boldsymbol{\Theta}) = \mathbb{E}\left[\boldsymbol{I}_c^{\boldsymbol{\alpha},\boldsymbol{\alpha}}(\boldsymbol{\Theta};\boldsymbol{N}) \mid \boldsymbol{\Theta}\right] = \boldsymbol{I}_c^{\boldsymbol{\alpha},\boldsymbol{\alpha}}(\boldsymbol{\Theta};\boldsymbol{N}).$$

By combing (5.16) and (5.17), evaluated at $\boldsymbol{\Theta} = \widehat{\boldsymbol{\Theta}}$, we thus find that

$$\begin{aligned}\boldsymbol{I}^{\boldsymbol{\alpha},\boldsymbol{\alpha}}(\widehat{\boldsymbol{\Theta}};\boldsymbol{N}^{\mathrm{R}}) &= \boldsymbol{\mathcal{I}}_c^{\boldsymbol{\alpha},\boldsymbol{\alpha}}(\widehat{\boldsymbol{\Theta}};\boldsymbol{N}^{\mathrm{R}}) - \boldsymbol{\mathcal{I}}_m^{\boldsymbol{\alpha},\boldsymbol{\alpha}}(\widehat{\boldsymbol{\Theta}};\boldsymbol{N}^{\mathrm{R}}) \\ &= \sum_{t=1}^{\tau} e_t \exp(\boldsymbol{x}_t'\widehat{\boldsymbol{\alpha}})\boldsymbol{x}_t\boldsymbol{x}_t' - \sum_{t=1}^{\tau} \widehat{\lambda}_t\, \widehat{p}_t^{\mathrm{IBNR}}\boldsymbol{x}_t\boldsymbol{x}_t' \\ &= \sum_{t=1}^{\tau} \widehat{\lambda}_t\, \widehat{p}_t^{\mathrm{R}}\boldsymbol{x}_t\boldsymbol{x}_t', \end{aligned} \tag{5.18}$$

which does not depend on the observed data $\boldsymbol{N}^{\mathrm{R}}$ and hence also equals $\boldsymbol{\mathcal{I}}^{\boldsymbol{\alpha},\boldsymbol{\alpha}}(\widehat{\boldsymbol{\Theta}})$. Its inverse estimates the asymptotic covariance matrix of the MLE $\widehat{\boldsymbol{\alpha}}$. The missing information principle applied to the parameters of the Poisson regression model for the daily claim occurrences has a very intuitive interpretation: the observed information (5.18) related to the observed daily claim counts N_t^{R} equals the complete information (5.17) related to the total daily claim counts N_t minus the missing information (5.16) related to the IBNR daily claim counts N_t^{IBNR}.

Covariance matrix with respect to $\boldsymbol{\beta}$ Similarly for $\boldsymbol{\beta}$, we use the relation

$$\begin{aligned}\boldsymbol{I}^{\boldsymbol{\beta},\boldsymbol{\beta}}(\boldsymbol{\Theta};\boldsymbol{N}^{\mathrm{R}}) &= -\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\log\mathcal{L}(\boldsymbol{\Theta};\boldsymbol{N}^{\mathrm{R}}) \\ &= \boldsymbol{\mathcal{I}}_c^{\boldsymbol{\beta},\boldsymbol{\beta}}(\boldsymbol{\Theta};\boldsymbol{N}^{\mathrm{R}}) - \boldsymbol{\mathcal{I}}_m^{\boldsymbol{\beta},\boldsymbol{\beta}}(\boldsymbol{\Theta};\boldsymbol{N}^{\mathrm{R}}). \end{aligned} \tag{5.19}$$

The score vector associated to $\boldsymbol{\beta}$ is given by

$$\begin{aligned}\boldsymbol{S}_c^{\boldsymbol{\beta}}(\boldsymbol{\Theta};\boldsymbol{N}) &= \frac{\partial}{\partial\boldsymbol{\beta}}\log\mathcal{L}_c(\boldsymbol{\Theta};\boldsymbol{N}) \\ &= \sum_{t=1}^{\tau}\sum_{w=0}^{\infty}\left(\sum_{d=0}^{6} N_{t,7w+d}\right)\frac{\partial}{\partial\boldsymbol{\beta}}\log\left(p_{tw}^{\mathrm{W}}\right)\end{aligned}$$

where

$$\frac{\partial}{\partial\boldsymbol{\beta}}\log\left(p_{tw}^{\mathrm{W}}\right) = \frac{\partial}{\partial\boldsymbol{\beta}}\left[w\boldsymbol{z}_t'\boldsymbol{\beta} - (\phi+w)\log(\phi+\exp(\boldsymbol{z}_t'\boldsymbol{\beta}))\right]$$

$$
\begin{aligned}
&= w\boldsymbol{z}_t - (\phi + w)\frac{\exp(\boldsymbol{z}_t'\boldsymbol{\beta})\boldsymbol{z}_t}{\phi + \exp(\boldsymbol{z}_t'\boldsymbol{\beta})} \\
&= \frac{\phi(w - \exp(\boldsymbol{z}_t'\boldsymbol{\beta}))}{\phi + \exp(\boldsymbol{z}_t'\boldsymbol{\beta})}\boldsymbol{z}_t\,.
\end{aligned}
$$

