**Fundação Getulio Vargas**
**Escola de Administração de Empresas de São Paulo**

**Alexandre Bassi Galdino**

# Nonparametric Extreme Value Mixture Models: Applications to Insurance Losses

São Paulo
2021

Alexandre Bassi Galdino

# Nonparametric Extreme Value Mixture Models: Applications to Insurance Losses

Dissertação apresentada à Escola de Administração de Empresas de São Paulo da Fundação Getulio Vargas (FGV-EAESP), como requisito parcial para a obtenção do grau de Mestre em Administração de Empresas.

**Área de Concentração**: Finanças

**Orientador**: Prof. Dr. Alan De Genaro

São Paulo
2021

Galdino, Alexandre Bassi.
Nonparametric extreme value mixture models: applications to insurance losses / Alexandre Bassi Galdino. - 2021.
99 f.

Orientador: Alan de Genaro.
Dissertação (mestrado CMAE) – Fundação Getulio Vargas, Escola de Administração de Empresas de São Paulo.

1. Teoria dos valores extremos. 2. Estatística não paramétrica. 3. Sinistro (Seguros). 4. Estatística - Análise. I. Genaro, Alan de. II. Dissertação (mestrado CMAE) – Escola de Administração de Empresas de São Paulo. III. Fundação Getulio Vargas. IV. Título.

CDU 368

Ficha Catalográfica elaborada por: Isabele Oliveira dos Santos Garcia CRB SP-010191/O
Biblioteca Karl A. Boedecker da Fundação Getulio Vargas - SP

Alexandre Bassi Galdino

# Nonparametric Extreme Value Mixture Models: Applications to Insurance Losses

Dissertação apresentada à Escola de Administração de Empresas de São Paulo da Fundação Getulio Vargas (FGV-EAESP), como requisito parcial para a obtenção do grau de Mestre em Administração de Empresas.

**Área de Concentração**: Finanças

**Data da Aprovação**: 28 de maio de 2021

**Banca Examinadora**:

---

Prof. Dr. Alan De Genaro Dario (Orientador)
FGV-EAESP

---

Prof. Dr. Rafael Felipe Schiozer
FGV-EAESP

---

Prof. Dr. Eduardo da Silva Flores
USP-FEA

## Agradecimentos

Gostaria de agradecer a todos que, de alguma forma, contribuíram para que esse trabalho fosse concluído. Em especial, agradeço à minha família e amigos pelo suporte, incentivo e compreensão constantes. Agradeço, também, ao Prof. Dr. Alan De Genaro pelas orientações e direcionamentos proporcionados durante o desenvolvimento dessa dissertação. Adicionalmente, agradeço aos professores Dr. Rafael Schiozer e Dr. Eduardo Flores, seus valiosos comentários enriqueceram e aumentaram significativamente a qualidade da pesquisa. Por fim, agradeço o apoio financeiro concedido pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) durante o período em que estive dedicado à elaboração desse trabalho.

**Resumo**

A modelagem da severidade de sinistros é um tópico desafiador para atuários e profissionais que atuam no mercado segurador. Modelos paramétricos comumente utilizados para aproximar as distribuições de severidade (Lognormal, Gama, Weibull, Burr Tipo XII, Gaussiana Inversa e Gamma Inversa) são capazes de fornecer um bom ajuste para os dados localizados no corpo das distribuições, mas falham ao descrever o comportamento das observações mais extremas. Uma abordagem popular empregada para superar essa limitação consiste em isolar a porção extrema das caudas e modelá-las separadamente utizando a célebre Teoria de Valores Extremos e a Distribuição Generalizada de Pareto, um método conhecido como *Peaks-Over-Threshold* (POT). Entretanto, na maioria das aplicações práticas, atuários estão interessados em obter um único modelo que proporcione um ajuste satisfatório em todo o suporte da distribuição. Nesta dissertação, consideramos uma mistura não-paramétrica de valores extremos capaz de modelar conjuntamente pequenas e grandes perdas. O modelo possui a vantagem de ser extremamente flexível devido ao seu componente não-paramétrico, evitando-se que seja necessário impor uma forma funcional para o corpo da distribuição, como na maioria dos modelos de mistura propostos na literatura atuarial. Adicionalmente, o estimador de densidade kernel tem apenas um único parâmetro adicional a ser estimado, superando o problema da complexidade computacional relacionado a modelos similares. Para demonstrar a aplicabilidade e efetividade do modelo proposto no contexto da modelagem da severidade de sinsitros, utilizamos três conjuntos de dados reais amplamente acessíveis e bastante explorados na literatura atuarial. Os resultados sugerem que o modelo analisado proporciona um ajuste superior quando comparado às outras alternativas existentes.

**Palavras-chave**: Severidade de Sinistros; Teoria de Valores Extremos; Métodos Não-Paramétricos; Densidade Kernel; Modelos de Mistura Finita.

**Abstract**

Modelling insurance losses is a challenging topic to actuaries and practitioners in the insurance industry. Commonly used loss models based on standard parametric density functions (Lognormal, Gamma, Weibull, Burr Type XII, Inverse Gaussian and Inverse Gamma) are often able to fit the bulk of the claim size distributions well but they fail to describe the behaviour of the most extremal observations. A popular approach used to overcome this limitation is to isolate the extreme data points and model them separately using Extreme Value Theory and the Generalized Pareto Distribution, an approach known as Peaks-Over-Threshold (POT) method. However, in most empirical applications, actuaries are interested in obtain a single model that provides a suitable global fit over the whole range of the distribution. In this thesis, we consider a nonparametric extreme value mixture model that is able to fit both small and large claims simultaneously. The model is extremely flexible due to its nonparametric component, avoiding the need to impose a functional form to the bulk of the loss distribution, as in most of the previous mixture approaches proposed in the actuarial literature. Further, the kernel density estimator has just a single extra parameter to be estimated, overcoming the problem of high computational burden related to other similar models. To illustrate the applicability and effectiveness of our model in the context of property and casualty losses, we consider three real data sets widely accessible and well-studied in the actuarial literature. The results suggest that the model provides a superior fit when compared with other existing alternatives.

**Keywords:** Insurance Losses; Extreme Value Theory; Nonparametric Methods; Kernel Density; Finite Mixture Models.

# Contents

**1 Introduction** **10**

**2 Nonparametric Extreme Value Mixture Models** **16**
2.1 Extreme Value Theory . . . . . 16
2.2 Nonparametric Kernel Density . . . . . 19
2.3 Nonparametric Extreme Value Mixture Model . . . . . 21
2.4 Statistical Inference . . . . . 23

**3 Applications to Insurance Data** **27**
3.1 Insurance Loss Data . . . . . 27
3.1.1 Danish Fire Insurance . . . . . 27
3.1.2 U.S. Automobile Insurance . . . . . 30
3.1.3 Australian Personal Injury Insurance . . . . . 33
3.2 Model Selection . . . . . 36
3.3 Results . . . . . 40
3.3.1 Threshold Selection . . . . . 41
3.3.2 Danish Fire Insurance . . . . . 46
3.3.3 U.S. Automobile Insurance . . . . . 51
3.3.4 Australian Personal Injury Insurance . . . . . 55
3.4 Robust Inference . . . . . 60
3.4.1 Method of Probability Weighted Moments . . . . . 62
3.4.2 Method of Trimmed Moments . . . . . 63
3.4.3 Method of Medians . . . . . 66
3.4.4 Method of Minimum Density Power Divergence . . . . 67
3.4.5 Robust Estimators . . . . . 70

**4 Concluding Remarks** **79**

**5 Bibliography** **81**

**A Table of Distribution Functions** **90**

**B Transformed Kernel Density** **94**

# 1 Introduction

In the context of non-life insurance, accurate statistical models for property and casualty losses are essential to a wide range of empirical applications. For instance, loss models are used for pricing insurance contracts, determine optimal retention levels in reinsurance operations, calculate risk measures, estimate Probable Maximum Losses (PML) and set the adequate level of risk-based capital requirements for solvency regulations.

Finding a good-fitting distribution for insurance loss data, however, is usually not an easy task. In particular, one of the main difficulties in modelling loss amounts is the inability to obtain a single theoretical model that provides a sufficiently accurate fit over the whole range of the distribution.

As highlighted by Bolancé, Guillen and Nielsen (2003), "Actuaries are interested in having good estimates for all the values in the domain range: small losses because they are very frequent, medium losses causing a dramatic increase of expenses (demanding liquidity) and large losses that may mean that reinsurance contracts must be reconsidered."

Distributions of insurance losses have positive support and, in most cases, are unimodal hump-shaped, right-skewed and extremely heavy-tailed (Punzo, Bagnato and Maruotti, 2018). Commonly used standard parametric models such as Gamma, Lognormal, Weibull and Inverse Gaussian are often able to fit the bulk of the loss distribution well, but they fail to capture adequately the behaviour of the most extreme observations, underestimating the probability of large claims.

As an alternative, empirical actuarial analyses isolate large losses and model them separately. Extreme Value Theory (EVT) and the Peaks-Over-Threshold (POT) methodology are used to describe the behaviour of these large claims. These methods assume that, for a sufficiently high threshold, the distribution of excesses may be well approximated by a Generalized Pareto Distribution (GPD). Successful applications of the EVT in the actuarial loss modelling context can be easily found in the literature (Rootzén and Tajdivi, 1997; McNeil, 1997; McNeil and Saladin, 1997; Resnick, 1997; Cebrián, Denuit and Lambert, 2003).

The Peaks-Over-Threshold (POT) approach, in turn, is only suitable to model large losses and no information is obtained from the small losses located below the threshold. McNeil (1997) suggests that "if smaller losses were also of interest we could in any case use a mixture model so that one model applied to the tail and another to the main body of the data."

Following McNeil's suggestion, one stream of literature in loss modelling discusses a wide variety of composite or splicing models. The purpose of these models is to combine a light-tailed distribution for the bulk, which covers small and moderate losses (also called attritional losses), and a heavy tail distribution for the tail to capture the behaviour of large losses (also known as catastrophic losses).

Cooray and Ananda (2005) introduced a two parameter composite Lognormal-Pareto model that is a Lognormal density up to a threshold (henceforth denominated as splicing point) and a Pareto distribution for the remainder density. The authors impose continuity and diferentiability conditions at the splicing point to ensure a continuous and smooth density and to reduce the number of parameters to be estimated. The resulting weighted density is similar in shape to a Lognormal, but its upper tail behaviour is heavier than the Lognormal and quite similar to the Pareto density.

Scollnik (2007) extended the Lognormal-Pareto model proposed by Cooray and Ananda (2005) by developing a composite model featuring mixing weights that are not fixed *a priori*. The author show that, under the continuity constraint, the mixing weights can be determined as a function of the model parameters. He also developed a second splicing model by replacing the classical single parameter Pareto distribution with the Pareto Type II (Lomax) distribution in the second piece of the model (*i.e.* the right tail of the loss distribution).

Many other composite models have been proposed in the literature, including Exponential-Pareto models (Klugman, Panjer and Willmot, 2008), Normal-Pareto models (Carreau and Bengio, 2009), Lognormal-Pareto models with random thresholds (Pigeon and Denuit, 2011), Weibull-Pareto models (Scollnik and Sun, 2012), Lognormal-Burr models (Nadarajah and Bakar, 2014), Weibull-Burr models (Bakar *et al.* 2015) and Lognormal/Weibull-Stoppa models (Calderín-Ojeda and Kwok, 2016). Brazauckas and Kleefeld

(2016), Grün and Miljokovic (2019) and Wang, Haff and Huseby (2020) compared the modelling performance of several of these splicing models on real insurance loss data.

The class of composite models, however, may not always be flexible enough in terms of the possible shapes of their components in order to obtain a satisfying fit, especially in the few cases where the loss distribution presents multimodality. For this reason, another stream of literature suggests the use of the broader class of finite mixture models to fit insurance losses. Mixture distributions allow for the flexibility to easily add additional components as compared to composite modelling that is limited to two distributions only (Miljkovic and Grün, 2016).

Keatinge (1999) introduced the class of mixture models in the actuarial literature by proposing finite mixtures of exponential distributions (also known as hyperexponential distributions) to model insurance losses. Hyperexponential distributions have decreasing failure rates, which means that their tails are thicker than the exponential tail, a desirable property in our context. On the other hand, it can be shown that mixtures of exponential distributions are completely monotonic, implying that they always have a zero mode.

To overcome this limitation, Klugman and Rioux (2006) suggested the augmented mixture of Exponential distributions. As defined by the authors, this class of finite mixture models includes not only exponential, but also Lognormal, Gamma and Pareto components, with the restriction, for purposes of parsimony, that the Gamma and Lognormal components are not simultaneously included in the mixture. The addition of either the Lognormal or Gamma components allows for nonzero interior modes and the Pareto component is included to add thickness to the right tail of the mixture.

Lee and Lin (2010) argued that even the augmented mixtures of exponential distributions might not be appropriate for insurance and actuarial applications, since they still have only three modes at most. Thereby, the authors suggested the class of mixtures of Erlang distributions with common scale parameter to model loss data. Erlang mixtures are dense in the space of positive continuous distributions, *i.e.* they converge weakly to any probability measure on $(0, \infty)$. This means that any positive continuous distribution

can be approximated by an Erlang mixture with enough components.

Verbelen *et al.* (2015) extended the Erlang mixture model presented by Lee and Lin (2010) to truncated and censored losses, a very common setting in many actuarial applications due to the presence of deductibles and coverage limits, which are included in most insurance policies.

Miljkovic and Grün (2016) proposed mixture models where the components are assumed to be from the same parametric family of distributions. They considered six non-Gaussian parametric families previously used in actuarial modelling: Burr Type XII, Inverse Burr Type XII, Gamma, Inverse Gamma, Lognormal and Weibull. Blostein and Miljkovic (2019) extended their approach to left-truncated data, allowing for mixture models with any combination of Gamma, Lognormal and Weibull distributions.

In general, when using finite mixtures of standard parametric distributions (*e.g.* Exponential, Erlang, Gamma and Normal) to approximate an unknown density function with extremely heavy tails, as in the case of loss distributions, a large number of mixture components might be needed. An alternative approach to obtain more parsimonious parametrizations is modelling the tail distribution using the classical Extreme Value Theory (EVT) and the bulk distribution using a standard parametric model or even a finite mixture. This approach is precisely what we refer here as extreme value mixture models.

The literature on extreme value mixtures is quite large. Frigessi, Haug and Rue (2002) suggested a dynamically weighted extreme value mixture model where the upper tail term is a GPD and the bulk is described by a light-tailed Weibull density function. The mixing function varies on $\mathbb{R}^+$ in such a way that for large values the GPD component is predominant and thus takes the role of threshold selection.

Mendes and Lopes (2004) proposed a extreme value mixture where both the right and left tails are GPD and the center is modelled by a Normal distribution. They suggest to initially perform robust standardization techniques to distinguish the bulk of the data from the tails, amplifying extreme data points. The parameters are then estimated by maximum likelihood procedures. Behrens, Lopes and Gamerman (2004) introduced a similar extreme

value mixture where the bulk distribution is assumed to follow a Gamma distribution and the upper tail is GPD. The authors also proposed a Bayesian inference method to account for threshold uncertainty.

In order to provide further flexibility to the extreme value mixture framework and avoid over-restrictive parametric assumptions, some authors have proposed the use of semiparametric and nonparametric methods to describe the observations located below the threshold.

Tancredi, Anderson and O'Hagan (2006) suggested a mixture of uniform densities to model the central part of the data. Nascimento, Gamerman and Lopes (2012) chose to employ a finite mixture of Gammas, as the Gamma belongs to the maximum domain of attraction of the Generalized Extreme Value (GEV) distribution. Patiño (2015) showed that a Dirichlet process mixture of Gammas adds further flexibility to the possible shapes assumed by the central distribution. Cabras and Castellanos (2011) employed, in turn, the Lindsey's semiparametric method (Lindsey, 1974a; Lindsey,1974b) to estimate the bulk density conditionally on the threshold value.

In the nonparametric setting, MacDonald *et al.* (2011) introduced the class of extreme value kernel mixtures, where the bulk distribution is described by a nonparametric kernel density estimator and the tails are GPD.

Despite the popularity of extreme value mixtures in many areas of statistics, these models are still mostly unexplored in the actuarial literature. To the best of our knowledge, Lee, Li and Wong (2012) and Reynkens *et al.* (2017) are the only two papers to date that employed extreme value mixtures to fit insurance loss data.

Li, Lee and Wong (2012) proposed an extreme value mixture in which the bulk distribution is assumed to follow a two component mixture of exponential distributions (hyperexponential) up to a fixed threshold and a Generalized Pareto Distribution (GPD) for the remainder density. One major drawback of their model, similar to Keating (1999), is the monotonically decreasing behaviour of the hyperexponential distribution. Hence, the fitted distributions will have no interior modes, a potential problem when dealing with property and casualty losses.

Reynkens *et al.* (2017) overcomes this limitation by replacing the two

component exponential mixture in the bulk distribution with a mixture of Erlangs with common scale parameter, allowing for further flexibility and interior modes.

In this thesis, we extend the actuarial literature on extreme value mixture models by investigating the performance of a nonparametric extreme value mixture, originally proposed by MacDonald *et al.* (2011), to fit insurance loss data.

The extreme value kernel mixture model is extremely flexible due to its nonparametric component, avoiding the need to impose a parametric/semiparametric form to the bulk distribution, as in the previous extreme value mixture approaches proposed in the actuarial literature. Further, the kernel density estimator has just a single extra parameter to be estimated (the bandwidth parameter), overcoming the problem of high computational burden related to other similar models.

To illustrate the effectiveness of the proposed model to describe the severity of property and casualty losses, we consider three real claims data sets widely accessible and well-studied in the actuarial science literature: (i) the Danish Fire Insurance Data (McNeil, 1997; Reisneck, 1997), (ii) the U.S. Automobile Insurance Data (Frees, 2010; Punzo, Bagnato and Maruotti, 2018), and (iii) the Australian Personal Injury Insurance Data (de Jong and Heller, 2008).

The remainder of the thesis is organized as follows. In section 2, we briefly outline the proposed nonparametric extreme value mixture as well as the inference procedure used to estimate its parameters. In section 3, we describe the real insurance loss data sets used in our empirical applications and the model selection criteria. In this section, we also present and analyse the obtained results. Finally, section 4 provides a general discussion of our findings and some concluding remarks are given.

# 2 Nonparametric Extreme Value Mixture Models

## 2.1 Extreme Value Theory

Classical Extreme Value Theory, as its name suggests, is frequently used to model data that present "unusual" or extreme behaviour. In the context of non-life insurance, particular interest lies in studying losses arising from extreme events, since a few collection of large claims often represents the greatest part of the indemnities paid by an insurance company. Understanding the behaviour of these large claims is essential for designing reinsurance contracts and to define the adequate level of capital requirements for solvency purposes.

Extreme value analysis was first developed as a robust alternative to standard statistical techniques, as they proved to be very inefficient to analyse observations located deep in the tails of the distributions. In fact, most traditional statistical theory typically examines only the "usual" or average behaviour of the underlying processes (*e.g.* Law of Large Numbers and Central Limit Theorem). Extreme Value Theory (EVT), on the other hand, provides sound probabilistic arguments and form the theoretical foundation for describing extremes of random phenomena (Embrechts, Klüppelberg and Mikosch, 1997).

Formally, let $\mathscr{X} = \{X_1, X_2, X_3, ...\}$ denote a collection of independent and identically distributed (*i.i.d.*) random variables with distribution function $F(x)$, perhaps a loss severity distribution observed by an insurance company. We will denote the maximum of the first $n$ observations as $M_n = \max(X_1, ..., X_n)$. Fisher and Tippett (1928) and Gnedenko (1943) showed that we can find a sequence of real numbers $a_n > 0$ and $b_n$, such that $\frac{(M_n - b_n)}{a_n}$ converges in distribution to the Generalized Extreme Value (GEV) distribution,

$$F(x|\mu,\sigma,\xi) = \begin{cases} e^{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{-\frac{1}{\xi}}}; & \xi \neq 0 \\ e^{-e^{-\left[\frac{x-\mu}{\sigma}\right]_+}}; & \xi = 0 \end{cases} \tag{1}$$

where $[y]_+ = \max(y, 0)$, $\sigma > 0$ and $\xi, \mu \in \mathbb{R}$. The parameters $\mu$, $\sigma$ and $\xi$ are the location, scale and shape parameters respectively. For $\xi = 0$ the distribution function is interpreted as the limit when $\xi \to 0$.

Just as the Normal distribution is a canonical limiting distribution in the study of sample means, the GEV distribution is an important distribution in the study of the limiting behaviour of sample extremes (McNeil, 1997).

However, as the reader may have noted in the above definition, modelling extremal observations using the GEV distribution is a very wasteful method, since the only sample information considered is the maximum of each sample block of $n$ observations. A more resourceful and convenient approach consists in modelling extreme events as those exceeding some high fixed threshold value $u$.

In other words, we are interested in finding a conditional distribution that describes the behaviour of sample extremes, given that they exceed the threshold value, $\mathbb{P}[X \leq x|X > u]$. Pickands (1975) and Balkema and de Haan (1974) proved that, under certain mild conditions, this conditional distribution converges asymptotically to a Generalized Pareto Distribution (GPD), with distribution function given by

$$G(x|u,\sigma,\xi) = \mathbb{P}[X \leq x|X > u] = \begin{cases} 1 - \left[1 + \xi\left(\frac{x-u}{\sigma}\right)\right]_+^{\frac{-1}{\xi}}; & \xi \neq 0 \\ 1 - \exp\left[-\left(\frac{x-u}{\sigma}\right)\right]_+; & \xi = 0 \end{cases} \tag{2}$$

where $x > u$, $[y]_+ = \max(y, 0)$ and $\xi \in \mathbb{R}, \sigma > 0$ are respectively the shape and scale parameters.

The corresponding density function can be written as

$$g(x|u,\sigma,\xi) = \begin{cases} \frac{1}{\sigma}\left[1+\xi\left(\frac{x-u}{\sigma}\right)\right]_+^{-1-\frac{1}{\xi}}; & \xi \neq 0 \\ \frac{1}{\sigma}\exp\left[-\left(\frac{x-u}{\sigma}\right)\right]_+; & \xi = 0 \end{cases} \tag{3}$$

The shape parameter $\xi$ determines the thickness of the tails, when

- $\xi < 0$: light tail with finite support and endpoint at $u - \frac{\sigma}{\xi}$;
- $\xi = 0$: exponential tail; and
- $\xi > 0$: heavy tail.

Fitting extreme observations using the GPD is known as Excess-Over-Threshold (EOT) or Peaks-Over-Threshold (POT) method. Successful applications of the POT methodology to actuarial loss modelling can be easily found in the literature (Rootzén and Tajvidi, 1997; McNeil, 1997; Resnick, 1997; McNeil and Saladin, 1997; and Cebrián, Denuit and Lambert, 2003).

As defined in equation (2), the GPD is a conditional distribution able to fit excess $(x-u)$. However, it is also possible to fit it to the tail of an unconditional distribution, assuming that $x$ is above the threshold $u$. More precisely, we can represent it as a truncated distribution with lower truncation point at the threshold value $u$.

Note that the conditional probability $\mathbb{P}[X \leq x | X > u]$ can be rewritten as

$$\mathbb{P}[X \leq x | X > u] = \frac{\mathbb{P}[u \leq X \leq x]}{1 - \mathbb{P}[X \leq u]} = \frac{F(x) - F(u)}{1 - F(u)}$$

Now let $\phi_u := \mathbb{P}[X > u]$. After some algebraic manipulations, the unconditional distribution function $F(x)$ can be written as

$$F(x) = \mathbb{P}[X \leq x] = (1 - \phi_u) + \phi_u \, \mathbb{P}[X \leq x | X > u], \;\; x > u$$

Therefore, for $x > u$, we can use the following expression to model the right tail of an unconditional distribution

$$F(x) = (1 - \phi_u) + \phi_u G(x|u, \sigma, \xi) \tag{4}$$

where $\phi_u$ is estimated by the sample proportion above the threshold (Patrik, 1980; p. 76).

## 2.2 Nonparametric Kernel Density

Nonparametric statistics is the branch of statistics devoted to develop methods that do not make any assumption about the functional form of the data generating process being analysed. The main advantage of this approach lies in the greater flexibility provided by these methods. Flexibility, however, does not come without a cost. By imposing less structure on the functional form of the analysed data generating process, nonparametric models require more data to achieve the same degree of precision as a corrected specified parametric model (Li and Racine, 2007).

The use of nonparametric statistics is suitable for modelling loss data and it is useful in actuarial practice since insurance portfolios are usually very large (Bolancé, Guillen and Nielsen, 2003). The actuarial literature, however, has not made extensive use of these methods to model insurance losses. A few exceptions are Bolancé, Guillen and Nielsen (2003), Buch-Kromann (2006) and Balasooriya and Low (2008). We believe that this lack of interest can be explained by the fact that nonparametric smoothing fails to provide consistent estimates at the tails of the distributions, a limitation that our model is able to overcome.

Let $\mathscr{X}_N = \{X_1, X_2, ..., X_N\}$ be a set of *i.i.d.* random variables with density function $f(x)$. Suppose we are interested in estimating the unknown density $f(x)$ based on an observed sample $\boldsymbol{X} = \{x_i : i = 1, ..., N\}$. The classical kernel density estimator is a consistent nonparametric estimator for $f(x)$ given by

$$\hat{f}(x) = \frac{1}{N\lambda} \sum_{i=1}^{N} K\left(\frac{x - x_i}{\lambda}\right) \tag{5}$$

where $K(.)$ is known as kernel function and $\lambda$ is a positive smoothing parameter often called bandwidth. Note that the kernel density estimator is nonparametric since it does not make any distributional assumption about the underlying density function.

We can also define the kernel density estimator in its scale notation (Wand and Jones, 1995)

$$\hat{f}(x) = N^{-1} \sum_{i=1}^{N} K_\lambda\left(x - x_i\right) \tag{6}$$

where $K_\lambda(x) = \frac{1}{\lambda} K(\frac{x}{\lambda})$.

Moreover, the kernel function $K : \mathbb{R} \to \mathbb{R}$ must attend the following conditions

$$\int K(x)dx = 1$$

and

$$\int x^2 K(x)dx > 0.$$

These conditions ensure that the resulting kernel density estimate is a valid probability density function and a consistent estimator for $f(x)$.

In addition, it is also common to define the kernel function as an unimodal function that is symmetric around zero, $K(x) = K(-x)$, thus ensuring that the average of the corresponding distribution is equal to that of the sample used. This can be summarized as an additional condition

$$\int xK(x)dx = 0$$

This last condition, however, can be relaxed. For details about nonparametric kernel density estimation using asymmetric kernel functions, refer to Abadir and Lawford (2004).

There are several possible choices for the kernel function. Some of the most commonly used are the Gaussian, Epanechnikov, Silverman, Uniform and triangular kernels.

Nevertheless, little guidance is given in the literature about the optimal choice of the kernel function. Wand and Jones (1995) show that the Epanechnikov kernel is the optimal kernel function since it is obtained from minimization of the Asymptotic Mean Squared Error (AMSE) over nonnegative kernels. Tsybakov (2009), on the other hand, argues that nonnegativity of density estimators is not necessarily achieved via nonnegative kernel functions. He shows that, when the nonnegativity constraint is relaxed, the Epanechnikov kernel is no longer the kernel function with smaller AMSE.

Because of this lack of consensus, most authors use the simple gaussian kernel as kernel function in nonparametric density estimation. We choose to follow this approach. The gaussian kernel function can be defined as

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}; \;\; -\infty < x < \infty \tag{7}$$

which is simply the density function of a standard Normal distribution.

## 2.3 Nonparametric Extreme Value Mixture Model

The class of nonparametric extreme value mixtures was first introduced by MacDonald *et al.* (2011). According to this model, observations below a given threshold $u$, where data are typically abundant, are described by a nonparametric kernel density $h(.|\lambda, \mathbf{X})$, where $\lambda$ is the bandwidth parameter and $\mathbf{X}$ is the vector of sample observations. Observations above the threshold $u$ (*i.e.* the upper tail of the distribution), on the other hand, are assumed to follow a Generalized Pareto Distribution (GPD) with support starting at the threshold value $u$.

Therefore, the nonparametric extreme value model is defined by the following distribution function

$$F(x|\lambda, u, \sigma, \xi, \mathbf{X}) = \begin{cases} (1-\phi_u)\frac{H(x|\lambda,\mathbf{X})}{H(u|\lambda,\mathbf{X})}; & x \leq u \\ (1-\phi_u) + \phi_u G(x|u,\sigma,\xi); & x > u \end{cases} \tag{8}$$

where $(1+\phi_u) + \phi_u G(.|u,\sigma,\xi)$ is the unconditional GPD distribution function given by equation (4).

The corresponding density function can be defined as

$$f(x|\lambda, u, \sigma, \xi, \boldsymbol{X}) = \begin{cases} (1-\phi_u)\frac{h(x|\lambda,\boldsymbol{X})}{H(u|\lambda,\boldsymbol{X})}; & x \leq u \\ \phi_u\, g(x|u,\sigma,\xi); & x > u \end{cases} \tag{9}$$

To facilitate the inference process, as described in the next section, it is convenient to represent the above nonparametric extreme value composite model as a more general finite mixture model. Mixtures of probability distributions are simply convex combinations of their individual components. Thus, a $K$-component finite mixture can be written as

$$f(x|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_k(x|\boldsymbol{\Psi}_k)$$

where $\boldsymbol{\theta} = (\boldsymbol{\pi}', \boldsymbol{\Psi}')' = (\pi_1, \pi_2, ..., \pi_{K-1}, \boldsymbol{\Psi}'_1, \boldsymbol{\Psi}'_2, ..., \boldsymbol{\Psi}'_K)'$ is a vector of unknown parameters, $\pi_k$ denotes the mixing weight of the $k$th component satisfying $0 < \pi_k < 1$, $\forall k \in \{1, ..., K\}$, and $\sum_{k=1}^{K} \pi_k = 1$, and $\boldsymbol{\Psi}_k$ are the parameters of the $k$th density function $f_k(.|\boldsymbol{\Psi}_k)$.

According to the above definition, we can reformulate our nonparametric extreme value composite model as a finite mixture of two components

$$f(x) = \pi f_1(x) + (1-\pi) f_2(x) \tag{10}$$

where the mixing weight is given by $\pi = (1 - \phi_u)$ and

$$f_1(x) = \frac{h(x|\lambda, \mathbf{X})}{H(u|\lambda, \mathbf{X})} \mathbb{1}_{(-\infty,u]}(x)$$
$$f_2(x) = g(x|u, \sigma, \xi) \mathbb{1}_{(u,\infty)}(x)$$

## 2.4 Statistical Inference

In the finite mixture models framework, we can think of our sample as incomplete data. More formally, let $\mathbf{X}$ be again a vector of random variables corresponding to our observed sample $\mathbf{X} = \{x_i : i = 1, ..., N\}$. Each one of these sample points is assumed to have originated from one of the mixture components. Thus, we may want to define a random variable $\mathbf{Z} = (z_{ik} \in \{0, 1\}, i = 1, ..., N; k = 1, ..., K)$ indicating for all sample observations from which component of the mixture they came from.

The complete data random vector is then defined as $\mathbf{C} = (\mathbf{X}, \mathbf{Z})$, that is, the observed sample information and the additional information indicating for each sample point their originating mixture component, $\{c_i\} = \{x_i, z_{i1}, z_{i2}, ..., z_{iK}\}$, where $z_{ij} = 1$ if observation $x_i$ comes from component $f_j(x|\mathbf{\Psi}_j)$ and zero otherwise.

To properly estimate the vector of unknown parameters associated with a finite mixture model via Maximum Likelihood Estimators (MLE), we need the likelihood function based on the complete data information vector $\mathbf{C}$. This complete data likelihood function is defined as

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{k=1}^{K} \left[\pi_k f_k(x_i|\mathbf{\Psi}_k)\right]^{z_{ik}} \tag{11}$$

The natural logarithm of the complete data likelihood function is then given by

$$l(\boldsymbol{\theta}) = \ln L(\boldsymbol{\theta}) = \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik} \log[\pi_k f_k(x_i|\boldsymbol{\Psi}_k)] \tag{12}$$

Using equation (11), the likelihood function for the nonparametric extreme value mixture model can be separated out into the contributions from the observations below the threshold (kernel density component) and those above the threshold (GPD tail model)

$$L(\boldsymbol{\theta}|u, \mathbf{X}) = L_K(\lambda|u, \mathbf{X}) L_{GPD}(\sigma, \xi|u, \mathbf{X}) \tag{13}$$

where the parameter vector is $\boldsymbol{\theta} = (\lambda, \sigma, \xi)$.

The $L_{GPD}$ likelihood function is obtained by straightforward application of the likelihood function using the GPD density function described in equation (3)

$$L_{GPD}(\sigma, \xi|u, \mathbf{X}) = \begin{cases} \phi_u^{|B|} \prod\limits_B \frac{1}{\sigma} \left[1 + \xi(\frac{x_i - u}{\sigma})\right]^{-1-\frac{1}{\xi}}; & \xi \neq 0 \\ \phi_u^{|B|} \prod\limits_B \frac{1}{\sigma} e^{-(\frac{x_i - u}{\sigma})}; & \xi = 0 \end{cases} \tag{14}$$

where $B = \{i : x_i > u\}$.

The likelihood function associated with the nonparametric component $L_K(\lambda|u, \mathbf{X})$, on the other hand, is much less intuitively constructed, since the estimation of the smoothing parameter $\lambda$ is not as direct as in the case of the GPD parameters.

In most empirical applications, the selection of the bandwidth often plays a much more important role in the accuracy of the kernel density estimation than the choice of kernel function $K(.)$ itself (Frees, 2010).

The actuarial literature typically estimates the optimal bandwidth parameter using the normal scale selection method, also known as Silverman's bandwidth selection method (Silverman, 1986), given by $\hat{\lambda} = \left(\frac{4\hat{\sigma}^5}{3N}\right)^{1/5}$, where

$\hat{\sigma}$ is the sample standard deviation.

This selection method, however, is based on the assumption that the data were generated from a Normal distribution, an assumption that is not feasible in our case. As an alternative, more sophisticated methods have been developed in the statistical and econometrics literature for global bandwidth selection in univariate nonparametric kernel density estimation (Li and Racine, 2007).

Likelihood inference is one of these methods. It was first proposed by Habbema, Hermans and van den Broek (1974) and Duin (1976) and treats the bandwidth as an additional parameter to be estimated. The corresponding likelihood function for the nonparametric kernel component can be written as the product

$$\prod_{i=1}^{N} N^{-1} \sum_{j=1}^{N} K_{\lambda}(x_i - x_j),$$

It can be shown that this likelihood function is unbounded as $\lambda \to 0$, as each sum term in the product of the above likelihood is infinite in the limit $\lambda \to 0$ because the term $(x_i - x_j)$ becomes zero when $i = j$ (Duin, 1976). To avoid this degeneracy is common practice to replace the likelihood function with the likelihood cross-validation (LCV) function

$$L(\lambda|u, \mathbf{X}) = \prod_{i=1}^{N} \frac{1}{(N-1)} \sum_{\substack{j=1 \\ j \neq i}}^{N} K_{\lambda}(x_i - x_j) \tag{15}$$

which is simply the leave-one-out density estimate. Bowman (1984) shows that maximizing (15) is equivalent to minimizing the Kullback-Leibler distance between the estimated nonparametric kernel density and the true underlying density function. This likelihood cross-validation method for the kernel bandwidth selection is used in the likelihood function for the proposed model.

It is worth mentioning that the described likelihood based kernel band-

width estimator has elicited little interest in the literature as it is extremely affected by the tail behaviour of the underlying density function $f(x)$. It may work well for short tailed distributions, but it drastically oversmooth heavy tailed distributions, yielding inconsistent kernel density estimates (Schuster and Gregory, 1981). In fact, the likelihood cross-validation based inference will tend to give smoothing parameters which are far too large not only for heavy tailed distributions, but also in situations where extremal observations (outliers) are present (Scott and Factor, 1981).

Within the proposed nonparametric extreme value mixture model, the right tail of the loss distribution is captured by the GPD component, so the inconsistency of the likelihood cross-validation based estimator for heavy-tailed distributions is overcome.

Finally, note that since all $N$ sample observations are used as kernel centres and only those below the threshold in the set $A = \{i : x_i \leq u\}$ contribute to the likelihood function, the kernel density needs to be rescaled to get the appropriate contribution to the likelihood, resulting in the following expression

$$L_K(\lambda|u, \mathbf{X}) = \left[ \frac{(1-\phi_u)}{\frac{1}{N}\sum_{j=1}^{N}\{\int_{-\infty}^{u} K_\lambda(z-x_j)dz\}} \right]^{|A|} \prod_{A} \frac{1}{(N-1)} \sum_{\substack{j=1 \\ j\neq i}}^{N} K_\lambda(x_i - x_j) \tag{16}$$

We apply the MLE method to estimate the parameter vector $\hat{\boldsymbol{\theta}} = (\hat{\lambda}, \hat{\sigma}, \hat{\xi})$. The MLE method chooses the vector $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ on the parameter space as the one that maximizes the logarithm of the likelihood function described in equation (13)

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \ln L(\boldsymbol{\theta}|u, \boldsymbol{X})$$

For large sample sizes and under certain regularity conditions, MLE estimators are consistent and asymptotically normally distributed

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{0}})$$

where $\boldsymbol{\Sigma}_{\mathbf{0}} = \boldsymbol{I}^{-1}$ and $\boldsymbol{I}$ is the Fisher information matrix given by

$$\boldsymbol{I}_{ij} = \mathbb{E}\left[-\frac{\partial^2 \ln L(\boldsymbol{\theta}|u, \boldsymbol{X})}{\partial\theta_i \partial\theta_j}\right].$$

# 3 Applications to Insurance Data

## 3.1 Insurance Loss Data

### 3.1.1 Danish Fire Insurance

The Danish Fire Insurance claim data set consists of 2,492 observations of fire loss claims in millions of Danish Kroner (DKK) adjusted by inflation to reflect 1985 values. The claims occurred in Copenhagen between 1980 and 1990, both years included. The loss figure includes damage to buildings, furniture and personal property as well as loss of profits. Claims settled with zero payment are not included. The data set was made available by the Copenhagen Reinsurance Company and can be found in the R **SMPraticals** add-on package.

These data are perhaps the most famous data set in the actuarial literature. It has been extensively analysed by several authors, including McNeil (1997), Cooray and Ananda (2005), Scollnik (2007), Pigeon and Denuit (2011), Scollnik and Sun (2012), Lee, Li and Wong (2012), Bakar *et al.* (2015) and Reynkens *et al.* (2017).

Descriptive statistics for the Danish Fire Insurance data set are presented in Table 1.

| | |
|---|---|
| Number of Obs. | 2,492 |
| Mean | 3.06 |
| Median | 1.63 |
| Standard Dev. | 7.97 |
| Skewness | 19.88 |
| Kurtosis | 549.13 |
| Minimum | 0.31 |
| Maximum | 263.25 |
| 25% Quantile | 1.15 |
| 75% Quantile | 2.64 |

**Table 1.** Summary Statistics - Danish Fire Insurance

The statistics above show that the empirical loss distribution is positive skewed (19.88) with a high excesses kurtosis coefficient (leptokurtosis) (549.13). These are common features in many observed loss distributions (Punzo, Bagnato and Maruotti, 2018) and are indicatives that the empirical distribution has a heavy right tail. Skewness and tail thickness can also be inferred from the distance between the quartiles and the maximum/minimum values. In our case, both the first (25% quantile) and the second quartiles (median) are relatively close to the minimum value, but the third quartile (75% quantile) is very far apart from the sample maximum.

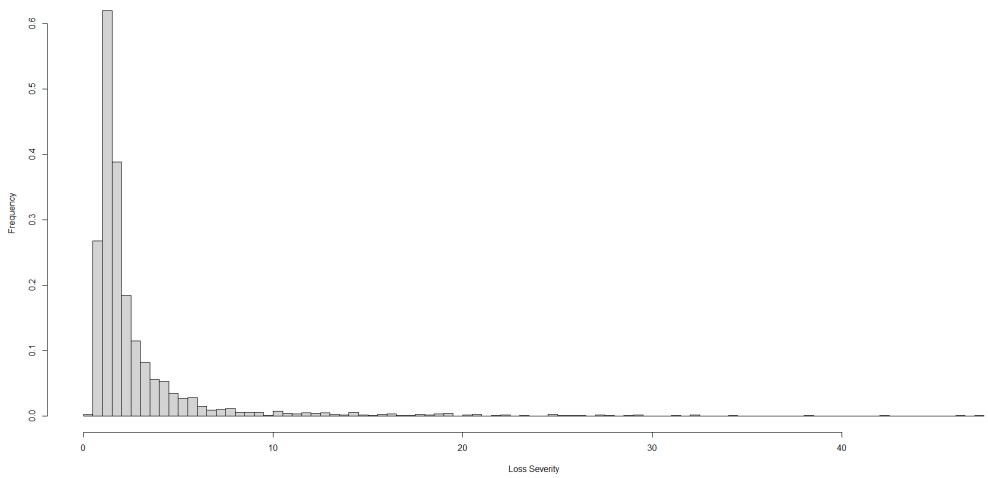
To further investigate the characteristics of the analysed data, Figure 1 presents the histogram, along with the box plot and the exponential QQ-plot for the Danish Fire Insurance data set.

The exponential QQ-plot is a very useful guide to the identification of heavy-tailed distributions and has been employed extensively in the extreme value analysis literature (McNeil, 1997; Cebrián, Denuit and Lambert, 2003). This plot is capable to examine, in a very intuitive way, the hypothesis that the loss data come from an exponential distribution. If the data have an exponential shape, data points should lie approximately in a 45 degree straight line. Concave departures indicates thin-tailed distributions (*i.e.* tails decaying faster than the exponential decay). Convex shapes, on the other hand, indicate heavy tails.

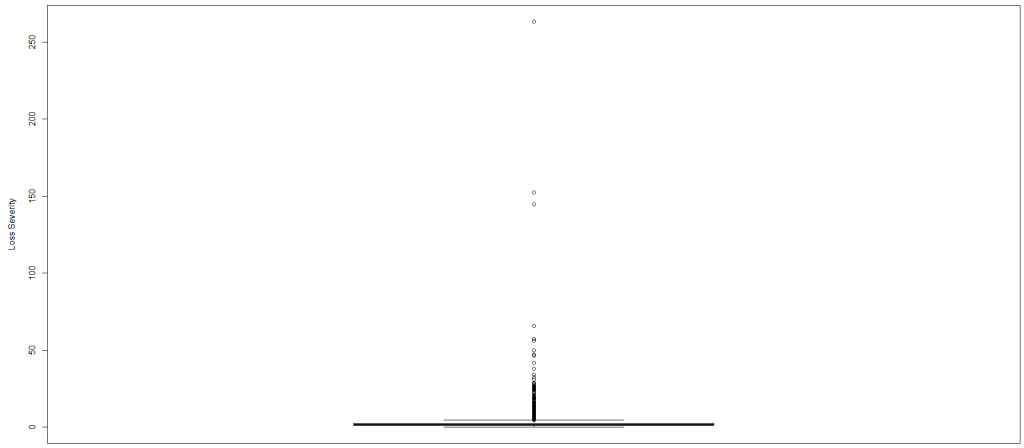
Is worth mentioning that the exponential QQ-plot, as any other type of

QQ-plot, has some caveats. Even data draw from an exponential distribution may show departures from the expected behaviour. However, for large sample sizes (over 2,000 sample points) it seems safe to rely on the conclusions draw from the plot (McNeil, 1997).

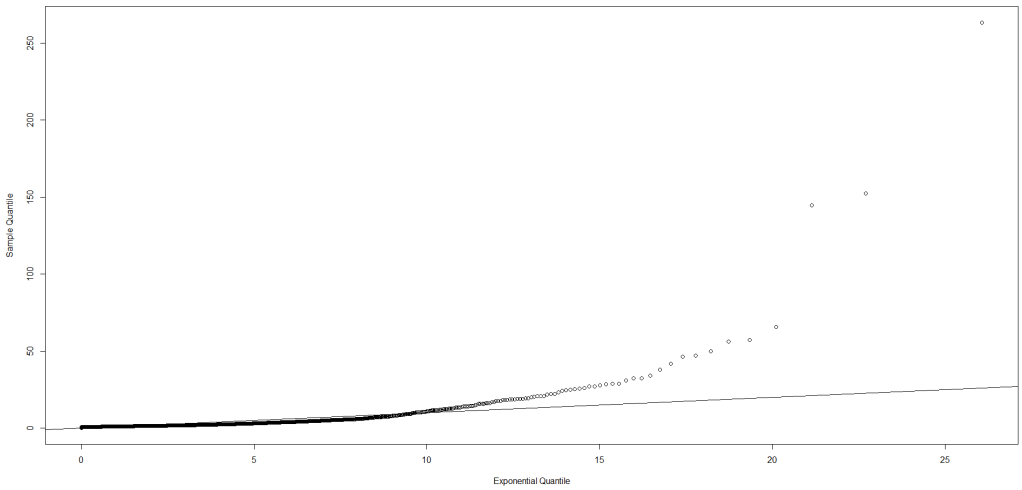
For the sake of readability, the histogram only displays values up to 50,000,000 DKK.




**(a)** Histogram




**(b)** Box Plot




**(c)** Exponential QQ-Plot

**Figure 1.** Panel (a) presents the histogram, panel (b) shows the box plot, and panel (c) presents the exponential QQ-plot for the Danish Fire Insurance data.

As can be seen in Figure 1, both the histogram and the box plot reveal a further very typical feature of insurance loss data, which is the large number of small losses around a unique mode and the lower number of large (extreme) losses very far away from the bulk data, resulting in a unimodal hump-shaped distribution.

Moreover, the convex shape of the exponential QQ-plot confirms the inference draw upon the descriptive statistics that the empirical loss distribution has a very heavy right tail decaying slowly than an exponential distribution.

### 3.1.2 U.S. Automobile Insurance

The second data set is composed of 6,773 claim amounts in U.S. Dollars (USD) paid by a large property and casualty Midwestern (U.S.) insurance company to settle and close claims for private passenger automobile policies. Again, claims settled with zero payment are not included. These data were analyzed, among others, by Frees (2010) and Punzo, Bagnato and Maruotti

(2018). The data set can be found in the R package **insuranceData**.

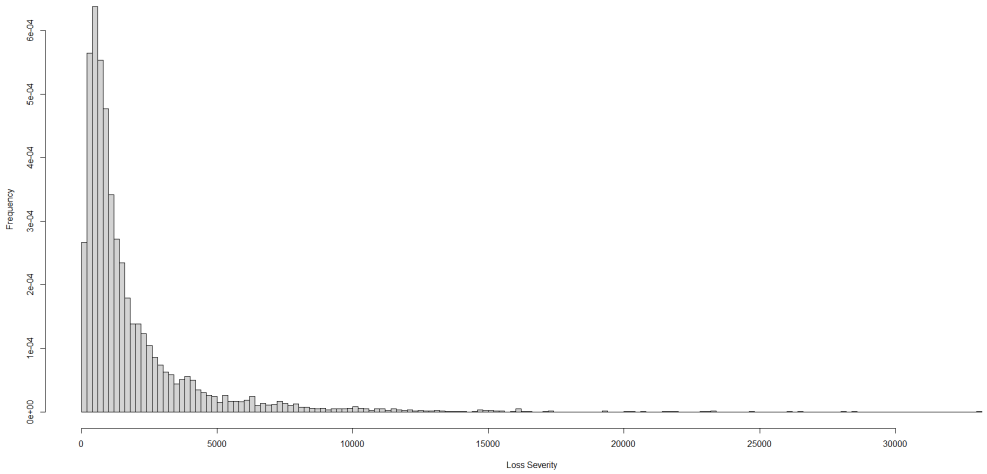
Summary statistics for the U.S. Automobile Insurance data set are available in Table 2.

| | |
|---|---|
| Number of Obs. | 6,773 |
| Mean | 1,853.03 |
| Median | 1,001.70 |
| Standard Dev. | 2,646.91 |
| Skewness | 6.23 |
| Kurtosis | 87.25 |
| Minimum | 9.50 |
| Maximum | 60,000.00 |
| 25% Quantile | 523.73 |
| 75% Quantile | 2,137.40 |

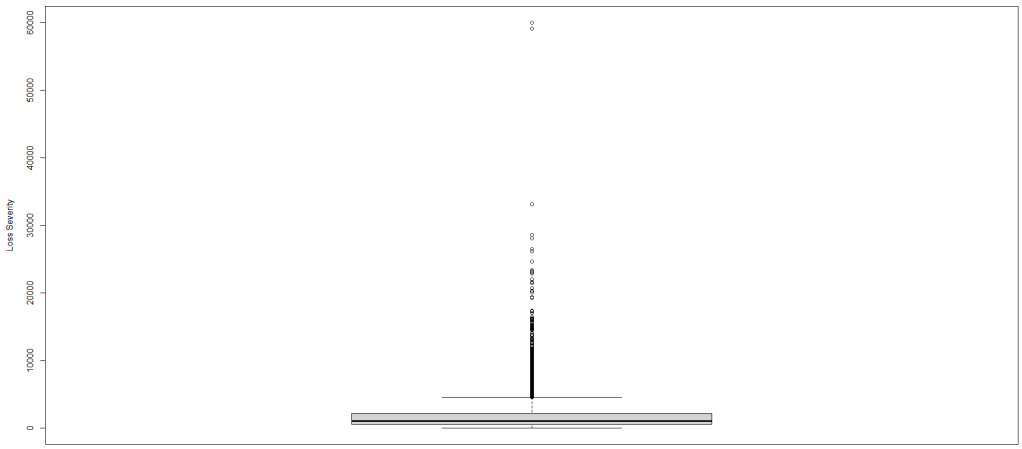
**Table 2.** Summary Statistics - U.S. Automobile Insurance

The empirical loss distribution related to the U.S. Automobile Insurance data, like the previous data set, is also positive skewed (6.23) with a high excess kurtosis coefficient (87.25), although much less leptokurtic than the Danish Fire Insurance data.