Its conditional covariance is the missing information matrix related to $\boldsymbol{\beta}$,

$$
\begin{aligned}
\boldsymbol{\mathcal{I}}_m^{\boldsymbol{\beta},\boldsymbol{\beta}}(\boldsymbol{\Theta}; \boldsymbol{N}^{\mathrm{R}}) &= \mathrm{Cov}\left[\boldsymbol{S}_c^{\boldsymbol{\beta}}(\boldsymbol{\Theta}; \boldsymbol{N}) \mid \boldsymbol{N}^{\mathrm{R}}; \boldsymbol{\Theta}\right] \\
&= \sum_{t=1}^{\tau}\left(\sum_{w=0}^{\infty}\left(\sum_{\substack{d=0,\ldots,6 \\ 7w+d>\tau-t}} \lambda_t p_{t,7w+d}\right)\frac{\phi^2(w - \exp(\boldsymbol{z}_t'\boldsymbol{\beta}))^2}{(\phi + \exp(\boldsymbol{z}_t'\boldsymbol{\beta}))^2}\right)\boldsymbol{z}_t\boldsymbol{z}_t'\,.
\end{aligned} \tag{5.20}
$$

Moreover, we calculate

$$
\begin{aligned}
\boldsymbol{I}_c^{\boldsymbol{\beta},\boldsymbol{\beta}}(\boldsymbol{\Theta}; \boldsymbol{N}) &= -\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\log\mathcal{L}(\boldsymbol{\Theta}; \boldsymbol{N}^{\mathrm{R}}) \\
&= -\frac{\partial}{\partial\boldsymbol{\beta}}\boldsymbol{S}_c^{\boldsymbol{\beta}}(\boldsymbol{\Theta}; \boldsymbol{N}) \\
&= -\sum_{t=1}^{\tau}\sum_{w=0}^{\infty}\left(\sum_{d=0}^{6} N_{t,7w+d}\right)\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\log\left(p_{tw}^{\mathrm{W}}\right)
\end{aligned} \tag{5.21}
$$

where

$$
\begin{aligned}
\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\log\left(p_{tw}^{\mathrm{W}}\right) &= \frac{\partial}{\partial\boldsymbol{\beta}}\left[\frac{\phi(w - \exp(\boldsymbol{z}_t'\boldsymbol{\beta}))}{\phi + \exp(\boldsymbol{z}_t'\boldsymbol{\beta})}\boldsymbol{z}_t\right] \\
&= \frac{-\phi(\phi + \exp(\boldsymbol{z}_t'\boldsymbol{\beta}))\exp(\boldsymbol{z}_t'\boldsymbol{\beta}) - \phi(w - \exp(\boldsymbol{z}_t'\boldsymbol{\beta}))\exp(\boldsymbol{z}_t'\boldsymbol{\beta})}{(\phi + \exp(\boldsymbol{z}_t'\boldsymbol{\beta}))^2}\boldsymbol{z}_t\boldsymbol{z}_t' \\
&= -\frac{\phi(\phi + w)\exp(\boldsymbol{z}_t'\boldsymbol{\beta})}{(\phi + \exp(\boldsymbol{z}_t'\boldsymbol{\beta}))^2}\boldsymbol{z}_t\boldsymbol{z}_t'\,.
\end{aligned}
$$

On substituting (5.20) and (5.21) into (5.19), we then have that

$$
\begin{aligned}
\boldsymbol{I}^{\boldsymbol{\beta},\boldsymbol{\beta}}(\widehat{\boldsymbol{\Theta}}; \boldsymbol{N}^{\mathrm{R}}) &= \boldsymbol{\mathcal{I}}_c^{\boldsymbol{\beta},\boldsymbol{\beta}}(\widehat{\boldsymbol{\Theta}}; \boldsymbol{N}^{\mathrm{R}}) - \boldsymbol{\mathcal{I}}_m^{\boldsymbol{\beta},\boldsymbol{\beta}}(\widehat{\boldsymbol{\Theta}}; \boldsymbol{N}^{\mathrm{R}}) \\
&= \sum_{t=1}^{\tau}\left(\sum_{w=0}^{\infty}\left(\sum_{d=0}^{6}\widehat{N}_{t,7w+d}\right)\frac{\widehat{\phi}(\widehat{\phi} + w)\exp(\boldsymbol{z}_t'\widehat{\boldsymbol{\beta}})}{(\widehat{\phi} + \exp(\boldsymbol{z}_t'\widehat{\boldsymbol{\beta}}))^2}\right)\boldsymbol{z}_t\boldsymbol{z}_t' \\
&\quad - \sum_{t=1}^{\tau}\left(\sum_{w=0}^{\infty}\left(\sum_{\substack{d=0,\ldots,6 \\ 7w+d>\tau-t}}\widehat{\lambda}_t\,\widehat{p}_{t,7w+d}\right)\frac{\widehat{\phi}^2(w - \exp(\boldsymbol{z}_t'\widehat{\boldsymbol{\beta}}))^2}{(\widehat{\phi} + \exp(\boldsymbol{z}_t'\widehat{\boldsymbol{\beta}}))^2}\right)\boldsymbol{z}_t\boldsymbol{z}_t'
\end{aligned}
$$

with $\widehat{N}_{t,7w+d}$ defined as in (5.8) using the MLE $\widehat{\boldsymbol{\Theta}}$. Its expectation with respect to the observed data $\boldsymbol{N}^{\mathrm{R}}$ is given by