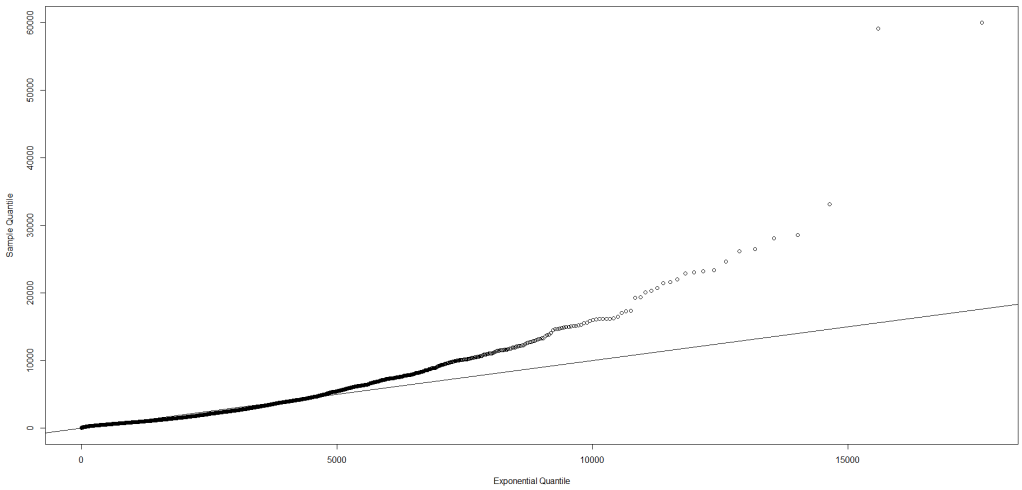
Figure 2 presents the histogram, the box plot, and the exponential QQ-plot for the U.S. Automobile Insurance data set. In order to facilitate the visualization, the histogram is censored at the value of 40,000 USD.




**(a)** Histogram




**(b)** Box Plot




**(c)** Exponential QQ-Plot

**Figure 2.** Panel (a) presents the histogram, panel (b) shows the box plot, and panel (c) presents the exponential QQ-plot for the U.S. Automobile Insurance data.

The usual unimodal hump-shaped curve typically seen in claim size distributions is once again represented by the histogram in panel (a), where the majority of sample points lie around a unique mode. The box plot in panel (b) shows the few observations far away from the bulk distribution, indicating the existence of extreme claim sizes. Finally, the heavy-tailed behaviour of the right tail is confirmed visually by the exponential QQ-plot in panel (c), where the empirical quantiles grow faster than the theoretical ones.

### 3.1.3 Australian Personal Injury Insurance

The third data set contains information on 22,036 settled personal injury insurance claims in thousands of Australian Dollars (AUD). These claims originated from personal injury accidents occurring in Australia from July 1989 to January 1999. Claims settled with zero payment are not included. These data were previously analyzed by de Jong and Heller (2008) and are freely available in R **CASdatasets** add-on package.

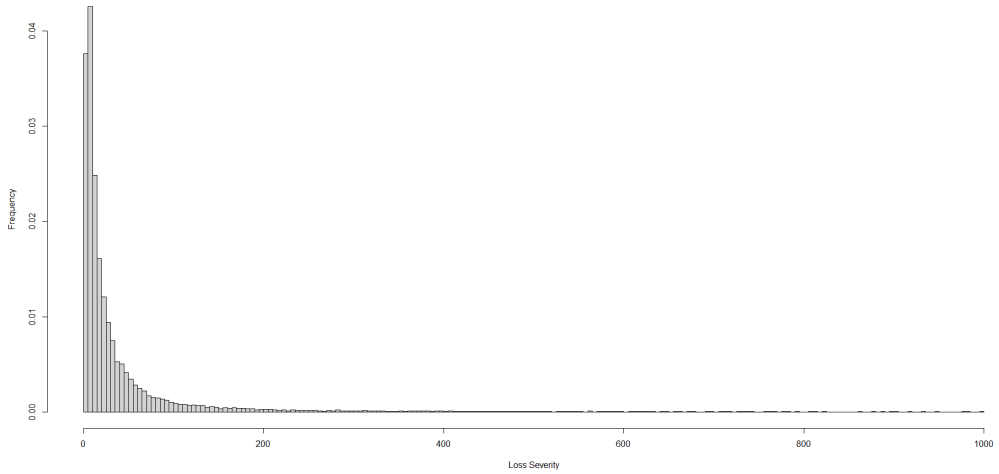
Table 3 presents descriptive statistics for the Australian Personal Injury Insurance data set.

| | |
|---|---|
| Number of Obs. | 22,036 |
| Mean | 38.37 |
| Median | 13.85 |
| Standard Dev. | 90.98 |
| Skewness | 12.74 |
| Kurtosis | 370.16 |
| Minimum | 0.01 |
| Maximum | 4,485.8 |
| 25% Quantile | 6.30 |
| 75% Quantile | 35.12 |

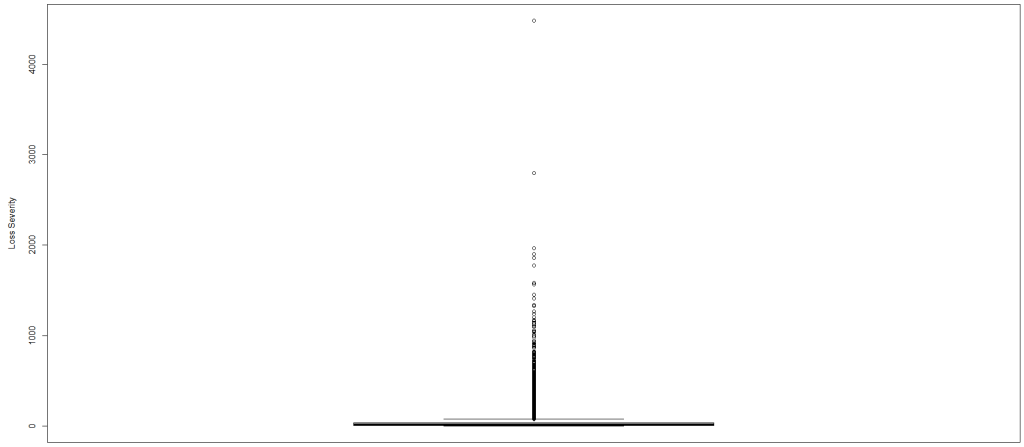
**Table 3.** Summary Statistics - Australian Personal Injury Insurance

It can be seen that the empirical loss distribution related to this data set is quite similar to the other two distributions previously analysed in this work, being positive skewed (12.74) and highly leptokuritc (370.16).

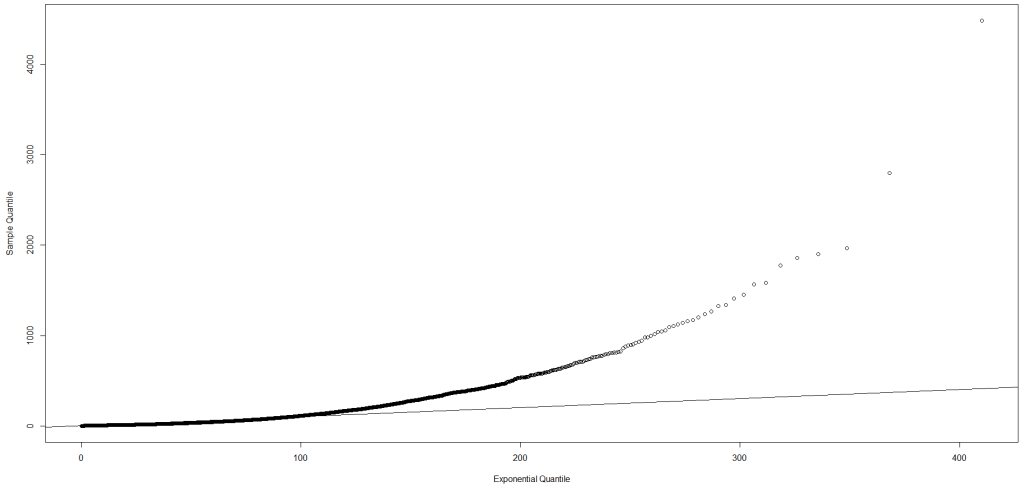
Figure 3 presents the graphical exploratory analysis for the Australian Personal Injury data set. For the sake of readability, the histogram is displayed only up to 1,000,000 AUD.




**(a)** Histogram




**(b)** Box Plot




**(c)** Exponential QQ-Plot

**Figure 3.** Panel (a) presents the histogram, panel (b) shows the box plot, and panel (c) presents the exponential QQ-plot for the Australian Personal Injury Insurance data.

Overall, Figure 3 shows that the Australian Personal Injury Insurance data set has the same typical features related to other claim size distributions, and already observed in the previous analysed data sets, which are, leptokurtosis, right-skewness and heavy-tails.

## 3.2 Model Selection

In order to compare the performance of our model to other existing loss models, we need model selection criteria. The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are the most popular information criteria used in the actuarial modelling literature for determining the best model among a set of competing candidates. Both of these criteria deal with the trade-off between goodness-of-fit and model complexity. We must choose the model that minimizes AIC and BIC values.

The AIC is defined as twice the negative log-likelihood function value

plus a penalty term, which is equal to twice the number of free parameters in the model (Akaike, 1974)

$$\text{AIC} = -2l(\boldsymbol{\theta}) + 2k$$

where $k$ is the length of the parameter vector $\boldsymbol{\theta}$.

Similar to AIC, the BIC selection criterion adjusts the log-likelihood by adding a penalty term consisting of the natural logarithm of the number of observations $n$ in the sample times the number of free parameters in the model (Schwarz, 1978)

$$\text{BIC} = -2l(\boldsymbol{\theta}) + k\ln(n)$$

The BIC criterion is often preferred in the literature as it gives more weight to the number of parameters in the penalty term.[1] That is, other things being equal, BIC will choose a more parsimonious model than the AIC.

Burnham and Anderson (2003) also recommended the use of the AIC and BIC differences, defined as $\Delta_{AIC_i} = AIC_i - AIC_{min}$ and $\Delta_{BIC_i} = BIC_i - BIC_{min}$, respectively, where $i = 1, ..., R$ accounts for all $R$ models under consideration. These differences estimate the relative expected Kullback-Leibler (K-L) distance between the candidate models, allowing for a more meaningful interpretation of the AIC/BIC criteria. More specifically, taking differences forces the best model to have $\Delta_{AIC_i} = 0$ and $\Delta_{BIC_i} = 0$, while the rest of the models have positive values for $\Delta_{AIC_i}$ and $\Delta_{BIC_i}$, representing the amount of lost information (estimated relative K-L distance) experienced if we have had chosen model $i$ for inference rather than the model with lowest AIC/BIC.

Although it can not be used as a model selection criterion, it is also very common in the actuarial literature to report the negative log-likelihood (NLL) function value. Let $l(\boldsymbol{\theta})$ denote the log-likelihood function for a given model, then the NLL is defined as

[1] Note that when $n \geq 8$, we have $\ln(n) > 2$.

$$\text{NLL} = -l(\boldsymbol{\theta})$$

The log-likelihood function monotonically increases as we increase the number of free parameters in the model. Therefore, since no penalty adjustment is made to account for model complexity, this measure is only suitable to compare models with the same number of parameters.

Finally, while the AIC and BIC criteria might be used to rank models, it could be the case that all models being compared fit the data very poorly. To account for this possibility, we also considered the nonparametric Kolmogorov-Smirnov (K-S) test for goodness-of-fit. The K-S test belongs to the class of EDF-based goodness-of-fit tests, a class that also includes the Anderson-Darling and Cramér-von Mises tests.

These tests assess fit quality by computing the distance between the null distribution function and the empirical distribution function (EDF). In the case of the K-S test, this distance is computed by the maximum absolute distance.

The K-S statistic is then defined as

$$D = \sup_x |\hat{F}_n(x) - F(x)|$$

where $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty,x)}(x_i)$ is the empirical distribution function for $n$ independent and identically distributed observations and $F(x)$ is the distribution specified in the null hypothesis.

For computational purposes, the following formula is more convenient

$$D = \max_{1\leq j\leq n}\left\{\left|F(x_{n,j}) - \frac{j-1}{n}\right|, \left|F(x_{n,j}) - \frac{j}{n}\right|\right\}$$

where $x_{n,1} < ... < x_{n,n}$ denote the order statistics of the sample vector $\mathbf{X} = \{x_1, x_2, ..., x_n\}$.

If the null hypothesis completely specifies the model, critical values are known. For instance, for $\alpha = 0.10$, the critical value is $1.22/\sqrt{n}$, for $\alpha = 0.05$, $1.36/\sqrt{n}$ and for $\alpha = 0.01$, $1.63/\sqrt{n}$. If the test statistic of the Kolmogorov-Smirnov goodness-of-fit test remains below the critical value, we cannot reject the hypothesis that the empirical distribution is similar to the distribution specified in the null.

However, when the parameters of the null distribution are estimated from the data, the test statistic tends to be smaller than it would have been had the parameters values been pre-specified. That is because the estimation method itself tries to choose parameter values that produce a distribution that is close to the observed data. Because rejection of the null hypothesis occurs for large values of the test statistic, this approximation tends to increase the probability of a type II error while lowering the probability of a type I error (Klugman, Panjer and Willmot, 2008).

In this case, the Kolmogorov-Smirnov test statistics are useful for ranking models, but are unreliable for statistical hypothesis testing, since the resulting $p$-values will be inaccurate. In order to obtain correct $p$-values, we employ the parametric bootstrap method described by Babu and Rao (2004) and Calderín-Ojeda and Kwok (2016). The steps of the proposed bootstrapping procedure are the following:

(1) Compute the Kolmogorov-Smirnov goodness-of-fit test statistic $D$.

(2) Simulate $M$ sets of resampled data from the fitted model, denote these sets as $\{x_1^{(i)}, ..., x_n^{(i)}\}$, $i = 1, ..., M$. Note that the resampled data have the same number of observations as the original sample.

(3) For each set of resampled data, fit the model and compute the K-S test statistics $D_i$, $i = 1, ..., M$.

(4) The $p$-value associated with the original test statistic is given by

$$\frac{\#\{i : D_i \geq D\}}{M}$$

In this work, $p$-values obtained through the above parametric bootstrap

method were computed using $M = 10,000$ simulations.

## 3.3 Results

Several standard parametric loss models commonly used in actuarial modelling were considered to compose the set of candidate distributions and asses the performance of the proposed model, including the Lognormal, Gamma, Weibull, Pareto Type I, Inverse Gaussian (Wald), Inverse Gamma (Vinci) and Burr Type XII (Singh–Maddala) distributions. For a comprehensive review about these parametric distributions and their properties, see Kleiber and Kotz (2003).

In addition to the above models, we follow Balasooriya and Low (2008) and also include the Generalized Lambda distribution (GLD) and the Transformed Kernel Density to the set of candidate distributions.

The GLD was first proposed by Ramberg and Schmeiser (1974) as a generalization of the well-known Tukey's Lambda distribution, a highly flexible three-parameter model typically used to approximate unimodal symmetric distributions (Ramberg and Schmeiser, 1972). With the addition of a fourth parameter, the GLD is also able to approximate many commonly used skewed (asymmetric) distributions. Due to its features, the GLD has been applied to various problems where a flexible parametric model for univariate asymmetric data is needed.

The Transformed Kernel Density, in turn, is based on the semiparametric transformation method to kernel smoothing. In a nutshell, this method first transforms the data using a parametric function and then applies the classical nonparametric kernel density estimator to the transformed data set. The final density estimator is obtained by scaling this kernel estimate by the gradient of the transformation function. If a transformation function is selected with transformed values, $T(x)$, converging to a finite value as $x$ goes to infinity, the Transformed Kernel Density ensures a stable tail behaviour of the final density estimator (Bolancé, Guillen and Nielsen, 2003).

The shifted power transformation was originally employed in the literature as the transformation function for most empirical applications of the

Transformed Kernel Density method (Wand, Marron and Ruppert, 1991). Here, we follow Buch-Kromman (2006) and employ the Modified Champernowne distribution function as the parametric transformation function. This *cdf* transformation is more intuitive and computationally less expensive when compared to the shifted power transformation. See appendix B for further details.

Finally, the statistical literature points to mixtures of Gamma distributions as reasonable models to fit insurance losses. The use of finite mixtures of Gamma distributions is sound from both a theoretical and practical point of view, since all distributions on the continuous positive domain can be approximated by a Gamma mixture with enough components (Wiper, Insua and Ruggeri, 2001).

The class of Gamma mixtures is more general than the Erlang mixtures previously employed in the literature (Lee and Lin, 2010; Verbelen *et al.*, 2015; Reynkens *et al.*, 2017), because the shape parameter of the Gamma distribution is not restricted to be an integer. As a consequence, we should typically obtain more parsimonious approximations than those provided by Erlang mixtures. They are also more flexible and computationally less expensive than the class of composite models (Cooray and Ananda, 2005; Scollnik, 2007; Pigeon and Denuit, 2011; Scollnik and Sun, 2012; Nadarajah and Bakar, 2014; Bakar *et al.*, 2015; Calderín-Ojeda and Kwok, 2016; Brazauskas and Kleefeld, 2016).

The number of Gamma components in the mixture was determined by the BIC statistics. More precisely, we add components until the BIC value is minimized. This approach to select the optimal number of components in a mixture model has already been applied by several authors (Lee and Lin, 2010; Verbelen *et al.*, 2015; Reynkens *et al.*, 2017). Expectation-Maximization (EM) algorithm was employed to find out the parameters associated with the Gamma mixture models (Arcidiacono and Jones, 2003).

### 3.3.1 Threshold Selection

The first step to fit our extreme value mixture to the data is to define the threshold level that separates small and large losses. Threshold selection is by

no means an easy task. If we choose a too low threshold we may get biased estimates since the POT approach is based on a limit theorem (Pickands-Balkema-De Haan Theorem), which applies only above high thresholds. On the other hand, if we set a too high threshold, we will have few data points and our estimates will be unstable and prone to high standard errors. This trade-off is known as bias-variance trade-off.

In classical extreme value theory, the threshold is often chosen graphically looking at the mean excess plot (Embrechts *et al.*, 1997), threshold stability plot and Gertensgarbe plot (Gertensgarbe and Werner, 1989). These methods often require a large amount of expertise about the phenomena being modelled. The analyst must carefully inspect the data and comprehend their main characteristics to find and justify a reasonable threshold choice that respects the bias-variance trade-off. This process may be very time-consuming. Another major drawback of these graphical tools is their high level of subjectivity, as different threshold choices are possible.

In light of these limitations, objective and automated methods have been proposed in the statistical literature to deal with threshold selection. Instead of a direct estimation of the threshold value $u$, these methods estimate the optimal number of exceedances $k$ to be used in tail estimation. The threshold value is then computed by $u = X_{n,\lfloor k \rfloor}$, where $X_{n,i}$ denotes the $i$th order statistics of $\{X_1, ..., X_n\}$ in descending order.

DuMouchel (1983) advocates the use of a heuristic rule where the threshold value is chosen as a high empirical quantile of the data $(1 - \epsilon)$, which means that $k = (1 - \epsilon)n$. He suggested the probability $\epsilon = 0.1$. Ferreira, de Haan and Peng (2003) proposed the squared root rule $k = \sqrt{n}$ and Loretan and Philips (1994) the empirically driven rule $k = \frac{n^{2/3}}{\ln[\ln(n)]}$.

In despite of their appealing simplicity, the above heuristic methods lack formal theoretical justification. More robust threshold selection methods can be obtained based on the well-known Hill estimator (Hill, 1975). Consider a set of *i.i.d.* random variables $\mathscr{X}_n = \{X_1, ..., X_n\}$ which has a power law form with regularly varying tails,

$$1 - F(x) \approx x^{-\alpha} L(x) \tag{17}$$

where $L(x)$ is a slowly varying function $\lim_{x\to\infty} \frac{L(tx)}{L(x)} = 1$, $t > 0$. Clearly, the above model does not have such a flexible tail behaviour as the GPD model described earlier in this work. However, this is an important special case in many empirical applications and the range of techniques developed for tail fraction estimation can be extrapolated to the GPD framework.

Define now the Hill estimator for the tail index ($\gamma = \alpha^{-1}$) based on the $k+1$ upper order statistics of $\mathscr{X}_n$ as

$$\gamma_n(k) = \frac{1}{k}\sum_{i=1}^{k} \log X_{n,i} - \log X_{n,k+1} \tag{18}$$

where $X_{n,1} \geq ... \geq X_{n,n}$ are the order statistics of $\mathscr{X}_n$ in descending order.

It is easily shown that if $X$ has tail behaviour as described by equation (17), log-differences between the order statistics should be exponentially distributed with rate $\gamma$. Hill (1975) suggested choosing the optimal sample fraction as the minimum $k$ such that log-differences $\{\ln X_{n,i} - \ln X_{n,k+1}, i = 1, ..., k\}$ does not fail a goodness-of-fit test considering the exponential distribution as the null. Guillou and Hall (2001) extended this idea to derive an estimator by applying a hypothesis test based on the accumulation of the above log-differences.

Moreover, various authors have investigated alternative procedures for the optimal choice of $k$ by minimizing the Asymptotic Mean Squared Error (AMSE) of $\gamma_n(k)$

$$k_0(n) = \underset{k}{\operatorname{argmin}}\, \mathrm{AMSE}(n,k) = \underset{k}{\operatorname{argmin}}\, \mathrm{Asy}\, \mathbb{E}[(\gamma_n(k) - \gamma)^2]$$

where $k_0(n)$ is the value which balances the asymptotic variance and bias components of $\mathbb{E}[(\gamma_n(k) - \gamma)^2]$.

Under certain regularity conditions, the $\mathrm{AMSE}(n,k)$ can be written as

$$\mathrm{AMSE}(n,k) = \gamma^2 \left( \frac{1}{k} + \frac{\beta^2}{(1-\rho)^2} \left( \frac{n}{k} \right)^{2\rho} \right) \tag{19}$$

where $(\rho, \beta)$ are second-order parameters. For further details, the reader may be referred to Gomes and Pestana (2007a).

The $k$ value that minimizes the $\mathrm{AMSE}(n,k)$ is

$$k_0(n) = \underset{k}{\operatorname{argmin}}\, \mathrm{AMSE}(n,k) = \left\lfloor \left( \frac{(1-\rho)^2 n^{-2\rho}}{-2\rho\beta^2} \right)^{1/(1-2\rho)} \right\rfloor \tag{20}$$

with $\lfloor x \rfloor$ denoting the integer part of $x$.

There are a great variety of estimators available in the literature for the second-order parameters $(\rho, \beta)$. See, for instance, the estimators proposed by Hall (1982), Hall and Welsh (1985) and Peng (1998), as well as the bootstrap based procedures in Hall (1990), Danielsson *et al.* (2001) and Caeiro and Gomes (2014).