$$\begin{aligned}\boldsymbol{\mathcal{I}}^{\boldsymbol{\beta},\boldsymbol{\beta}}(\widehat{\boldsymbol{\Theta}}) &= \boldsymbol{\mathcal{I}}_c^{\boldsymbol{\beta},\boldsymbol{\beta}}(\widehat{\boldsymbol{\Theta}}) - \mathbb{E}\left[\boldsymbol{\mathcal{I}}_m^{\boldsymbol{\beta},\boldsymbol{\beta}}(\widehat{\boldsymbol{\Theta}};\boldsymbol{N}^{\mathrm{R}})\middle|\,\widehat{\boldsymbol{\Theta}}\right] \\ &= \sum_{t=1}^{\tau}\left(\sum_{w=0}^{\infty}\widehat{\lambda}_t\,\widehat{p}_{tw}^{\mathrm{W}}\,\frac{\widehat{\phi}(\widehat{\phi}+w)\exp(\boldsymbol{z}_t'\widehat{\boldsymbol{\beta}})}{(\widehat{\phi}+\exp(\boldsymbol{z}_t'\widehat{\boldsymbol{\beta}}))^2}\right)\boldsymbol{z}_t\boldsymbol{z}_t' \\ &\quad -\sum_{t=1}^{\tau}\left(\sum_{w=0}^{\infty}\left(\sum_{\substack{d=0,\ldots,6\\ 7w+d>\tau-t}}\widehat{\lambda}_t\,\widehat{p}_{t,7w+d}\right)\frac{\widehat{\phi}^2(w-\exp(\boldsymbol{z}_t'\widehat{\boldsymbol{\beta}}))^2}{(\widehat{\phi}+\exp(\boldsymbol{z}_t'\widehat{\boldsymbol{\beta}}))^2}\right)\boldsymbol{z}_t\boldsymbol{z}_t'\,.\end{aligned}$$

Either the inverse of $\boldsymbol{I}^{\boldsymbol{\beta},\boldsymbol{\beta}}(\widehat{\boldsymbol{\Theta}};\boldsymbol{N}^{\mathrm{R}})$ or the inverse of $\boldsymbol{\mathcal{I}}^{\boldsymbol{\beta},\boldsymbol{\beta}}(\widehat{\boldsymbol{\Theta}})$ can be used to estimated the asymptotic covariance matrix of the MLE $\widehat{\boldsymbol{\alpha}}$.

Chapter 6

Outlook

By careful analysis of the available data using proper statistical techniques, insurance companies can improve the predictive power of their pricing and reserving tools and achieve a better understanding, measurement and management of the risks they are exposed to. We strongly believe that our proposed techniques in the context of loss reserving, telematics insurance and claims reserving may lead to better actuarial practices. This chapter concludes our work by presenting several suggestions for future research related to these topics.

The developed methodology can also be applied to other areas where similar data are collected and analyzed. The expected impact is broader than only on actuarial science with potential applications in, for instance, econometrics (e.g. the unemployment duration data from Section 2.5.2), geology (e.g. the Old faithful geyser data from Section 3.5.2 and the use of compositional predictors in Chapter 4) and biostatistics (e.g. the mastitis study from Section 3.5.3 and the modeling of reporting delays in infectious disease data using the approach for IBNR claims reserving of Chapter 5).

6.1 Further developments in loss modeling

In Chapter 2, we develop an estimation procedure using the EM algorithm that is able to fit a mixture of Erlang distributions to censored and truncated data. The flexibility of mixtures of Erlang distributions and the effectiveness of the proposed fitting algorithm is demonstrated using several simulated and real data sets. In particular, for the left truncated Secura Re data set a mixture of two Erlang components adequately represents the moderately heavy-tailed claim sizes. In the

example of Section 2.5.4, we illustrate its limitations when data are generated from a generalized Pareto distribution with extreme value index equal to one, which expresses a very heavy tail. Because mixtures of Erlangs have asymptotically exponential tails, which are lighter, in such example there is no parsimonious model possible using such mixtures. Erlang components in a mixture are not able to extrapolate the heaviness in the tail and instead behave similar to an empirical distribution in the upper tail. However, this behavior might be undesirable from a risk measurement perspective. An accurate description of the upper tail of the claim size distribution is important to safeguard the insurance company against extreme losses that might jeopardize its solvency. Reynkens et al. (2016) address this issue by considering a *splicing model* where a mixture of Erlang distributions is used for the body of the distribution and a Pareto distribution for the tail. A global fit results, which combines the flexibility of the mixture of Erlangs distribution to model light and moderate losses with the ability of the Pareto distribution to model extreme values.

This idea can be extended to the multivariate setting from Chapter 3 in order to provide a global fit strategy for *heavy-tailed, dependent losses*. The most promising approach is to combine the multivariate mixtures of Erlang distributions (MME) with the multivariate generalized Pareto distribution (MGPD). The MGPD class is proposed in Rootzén and Tajvidi (2006) and combines univariate generalized Pareto distributions using a dependence structure to model the tail regions where at least one component of the vector is large. In such a multivariate splicing model, an MME would be used to represent losses below a d-dimensional splicing point and an MGPD for losses that exceed the splicing point in at least one dimension.

The definition of a (univariate and multivariate) mixture of Erlang distributions can readily be generalized to include a discrete point mass at zero. This accommodates common situations in practice when one explicitly wants to model the positive probability of a zero loss, i.e. when no loss has been incurred. The presented EM algorithm in Chapters 2 and 3 can straightforwardly be adapted to this extension.

The selection of shape parameters for mixtures of Erlangs is based on iteratively using the EM algorithm and comparing the fits based on AIC or BIC. The algorithms used to initialize, adjust and reduce the shape parameters perform well, but the strategy is computationally intensive and depends on the values of the initializing parameters M and s. It would be interesting to look into alternative ways to approach the *choice of the shapes* which forms a computationally

unsolvable optimization problem over $\mathbb{N}^M$. In a recent effort, Yin and Lin (2016) propose the use of regularization techniques for univariate mixtures of Erlangs, inspired by the MSCAD penalized likelihood of Chen and Khalili (2008). However, it still requires an initial choice for the shape set and introduces new tuning parameters.