Fraga Alves, Gomes and de Haan (2003) proposed the following consistent estimator for $\rho$

$$\hat{\rho}_\tau(k) \equiv \hat{\rho}_n^{(\tau)}(k) = -\left| \frac{3(T_n^{(\tau)}(k) - 1)}{T_n^{(\tau)}(k) - 3} \right|; \quad \tau \in \mathbb{R}^+$$

derived from the statistics

$$T_n^{(\tau)}(k) = \begin{cases} \frac{[M_n^{(1)}(k)]^\tau - [M_n^{(2)}(k)/2]^{\tau/2}}{[M_n^{(2)}(k)/2]^{\tau/2} - [M_n^{(3)}(k)/6]^{\tau/3}}; & \tau > 0 \\ \\ \frac{\ln[M_n^{(1)}(k)] - \frac{1}{2}\ln[M_n^{(2)}(k)/2]}{\frac{1}{2}\ln[M_n^{(2)}(k)/2] - \frac{1}{3}\ln[M_n^{(3)}(k)/6]}; & \tau = 0 \end{cases}$$

where $M_n^{(j)}(k)$ are the moment estimators

$$M_n^{(j)}(k) = \frac{1}{k}\sum_{i=1}^{k}(\log X_{n,i} - \log X_{n,k+1})^j; \;\; j \geq 1; \;\; [\gamma_n(k) \equiv M_n^{(1)}(k)]$$

Gomes and Pestana (2007b) suggested the values $\kappa = \{\lfloor n^{0.995} \rfloor, \lfloor n^{0.999} \rfloor\}$ to compute the tuning parameter $\tau^*$, defined as

$$\tau^* = \underset{\tau}{\operatorname{argmin}} \sum_{k \in \kappa} (\hat{\rho}_\tau(k) - \chi_\tau)^2; \;\; \tau \in \{0, 1\}$$

where $\chi_\tau = \text{median}\{\rho_\tau(k)\}_{k \in \kappa}$. Note that, although the parameter $\tau$ can assume any nonnegative real value ($\tau \in \mathbb{R}^+$), for simplicity, it is restricted here to the set $\{0, 1\}$.

The final estimator for the second-order "shape" parameter $\rho$ is given by[2]

$$\hat{\rho} = \hat{\rho}_{\tau^*}(\lfloor n^{0.999} \rfloor) \tag{21}$$

The second-order "scale" parameter $\beta$ was consistently estimated by Gomes and Martins (2002) as

$$\hat{\beta} = \left(\frac{k}{n}\right)^{\hat{\rho}} \frac{d_{\hat{\rho}}(k) D_0(k) - D_{\hat{\rho}}(k)}{d_{\hat{\rho}}(k) D_{\hat{\rho}}(k) - D_{2\hat{\rho}}(k)} \tag{22}$$

where, for $\alpha \leq 0$, $d_\alpha(k) = \frac{1}{k}\sum_{i=1}^{k}(i/k)^{-\alpha}$ and $D_\alpha(k) = \frac{1}{k}\sum_{i=1}^{k}(i/k)^{-\alpha} U_i$ with $U_i$ being the scaled log-differences $U_i = i(\ln X_{n,i} - \ln X_{n,k+1})$, $1 \leq i \leq k < n$.

[2]Alternatively, Gomes, Caeiro and Figueiredo (2004) advise the consideration of the value $k = \min\left(n-1, \frac{2n}{\ln \ln n}\right)$ and a tuning parameter of $\tau = 0$ for the region $\rho \in [-1, 0)$ and $\tau = 1$ for the region $\rho \in (-\infty, -1)$. Gomes and Pestana (2007a) decided for the choice $k = \min\left(n-1, \frac{2n^{0.995}}{\ln \ln n}\right)$.

On the basis of $(\hat{\beta}, \hat{\rho})$, we are able to find the optimal sample fraction $\hat{k}_0$ using equation (20). The threshold value is then computed as $u = X_{n,\hat{k}_0}$.

### 3.3.2 Danish Fire Insurance

In order to unambiguously define the fitted distributions, Maximum Likelihood Estimators (MLE) for the Danish Fire Insurance data are reported in Table 4 for each fitted model. Reported ML estimates were performed in R using **fBasics** and **mixtools** add-on packages and the **nlm** function.

| Distribution | Parameters |
|---|---|
| Lognormal | $\hat{\mu} = 0.672$; $\hat{\sigma} = 0.732$ |
| Gamma | $\hat{\alpha} = 1.258$; $\hat{\beta} = 2.434$ |
| Weibull | $\hat{\alpha} = 0.948$; $\hat{\beta} = 2.952$ |
| Pareto | $\hat{\theta} = 0.313$; $\hat{\alpha} = 0.546$ |
| Inverse Gaussian | $\hat{\mu} = 3.063$; $\hat{\lambda} = 3.417$ |
| Inverse Gamma | $\hat{\alpha} = 2.753$; $\hat{\beta} = 4.447$ |
| Burr Type XII | $\hat{\lambda} = 14.928$; $\hat{k} = 0.088$; $\hat{\alpha} = 0.921$ |
| Generalized Lambda | $\hat{\lambda}_1 = 0.975$; $\hat{\lambda}_2 = -0.949$; $\hat{\lambda}_3 = -0.031$; $\hat{\lambda}_4 = -0.695$ |
| Gamma Mixture | $\hat{\alpha}_1 = 11.616$; $\hat{\beta}_1 = 15.937$; $\hat{\alpha}_2 = 1.590$; $\hat{\beta}_2 = 7.626$; $\hat{\alpha}_3 = 17.827$; $\hat{\beta}_3 = 0.085$; $\hat{\alpha}_4 = 106.840$; $\hat{\beta}_4 = 0.009$; $\hat{\alpha}_5 = 5.222$; $\hat{\beta}_5 = 0.609$; $\hat{\pi}_1 = 0.001$; $\hat{\pi}_2 = 0.088$; $\hat{\pi}_3 = 0.441$; $\hat{\pi}_4 = 0.181$; $\hat{\pi}_5 = 0.289$ |
| Extreme Value Kernel Mixture | $u = 2.456$; $\hat{\lambda} = 0.038$; $\hat{\sigma} = 1.868$; $\hat{\xi} = 0.659$ |

**Table 4.** Maximum Likelihood Estimates - Danish Fire Insurance

Table 5 provides the goodness-of-fit measures for the Danish Fire Insurance data, including the NLL, AIC, BIC, $\Delta_{AIC}$ and $\Delta_{BIC}$ statistics evaluated at the MLE estimators and the Kolmogorov-Sminorv (K-S) test statistic. Corrected $p$-values for the K-S test computed through the parametric bootstrap procedure described in section 3.2 are given inside the brackets.

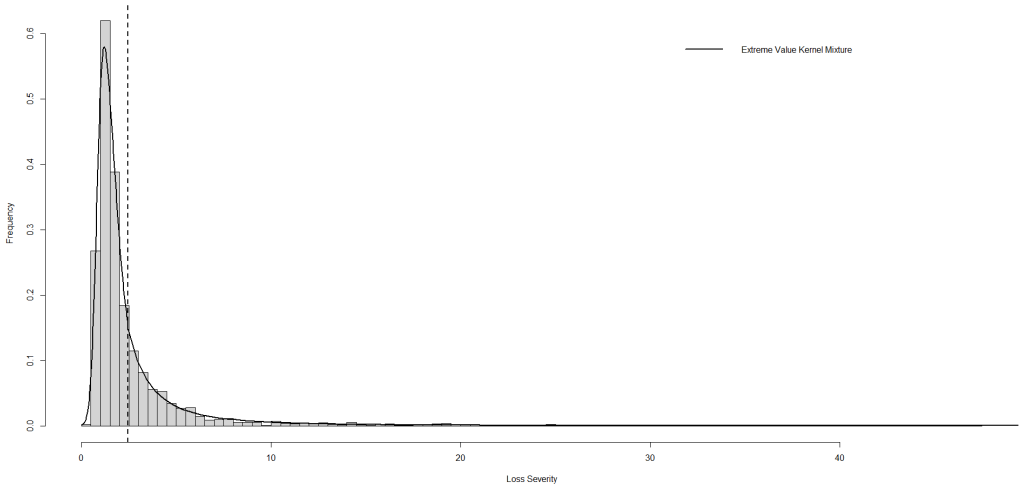
| Distribution | NLL | AIC | BIC | $\Delta_{AIC}$ | $\Delta_{BIC}$ | K-S |
|---|---|---|---|---|---|---|
| Lognormal | 4,433.89 | 8,871.78 | 8,883.42 | 1,266.09 | 1,260.27 | 0.127 (0.0000) |
| Gamma | 5,243.03 | 10,490.05 | 10,501.69 | 2,884.36 | 2,878.54 | 0.201 (0.0000) |
| Weibull | 5,270.47 | 10,544.94 | 10,556.58 | 2,939.25 | 2,933.43 | 0.255 (0.0000) |
| Pareto | 5,675.09 | 11,354.18 | 11,365.83 | 3,748.50 | 3,742.68 | 0.408 (0.0000) |
| Inverse Gaussian | 4,516.31 | 9,036.61 | 9,048.25 | 1,430.92 | 1,425.10 | 0.172 (0.0000) |
| Inverse Gamma | 4,097.88 | 8,199.75 | 8,211.40 | 594.07 | 588.25 | 0.087 (0.0000) |
| Burr Type XII | 3,835.12 | 7,676.24 | 7,693.70 | 70.55 | 70.55 | 0.038 (0.0000) |
| Generalized Lambda | 3,813.85 | 7,635.70 | 7,658.99 | 30.01 | 35.84 | 0.020 (0.0929) |
| Gamma Mixture | 3,803.70 | 7,635.40 | 7,716.89 | 29.71 | 93.74 | 0.015 (0.6010) |
| Extreme Value Kernel Mixture | 3,799.84 | **7,605.69** | **7,623.15** | - | - | 0.007 (0.9998) |
| Transformed Kernel | | | | | | 0.094 (0.0000) |

**Table 5.** Goodness-of-Fit Measures - Danish Fire

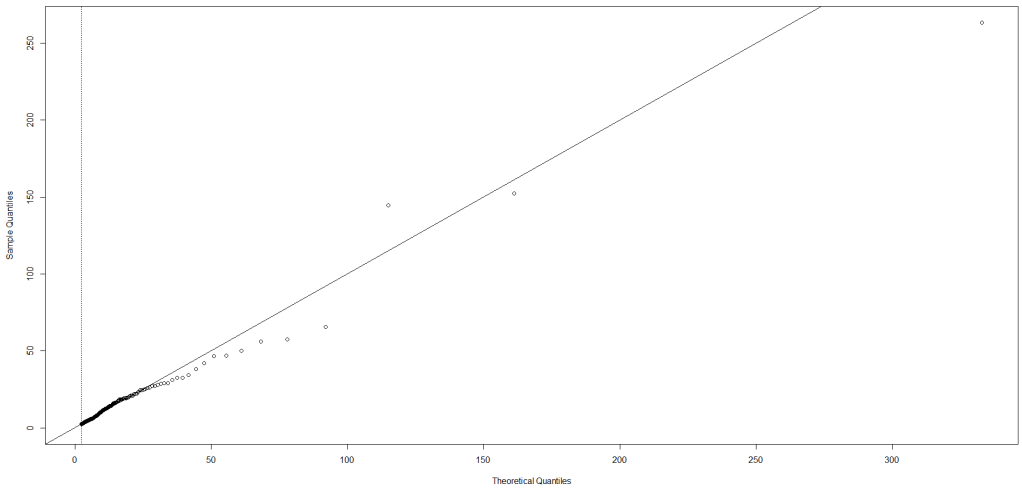
Overall, our results show that the Extreme Value Kernel Mixture was able to provide a superior fit than classical loss distributions and other alternative flexible models recently proposed in the actuarial literature. It can be seen that the model consistently outperformed the other candidate distributions across the different goodness-of-fit measures. $\Delta_{AIC}$ and $\Delta_{BIC}$ values also indicate that the competing models are quite distant (in a relative expected Kullback-Leibler distance sense) from the Extreme Value Kernel Mixture, meaning that a significant amount of information is lost when dealing with these models rather than the proposed model.

In panel (a) of Figure 4, we present the histogram for the Danish Fire Insurance data with the Extreme Value Kernel Mixture density overlaid in a solid curve. The dashed vertical line indicates the threshold value $u$ that splices the mixture components computed through the selection method described in section 3.3.1. The visual comparison of the fitted distribution and the empirical histogram provides further evidence that the fitted extreme value mixture density is a very good approximation for the analysed loss data.

To better investigate the fit in the tails, panel (b) shows the quantile plot for the Generalized Pareto component. As one can see, all but the last few sample points are reasonably close to the diagonal line, suggesting that the model is able to capture adequately the behaviour of the most extremal observations through its extreme value component.




**(a)** Histogram and Superimposed Density




**(b)** Generalized Pareto QQ-Plot

**Figure 4.** Danish Fire Insurance. Panel (a) shows the histogram with the Extreme Value Kernel Mixture density overlaid and panel (b) presents the Generalized Pareto QQ-plot.

For the sake of visual comparison, quantile plots for each fitted model are also given in Figure 5. As usual, the estimated theoretical quantiles $\hat{Q}(.)$

are plotted on $x$-axis against the ordered observations, $x_{n,1}, x_{n,2}, ..., x_{n,n}$, on $y$-axis. Here, $\hat{Q}(p_k)$ is the estimated $k^{th}$ quantile from the fitted model and $p_k = \frac{(k-0.5)}{n}$; $k = 1, 2, 3, ..., n$.

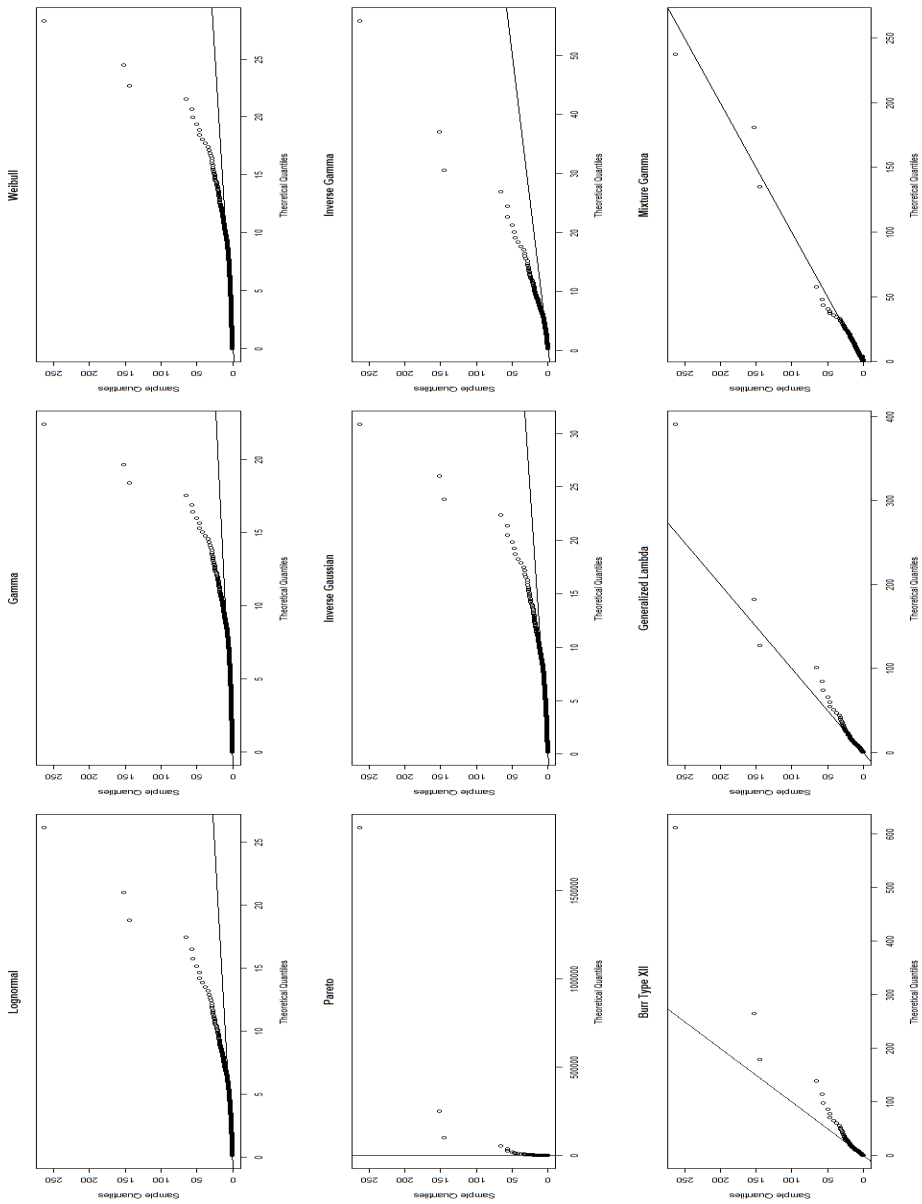



**Figure 5.** Quantile Plots - Danish Fire Insurance

From Figure 5, one can clearly see that none of the considered distributions seem to provide a better global fit than the Extreme Value Kernel Mixture model, the only exception being the Gamma mixture model that also shows a pretty good fit in the tail area.

Finally, as highlighted before in this thesis, it is often convenient for practitioners in the insurance industry to obtain reliable information about the tail area of the loss distribution. A measure that yields valuable knowledge about the right tail of a model is the high quantiles (Calderín-Ojeda and Kwok, 2016). Empirical and fitted quantiles in the extreme portion of the tail for the fitted models are given in Table 6. The empirical quantiles have been computed using the Type 8 quantile algorithm suggested in the statistical literature (Hyndman and Fan, 1996), where $p_k = \frac{k-1/3}{n+1/3}$, $k = 1, 2, 3, ..., n$. Quantiles estimated using this algorithm are approximately median unbiased.

| Distribution | 90% | 95% | 97.5% | 99% | 99.5% | 99.9% | 99.95% | 99.99% |
|---|---|---|---|---|---|---|---|---|
| Empirical | 5.08 | 8.46 | 14.50 | 24.87 | 32.81 | 146.01 | 199.02 | 263.25 |
| Lognormal | 5.00 | 6.53 | 8.22 | 10.76 | 12.91 | 18.82 | 21.79 | 29.82 |
| Gamma | 6.66 | 8.47 | 10.25 | 12.59 | 14.35 | 18.41 | 20.14 | 24.17 |
| Weibull | 7.12 | 9.40 | 11.71 | 14.79 | 17.15 | 22.70 | 25.11 | 30.75 |
| Pareto | 21.29 | 75.81 | 269.94 | 1,446.57 | 5,150.64 | $9.83 \times 10^3$ | $3.50 \times 10^5$ | $6.68 \times 10^6$ |
| Inverse Gaussian | 6.44 | 8.66 | 11.06 | 14.46 | 17.18 | 23.84 | 26.84 | 34.01 |
| Inverse Gamma | 4.66 | 6.41 | 8.63 | 12.55 | 16.51 | 30.60 | 39.71 | 72.20 |
| Burr Type XII | 5.34 | 9.07 | 15.40 | 30.99 | 52.61 | 179.78 | 305.19 | 1,042.93 |
| Generalized Lambda | 5.14 | 8.37 | 13.59 | 25.77 | 41.76 | 127.95 | 207.17 | 634.05 |
| Gamma Mixture | 4.94 | 8.71 | 15.81 | 24.92 | 32.28 | 135.36 | 192.30 | 264.78 |
| Extreme Value Kernel Mixture | 5.17 | 8.39 | 13.47 | 24.95 | 39.63 | 115.22 | 182.19 | 527.20 |
| Transformed Kernel | 4.64 | 6.79 | 9.83 | 16.08 | 23.48 | 56.06 | 79.89 | 158.27 |

**Table 6.** Point Estimates of High Quantiles - Danish Fire Insurance

The reader may observe that the Lognormal, Gamma, Weibull, Inverse Gaussian and Inverse Gamma distributions significantly underestimate the quantiles in the extreme proportion of the tail, whereas the Burr Type XII and Pareto Type I distributions largely overestimates them.

The only distributions to provide reasonable estimates are the Generalized Lambda, Gamma Mixture and Extreme Value Kernel Mixture. These models were able to approximate the extreme portion of the tails very well, with the Gamma Mixture being superior at the 99.99% probability. We remark, however, that the 99.99% quantile represents an event that occurs only 1 in 10,000 times. In a sample of 2,492 observations, as the one being analysed,

it is likely that we have not seen an event of this magnitude. Thus, we must remain cautious in making inferences based on those extreme empirical quantiles.

Moreover, Gamma Mixture models (or any other kind of finite mixture where the components are standard parametric distributions) can cover adequately the data span, approximating empirical quantiles very closely, but are not designed for handling extrapolations towards the tails, where little or no data is available. Extreme value mixtures, on the other hand, are designed precisely to this end and tends to provide much more accurate descriptions of the tails.

### 3.3.3 U.S. Automobile Insurance

Turning now to the U.S. Automobile Insurance data set, Maximum Likelihood Estimators (MLE) for each fitted model are presented in Table 7.

| Distribution | Parameters |
|---|---|
| Lognormal | $\hat{\mu} = 6.956$; $\hat{\sigma} = 1.071$ |
| Gamma | $\hat{\alpha} = 1.013$; $\hat{\beta} = 1,829.303$ |
| Weibull | $\hat{\alpha} = 0.938$; $\hat{\beta} = 1,789.272$ |
| Pareto | $\hat{\theta} = 9.500$; $\hat{\alpha} = 0.213$ |
| Inverse Gaussian | $\hat{\mu} = 1,852.336$; $\hat{\lambda} = 802.422$ |
| Inverse Gamma | $\hat{\alpha} = 0.928$; $\hat{\beta} = 519.347$ |
| Burr Type XII | $\hat{\lambda} = 1.669$; $\hat{k} = 0.984$; $\hat{\alpha} = 1,028.893$ |
| Generalized Lambda | $\hat{\lambda}_1 = 9.500$; $\hat{\lambda}_2 = -0.000153$; $\hat{\lambda}_3 = 0$; $\hat{\lambda}_4 = -0.218$ |
| Gamma Mixture | $\hat{\alpha}_1 = 2.286$; $\hat{\beta}_1 = 369.100$; $\hat{\alpha}_2 = 17.748$; $\hat{\beta}_2 = 140.532$; $\hat{\alpha}_3 = 127.264$; $\hat{\beta}_3 = 31.461$ $\hat{\alpha}_4 = 1.003$; $\hat{\beta}_4 = 3,800.170$; $\hat{\pi}_1 = 0.630$; $\hat{\pi}_2 = 0.070$; $\hat{\pi}_3 = 0.018$; $\hat{\pi}_4 = 0.282$ |
| Extreme Value Kernel Mixture | $u = 6,750.860$; $\hat{\lambda} = 31.500$; $\hat{\sigma} = 3,049.99$; $\hat{\xi} = 0.245$ |

**Table 7.** Maximum Likelihood Estimates - U.S. Automobile Insurance

Table 8 illustrates the goodness-of-fit measures evaluated at the ML estimators presented above. Corrected $p$-values for the Kolmogorov-Smirnov test (K-S) are again given in brackets.

| Distribution | NLL | AIC | BIC | $\Delta_{AIC}$ | $\Delta_{BIC}$ | K-S |
|---|---|---|---|---|---|---|
| Lognormal | 57,185.11 | 114,374.21 | 114,387.85 | 89.58 | 82.76 | 0.021 (0.0177) |
| Gamma | 57,736.62 | 115,477.24 | 115,490.88 | 1,192.61 | 1,185.79 | 0.096 (0.0000) |
| Weibull | 57,707.94 | 115,419.88 | 115,433.52 | 1,135.25 | 1,128.43 | 0.078 (0.0000) |
| Pareto | 64,371.21 | 128,746.42 | 128,760.07 | 14,461.79 | 14,454.98 | 0.429 (0.0000) |
| Inverse Gaussian | 57,629.71 | 115,263.41 | 115,277.05 | 978.78 | 971.96 | 0.077 (0.0000) |
| Inverse Gamma | 58,124.31 | 116,252.62 | 116,266.26 | 1,967.99 | 1,961.17 | 0.098 (0.0000) |
| Burr Type XII | 57,178.08 | 114,362.16 | 114,382.62 | 77.53 | 77.53 | 0.019 (0.0636) |
| Generalized Lambda | 57,455.38 | 114,918.76 | 114,946.04 | 634.13 | 640.95 | 0.080 (0.0000) |
| Gamma Mixture | 57,132.99 | 114,287.99 | 114,363.02 | 3.36 | 57.92 | 0.009 (0.6738) |
| Extreme Value Kernel Mixture | 57,139.32 | **114,284.63** | **114,305.09** | - | - | 0.005 (0.9984) |
| Transformed Kernel | | | | | | 0.043 (0.0000) |

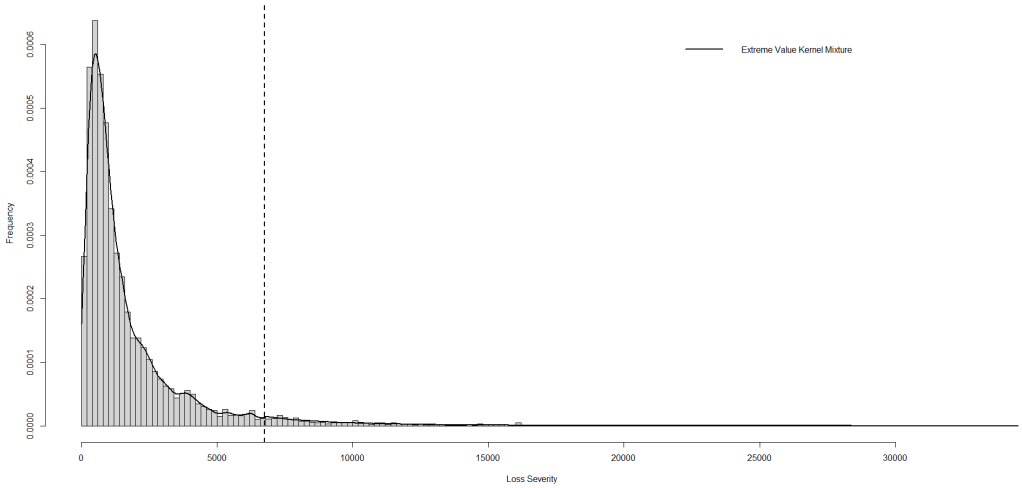
**Table 8.** Goodness-of-Fit Measures - U.S. Automobile Insurance

Once more, it is apparent that the Extreme Value Kernel Mixture provides a better fit than any of the other considered models. Note from the table that the nonparametric extreme value mixture consistently outperforms the other models in all of the goodness-of-fit measures (NLL, AIC, BIC and K-S), even when allowances are made for the larger number of parameters in some models (*e.g.* Generalized Lambda and Gamma Mixture).

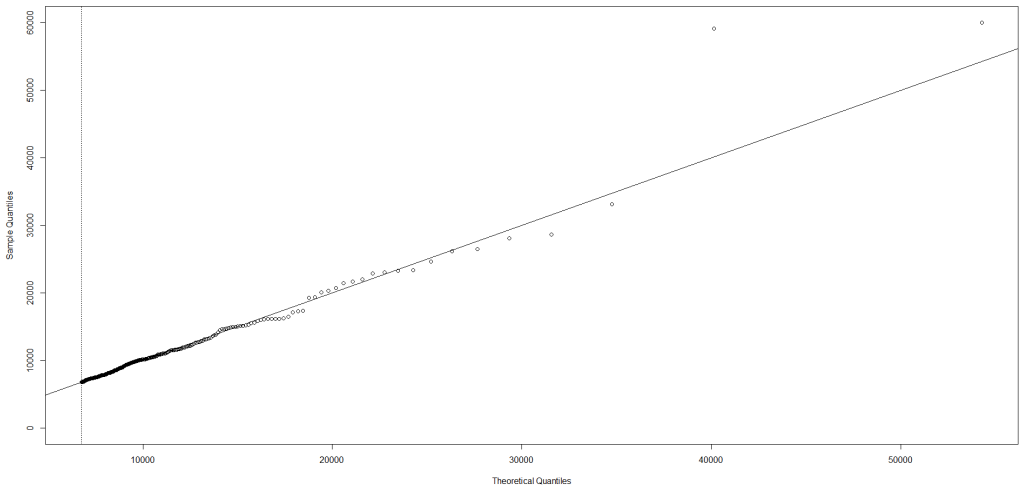
It can also be seen that the differences $\Delta_{AIC}$ and $\Delta_{BIC}$ are also very significant, indicating that the Extreme Value Kernel Mixture is the model that losses the least information when approximating the given loss data relative to the other models in the candidate set.

To investigate visually the fit for the U.S. Automobile Insurance data, panel (a) of Figure 6 presents the histogram with the non-parametric extreme value mixture superimposed density. The dashed vertical line represents the chosen threshold value that splices the mixture components. As suggested by the graph, the model provides an excellent global fit.

The quantile plot for the Generalized Pareto distribution, presented in panel (b), shows that the tail fit is also quite good. It is possible to see that only the last two data points are relatively distant from the diagonal line.




**(a)** Histogram and Superimposed Density




**(b)** Generalized Pareto QQ-Plot

**Figure 6.** U.S. Automobile Insurance. Panel (a) shows the histogram with the Extreme Value Kernel Mixture density overlaid and panel (b) represents the Generalized Pareto QQ-plot.

For the sake of visual comparison, Figure 7 presents the quantile plots for the other fitted models. According to these graphs, we can observe that

none of the represented models seem to provide a superior fit for the given data, verifying the conclusions draw from Table 8.

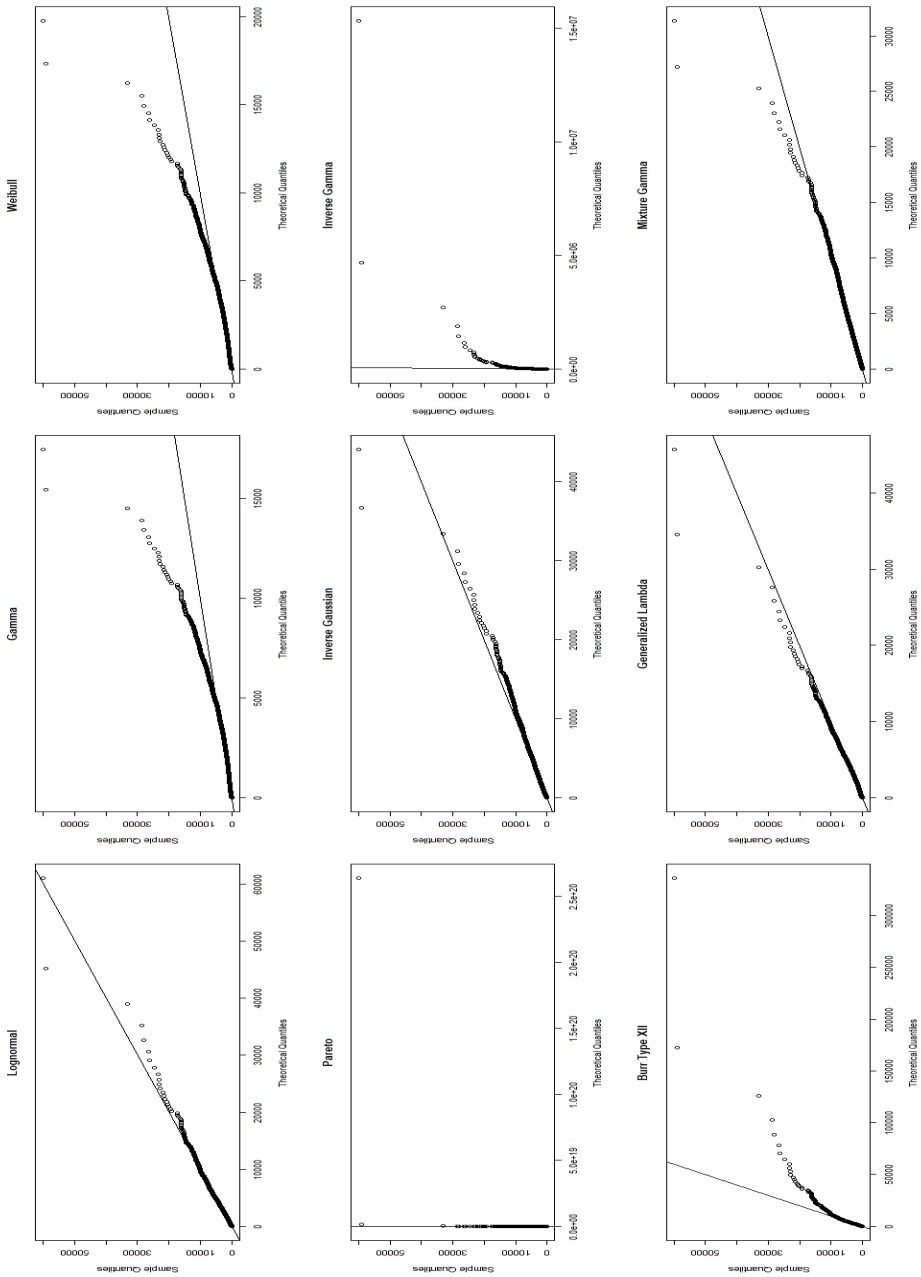



**Figure 7.** Quantile Plots - U.S. Automobile Insurance

Closing our analysis of the U.S. Automobile Insurance data, Table 9 gives estimated quantiles in the extreme portion of the tails for each fitted model.

| Distribution | 90% | 95% | 97.5% | 99% | 99.5% | 99.9% | 99.95% | 99.99% |
|---|---|---|---|---|---|---|---|---|
| Empirical | 4,171.19 | 6,356.90 | 8,725.65 | 12,087.85 | 15,449.25 | 25,983.84 | 29,865.81 | 59,990.55 |
| Lognormal | 4,138.52 | 6,106.89 | 8,558.16 | 12,670.36 | 16,551.03 | 28,713.00 | 35,582.57 | 56,303.04 |
| Gamma | 4,253.23 | 5,525.98 | 6,797.89 | 8,478.41 | 9,749.21 | 12,698.84 | 13,968.82 | 16,917.07 |
| Weibull | 4,353.43 | 5,763.32 | 7,195.13 | 9,115.03 | 10,584.61 | 14,043.90 | 15,551.09 | 19,084.65 |
| Pareto | $4.81 \times 10^5$ | $1.25 \times 10^7$ | $3.27 \times 10^8$ | $2.43 \times 10^{10}$ | $6.35 \times 10^{11}$ | $1.23 \times 10^{15}$ | $3.21 \times 10^{16}$ | $6.24 \times 10^{19}$ |
| Inverse Gaussian | 4,423.25 | 6,822.91 | 9,678.87 | 14,035.48 | 17,689.45 | 27,069.84 | 31,408.17 | 41,988.57 |
| Inverse Gamma | 6,138.14 | 13,263.88 | 28,303.45 | 76,455.62 | $1.62 \times 10^5$ | $9.18 \times 10^5$ | $1.94 \times 10^6$ | $1.09 \times 10^7$ |
| Burr Type XII | 3,935.27 | 6,193.69 | 9,588.76 | 16,899.22 | 25,847.08 | 69,027.27 | $1.05 \times 10^5$ | $2.81 \times 10^5$ |
| Generalized Lambda | 4,284.65 | 6,052.48 | 8,109.15 | 11,352.45 | 14,275.07 | 23,037.13 | 27,868.90 | 42,354.59 |
| Gamma Mixture | 4,184.25 | 6,583.52 | 9,220.61 | 12,705.78 | 15,341.77 | 21,461.45 | 24,096.78 | 30,215.42 |
| Extreme Value Kernel Mixture | 4,175.02 | 6,357.81 | 8,704.59 | 12,329.28 | 15,665.63 | 25,990.18 | 31,854.27 | 50,001.09 |
| Transformed Kernel | 3,360.40 | 5,018.08 | 7,479.73 | 12,614.83 | 18,405.88 | 37,359.01 | 45,306.60 | 56,102.65 |

**Table 9.** Point Estimates of High Quantiles - U.S. Automobile Insurance

Overall, the proposed extreme value mixture provided the best point estimates of the high quantiles. Gamma and Weibull distributions largely underestimated the quantile values, whereas Pareto, Inverse Gamma and Burr Type XII distributions overestimated them. Lognormal, Inverse Gaussian and Gamma Mixture distributions provided reasonable fits, although not as good as the Extreme Value Kernel Mixture, as a whole.

### 3.3.4 Australian Personal Injury Insurance

To conclude this section, we analyse next the obtained results for the Australian Personal Injury Insurance data. Maximum Likelihood Estimators (MLE) for each fitted model are reported in Table 10.

| Distribution | Parameters |
|---|---|
| Lognormal | $\hat{\mu} = 2.649$; $\hat{\sigma} = 1.462$ |
| Gamma | $\hat{\alpha} = 0.616$; $\hat{\beta} = 62.257$ |
| Weibull | $\hat{\alpha} = 0.709$; $\hat{\beta} = 28.767$ |
| Pareto | $\hat{\theta} = 0.010$; $\hat{\alpha} = 0.138$ |
| Inverse Gaussian | $\hat{\mu} = 38.384$; $\hat{\lambda} = 3.250$ |
| Inverse Gamma | $\hat{\alpha} = 0.421$; $\hat{\beta} = 1.260$ |
| Burr Type XII | $\hat{\lambda} = 1.181$; $\hat{k} = 1.205$; $\hat{\alpha} = 18.196$ |
| Generalized Lambda | $\hat{\lambda}_1 = 0.010$; $\hat{\lambda}_2 = -0.031$; $\hat{\lambda}_3 = 0$; $\hat{\lambda}_4 = -0.553$ |
| Gamma Mixture | $\hat{\alpha}_1 = 0.639$; $\hat{\beta}_1 = 488.357$; $\hat{\alpha}_2 = 0.720$; $\hat{\beta}_2 = 114.174$; $\hat{\alpha}_3 = 5.314$; $\hat{\beta}_3 = 1.507$; $\hat{\alpha}_4 = 0.952$; $\hat{\beta}_4 = 21.553$; $\hat{\alpha}_5 = 522.194$; $\hat{\beta}_5 = 1.078$ ; $\hat{\pi}_1 = 0.020$; $\hat{\pi}_2 = 0.225$; $\hat{\pi}_3 = 0.175$; $\hat{\pi}_4 = 0.579$; $\hat{\pi}_5 = 0.001$ |
| Extreme Value Kernel Mixture | $u = 228.408$; $\hat{\lambda} = 0.201$; $\hat{\sigma} = 148.448$; $\hat{\xi} = 0.282$ |

**Table 10.** Maximum Likelihood Estimates - Australian Personal Injury Insurance

In Table 11, we provide the goodness-of-fit measures (NLL, AIC, BIC, $\Delta_{AIC}$, $\Delta_{BIC}$ and K-S) evaluated at the Maximum Likelihood Estimators (MLE).

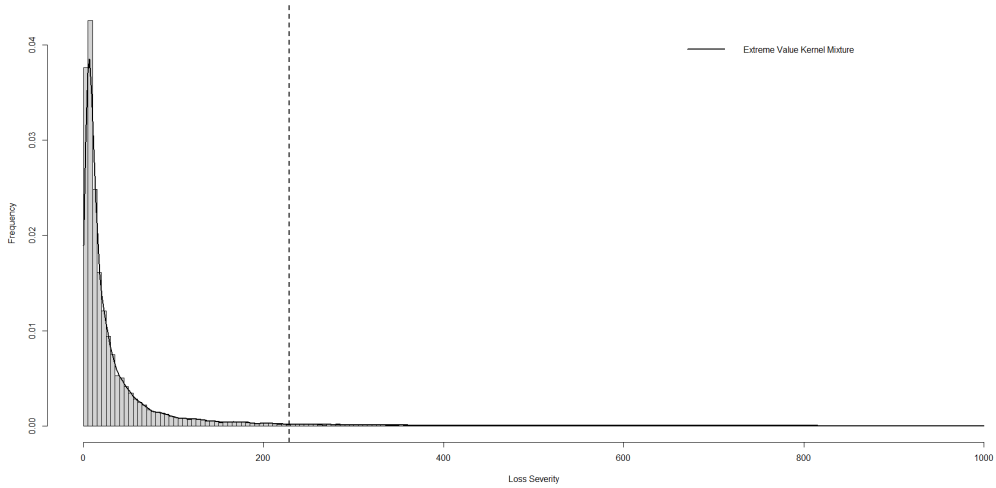
| Distribution | NLL | AIC | BIC | $\Delta_{AIC}$ | $\Delta_{BIC}$ | K-S |
|---|---|---|---|---|---|---|
| Lognormal | 98,008.40 | 196,020.80 | 196,036.80 | 1,251.98 | 1,243.97 | 0.053 (0.0000) |
| Gamma | 100,313.30 | 200,630.59 | 200,646.60 | 5,861.77 | 5,853.77 | 0.117 (0.0000) |
| Weibull | 99,087.38 | 198,178.76 | 198,194.76 | 3,409.94 | 3,401.93 | 0.077 (0.0000) |
| Pareto | 124,077.97 | 248,159.94 | 248,175.94 | 53,391.12 | 53,383.11 | 0.449 (0.0000) |
| Inverse Gaussian | 105,829.31 | 211,662.61 | 211,678.61 | 16,893.79 | 16.885,78 | 0.270 (0.0000) |
| Inverse Gamma | 106,464.42 | 212,932.84 | 212,948.84 | 18,164.02 | 18,156.01 | 0.226 (0.0000) |
| Burr Type XII | 97,488.87 | 194,983.75 | 195,007.75 | 214.93 | 214.93 | 0.027 (0.0074) |
| Generalized Lambda | 97,602.69 | 195,213.38 | 195,245.38 | 444.55 | 452.55 | 0.048 (0.0000) |
| Gamma Mixture | 97,394.75 | 194,817.50 | 194,929.51 | 48.68 | 136.68 | 0.007 (0.1876) |
| Extreme Value Kernel Mixture | 97,381.41 | **194,768.82** | **194,792.83** | - | - | 0.004 (0.7512) |
| Transformed Kernel | | | | | | 0.031 (0.0000) |

**Table 11.** Goodness-of-Fit Measures - Australian Personal Injury Insurance

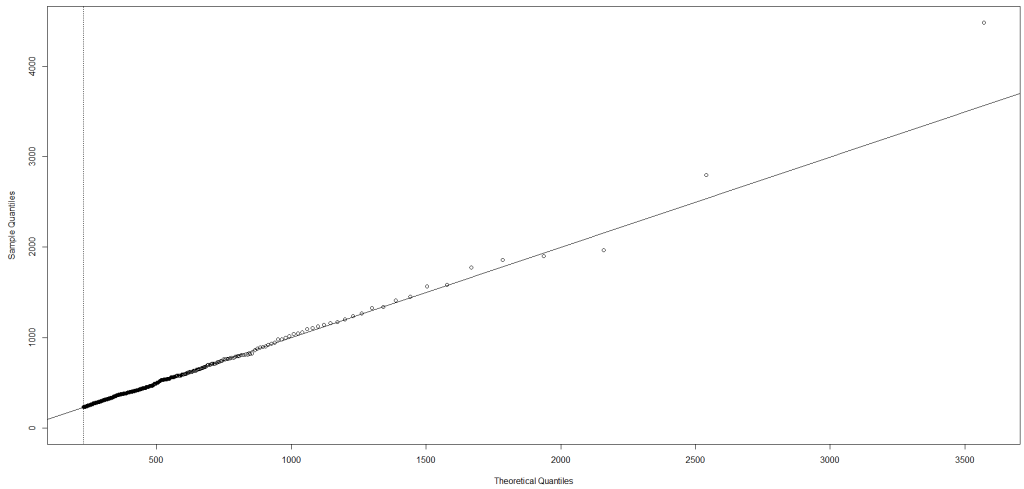
Similarly to the previously analyzed data sets, results from Table 11 show that the Extreme Value Kernel Mixture provided the best fit for the Australian Personal Injury Insurance data. It can be seen that the proposed model was able to minimize all the considered goodness-of-fit measures. Additionally, the $\Delta_{AIC}$ and $\Delta_{BIC}$ differences show that the candidate models

are quite distant from the extreme value mixture in the Kullback-Leibler relative distance sense.

Figure 8 confirms visually the conclusions draw from the table above. Panel (a) illustrates an excellent global fit for the empirical loss distribution, while the GPD quantile plot in panel (b) shows a good tail fit, as all data points are reasonably close to the diagonal line.




**(a)** Histogram and Superimposed Density




**(b)** Generalized Pareto QQ-Plot

**Figure 8.** Australian Personal Injury Insurance. Panel (a) shows the histogram with the Extreme Value Kernel Mixture density overlaid and panel (b) presents the Generalized Pareto QQ-plot.

Quantile plots for the competing models are given in Figure 9. Through visual inspection, it is quite clear that, among of them, only the Gamma

Mixture provided a satisfying fit for the given data.

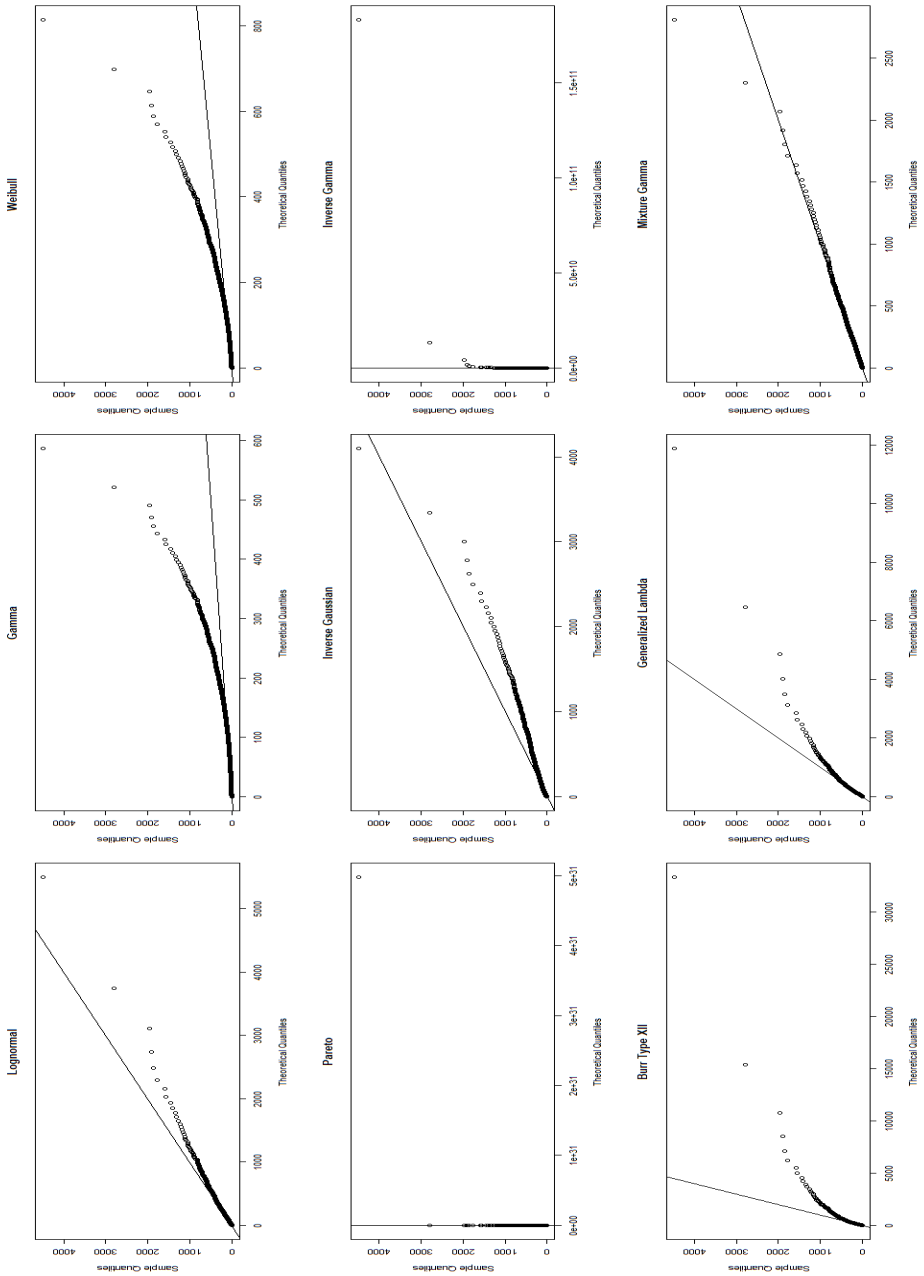



**Figure 9.** Quantile Plots - Australian Personal Injury Insurance

Finally, table 12 presents point estimates of high quantiles for the Australian Personal Injury Insurance data.

| Distribution | 90% | 95% | 97.5% | 99% | 99.5% | 99.9% | 99.95% | 99.99% |
|---|---|---|---|---|---|---|---|---|
| Empirical | 85.33 | 148.87 | 237.98 | 394.30 | 543.75 | 1,051.64 | 1,335.99 | 2,352.41 |
| Lognormal | 92.09 | 156.66 | 248.37 | 424.42 | 611.28 | 1,297.04 | 1,738.45 | 3,253.16 |
| Gamma | 99.19 | 136.73 | 175.30 | 227.34 | 267.27 | 361.29 | 402.20 | 497.85 |
| Weibull | 93.30 | 135.25 | 181.40 | 248.07 | 302.33 | 439.54 | 503.02 | 659.57 |
| Pareto | $1.80 \times 10^5$ | $2.76 \times 10^7$ | $4.22 \times 10^9$ | $3.26 \times 10^{12}$ | $4.99 \times 10^{14}$ | $5.91 \times 10^{19}$ | $9.05 \times 10^{21}$ | $1.07 \times 10^{27}$ |
| Inverse Gaussian | 77.16 | 168.27 | 314.63 | 597.16 | 872.67 | 1,674.64 | 2,073.75 | 3,088.45 |
| Inverse Gamma | 399.12 | 2,076.92 | 10,792.15 | 95,285.85 | $4.94 \times 10^5$ | $2.27 \times 10^7$ | $1.18 \times 10^8$ | $5.41 \times 10^9$ |
| Burr Type XII | 80.17 | 138.85 | 233.60 | 454.66 | 746.23 | 2,331.07 | 3,799.09 | 11,789.50 |
| Generalized Lambda | 82.59 | 136.19 | 214.86 | 377.94 | 569.64 | 1,434.23 | 2,119.86 | 5,212.03 |
| Gamma Mixture | 84.58 | 151.08 | 238.98 | 388.24 | 546.88 | 1,099.57 | 1,402.31 | 2,127.81 |
| Extreme Value Kernel Mixture | 85.38 | 148.82 | 238.17 | 396.28 | 546.21 | 1,031.39 | 1,318.57 | 2,247.87 |
| Transformed Kernel | 67.08 | 110.04 | 180.97 | 353.38 | 584.70 | 1,679.10 | 2,375.75 | 3,767.32 |

**Table 12.** Point Estimates of High Quantiles - Australian Personal Injury Insurance

Once again, the Extreme Value Kernel Mixture outperformed the other candidate models, providing excellent point estimates for the quantiles located in the extreme proportion of the tails.