In future research we aim to extend the mixtures of Erlangs framework towards the inclusion of predictor variables and introduce the flexibility of this approach in a *regression* context. Our idea is to translate the covariate information into the mixing weights of the components in the mixture of Erlang distributions. The common scale parameter then remains the same for all observations, but the weights in the mixture become a function of the linear predictor including the available covariates. In the univariate setting, we suggest to model the weights using a cumulative logit model (also called a proportional odds model). This model is used for ordinal dependent variables, which makes it a promising avenue in this setting since the mixing components have a natural ordering based on the value of the corresponding shape parameter. Introducing such a model for the mixing weights leads to solving an additional regression model in each M-step of the EM algorithm in order to update the corresponding parameters used in the weight specifications. In the multivariate setting, there is no natural ordering of the shape parameter vectors and a multinomial logit model for the mixing weights could be considered instead.

6.2 Further developments in telematics insurance

In Chapter 4, we investigate a Belgian telematics car insurance data set. The goal is to incorporate the telematics information to make better predictions on the number of claims and to identify the relationship between the driving habits and the accident risk. Compositional predictors are introduced to quantify and interpret the effect of the driving habits on the riskiness. The analysis shows that the use of this new type of data collected through telematics technology leads to improved predictions in actuarial pricing. Moreover, moving towards car insurance rating based on individual driving habits and style can resolve possible discrimination of basing the premium on proxies such as gender.

The novelty of telematics insurance calls for future research and requires an interdisciplinary approach. *From a business perspective*, it would be interesting to evaluate how the proposed prediction model using telematics variables can impact the pricing strategies and profitability of insurance companies. The cost

effectiveness of usage-based insurance could be assessed, taking into account the implementation cost of black box devices and related data management. Business models need to be designed that generate value from pay-as-you-drive insurance for both individuals and firms.

From the perspective of actuarial science and econometrics, telematics insurance is at the cross-road of a priori and a posteriori rating and demands a rethinking of common practices in both activities. This opens up new possibilities for future research on competitive adaptive pricing strategies. Premium structures can be developed to more closely reflect the actual risk exposure and to adapt over time based on the observed driving behavior after the underwriting of the policy. Financial rewards along with personalized driving style feedback will give policyholders a high incentive to drive more responsible, thus minimizing risk and improving road safety.

From a statistics and machine learning perspective, the most exciting future challenges lie in the analysis of telematics data on a more granular level. Telematics technology offers the possibility to collect real-time driving data via the black box device installed in a car. Insurers however partner with telematics data providers who process the raw telematics data, enrich these using external data sources (e.g. road maps) and deliver structured, aggregated telematics information. The daily summarized data we analyzed in Chapter 4 on how much, where and when the vehicle is driven forms a typical data setup in which insurance companies receive telematics data from such data providers. More extensive data formats also include certain driving style scores based on speeding violations, harsh braking, excessive accelerating, and cornering style. These kind of UBI driving scores can be easily incorporated in the presented framework.

To obtain a more comprehensive view on the driving style, it is desirable to have the raw telematics data in the form of streams of coordinates available (which is not the case in our setting). Statistical analysis of these spatiotemporal data is a highly relevant direction for future research. The main difficulty is to transform this high-frequency GPS location data into interpretable covariates describing complex driving patterns. Basic features that can be derived from GPS data at every time point are the speed, difference in speed, acceleration, difference in acceleration and angular speed. It is important to also account for specific driving contexts, such as road type, traffic situations and weather conditions. Using techniques from unsupervised learning and pattern recognition the goal is to classify different driving styles. In a next step, these driving style classes can be used as risk factors in claim count regression models to evaluate the effectiveness

of the classification in assessing the accident risk. Insurers must carefully consider which of these sensor data-derived classifications constitute suitable rating factors in usage-based car insurance pricing and to what extent they improve the quality of predictions.

6.3 Further developments in claims reserving

In Chapter 5, we contribute to the micro-level loss reserving literature by formulating a regression framework to model the claims arrival process along with its reporting delays on a daily level. The model can be used to predict the number of daily future claim counts in order to set up adequate IBNR claim reserves. The proposed methodology can be further developed and applied to multiple other case-studies, from different lines of business. The presented estimation framework using the EM algorithm can also be employed in alternative claims reserving models to obtain a joint estimation of both the occurrence and the reporting delay model parameters.

Several directions exist for future research in micro-level claims reserving. The most important path is to focus on the *payment process*, from reporting until settlement of a claim. Modeling the dynamics of the individual development of claims forms the next necessary building block to extend our micro-level loss reserving technique and to estimate future cash flows.

Arjas (1989), Norberg (1993) and Norberg (1999) developed a mathematical, probabilistic framework for the development of individual claims in continuous time. More recently, Antonio and Plat (2014) make this theory accessible to reserving practice by translating these probabilistic ideas to a statistical model in which estimation, inference and prediction is demonstrated on a real life data set. In their approach, hazard rates drive the time to events in the development of a claim (e.g. a payment, or settlement) and a lognormal regression is used to model intermediate payments corrected for inflation using a consumer price index.

Building upon the work of Antonio and Plat (2014), it would be interesting to relax the distributional assumptions made and to *incorporate claim-specific information* as covariates. Insurers' data base systems contain detailed information on open claims and their ongoing development: characteristics of the policy(holder), the accident, the (initial) case estimate (i.e. an expert judgment of the final claim amount), the reporting delay, the cumulative amount paid so far, etc. Traditional reserving methods compress these large data sets into small run-off triangles and hereby ignore this detailed information. Micro-level loss reserving offers the op-

portunity to instead use these claim-specific characteristics as predictive variables. This allows for a more realistic modeling of the development process which is expected to result in more accurate estimates and forecasts.

List of Figures

List of Tables

Bibliography

Aitchison, J. (1986). *The statistical analysis of compositional data*. Chapman and Hall London.