## 3.4 Robust Inference

Classical inference statistics, as the one applied so far in this work, is based only in part upon the available data. Prior assumptions about the data generating processes being analysed play an equally important role in the standard statistical inference framework (Huber and Ronchetti, 2009). Assumptions like randomness, linearity, independence and general assumptions about the distributional models are made either implicitly or explicitly, even in the simplest statistical procedures.

These assumptions, however, are not necessarily met in most practical situations. They are merely mathematical simplifications justified by the stability principle: a minor error in the mathematical theoretical model should cause only a small deviation in the final conclusions.

Unfortunately, statistical theory has shown that this vague principle does

not hold in most cases (and in particular for those inference procedures based on the Normal distribution). Even minor deviations from model assumptions cause very significant effects in the final results. In the specific case of the Maximum Likelihood Estimator (MLE), for example, it is well documented that a slightly deviation from the assumed underlying distribution quickly destroys the desirable asymptotic properties of the MLE (Brazauskas and Serfling, 2000).

Since we can not rule out the possibility of anomalies or "oddities" that might be present in our insurance data (*e.g.* gross errors of measurement and other types of contamination introduced by deductibles and retention limits), it is useful to complement our analysis with procedures that are still reliable and reasonably efficient under small deviations from the assumed parametric model — in our case, the nonparametric extreme value mixture.

Robust statistics were developed precisely to provide this kind of inference methods. In fact, robustness, in its narrow statistical sense, is formally defined as insensitivity to small deviations (Huber and Ronchetti, 2009). Here, we are primarily concerned with distributional robustness, *i.e.* situations where the shape of the true underlying distribution slightly deviates from the assumed parametric model.

From the point of view of their statistical properties, robust estimators are consistent estimators of the unknown parameters. Efficiency, on the other hand, is partially sacrificed in order to achieve the desired degree of robustness. The more robust, the less efficient an estimator tends to be relative to the MLE.[3] Thus, in most practical situations, efficiency and robustness trade-off against each other.

If the assumptions of our idealized model are true, robust estimates will not deviate much from traditional MLE estimates. Otherwise, outside the model framework, they will behave as limiting values of the true parameters estimates.

Hereafter, we will focus our attention on the parametric component of our mixture model — the extreme value GPD component. Robust estimates

[3]The MLE provides a quantitative benchmark for efficiency considerations because of its optimum asymptotic variance.

for the scale and shape $(\sigma, \xi)$ parameters will be computed and subsequently compared with MLE estimates in order to asses distributional robustness. This assessment will be performed through the analysis of visual (PR plots) and quantitative (tMAD) metrics of goodness-of-fit. These tools are taken from Brazauskas (2009) and defined precisely further in this section.

In this work, we have considered four robust inference methods well-accepted in the literature: (i) Method of Probability Weighted Moments, (ii) Method of Trimmed Moments, (iii) Method of Medians and (iv) Method of Minimum Density Power Divergence Estimator. Each of these methods is briefly described below.

### 3.4.1 Method of Probability Weighted Moments

Hosking and Wallis (1987) proposed an estimator for the GPD parameters $(\sigma, \xi)$ based on probability weighted moments (PWM), defined as

$$M_{p,r,s} = \mathbb{E}[X^p\{F(x)\}^r\{(1 - F(x))\}^s]$$

where $p, r, s \in \mathbb{R}^+$.

The authors show that the relationship between the GPD parameters and the PWM's $M_{1,0,s}$ is simpler than the relationship between the GPD parameters and the conventional moments $M_{p,0,0}$. Formally,

$$a_s = M_{1,0,s} = \mathbb{E}[X\{1 - F(x)\}^s] = \frac{\sigma}{(s+1)(s+1-\xi)}.$$

Thus, provided that $\xi < 1$, PWM estimators for the scale $\sigma$ and shape $\xi$ parameters of the GPD are given by

$$\hat{\sigma}_{PWM} = \frac{2a_0a_1}{a_0 - 2a_1} \tag{23}$$

and

$$\hat{\xi}_{PWM} = 2 - \frac{a_0}{a_0 - 2a_1} \tag{24}$$

where, for an observed sample of size $n$,

$$a_r = n^{-1} \sum_{j=1}^{n} \left(1 - \frac{j+\gamma}{n+\delta}\right)^r (X_{n,j} - u).$$

Hosking and Wallis (1987) recommended the use of $\gamma = -0.35$ and $\delta = 0$. It may be shown that, for $\xi < 1/2$, the PWM estimators are asymptotically normally distributed

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{\text{PWM}} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma_1}) \tag{25}$$

with

$$\boldsymbol{\Sigma_1} = \frac{1}{(1-2\xi)(3-2\xi)} \begin{bmatrix} \sigma^2(7 - 18\xi + 11\xi^2 - 2\xi^3) & \sigma(2-\xi)(2 - 6\xi + 7\xi^2 - 2\xi^3) \\ \sigma(2-\xi)(2 - 6\xi + 7\xi^2 - 2\xi^3) & (1-\xi)(2-\xi)^2(1 - \xi + 2\xi^2) \end{bmatrix}.$$

Although it can not be formally classified as a robust inference method, PWM estimators are fairly simple and require less assumptions about the underlying data being analysed, while maintaining a great degree of asymptotic efficiency relative to MLE.

### 3.4.2 Method of Trimmed Moments

Brazauskas and Kleefeld (2009) proposed a robust $L$-estimator for the GPD scale and shape parameters.[4] The Method of Trimmed Moments

[4]For further details about $L$-estimators, see Huber and Ronchetti (2009).

(MTM) estimates the parameters $(\sigma, \xi)$ by matching theoretical and empirical trimmed moments (in this case, trimmed means), and then solving the resulting equation with respect to $(\sigma, \xi)$.

Assuming a sample of size $n$, sample trimmed means may be easily computed by the following equation

$$\hat{\mu}_j = \frac{1}{n - m_n(j) - m_n^*(j)} \sum_{i=m_n(j)+1}^{n-m_n^*(j)} (X_{n,i} - u); \;\; j = 1, 2. \tag{26}$$

where $m_n(j)$ and $m_n^*(j)$ are integers such that $0 \leq m_n(j) \leq m_n^*(j) \leq n$, and $m_n(j)/n \to a_j$ and, $m_n^*(j)/n \to b_j$ as $n \to \infty$. The trimming proportions must be chosen by the analyst in view of the trade-off between efficiency and robustness. For $a_j > 0$, $b_j > 0$ and $(0 < a_j + b_j < 1)$, $j \in \{1, 2\}$, the resulting MTM estimators are robust with lower and upper breakdown points[5] given by LBP $= \min\{a_1, a_2\}$ and UBP $= \min\{b_1, b_2\}$.

In fact, the robustness of the MTM estimators against extremely small or large observations comes from the fact that order statistics with index less than $n \times$ LBP or more than $n \times (1 -$ LBP$)$ are simply not included in the estimates.

The corresponding population trimmed mean can be written as

$$\mu_j = \frac{1}{1 - a_j - b_j} \int_{a_j}^{1-b_j} (F^{-1}(x) - u) dx$$

$$= \sigma \begin{cases} -1 + \frac{1}{1-a_j-b_j} \log\left(\frac{1-a_j}{b_j}\right); & \xi = -1 \\ 1 + \frac{b_j \log(b_j) - (1-a_j)\log(1-a_j)}{1-a_j-b_j}; & \xi = 0 \\ (1/\xi)\left(1 - \frac{(1-a_j)^{\xi+1} - b_j^{\xi+1}}{(\xi+1)(1-a_j-b_j)}\right); & \text{otherwise} \end{cases} \tag{27}$$

[5]Breakdown Points are popular criterion of robustness loosely characterized as the largest proportion of sample observations that can be corrupted without the estimator itself becoming corrupted.

Therefore, MTM estimators for the scale $\sigma$ and shape $\xi$ parameters of the GPD can be computed by

$$\hat{\sigma}_{MTM} = \hat{\mu}_1 \begin{cases} -\left(1 - \frac{\log(1-a_1) - \log b_1}{1-a_1-b_1}\right)^{-1}; & \hat{\xi}_{MTM} = -1 \\ \left(1 - \frac{(1-a_1)\log(1-a_1) - b_1 \log b_1}{1-a_1-b_1}\right)^{-1}; & \hat{\xi}_{MTM} = 0 \\ \hat{\xi}_{MTM}\left(1 - \frac{(1-a_1)^{\hat{\xi}_{MTM}+1} - b_1^{\hat{\xi}_{MTM}+1}}{(\hat{\xi}_{MTM}+1)(1-a_1-b_1)}\right)^{-1}; & \text{otherwise} \end{cases} \tag{28}$$

where $\hat{\xi}_{MTM}$ is found numerically by solving the following equation with respect to $\xi$

$$\hat{\mu}_1/\hat{\mu}_2 - \mu_1/\mu_2 = 0. \tag{29}$$

Brazauskas and Kleefeld (2009) show that the MTM estimators are consistent and asymptotically normal

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{\mathrm{MTM}} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma_2}) \tag{30}$$

where $\boldsymbol{\Sigma_2} = \boldsymbol{D}\boldsymbol{\Sigma_{**}}\boldsymbol{D}^{\mathrm{T}}$, with $\boldsymbol{\Sigma_{**}} = [\sigma_{ij}^2]_{i,j=1}^2$

$$\sigma_{ij}^2 = \frac{1}{(1-a_i-b_i)} \times \int_{a_i}^{1-b_i} \int_{a_j}^{1-b_j} (\min\{u,v\} - uv) dF^{-1}(v) dF^{-1}(u)$$

and $\boldsymbol{D} = [d_{ij}]_{i,j=1}^2$ with entries

$$\boldsymbol{D} = \begin{cases} d_{11} = \sigma[(\mu_1^{'}\mu_2 - \mu_2^{'}\mu_1) - \mu_1^{'}\mu_2]/[\mu_1(\mu_1^{'}\mu_2 - \mu_2^{'}\mu_1)] \\ d_{12} = \sigma\mu_1^{'}/(\mu_1^{'}\mu_2 - \mu_2^{'}\mu_1) \\ d_{21} = \mu_2/(\mu_1^{'}\mu_2 - \mu_2^{'}\mu_1) \\ d_{22} = -\mu_1/(\mu_1^{'}\mu_2 - \mu_2^{'}\mu_1) \end{cases}.$$

Here $\mu_j^{'}$ is the derivative of $\mu_j$ with respect to $\xi$

$$\mu_j^{'} = -\sigma \times \begin{cases} 1 - \frac{\log(1-a_j) - \log b_j}{1-a_j-b_j} - \frac{\log^2(1-a_j) - \log^2 b_j}{2(1-a_j-b_j)}; & \xi = 1 \\ 1 - \frac{(1-a_j)\log(1-a_j) - b_j \log b_j}{1-a_j-b_j} + \frac{(1-a_j)\log^2(1-a_j) - b_j \log^2 b_j)}{2(1-a_j-b_j)}; & \xi = 0 \\ \frac{1}{\xi(\xi+1)}\left[(2\xi+1)(\mu_j/\sigma) - 1 + \frac{(1-a_j)^{\xi+1}\log(1-a_j) - b_j^{\xi+1}\log b_j}{1-a_j-b_j}\right]; & \text{otherwise} \end{cases}.$$

### 3.4.3 Method of Medians

Peng and Welsh (2001) proposed a robust $M$-estimator for the GPD parameters which they called the Method of Medians (MM).[6] The basic idea is to equate the sample median of each component of the score function to the corresponding population median.

For the GPD scale and shape parameters $(\sigma, \xi)$, MM estimators are written as

$$\hat{\sigma}_{MM} = \frac{\hat{\xi}_{MM}}{2^{\hat{\xi}_{MM}} - 1}\text{Median}(X_i - u) \tag{31}$$

The shape parameter $\hat{\xi}_{MM}$ is found numerically by solving the below equation with respect to $\xi$

$$\text{Median}\left(\frac{\log(1+\xi(X_i-u)/\sigma)}{\xi^2} - \frac{(1+\xi)(X_i-u)}{\sigma\xi + \xi^2(X_i-u)}\right) = -\frac{\log y_1}{\xi} - \frac{1+\xi}{\xi^2}(1-y_1^{\xi}) \tag{32}$$

where $y_1$ is computed as the solution of

$$-\frac{\log y_1}{\xi} - \frac{1+\xi}{\xi^2}(1-y_1^{\xi}) = -\frac{\log(y_1+1/2)}{\xi} - \frac{1+\xi}{\xi^2}(1-(y_1+1/2)^{\xi}).$$

[6]For further details about $M$-estimators, see Huber and Ronchetti (2009).

As shown by the authors, the method of medians (MM) estimator is consistent and asymptotically normally distributed

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{\mathrm{MM}} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma_3}) \tag{33}$$

where $\boldsymbol{\Sigma_3} = \boldsymbol{V}^{-1}\boldsymbol{B}\boldsymbol{V}^{-\mathrm{T}}$, with

$$\boldsymbol{V} = \begin{cases} v_{11} = \frac{2^{-\xi}-1}{\sigma\xi} sgn(\xi+1) \\ v_{12} = \frac{2^{-\xi}+\xi\log 2-1}{\xi^2} sgn(\xi+1) \\ v_{21} = \frac{2}{\xi\sigma}\{(y_1+1/2)^{\xi+1} - y_1^{\xi+1}\} - \frac{1}{\xi\sigma} \\ v_{22} = -\frac{2y_1\log y_1}{\xi} + \frac{2(y_1+1/2)\log(y_1+1/2)}{\xi} + \frac{2y_1^{\xi+1}}{\xi^2} - \frac{2(y_1+1/2)^{\xi+1}}{\xi^2} + \frac{1}{\xi^2} \end{cases}$$

and

$$\boldsymbol{B} = \begin{cases} b_{11} = 1 \\ b_{12} = \mathbb{E}\, sgn\left[\{\frac{(1+\xi)(X_1-u)}{1+\xi(X_1-u)/\sigma} - \frac{\sigma(1+\xi)}{\xi}(1-2^{-\xi})\}\{\frac{\log(1+\xi(X_1-u)/\sigma)}{\xi^2} - \frac{(1+\xi)(X_1-u)}{\sigma\xi+\xi^2(X_1-u)}\right. \\ \left. +\frac{\log y_1}{\xi} + \frac{1+\xi}{\xi^2}(1-y_1^{\xi})\}\right] \\ b_{21} = b_{12} \\ b_{22} = 1 \end{cases}.$$

### 3.4.4 Method of Minimum Density Power Divergence

Minimum Density Power Divergence Estimators (MDPDE) for the shape and scale parameters of the GPD were first implemented by Juaréz and Schucany (2004). The MDPDE is a robust $M$-estimator indexed by a constant $\alpha \geq 0$ that controls the trade-off between robustness and efficiency. As $\alpha$ increases, robustness increases and efficiency decreases. For $\alpha = 0$, the MDPDE is equivalent to the MLE. Good asymptotic relative efficiencies (above 90%) are usually maintained for $\alpha \leq 0.20$.

Let $\mathscr{X}_n = X_1, ..., X_n$ denote a random sample of excesses over some fixed threshold $u$, the density function of the GPD can be written as

$$g(x|\sigma,\gamma) = \frac{1}{\sigma}\left(1-\gamma\frac{x}{\sigma}\right)^{1/\gamma-1}$$

Note that we have omitted the location parameter $u$ in the above equation as the random sample is already defined as excesses above the threshold $u$. Moreover, the shape parameter $\gamma$ is defined here as the negative value of the shape parameter ($\gamma = -\xi$) in equation (3).

For a fixed $\alpha > 0$, the MDPDE for the GPD parameters $(\hat{\sigma}_\alpha, \hat{\gamma}_\alpha)$ are the values that minimizes the following equation

$$H_\alpha(\sigma,\gamma) = \frac{1}{\sigma^\alpha(1+\alpha-\alpha\gamma)} - \left(1+\frac{1}{\alpha}\right)\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sigma^\alpha}\left(1-\gamma\frac{X_i}{\sigma}\right)^{(1/\gamma-1)\alpha} \tag{34}$$

over the support

$$\{(\gamma,\sigma)\in\boldsymbol{\theta} : \sigma > 0, \max_{1\le i\le n}\{X_i\}\gamma < \sigma, -\infty < \gamma < 0, \text{ and } 0 \le \gamma \le (1+\alpha)/\alpha\}.$$

Juaréz and Schucany (2004) show that the MDPDE are consistent and asymptotically normal

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{\text{MDPDE}} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma_4}) \tag{35}$$

where $\boldsymbol{\Sigma_4} = \boldsymbol{J}_\alpha^{-1}\boldsymbol{K}_\alpha\boldsymbol{J}_\alpha^{-T}$, with

$$\boldsymbol{K}_\alpha = \frac{1}{n}\sum_{i=1}^{n} S(X_i|\sigma,\gamma)S^T(X_i|\sigma,\gamma)g^{2\alpha}(X_i|\sigma,\gamma) - \frac{1}{n^2}\Big\{\sum_{i=1}^{n} S(X_i|\sigma,\gamma)g^{\alpha}(X_i|\sigma,\gamma)\Big\}$$

$$\Big\{\sum_{i=1}^{n} S(X_i|\sigma,\gamma)g^{\alpha}(X_i|\sigma,\gamma)\Big\}^T$$

and

$$\boldsymbol{J}_{\alpha} = (1+\alpha)\int_{D(\gamma,\sigma)} S(x|\sigma,\gamma)S^T(x|\sigma,\gamma)g^{\alpha+1}(x|\sigma,\gamma)dx - \int_{D(\gamma,\sigma)} i(x|\sigma,\gamma)g^{\alpha+1}(x|\sigma,\gamma)dx$$

$$+\frac{1}{n}\sum_{i=1}^{n}\{i(X_i|\sigma,\gamma) - \alpha S(X_i|\sigma,\gamma)S^T(X_i|\sigma,\gamma)\}g^{\alpha}(X_i|\sigma,\gamma)$$

The integration limit is represented by $D(\gamma,\sigma) = [0,\infty)$ if $\gamma \leq 0$, and $D(\gamma,\sigma) = [0,\sigma/\gamma]$ if $\gamma > 0$.

Finally, the components of the score function and the information matrix,

$$S(x|\sigma,\gamma) = \begin{bmatrix} S_{\gamma}(x|\sigma,\gamma) \\ S_{\sigma}(x|\sigma,\gamma) \end{bmatrix} \qquad i(x|\sigma,\gamma) = \begin{bmatrix} i_{\gamma}(x|\sigma,\gamma) & i_{\gamma\sigma}(x|\sigma,\gamma) \\ i_{\xi\sigma}(x|\sigma,\gamma) & i_{\sigma}(x|\sigma,\gamma) \end{bmatrix}$$

are given by

$$S_{\gamma}(x|\sigma,\gamma) = \frac{\partial}{\partial\gamma}\log g(x|\sigma,\gamma) = -\frac{1}{\gamma^2}\log\left(1-\gamma\frac{x}{\sigma}\right) + \frac{1}{\gamma}\left(\frac{1}{\gamma}-1\right)\left\{1-\left(1-\gamma\frac{x}{\sigma}\right)^{-1}\right\},$$

$$S_{\sigma}(x|\sigma,\gamma) = \frac{\partial}{\partial\sigma}\log g(x|\sigma,\gamma) = -\frac{1}{\gamma\sigma} + \frac{1}{\sigma}\left(\frac{1}{\gamma}-1\right)\left(1-\gamma\frac{x}{\sigma}\right)^{-1},$$

$$i_{\gamma}(x|\sigma,\gamma) = -\frac{\partial^2}{\partial\gamma^2}\log g(x|\sigma,\gamma) = -\frac{2}{\gamma^2}\log\left(1-\gamma\frac{x}{\sigma}\right) + \frac{3-\gamma}{\gamma^3} - \frac{2(2-\gamma)}{\gamma^3}\left(1-\gamma\frac{x}{\sigma}\right)^{-1}$$

$$+\frac{1-\gamma}{\gamma^3}\left(1-\gamma\frac{x}{\sigma}\right)^{-2},$$

$$i_\sigma(x|\sigma,\gamma) = -\frac{\partial^2}{\partial\sigma^2}\log g(x|\sigma,\gamma) = -\frac{1}{\sigma^2\gamma} + \frac{1}{\sigma^2}\left(\frac{1}{\gamma}-1\right)\left(1-\gamma\frac{x}{\sigma}\right)^{-2},$$

$$i_{\gamma\sigma}(x|\sigma,\gamma) = -\frac{\partial^2}{\partial\gamma\partial\sigma}\log g(x|\sigma,\gamma) = -\frac{1}{\sigma\gamma^2}+\frac{2-\gamma}{\sigma\gamma^2}\left(1-\gamma\frac{x}{\sigma}\right)^{-1}-\frac{1-\gamma}{\sigma\gamma^2}\left(1-\gamma\frac{x}{\sigma}\right)^{-2}.$$

### 3.4.5 Robust Estimators

As mentioned at the beginning of this section, efficiency and robustness tend to trade-off against each other. The literature points out that results drawn form highly inefficient estimators should be analysed with great caution. For instance, Brazauskas (2009) remarks that a robust but inefficient procedure tends to capture spurious patterns in the sample, leading to misleading conclusions. Thus, in our subsequent analysis, asymptotic efficiency will be considered in conjunction with robustness in order to avoid spurious conclusions.

The Asymptotic Relative Efficiency (ARE) of one estimator with respect to another is defined as the ratio of their asymptotic variances. For an estimator in the parameter space $\boldsymbol{\theta} \in \mathbb{R}^k$, the two variances are replaced by the corresponding generalized variances, and the ratio is then raised to the power $1/k$. Thus, for the two parameter GPD $(\sigma, \xi)$, we have that

$$\mathrm{ARE}(\boldsymbol{i}, \mathrm{MLE}) = \left(\frac{|\boldsymbol{\Sigma_0}|}{|\boldsymbol{\Sigma_i}|}\right)^{1/2} \tag{36}$$

where $|\boldsymbol{\Sigma_i}|$ denotes the determinant of the asymptotic covariance matrix $\boldsymbol{\Sigma_i}$ (generalized variance) related to method $\boldsymbol{i} = \mathbf{0, 1, 2, 3, 4}$.

To asses goodness-of-fit, the trimmed Mean Absolute Deviation (tMAD) is employed. The tMAD measures the absolute distance between the fitted GPD quantiles and the observed data. It can be defined as

$$\text{tMAD}_\delta = \frac{1}{\lfloor n\delta \rfloor} \sum_{i=1}^{\lfloor n\delta \rfloor} b_{n,i} \tag{37}$$

where $b_{n,i}$ denotes the $i$th smallest distance among

$$\left| X_{n,j} - \hat{G}^{-1}\left( \frac{j-0.5}{n} \middle| u, \sigma, \xi \right) \right| ; \;\; j = 1, ..., n.$$

Following Brazauskas and Kleefeld (2009), we use the following values of $\delta$: $0.50, 0.75, 0.90, 0.95, 1$. The choice of $\delta = 0.90$, for instance, indicates how far, on the average, are the 90% closest observations from their corresponding fitted quantiles.

Tables 13 to 18 report ARE and tMAD measures, as well as parameter estimates, for each analysed estimator, namely: Maximum Likelihood (MLE), Probability Weighted Moments (PWM), Method of Trimmed Moments ($\text{MTM}_1$) with ($a_1$, $b_1$) = (0.30, 0.50), ($a_2$, $b_2$) = (0.70, 0.15), Method of Trimmed Moments ($\text{MTM}_2$) with ($a_1$, $b_1$) = (0.10, 0.55), ($a_2$, $b_2$) = (0.70, 0.05), Method Medians (MM), Method of Minimum Density Power Divergence ($\text{MDPDE}_1$) with $\alpha = 0.10$ and Method of Minimum Density Power Divergence ($\text{MDPDE}_2$) with $\alpha = 0.15$.

| Inference Method | MLE | PWM | $\text{MTM}_1$ | $\text{MTM}_2$ | MM | $\text{MDPDE}_1$ | $\text{MDPDE}_2$ |
|---|---|---|---|---|---|---|---|
| ARE | 1.000 | - [7] | 0.559 | 0.740 | 0.506 | 0.986 | 0.921 |

**Table 13.** Asymptotic Relative Efficiency - Danish Fire Insurance

| | Parameters | | tMAD | | | | |
|---|---|---|---|---|---|---|---|
| Inference Method | $\hat{\sigma}$ | $\hat{\xi}$ | $\delta = 0.50$ | $\delta = 0.75$ | $\delta = 0.90$ | $\delta = 0.95$ | $\delta = 1$ |
| MLE | 1.868 (0.130) | 0.659 (0.063) | 0.013 | 0.038 | 0.104 | 0.160 | 0.541 |
| PWM | 1.955 (–)[7] | 0.599 (–)[7] | 0.015 | 0.045 | 0.134 | 0.207 | 0.452 |
| $MTM_1$ | 1.962 (0.166) | 0.583 (0.109) | 0.015 | 0.041 | 0.133 | 0.219 | 0.514 |
| $MTM_2$ | 1.899 (0.152) | 0.689 (0.080) | 0.012 | 0.063 | 0.126 | 0.158 | 0.794 |
| MM | 2.018 (0.140) | 0.484 (0.105) | 0.018 | 0.031 | 0.196 | 0.383 | 1.095 |
| $MDPDE_1$ | 1.856 (0.128) | 0.670 (0.064) | 0.014 | 0.040 | 0.105 | 0.157 | 0.600 |
| $MDPDE_2$ | 1.853 (0.129) | 0.673 (0.069) | 0.014 | 0.040 | 0.105 | 0.156 | 0.617 |

**Table 14.** Robust Estimators - Danish Fire Insurance

| Inference Method | MLE | PWM | $MTM_1$ | $MTM_2$ | MM | $MDPDE_1$ | $MDPDE_2$ |
|---|---|---|---|---|---|---|---|
| ARE | 1.000 | 0.943 | 0.433 | 0.602 | 0.353 | 0.925 | 0.889 |

**Table 15.** Asymptotic Relative Efficiency - U.S. Automobile Insurance

| | Parameters | | tMAD | | | | |
|---|---|---|---|---|---|---|---|
| Inference Method | $\hat{\sigma}$ | $\hat{\xi}$ | $\delta = 0.50$ | $\delta = 0.75$ | $\delta = 0.90$ | $\delta = 0.95$ | $\delta = 1$ |
| MLE | 3,049.987 (269.304) | 0.245 (0.069) | 37.612 | 71.098 | 101.484 | 120.883 | 249.108 |
| PWM | 3,058.232 (276.951) | 0.246 (0.072) | 39.760 | 72.910 | 104.559 | 123.132 | 252.087 |
| $MTM_1$ | 3,054.937 (366.080) | 0.202 (0.155) | 35.748 | 60.494 | 95.800 | 122.452 | 299.819 |
| $MTM_2$ | 2,996.625 (341.139) | 0.238 (0.114) | 33.979 | 58.816 | 94.133 | 116.693 | 247.581 |
| MM | 3,256.982 (311.867) | 0.271 (0.143) | 58.750 | 133.276 | 238.904 | 295.256 | 483.717 |
| $MDPDE_1$ | 3,074.402 (272.698) | 0.234 (0.074) | 38.098 | 69.652 | 98.848 | 120.744 | 249.651 |
| $MDPDE_2$ | 3,088.424 (278.399) | 0.228 (0.077) | 39.446 | 70.152 | 98.602 | 119.976 | 251.597 |

**Table 16.** Robust Estimators - U.S. Automobile Insurance

| Inference Method | MLE | PWM | $MTM_1$ | $MTM_2$ | MM | $MDPDE_1$ | $MDPDE_2$ |
|---|---|---|---|---|---|---|---|
| ARE | 1.000 | 0.934 | 0.441 | 0.606 | 0.395 | 0.965 | 0.927 |

**Table 17.** Asymptotic Relative Efficiency - Australian Personal Injury Insurance

[7] Asymptotic standard errors are not available for $\hat{\xi} \geq 1/2$.

| | Parameters | | tMAD | | | | |
|---|---|---|---|---|---|---|---|
| Inference Method | $\hat{\sigma}$ | $\hat{\xi}$ | $\delta = 0.50$ | $\delta = 0.75$ | $\delta = 0.90$ | $\delta = 0.95$ | $\delta = 1$ |
| MLE | 148.448 (9.650) | 0.282 (0.052) | 1.024 | 2.107 | 3.243 | 3.774 | 7.225 |
| PWM | 148.797 (9.884) | 0.283 (0.055) | 0.970 | 2.048 | 3.273 | 3.829 | 7.196 |
| $MTM_1$ | 154.214 (13.372) | 0.211 (0.112) | 0.854 | 2.230 | 3.830 | 5.303 | 15.685 |
| $MTM_2$ | 155.140 (12.750) | 0.229 (0.082) | 0.811 | 1.796 | 3.448 | 4.257 | 11.991 |
| MM | 156.645 (10.646) | 0.228 (0.101) | 0.866 | 2.008 | 3.726 | 4.533 | 11.888 |
| $MDPDE_1$ | 149.482 (9.561) | 0.272 (0.053) | 0.947 | 2.027 | 3.237 | 3.801 | 7.896 |
| $MDPDE_2$ | 149.819 (9.687) | 0.268 (0.056) | 0.898 | 2.001 | 3.228 | 3.807 | 8.307 |

**Table 18.** Robust Estimators - Australian Personal Injury Insurance

After examining the results, we make the following considerations. All estimators were able to maintain a reasonable degree of efficiency. PWM, $MDPDE_1$ and $MDPDE_2$ presented AREs well above 90%, on average. The high asymptotic efficient of the PWM estimator was already expected since it is only marginally robust when compared to the MLE. The high values of ARE associated with MDPDE estimators, on the hand, corroborates the claims of Juárez and Schucany (2004) that the MDPDE estimator is able to maintain good asymptotic relative efficiency for small values of the tuning parameter ($\alpha < 0.2$).

Moreover, $MTM_1$ and $MTM_2$ presented relative efficiencies between 40% and 55%, and 60% and 75%, respectively. Note that the $MTM_1$ has lower ARE ratios because of its higher degree of robustness. In fact, UBP and LBP for $MTM_1$ are 0.30 and 0.85, while for $MTM_2$ the same measures corresponds to the quantiles 0.10 and 0.95. Finally, MM estimators showed the worst asymptotic efficient among the analysed estimators (ARE between 35% and 50%). The lack of asymptotic efficiency of the MM estimators was already pointed out by He and Fung (1999). We believe, however, that it will not compromise significantly our conclusions.

Turning now to the tMAD goodness-of-fit measure, MLE and PWM fits are best for all data under consideration ($\delta = 1$). This should not be surprising since these procedures are designed to fit all data points (*i.e.* no trimming is used whatsoever). But if we look to the other values of $\delta$ (which reflect the fit for most but not all observations), we see that the MLE/PWM and robust fits are very similar and fairly close to the actual data. Thus, so far the robust procedures have not offered any significant improvements over the MLE.

One aspect that is not addressed by the tMAD measure is the variability of the estimated parameters. Since our robust estimators have different degrees of asymptotic efficient, this might be an important aspect to be incorporated. Thus, to further evaluate the goodness-of-fit of each estimator considering their asymptotic variances, we employed a graphical tool known as percentile-residuals (PR) plots. PR plots are constructed by plotting the empirical percentile levels, $(j/n)100\%$, versus standardized residuals (absolute residuals divided by their asymptotic standard errors)

$$R_{n,j} = \frac{X_{n,j} - G^{-1}\left(\frac{j-0.5}{n}\middle| u, \hat{\sigma}, \hat{\xi}\right)}{\text{standard deviation of } G^{-1}\left(\frac{j-0.5}{n}\middle| u, \hat{\sigma}, \hat{\xi}\right)}; \;\; j = 1, ..., n. \tag{38}$$

In the PR plot, the horizontal line at 0 represents the estimated quantiles, and the $\pm 2.5$ lines are the tolerance limits. A good fit would be the one for which the majority of points are scattered between the tolerance limits (A random variable following a standard Normal distribution exceeds 2.5 with probability 0.0062). These same limits were employed before by Hubert, Rousseeuw and Van Aelst (2004) and Brazauskas (2009). The denominator of the above equation is estimated by the delta method in conjunction with the corresponding asymptotic covariance matrix.

Figures 10, 11 and 12 present the PR plots for MLE, PWM, $\text{MTM}_1$, $\text{MTM}_2$, MM, $\text{MDPDE}_1$ and $\text{MDPDE}_2$ fits based on the three data sets analysed in this thesis.

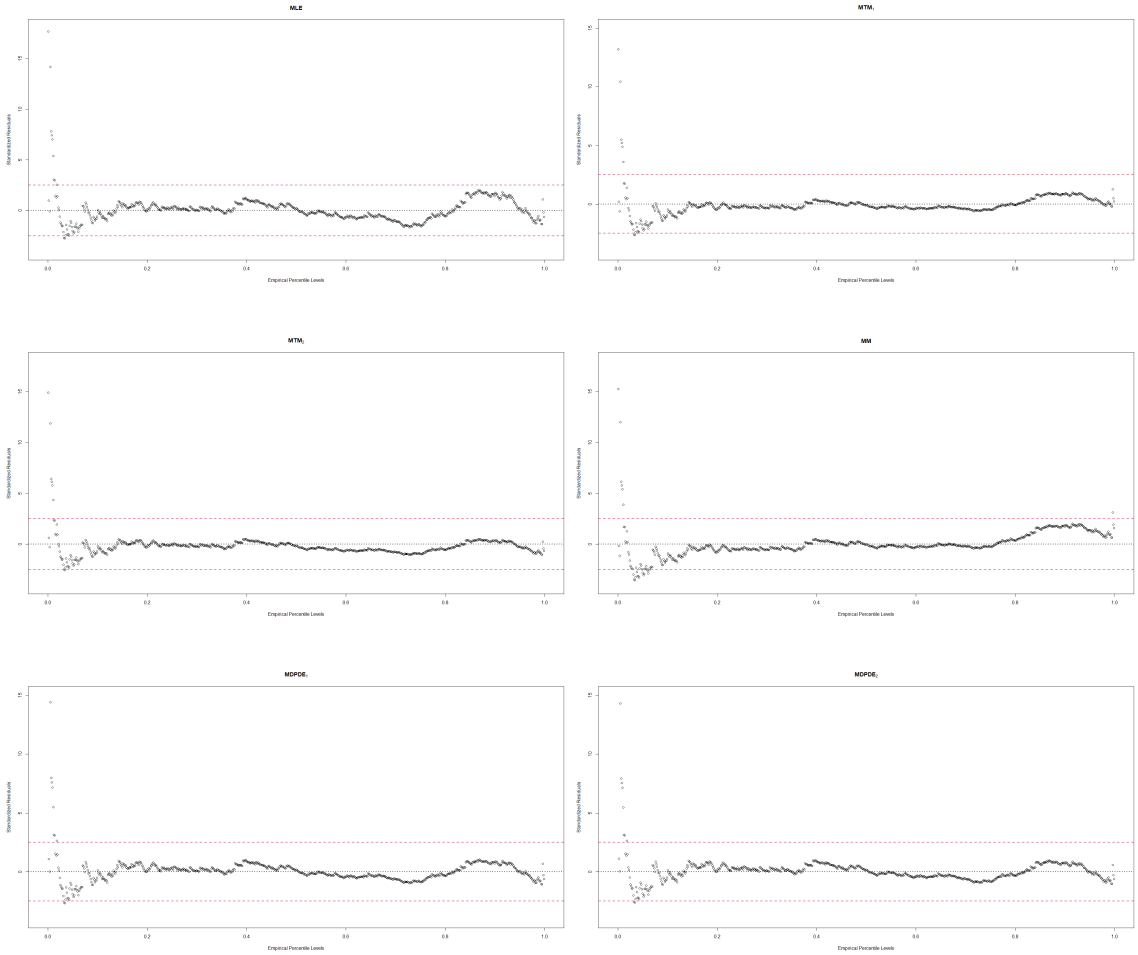



**Figure 10.** Percentile-Residual Plots - Danish Fire Insurance

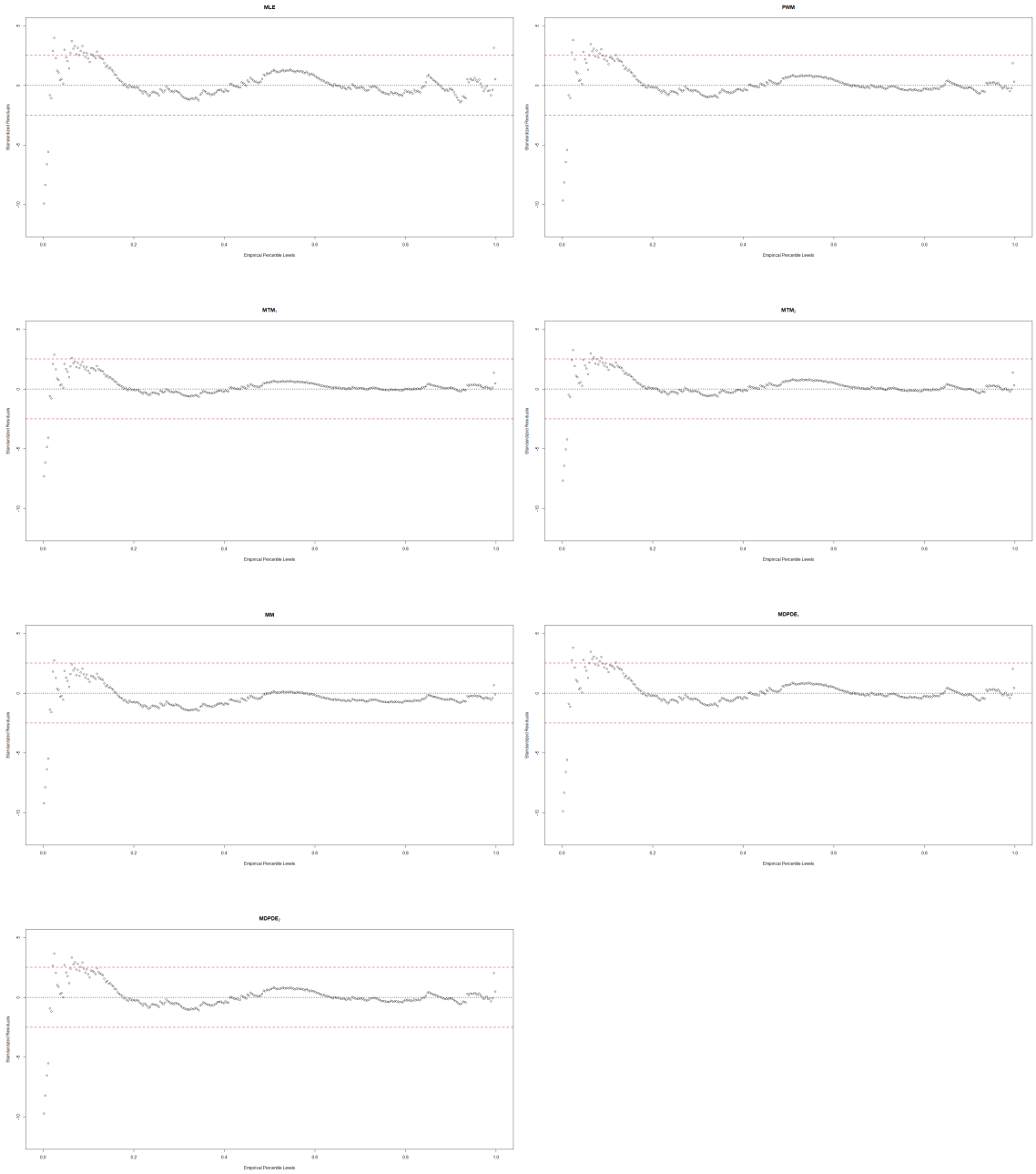



**Figure 11.** Percentile-Residual Plots - U.S. Automobile Insurance

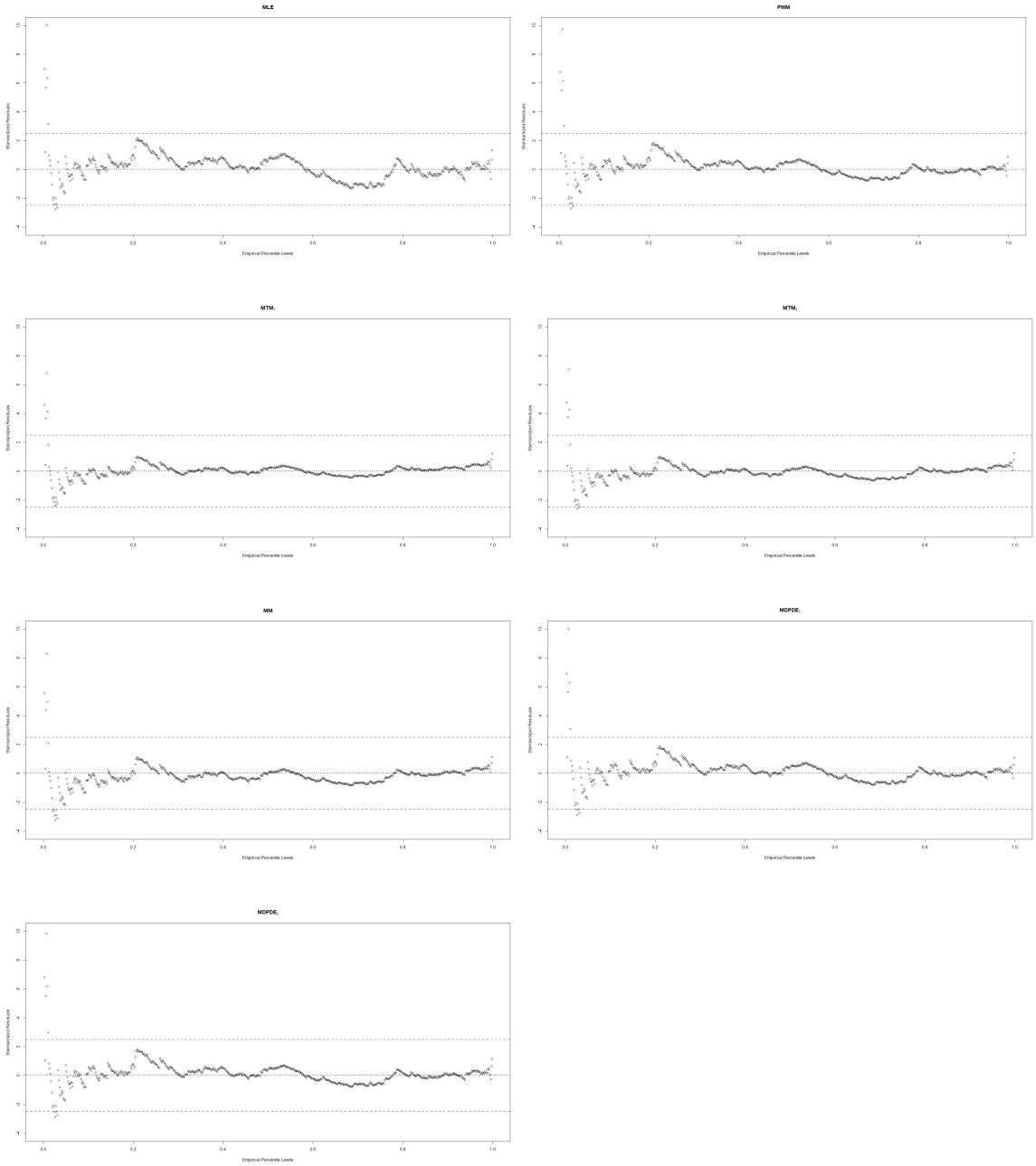



**Figure 12.** Percentile-Residual Plots - Australian Personal Injury Insurance

As one can see from the above figures, all estimation methods do a very poor job at fitting small losses, *i.e.* those losses slightly above the threshold $u$, but they perform reasonably well for medium and large losses. This is an expected behaviour, since the GPD distribution is designed precisely to fit the most extreme observations.

Among the PR plots, the MLE fit looks worst. One should keep in mind, however, that the standardized deviations in these plots depend on the efficiency of the estimator. Less efficient estimators will have larger standard deviations of $G^{-1}\left(.|u,\hat{\sigma},\hat{\xi}\right)$ and, consequently, smaller standardized residuals. Thus, the same value of absolute residual will appear as significantly larger on the MLE's PR plot that it will on the other less efficient robust estimators.

Overall, we can conclude from the PR plots that the robust estimators do not improve significantly the goodness-of-fit of our GPD extreme value component, since all methods (including MLE) accommodate the majority of data points between the tolerance limits. In other words, MLE estimators seems to already provide a reasonable good fit for the analysed insurance data, corroborating the conclusions that emerged form the analysis of the tMAD measures. The robust estimators were only able to marginally improve the fit and their use is not justified given their lower asymptotic relative efficiency (ARE).

Taken simultaneously, the results in this section lead us to conclude that our results are in fact distributional robust, *i.e.* there is no evidence that our distributional assumptions are incorrect in such a way that would invalidate our results.

# 4 Concluding Remarks

In this thesis, we suggested a simple and flexible nonparametric extreme value mixture model for insurance loss data. This new development is obtained by allowing the bulk of the loss distribution (*i.e.* observations below some threshold) to be described by a nonparametric kernel density estimator, while the excess above the threshold are assumed to follow a Generalized Pareto distribution.

The model has the advantage of being extremely flexible due to its nonparametric component, avoiding the need to impose a functional form to the bulk of the loss distribution, as in the previous extreme value mixture approaches proposed in the actuarial literature. Moreover, the kernel density estimator has just a single extra parameter to be estimated, overcoming the problem of high computational burden related to other similar models.

A further advantage of the model is the ability to infer the bandwidth parameter related to the nonparametric component automatically through a data-driven procedure. The problem of inconsistency (*i.e.* high sensitivity to extreme observations and tail heaviness) of the traditional likelihood cross-validation (LCV) based estimators is overcame by capturing the tails of the loss distribution through the GPD component, yielding a more robust inference method for bandwidth selection than the "rule-of-thumb" methods typically applied in the actuarial literature.

To illustrate the effectiveness of the proposed model in the context of casualty and property losses, we considered three real insurance loss data sets widely available and well-studied in the loss modelling literature. Overall, the results indicated that the nonparametric extreme value mixture model was able to provide a better fit to the empirical loss distributions than classical loss models and other flexible alternative models recently proposed in the literature, even when allowances are made for the larger number of parameters in some models.

It is also worth mentioning that the proposed model may be very useful for practitioners in the actuarial and insurance industry. As noted by Buch-Kromann (2006), actuaries and statisticians are "spending too much time trying to develop parametric models of losses (...) as no single parametric

model seemed to fit both small and large losses". Practitioners who encounter smaller data with higher frequencies as well as occasional larger data with lower frequencies are now exposed to a flexible semiparametric alternative for their research work.

# 5 Bibliography


[1] Abadir, K. M., Lawford, S. (2004). Optimal Asymmetric Kernels. *Economics Letters*, 83(1), pp. 61-68.

[2] Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 19(6), pp. 716-723.

[3] Arcadiacono, P., Jones, J. B. (2003). Finite Mixture Distributions, Sequential Likelihood and the EM Algorithm. *Econometrica*, 71(3), pp. 933-946.

[4] Babu, G. J., Rao, C. R. (2004). Goodness-of-Fit Tests when Parameters are Estimated. *Indian Journal of Statistics*, 66(1), pp. 63-74.

[5] Bakar, S. A. A., Hamzah, N. A., Maghsoudi, M., Nadarajah, S. (2015). Modeling Loss Data Using Composite Models. *Insurance: Mathematics and Economics*, 61, pp. 146-154.