Aitchison, J. and Kay, J. W. (2003). Possible solution of some essential zero problems in compositional data analysis. In Thió-Henestrosa, S. and Martín-Fernández, J. A., editors, *Proceedings of CoDaWork'03, The 1st Compositional Data Analysis Workshop*. University of Girona.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Ampe, B., Goethals, K., Laevens, H., and Duchateau, L. (2012). Investigating clustering in interval-censored udder quarter infection times in dairy cows using a gamma frailty model. *Preventive veterinary medicine*, 106(3):251–257.

Antonio, K., Godecharle, E., and Oirbeek, R. V. (2016). A multi-state approach and flexible payment distributions for micro-level reserving in general insurance. *Working paper AFI16106*.

Antonio, K. and Plat, R. (2014). Micro–level stochastic loss reserving for general insurance. *Scandinavian Actuarial Journal*, 7:649–669.

Arjas, E. (1989). The claims reserving problem in non–life insurance: some structural ideas. *ASTIN Bulletin*, 19(2):139–152.

Asmussen, S., Nerman, O., and Olsson, M. (1996). Fitting phase-type distributions via the EM algorithm. *Scandinavian Journal of Statistics*, pages 419–441.

Assaf, D., Langberg, N. A., Savits, T. H., and Shaked, M. (1984). Multivariate phase-type distributions. *Operations Research*, 32(3):688–702.

Avanzi, B., Wong, B., and Yang, X. (2016). A micro-level claim count model with overdispersion and reporting delays. *Insurance: Mathematics and Economics*, 71:1–14.

Ayuso, M., Guillén, M., and Pérez-Marín, A. M. (2014). Time and distance to first accident and driving patterns of young drivers with pay-as-you-drive insurance. *Accident Analysis & Prevention*, 73:125–131.

Ayuso, M., Guillén, M., and Pérez-Marín, A. M. (2016a). Telematics and gender discrimination: Some usage-based evidence on whether men's risk of accidents differs from women's. *Risks*, 4(2):10.

Ayuso, M., Guillén, M., and Pérez-Marín, A. M. (2016b). Using {GPS} data to analyse the distance travelled to the first accident at fault in pay-as-you-drive insurance. *Transportation Research Part C: Emerging Technologies*, 68:160 – 167.

Azzalini, A. and Bowman, A. (1990). A look at some data on the old faithful geyser. *Applied Statistics*, pages 357–365.

Bacon Shone, J. (2003). Modelling structural zeros in compositional data. In Thió-Henestrosa, S. and Martín-Fernández, J. A., editors, *Proceedings of CoDaWork'03, The 1st Compositional Data Analysis Workshop*. University of Girona.

Badescu, A., Gong, L., Lin, X. S., and Tang, D. (2015). Modeling correlated frequencies with applications in operational risk management. *Journal of Operational Risk*. forthcoming.

Badescu, A. L., Lin, X. S., and Tang, D. (2016a). A marked Cox model for the number of IBNR claims: Estimation and application. *Available at SSRN 2747223*.

Badescu, A. L., Lin, X. S., and Tang, D. (2016b). A marked Cox model for the number of IBNR claims: Theory. *Insurance: Mathematics and Economics*, 69:29–37.

Barceló-Vidal, C., Martín-Fernández, J. A., and Mateu-Figueras, G. (2011). Compositional differential calculus on the simplex. In Pawlowsky-Glahn, V. and Buccianti, A., editors, *Compositional Data Analysis: Theory and Applications*. John Wiley & Sons.

Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J., De Waal, D., and Ferro, C. (2004). *Statistics of Extremes: Theory and Applications*. Wiley Series in Probability and Statistics. Wiley.

Bolancé, C., Guillén, M., Gustafsson, J., and Nielsen, J. P. (2012). *Quantitative operational risk models*. CRC Press.

Bordoff, J. E. and Noel, P. J. (2008). Pay-as-you-drive auto insurance: A simple way to reduce driving-related harms and increase equity. The Brookings Institution. Discussion Paper.

Boucher, J.-P. and Charpentier, A. (2014). General insurance pricing. In *Computational Actuarial Science with R*, pages 475–510. Chapman and Hall/CRC.

Boucher, J.-P., Pérez-Marín, A. M., and Santolino, M. (2013). Pay-as-you-drive insurance: the effect of the kilometers on the risk of accident. *Anales del Instituto de Actuarios Españoles, 3ª época*, 19:135–154.

Butler, P. (1993). Cost-based pricing of individual automobile risk transfer: Carmile exposure unit analysis. *Journal of Actuarial Practice*, 1(1):51–84.

Cameron, A. and Trivedi, P. (2005). *Microeconometrics: Methods and applications*. Cambridge University Press.

Chen, J. and Khalili, A. (2008). Order selection in finite mixture models with a nonsmooth penalty. *Journal of the American Statistical Association*, 103(484):1674–1683.

Chernobai, A., Rachev, S., and Fabozzi, F. (2014). Composite goodness-of-fit tests for left-truncated loss samples. In Lee, C.-F. and Lee, J. C., editors, *Handbook of Financial Econometrics and Statistics*, pages 575–596. Springer New York.

Clauset, A., Shalizi, C. R., and Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM review*, 51(4):661–703.

Cossette, H., Côté, M.-P., Marceau, E., and Moutanabbir, K. (2013a). Multivariate distribution defined with Farlie–Gumbel–Morgenstern copula and mixed Erlang marginals: Aggregation and capital allocation. *Insurance: Mathematics and Economics*, 52(3):560–572.

Cossette, H., Mailhot, M., Marceau, E., and Mesfioui, M. (2013b). Bivariate lower and upper orthant Value-at-Risk. *European Actuarial Journal*, 3(2):321–357.

Czado, C., Gneiting, T., and Held, L. (2009). Predictive model assessment for count data. *Biometrics*, 65(4):1254–1261.

de Jong, P. and Heller, G. (2008). *Generalized linear models for insurance data*. Cambridge.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Denuit, M. and Charpentier, A. (2005). *Mathématiques de l'assurance non-vie. Tome II: Tarification et provisionnement*. Collection "Economie et statistiques avancées". Economica.

Denuit, M. and Lang, S. (2004). Non-life ratemaking with Bayesian GAMs. *Insurance: Mathematics and Economics*, 35(3):627–647.

Denuit, M., Marechal, X., Pitrebois, S., and Walhin, J. (2007). *Actuarial modelling of claim counts: risk classification, credibility and bonus-malus systems*. Wiley.

Desyllas, P. and Sako, M. (2013). Profiting from business model innovation: Evidence from pay-as-you-drive auto insurance. *Research Policy*, 42(1):101–116.

Dhaene, J., Tsanakas, A., Valdez, E. A., and Vanduffel, S. (2012). Optimal capital allocation principles. *Journal of Risk and Insurance*, 79(1):1–28.

Dufour, R. and Maag, U. (1978). Distribution results for modified kolmogorov-smirnov statistics for truncated or censored. *Technometrics*, 20(1):29–32.

Efron, B. and Tibshirani, R. (1994). *An Introduction to the Bootstrap*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Egozcue, J. J., Barceló-Vidal, C., Martín-Fernández, J. A., Jarauta-Bragulat, E., Díaz-Barrero, J. L., and Mateu-Figueras, G. (2011). Elements of simplicial linear algebra and geometry. In Pawlowsky-Glahn, V. and Buccianti, A., editors, *Compositional Data Analysis: Theory and Applications*. John Wiley & Sons.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barcelo-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3):279–300.

Eisele, K.-T. (2005). EM algorithm for bivariate phase distributions. In *ASTIN Colloquium, Zurich, Switzerland. http://www. actuaries. org/ASTIN/Colloquia/Zurich/Eisele. pdf.*

England, P. D. and Verrall, R. J. (2002). Stochastic claims reserving in general insurance. *British Actuarial Journal*, 8(3):443–518.

Ferreira, J. and Minikel, E. (2010). Pay-As-You-Drive Auto Insurance In Massachusetts: A Risk Assessment And Report On Consumer. `http://mit.edu/jf/www/payd/PAYD_CLF_Study_Nov2010.pdf`.

Filipova-Neumann, L. and Welzel, P. (2010). Reducing asymmetric information in insurance markets: Cars with black boxes. *Telematics and Informatics*, 27(4):394–403.

Frees, E. W. (2014). Frequency and severity models. In Frees, E. W., Derrig, R. A., and Meyers, G., editors, *Predictive modeling applications in actuarial science*, volume 1. Cambridge University Press.

Frees, E. W., Carriere, J., and Valdez, E. (1996). Annuity valuation with dependent mortality. *Journal of Risk and Insurance*, pages 229–261.

Frees, E. W. and Valdez, E. A. (2008). Hierarchical insurance claims modeling. *Journal of the American Statistical Association*, 103(484):1457–1469.

Gelman, A. and Hill, J. (2007). *Applied Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge.

Georges, P., Lamy, A.-G., Nicolas, E., Quibel, G., and Roncalli, T. (2001). Multivariate survival modelling: A unified approach with copulas. *Unpublished paper, Groupe de Recherche Operationnelle, Credit Lyonnais, France*.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Godecharle, E. and Antonio, K. (2015). Reserving by conditioning on markers of individual claims: a case study using historical simulation. *North American Actuarial Journal*, 19(4):273–288.

Goethals, K., Ampe, B., Berkvens, D., Laevens, H., Janssen, P., and Duchateau, L. (2009). Modeling interval-censored, clustered cow udder quarter infection

times through the shared gamma frailty model. *Journal of agricultural, biological, and environmental statistics*, 14(1):1–14.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric regression and generalized linear models: a roughness penalty approach*. Chapman and Hall.

Greenberg, A. (2009). Designing pay-per-mile auto insurance regulatory incentives. *Transportation Research Part D: Transport and Environment*, 14(6):437–445.

Guilbaud, O. (1988). Exact kolmogorov-type tests for left-truncated and/or right-censored data. *Journal of the American Statistical Association*, 83(401):213–221.

Härdle, W. (1991). *Smoothing techniques: with implementation in S*. Springer.

Harris, J. E. (1990). Reporting delays and the incidence of aids. *Journal of the American Statistical Association*, 85(412):915–924.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, 1(3):297–318.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall.

Hastie, T. J., Tibshirani, R. J., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer-Verlag, Heidelberg, second edition.

Henckaerts, R., Antonio, K., Clijsters, M., and Verbelen, R. (2017). A data driven binning strategy for the construction of risk classes. *Working paper*.

Hron, K., Filzmoser, P., and Thompson, K. (2012). Linear regression with compositional explanatory variables. *Journal of Applied Statistics*, 39(5):1115–1128.

Husnjak, S., Peraković, D., Forenbacher, I., and Mumdziev, M. (2015). Telematics system in usage based motor insurance. *Procedia Engineering*, 100:816–825.

Joe, H. (1997). *Multivariate models and multivariate dependence concepts*, volume 73. CRC Press.

Jung, R. C. and Tremayne, A. R. (2011). Useful models for time series of counts or simply wrong ones? *AStA Advances in Statistical Analysis*, 95(1):59–91.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481.

Klein, J. and Moeschberger, M. (2003). *Survival Analysis: Techniques for censored and truncated data*. Statistics for Biology and Health. Springer, second edition.

Klein, N., Denuit, M., Lang, S., and Kneib, T. (2014). Nonlife ratemaking and risk management with Bayesian generalized additive models for location, scale, and shape. *Insurance: Mathematics and Economics*, 55:225 – 249.

Klugman, S. and Rioux, J. (2006). Toward a unified approach to fitting loss models. *North American Actuarial Journal*, 10(1):63–83.

Klugman, S. A., Panjer, H. H., and Willmot, G. E. (2012). *Loss models: from data to decisions*, volume 715. Wiley.

Klugman, S. A., Panjer, H. H., and Willmot, G. E. (2013). *Loss models: Further topics*. John Wiley & Sons.

Laevens, H., Deluyker, H., Schukken, Y., De Meulemeester, L., Vandermeersch, R., De Muelenaere, E., and De Kruif, A. (1997). Influence of parity and stage of lactation on the somatic cell count in bacteriologically negative dairy cows. *Journal of Dairy Science*, 80(12):3219–3226.

Lancaster, P. and Salkauskas, K. (1986). *Curve and surface fitting: An introduction*. London: Academic Press.

Lawless, J. (1994). Adjustments for reporting delays and the prediction of occurred but not reported events. *Canadian Journal of Statistics*, 22(1):15–31.

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.

Lee, G. and Scott, C. (2012). EM algorithms for multivariate Gaussian mixture models with truncated and censored data. *Computational Statistics & Data Analysis*, 56(9):2816 – 2829.

Lee, S. and McLachlan, G. J. (2014). Finite mixtures of multivariate skew t-distributions: some recent and new results. *Statistics and Computing*, 24(2):181–202.

Lee, S. C. and Lin, X. S. (2010). Modeling and evaluating insurance losses via mixtures of Erlang distributions. *North American Actuarial Journal*, 14(1):107–130.

Lee, S. C. and Lin, X. S. (2012). Modeling dependent risks with multivariate Erlang mixtures. *ASTIN Bulletin*, 42(1):153–180.

Lemaire, J. (1995). *Bonus–malus systems in automobile insurance*. Springer–Verlag, New York.

Lemaire, J., Park, S. C., and Wang, K. C. (2016). The use of annual mileage as a rating variable. *ASTIN Bulletin*, 46:39–69.

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.

Li, Y., Gillespie, B. W., Shedden, K., and Gillespie, J. A. (2015). Calculating profile likelihood estimates of the correlation coefficient in the presence of left, right or interval censoring and missing data. *Working paper.*

Lin, T. I. (2009). Maximum likelihood estimation for multivariate skew normal mixture models. *Journal of Multivariate Analysis*, 100(2):257 – 265.

Litman, T. (2011). Distance-based vehicle insurance feasibility, costs and benefits. Victoria Transport Policy Institute. `http://www.vtpi.org/dbvi_com.pdf`.

Litman, T. (2015). Pay-As-You-Drive Vehicle Insurance: Converting Vehicle Insurance Premiums Into Use-Based Charges. Victoria Transport Policy Institute. `http://www.vtpi.org/tdm/tdm79.htm`.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):226–233.

Mack, T. (1993). Distribution-free calculation of the standard error of chain ladder reserve estimates. *Astin bulletin*, 23(02):213–225.

Mailhot, M. (2012). *Mesures de risque et dépendance*. PhD thesis, Université Laval.

Marra, G. and Wood, S. N. (2012). Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics*, 39(1):53–74.

Martínez Miranda, M. D., Nielsen, J. P., Sperlich, S., and Verrall, R. (2013). Continuous chain ladder: Reformulating and generalizing a classical insurance problem. *Expert Systems with Applications*, 40(14):5588 – 5603.

Martín-Fernández, J. A., Palarea-Albaladejo, J., and Olea, R. A. (2011). Dealing with zeros. In Pawlowsky-Glahn, V. and Buccianti, A., editors, *Compositional Data Analysis: Theory and Applications*, pages 43–58. John Wiley & Sons.

Mateu-Figueras, G., Pawlowsky-Glahn, V., and Egozcue, J. J. (2011). The principle of working on coordinates. In Pawlowsky-Glahn, V. and Buccianti, A., editors, *Compositional Data Analysis: Theory and Applications*. John Wiley & Sons.

McCall, B. P. (1996). Unemployment insurance rules, joblessness, and part-time work. *Econometrica*, 64(3):647–82.

McCullagh, P. and Nelder, J. (1989). *Generalized linear models*. Chapman and Hall, New York, second edition.

McLachlan, G. and Jones, P. (1988). Fitting mixture models to grouped and truncated data via the EM algorithm. *Biometrics*, pages 571–578.

McLachlan, G. and Krishnan, T. (2008). *The EM algorithm and extensions*. Wiley series in probability and statistics. Wiley-Interscience.

McLachlan, G. and Peel, D. (2001). *Finite mixture models*. Wiley.

Midthune, D. N., Fay, M. P., Clegg, L. X., and Feuer, E. J. (2005). Modeling reporting delays and reporting corrections in cancer registry data. *Journal of the American Statistical Association*, 100(469):61–70.

Nair, V. N. (1984). Confidence bands for survival functions with censored data: a comparative study. *Technometrics*, 26(3):265–275.

Nelsen, R. B. (2006). *An Introduction to Copulas*. Springer, 2nd edition.

Neuts, M. F. (1981). *Matrix-geometric solutions in stochastic models: an algorithmic approach*. The John Hopkins University Press.

Norberg, R. (1993). Prediction of outstanding liabilities in non-life insurance. *ASTIN Bulletin*, 23(1):95–115.

Norberg, R. (1999). Prediction of outstanding liabilities II. Model variations and extensions. *ASTIN Bulletin*, 29(1):5–27.

Noufaily, A., Farrington, P., Garthwaite, P., Enki, D. G., Andrews, N., and Charlett, A. (2016). Detection of infectious disease outbreaks from laboratory data with reporting delays. *Journal of the American Statistical Association*, 111(514):488–499.

Noufaily, A., Ghebremichael-Weldeselassie, Y., Enki, D. G., Garthwaite, P., Andrews, N., Charlett, A., and Farrington, P. (2015). Modelling reporting delays for outbreak detection in infectious disease data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(1):205–222.

Olsson, M. (1996). Estimation of phase-type distributions from censored data. *Scandinavian journal of statistics*, pages 443–460.

Paefgen, J., Staake, T., and Fleisch, E. (2014). Multivariate exposure modeling of accident risk: Insights from pay-as-you-drive insurance data. *Transportation Research Part A: Policy and Practice*, 61:27 – 40.

Pagano, M., Tu, X. M., De Gruttola, V., and MaWhinney, S. (1994). Regression analysis of censored and truncated data: estimating reporting-delay distributions and aids incidence from surveillance data. *Biometrics*, pages 1203–1214.

Parry, I. W. H. (2005). Is pay-as-you-drive insurance a better way to reduce gasoline than gasoline taxes? *American Economic Review*, 95(2):288–293.

Pawlowsky-Glahn, V., Egozcue, J. J., and Tolosana-Delgado, R. (2015). *Modeling and Analysis of Compositional Data*. John Wiley & Sons.

Peel, D. and McLachlan, G. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing*, 10(4):339–348.

Pigeon, M., Antonio, K., and Denuit, M. (2013). Individual loss reserving with the multivariate skew normal distribution. *ASTIN Bulletin*, 43:399–428.

Pigeon, M., Antonio, K., and Denuit, M. (2014). Individual loss reserving using paid–incurred data. *Insurance: Mathematics and Economics*, 58:121–131.

Pigeon, M. and Denuit, M. (2011). Composite lognormal-Pareto model with random threshold. *Scandinavian Actuarial Journal*, 2011(3):177–192.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reynkens, T., Verbelen, R., Beirlant, J., and Antonio, K. (2016). Modeling censored losses using splicing: a global fit strategy with mixed erlang and extreme value distributions. *arXiv:1608.01566*.

Rootzén, H. and Tajvidi, N. (2006). Multivariate generalized pareto distributions. *Bernoulli*, 12(5):917–930.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression*. Cambridge: Cambridge University Press.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. Chapman & Hall.

Taylor, G. (2000). *Loss reserving: an actuarial perspective*. Kluwer Academic Publishers.

Tijms, H. C. (1994). *Stochastic models: an algorithmic approach*. Wiley.

Toledo, T., Musicant, O., and Lotan, T. (2008). In-vehicle data recorders for monitoring and feedback on drivers' behavior. *Transportation Research Part C: Emerging Technologies*, 16(3):320 – 331.

Tselentis, D. I., Yannis, G., and Vlahogianni, E. I. (2016). Innovative insurance schemes: Pay as/how you drive. *Transportation Research Procedia*, 14:362 – 371.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 290–295.

Van den Boogaart, K. G. and Tolosana-Delgado, R. (2013). *Analyzing compositional data with R*. Springer.

Verbelen, R., Antonio, K., and Claeskens, G. (2016a). Multivariate mixtures of Erlangs for density estimation under censoring. *Lifetime Data Analysis*, 22(3):429–455.

Verbelen, R., Antonio, K., and Claeskens, G. (2016b). Unraveling the predictive power of telematics data in car insurance pricing. *FEB Research Report KBI 1624*.

Verbelen, R., Antonio, K., Claeskens, G., and Crèvecoeur, J. (2017). Predicting daily IBNR claim counts using a regression approach for the occurrence of claims and their reporting delay. *Working paper*.

Verbelen, R., Gong, L., Antonio, K., Badescu, A., and Lin, X. S. (2015). Fitting mixtures of Erlangs to censored and truncated data using the EM algorithm. *ASTIN Bulletin*, 45(3):729–758.

Verrall, R. J. and Wüthrich, M. V. (2016). Understanding reporting delay in general insurance. *Risks*, 4(3):25.

Wahba, G. (1981). Spline interpolation and smoothing on the sphere. *SIAM Journal on Scientific and Statistical Computing*, 2(1):5–16.

Weiss, J. and Smollik, J. (2012). Beginner's roadmap to working with driving behavior data. *Casualty Actuarial Society E-Forum*, 2:1–35.

Willmot, G. E. and Lin, X. S. (2011). Risk modelling with the mixed Erlang distribution. *Applied Stochastic Models in Business and Industry*, 27(1):2–16.

Willmot, G. E. and Woo, J.-K. (2007). On the class of Erlang mixtures with risk theoretic applications. *North American Actuarial Journal*, 11(2):99–115.

Willmot, G. E. and Woo, J.-K. (2015). On some properties of a class of multivariate Erlang mixtures with insurance applications. *ASTIN Bulletin*, 45(01):151–173.

Wood, S. (2006). *Generalized additive models: an introduction with R*. Chapman and Hall/CRC Press.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1):3–36.

Wood, S. N. (2013). A simple test for random effects in regression models. *Biometrika*, 100(4):1005–1010.

Wüthrich, M. V. and Merz, M. (2008). *Stochastic claims reserving methods in insurance*, volume 435 of *Wiley Finance*. John Wiley & Sons.

Yin, C. and Lin, X. S. (2016). Efficient estimation of Erlang mixtures using iSCAD penalty with insurance application. *ASTIN Bulletin*, 46(3):779–799.

Zadeh, A. H. and Bilodeau, M. (2013). Fitting bivariate losses with phase-type distributions. *Scandinavian Actuarial Journal*, 2013(4):241–262.

Doctoral dissertations of the Faculty of Economics and Business

A list of doctoral dissertations from the Faculty of Economics and Business can be found at the following website:
http://www.kuleuven.be/doctoraatsverdediging/archief.htm.