[6] Balasooriya, U., Low, C. (2008). Modeling Insurance Claims with Extreme Value Observations: Transformed Kernel Density and Generalized Lambda Distribution. *North American Actuarial Journal*, 12(2), pp. 129-142.

[7] Balkema, A., de Haan, L. (1974). Residual Life Time at Great Age. *Annals of Probability*, 2(5), pp. 792-804.

[8] Behrens, C. N., Lopes, H. F., Gamerman, D. (2004). Bayesian Analysis of Extreme Events with Threshold Estimation. *Statistical Modelling*, 4(3), pp. 227-244.

[9] Blostein, M., Miljkovic, T. (2019). On Modeling Left-Truncated Loss Data Using Mixtures of Distributions. *Insurance: Mathematics and Economics*, 85, pp. 35-46.

[10] Bolancé, C., Guillen, M., Nielsen, J. P. (2003). Kernel Density Estimation of Actuarial Loss Functions. *Insurance: Mathematics and Economics*, 32, pp. 19-36.

[11] Bowman, A. (1984). An Alternative Method of Cross-Validation for the Smoothing of Density Estimates. *Biometrika*, 71(2), pp.363-360.

[12] Brazauskas, V. (2009). Robust and Efficient Fitting of Loss Models: Diagnostics Tools and Insights. *North American Actuarial Journal*, 13(3), pp. 356-369.

[13] Brazauskas, V., Kleefeld, A. (2009). Robust and Efficient Fitting of the Generalized Pareto Distribution with Actuarial Applications in View. *Insurance: Mathematics and Economics*, 45, pp. 424-435.

[14] Brazauskas, V., Kleefeld, A. (2016). Modeling Severity and Measuring Tail Risk of Norwegian Fire Claims. *North American Actuarial Journal*, 20(1), pp. 1-16.

[15] Brazauskas, V., Serfling, R. (2000). Robust and Efficient Estimation of the Tail Index of A Single-Parameter Pareto Distribution. *North American Actuarial Journal*, 4(4), pp. 12-27.

[16] Buch-Kromann, T. (2006). Estimation of Large Insurance Losses: A Case Study. *Journal of Actuarial Pratice*, 13, pp. 191-211.

[17] Buch-Larsen, T., Nielsen, J. P., Guillén, M., Bolancé, C. (2005). Kernel Density Estimation for Heavy-Tailed Distributions using the Champernowne Transformation. *Statistics*, 39(6), pp. 503-518.

[18] Burnham, K. P., Anderson, D. R. (2003). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer.

[19] Cabras, S., Castellanos, M. E. (2011). A Bayesian Approach for Estimating Extreme Quantiles Under a Semiparametric Mixture Model. *ASTIN Bulletin*, 41(1), pp. 87-106.

[20] Caeiro, F., Gomes, M. I. (2014). On the Bootstrap Methodology for the Estimation of the Tail Sample Fraction. *Proceedings of COMPSTAT 2014*, pp. 545-552.

[21] Calderín-Ojeda, E., Kwok, C. F. (2016). Modeling Claims Using Composite Stoppa Models. *Scandinavian Actuarial Journal*, 2016(9), pp. 817-836.

[22] Carreau, J., Bengio, Y. (2009). A Hybrid-Pareto Model for Asymmetric Fat-Tailed Data: the Univariate Case. *Extremes*, 12, pp. 53-76.

[23] Cebrián, A. C., Denuit, M., Lambert, P. (2003). Generalized Pareto Fit to the Society of Actuaries' Large Claims Database. *North American Actuarial Journal*, 7(3), pp. 18-36.

[24] Clements, A., Hurn, S., Lindsay, K. (2003). Moibus-like Mappings and Their Use in Kernel Density Estimation. *Journal of the American Statistical Association*, 98, pp. 993-1000.

[25] Cooray, K., Ananda, M. M. A. (2005). Modeling Actuarial Data with a Composite Lognormal-Pareto Model. *Scandinavian Actuarial Journal*, 2005(5), pp. 321-334.

[26] Danielsson, J., de Haan, L., Peng, L., Vries, C. G. (2001). Using a Bootstrap Method to Choose the Sample Fraction in Tail Index Estimation. *Journal of Multivariate Statistics*, 76(2), pp. 226-248.

[27] Duin, R. P. W. (1976). On the Choice of Smoothing Parameters for Parzen Estimators of Probability Density Functions. *IEEE Transactions on Computers*, C 25, pp. 1175-1179.

[28] DuMouchel, W. H. (1983). Estimating the Stable Index $\alpha$ in Order to Measure Tail Thickness: A Critique. *Annals of Statistics*, 11(4), pp. 1019-1031.

[29] Embrechts, P., Klüppelberg, C., Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*. Applications of Mathematics. Springer-Verlag.

[30] Ferreira, A., de Haan, L., Peng, L. (2003). On Optimising the Estimation of High Quantiles of a Probability Distribution. *Statistics*, 37(5), pp. 401-434.

[31] Fisher, R., Tippett, L. (1928). Limiting Forms of the Frequency Distribution of the Largest or Smallest Member of a Sample. *Proceedings of the Cambridge Philosophical Society*, 24(2), pp. 180-190.

[32] Fraga Alves, M. I., Gomes, M. I., de Haan, L. (2003). A New Class of Semi-Parametric Estimators of the Second Order Parameter. *Potugaliae Mathematica*, 60(2), pp. 193-214.

[33] Frees, E. (2010). *Regression Modeling with Actuarial and Financial Applications*. International Series on Actuarial Science. Cambridge.

[34] Frigessi, A., Haug, O., Rue, H. (2002). A Dynamic Mixture Model for Unsupervised Tail Estimation without Threshold Selection. *Extremes*, 5, pp. 219-235.

[35] Gertensgarbe, F. W., Werner, P. C. (1989). A Method for the Statistical Definition of Extreme-Value Regions and Their Application to Meteorological Time Series. *Zeitschrift für Meteorologie*, 39(4), pp. 224-226.

[36] Gnedenko, B. (1943). On the Limiting Distribution of the Maximum Term in a Random Series. *Annals of Mathematics*, 44, pp. 423-453.

[37] Gomes, M. I., Caeiro, F., Figueiredo, F. (2004). Bias Reduction of a Tail Index Estimator Through an External Estimation of the Second-Order Parameter. *Statistics*, 38(6), pp. 497-510.

[38] Gomes, M. I., Martins, M. J. (2002). "Asymptotically Unbiased" Estimators of the Tail Index Based on External Estimation of the Second Order Parameter. *Extremes*, 5(1), pp. 5-31.

[39] Gomes, M. I., Pestana, D. (2007a). A Simple Second-Order Reduced Bias' Tail Index Estimator. *Journal of Statistical Computation and Simulation*, 77(6), pp. 487-504.

[40] Gomes, M. I., Pestana, D. (2007b). A Sturdy Reduced-Bias Extreme Quantile ($VaR$) Estimator. *Journal of the American Statistical Association*, 102, pp. 280-292.

[41] Grün, B., Miljkovic, T. (2019). Extending Composite Loss Models Using a General Framework of Advanced Computational Tools. *Scandinavian Actuarial Journal*, 2019(8), pp. 642-660.

[42] Guillou, A., Hall, P. G. (2001). A Diagnostic for Selecting the Threshold in Extreme Value Analysis. *Journal of the Royal Statistical Society, Series B*, 63(2), pp. 293-305.

[43] Habbema, J., Hermans, J., van den Broek, K. (1974). A Stepwise Discriminat Analysis Program using Density Estimation. In: Bruckmann, G. (Ed.) *Compstat 1974: Proceedings in Computational Statistics*. Physica-Verlang.

[44] Hall, P. (1982). On Some Simple Estimates of an Exponent of Regular Variation. *Journal of the Royal Statistical Society, Series B*, 44(1), pp. 37-42.

[45] Hall, P. (1990). Using the Bootstrap to Estimate Mean Squared Error and Select Smoothing Parameter in Nonparametric Problems. *Journal of Multivariate Analysis*, 32(2), pp. 177-203.

[46] Hall, P., Welsh, A. (1985). Adaptive Estimates of Parameters of Regular Variation. *The Annals of Statistics*, 13(1), pp. 331-341.

[47] He, X., Fung, W. K. (1999). Method of Medians for Lifetime Data with Weibull Models. *Statistics in Medicine*, 18(15), pp. 1993-2009.

[48] Hill, B. (1975). A Simple Approach to Inference About the Tail of a Distribution. *Annals of Statistics*, 3(5), pp. 1163-1174.

[49] Hosking, J. R. M., Wallis, J. R. (1987). Parameter and Quantile Estimation for the Generalized Pareto Distribution. *Technometrics*, 29(3), pp. 339-349.

[50] Huber, P. J., Ronchetti, E. M. (2009). *Robust Statistics*. Wiley Series in Probability and Statistics. Wiley.

[51] Hubert, M., Rousseeuw, P. J., Van Aelst, S. (2004). Robustness. *Encyclopedia of Actuarial Science*, 3, pp. 1515–29, Wiley.

[52] Hyndman, R. J., Fan, Y. (1996). Sample Quantiles in Statistical Packages. *American Statistician*, 50(4), pp. 361–365.

[53] Juárez, S. F., Schucany, W. R. (2004). Robust and Efficient Estimation for the Generalized Pareto Distribution. *Extremes*, 7(3), pp. 237-251.

[54] de Jong, P., Heller, G. Z. (2008). *Generalized Linear Models for Insurance Data*. International Series on Actuarial Science. Cambridge.

[55] Keatinge, C. L. (1999). Modeling Losses with the Mixed Exponential Distribution. *Proceedings of the Casualty Actuarial Society*, 86, pp. 654-698.

[56] Kleiber, C., Kotz, S. (2003). *Statistical Size Distributions in Economics and Actuarial Sciences*. Wiley Series in Probability and Statistics. Wiley.

[57] Klugman, S., Panjer, H. H., Willmot, G. E. (2008). *Loss Models: From Data to Decisions*. Wiley Series in Probability and Statistics. Wiley.

[58] Klugman, S., Rioux, J. (2006). Toward a Unified Approach to Fitting Loss Models. *North American Actuarial Journal*, 10(1), pp. 63-83.

[59] Lee, S. C. K., Lin, X. S. (2010). Modeling and Evaluating Insurance Losses Via Mixtures of Erlang Distributions. *North American Actuarial Journal*, 14(1), pp. 107-130.

[60] Lee, D., Li, W. K., Wong, T. S. T. (2012). Modeling Insurance Claims Via a Mixture of Exponential Model Combined with Peaks-Over-Threshold Approach. *Insurance: Mathematics and Economics*, 51, pp. 538-550.

[61] Li, Q., Racine, J. S. (2007). *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.

[62] Lindsey, J. K. (1974a). Comparison of Probability Distributions. *Journal of the Royal Statistical Society, Series B*, 36(1), pp. 38-47.

[63] Lindsey, J. K. (1974b). Construction and Comparison of Statistical Models. *Journal of the Royal Statistical Society, Series B*, 36(3), pp. 418-425.

[64] Loretan, M., Philips, P. C. B. (1994). Testing the Covariance Stationarity of Heavy Tailed Time Series: An Overview of the Theory with Applications to Several Financial Datasets. *Journal of Empirical Finance*, 1(2), pp. 211-248.

[65] MacDonald, A., Scarrott, C. J., Lee, D., Darlow, B., Reale, M., Russell, G. (2011). A Flexible Extreme Value Mixture Model. *Computational Statistics and Data Analysis*, 55(6), pp. 2137-2157.

[66] McNeil, A. J. (1997). Estimating the Tails of Loss Severity Distributions Using Extreme Value Theory. *ASTIN Bulletin*, 27(1), pp. 117-137.

[67] McNeil, A. J., Saladin, T. (1997). The Peaks Over Threshold Method for Estimating High Quantiles of Loss Distributions. *Proceedings of the 28th International ASTIN Colloquium*, pp. 23-43.

[68] Mendes, B. V. M., Lopes, H. F. (2004). Data Driven Estimates for Mixtures. *Computational Statistics and Data Analysis*, 47(3), pp. 583-598.

[69] Miljkovic, T., Grün, B. (2016). Modeling Loss Data Using Mixtures of Distributions. *Insurance: Mathematics and Economics*, 70, pp. 387-396.

[70] Nadarajah, S., Bakar, S. A. A. (2014). New Composite Models for the Danish Fire Insurance Data. *Scandinavian Actuarial Journal*, 2014(2), pp. 180-187.

[71] Nascimento, F. F., Gamerman, D., Lopes, H. F. (2012). A Semiparametric Bayesian Approach to Extreme Value Estimation. *Statistics and Computing*, 22(2), pp. 661-675.

[72] Patiño, J. A. F. (2015). A Semi-Parametric Bayesian Extreme Value Model using a Dirichlet Process Mixture of Gamma Densities. *Journal of Applied Statistics*, 42(2), pp. 267-280.

[73] Patrik, G. (1980). Estimating Casualty Insurance Loss Amount Distributions. *Proceedings of the Casualty Actuarial Society*, 67, pp. 57-109.

[74] Peng, L. (1998). Asymptotically Unbiased Estimator for the Extreme-Value Index. *Statistics and Probability Letters*, 38(2), pp. 107-115.

[75] Peng, L., Welsh, A. H. (2001). Robust Estimation of the Generalized Pareto Distribution. *Extremes*, 4(1), pp. 53-65.

[76] Pickands, J. (1975). Statistical Inference using Extreme Order Statistics. *Annals of Statistics*, 3(1), pp. 119-131.

[77] Pigeon, M., Denuit, M. (2011). Composite Lognormal-Pareto Model with Random Threshold. *Scandinavian Actuarial Journal*, 2011(3), pp. 177-192.

[78] Punzo, A., Bagnato, L., Maruotti, A. (2018). Compound Unimodal Distributions for Insurance Losses. *Insurance: Mathematics and Economics*, 81, pp. 95-107.

[79] Ramberg, J. S., Schmeiser, B. W. (1972). An Approximate Method for Generating Symmetric Random Variables. *Communications of the ACM*, 15(11), pp. 987-990.

[80] Ramberg, J. S., Schmeiser, B. W. (1974). An Approximate Method for Generating Asymmetric Random Variables. *Communications of the ACM*, 17(2), pp. 78-82.

[81] Resnick, S. I. (1997). Discussion of the Danish Data on Large Fire Insurance Losses. *ASTIN Bulletin*, 27(1), pp. 139-151.

[82] Reynkens, T., Verbelen, R., Beirlant, J., Antonio, K. (2017). Modelling Censored Losses using Splicing: A Global Fit Strategy with Mixed Erlang and Extreme Value Distributions. *Insurance: Mathematics and Economics*, 77, pp. 65-77.

[83] Rootzén, H., Tajvidi, N. (1997). Extreme Value Statistics and Wind Storm Losses: A Case Study. *Scandinavian Actuarial Journal*, 1997(1), pp. 70-94.

[84] Schuster, E. F., Gregory, G. G. (1981). On the Nonconsistency of Maximum Likelihood Nonparametric Density Estimators. *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, Springer-Verlang, pp. 295-298.

[85] Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics*, 6(2), pp. 461-464.

[86] Scollnik, D. P. M. (2007). On Composite Lognormal-Pareto Models. *Scandinavian Actuarial Journal*, 2007(1), pp. 20-33.

[87] Scollnik, D. P. M., Sun, C. (2012). Modeling with Weibull-Pareto Models. *North American Actuarial Journal*, 16(2), pp. 260-272.

[88] Scott, D. W., Factor, L. E. (1981). Monte Carlo Study of Three Data-Based Nonparametric Probability Density Estimators. *Journal of the American Statistical Association*, 76, pp. 9-15.

[89] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Monographs on Statistics and Applied Probability, Chapman & Hall.

[90] Tancredi, A., Anderson, C., O'Hagan, A. (2006). Accounting for Threshold Uncertainty in Extreme Value Estimation. *Extremes*, 9, pp. 87-106.

[91] Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*, Springer Science.

[92] Verbelen, R., Gong, L., Antonio, K., Badescu, A., Lin, S. (2015). Fitting Mixtures of Erlangs to Censored and Truncated Data Using the EM Algorithm. *ASTIN Bulletin*, 45(3), pp. 729-758.

[93] Wand, M. P., Jones, M. C. (1995). *Kernel Smoothing*, Chapman & Hall/CRC.

[94] Wand, M. P., Marron, J. S., Ruppert, D. (1991). Transformations in Density Estimation. *Journal of the American Statistical Association*, 86, pp. 343-353.

[95] Wang, Y., Haff, I. H., Huseby, A. (2020). Modelling Extreme Claims via Composite Models and Threshold Selection Methods. *Insurance: Mathematics and Economics*, 91, pp. 257-268.

[96] Wiper, M., Insua, D. R., Ruggeri, F. (2001). Mixtures of Gamma Distributions with Applications. *Journal of Computational and Graphical Statistics*, 10(3), pp. 440-454.

[97] Wu, X. (2019). Robust Likelihood Cross-Validation for Kernel Density Estimation. *Journal of Business & Economic Statistics*, 37(4), pp. 761-770.

# A Table of Distribution Functions

**Gamma distribution: $\boldsymbol{\Gamma(\alpha, \beta)}$**

Parameters: $\alpha > 0$; $\beta > 0$

Distribution Function: $F(x) = --$

Density: $f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)}$; $x > 0$

Expected Value: $\mathbb{E}[X] = \alpha\beta$

Variance: $\text{Var}[X] = \alpha\beta^2$

Moment Generating Function: $M_X(r) = (1 - \beta r)^{-\alpha}$

**Weibull distribution: Wei $\boldsymbol{(\alpha, \beta)}$**

Parameters: $\alpha > 0$; $\beta > 0$

Distribution Function: $F(x) = 1 - e^{-\left(\frac{x}{\beta}\right)^a}$; $x > 0$

Density: $f(x) = \frac{\alpha}{\beta^\alpha} x^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^\alpha}$; $x > 0$

Expected Value: $\mathbb{E}[X] = \beta\Gamma(1 + \frac{1}{\alpha})$

Variance: $\text{Var}[X] = \beta^2[\Gamma(1 + \frac{2}{\alpha}) - \Gamma^2(1 + \frac{1}{\alpha})]$

Moment Generating Function: $M_X(r) = --$

**Pareto distribution: Pa $\boldsymbol{(\alpha, \theta)}$**

Parameters: $\alpha > 0$; $\theta > 0$

Distribution Function: $F(x) = 1 - \left(\frac{\theta}{x}\right)^\alpha$; $x > \theta$

Density: $f(x) = \frac{\alpha\theta^\alpha}{x^{\alpha+1}}$; $x > \theta$

Expected Value: $\mathbb{E}[X] = \frac{\alpha\theta}{\alpha-1}$; $\alpha > 1$

Variance: $\mathrm{Var}[X] = \frac{\theta^2\alpha}{(\alpha-1)^2(\alpha-2)}$; $\alpha > 2$

Moment Generating Function: $M_X(r) = --$

**Lognormal distribution: LN $(\boldsymbol{\mu}, \boldsymbol{\sigma^2})$**

Parameters: $\mu \in \mathbb{R}$; $\sigma^2 > 0$

Distribution Function: $F(x) = \Phi(\sigma^{-1}(\ln x - \mu))$; $x > 0$

Density: $f(x) = \frac{1}{x}\frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-(\ln x-\mu)^2}{2\sigma^2}}$; $x > 0$

Expected Value: $\mathbb{E}[X] = e^{\frac{\sigma^2}{2}+\mu}$

Variance: $\mathrm{Var}[X] = e^{\sigma^2+2\mu}(e^{\sigma^2} - 1)$

Moment Generating Function: $M_X(r) = --$

**Inverse Gaussian distribution: IG $(\boldsymbol{\mu}, \boldsymbol{\lambda})$**

Parameters: $\mu > 0$; $\lambda > 0$

Distribution Function:

$$F(x) = \Phi\left(\sqrt{\frac{\lambda}{x}}\left(\frac{x}{\mu} - 1\right)\right) + \exp\left(\frac{2\lambda}{\mu}\right)\Phi\left(-\sqrt{\frac{\lambda}{x}}\left(\frac{x}{\mu} + 1\right)\right);\ x > 0$$

Density: $f(x) = \sqrt{\frac{\lambda}{2\pi x^3}}\exp\left[\frac{-\lambda(x-\mu)^2}{2\mu^2 x}\right]$; $x > 0$

Expected Value: $\mathbb{E}[X] = \mu$

Variance: $\text{Var}[X] = \frac{\mu^3}{\lambda}$

Moment Generating Function: $M_x(r) = e^{\frac{\lambda}{\mu}}\left(1 - \sqrt{1 - \frac{2\mu^2 r}{\lambda}}\right)$; $r < \frac{\lambda}{2}$

**Inverse Gamma distribution: $\mathbf{I\Gamma(\alpha, \beta)}$**

Parameters: $\alpha > 0$; $\beta > 0$

Distribution Function: $F(x) = --$

Density: $f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\left(\frac{\beta}{x}\right)}$; $x > 0$

Expected Value: $\mathbb{E}[X] = \frac{\beta}{\alpha-1}$; $\alpha > 1$

Variance: $\text{Var}[X] = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$; $\alpha > 2$

Moment Generating Function: $M_x(r) = --$

**Burr Type XII distribution: BTXII $(\boldsymbol{\lambda, k, \alpha})$**

Parameters: $\lambda > 0$; $k > 0$; $\alpha > 0$

Distribution Function: $F(x) = 1 - \frac{1}{\left(1+\left(\frac{x}{\alpha}\right)^\lambda\right)^k}$; $x > 0$

Density: $f(x) = \frac{\frac{k\lambda}{\alpha}\left(\frac{x}{\alpha}\right)^{\lambda-1}}{\left(1+\left(\frac{x}{\alpha}\right)^\lambda\right)^{k+1}}$; $x > 0$

Expected Value: $\mathbb{E}[X] = \frac{\alpha\Gamma(1+\frac{1}{\lambda})\Gamma(k-\frac{1}{\lambda})}{\Gamma(k)}$

Variance: $\text{Var}[X] = \frac{\alpha^2\{\Gamma(k)\Gamma(1+\frac{2}{\lambda})\Gamma(k-\frac{2}{\lambda}) - \Gamma^2(1+\frac{1}{\lambda})\Gamma^2(k-\frac{1}{\lambda})\}}{\Gamma^2(k)}$

Moment Generating Function: $M_X(r) = --$

**Generalized Lambda distribution**: $\mathbf{GLD(\lambda_1, \lambda_2, \lambda_3, \lambda_4)}$

Parameters: $\lambda_1 > 0$; $\lambda_2 \in \mathbb{R}$; $\lambda_3 \in \mathbb{R}$; $\lambda_4 \in \mathbb{R}$

Distribution Function: $F(x) = --$

Density: $f(x) = \frac{\lambda_2}{\lambda_3 y^{\lambda_3 - 1} + \lambda_4 (1-y)^{\lambda_4 - 1}}$; $x = Q(y)$

$Q(x) = \lambda_1 + \frac{y^{\lambda_3} - (1-y)^{\lambda_4}}{\lambda_2}$; $0 \leq y \leq 1$

Expected Value: $\mathbb{E}[X] = \lambda_1 + \frac{1/(\lambda_3+1) - 1/(\lambda_4+1)}{\lambda_2}$

Variance:

$\mathrm{Var}[X] = \frac{\left[\frac{1}{(2\lambda_3+1)} - 2B(\lambda_3+1, \lambda_4+1) + \frac{1}{(2\lambda_4+1)}\right] - \left[\frac{1}{(\lambda_3+1)} - \frac{1}{(\lambda_4+1)}\right]^2}{\lambda_2^2}$

Moment Generating Function: $M_X(r) = --$

# B Transformed Kernel Density

Semiparametric transformed kernel density estimators were introduced in the statistics literature by Wand, Marron and Ruppert (1991). They demonstrate that the classical kernel density estimator can be improved by transforming the data set with a monotonic transformation. More specifically, transforming the data set overcomes the inconsistency problem related to the nonparametric kernel estimators when applied to heavy-tailed distributions. Bolancé, Guillén and Nielsen (2003) applied this semiparametric approach to model insurance losses in the actuarial context.

Let $\mathscr{X}_N = \{X_1, ..., X_N\}$ be a collection of *i.i.d.* random variables having density $f_X$ with support $S(f_X)$. We shall now define $G = \{\tilde{g}_\theta | \theta \in \Theta\}$ to be a set of twice continuously differentiable transformations that map $S(f_X)$ into the real line. The function $\tilde{g}_\theta(.)$ could be, for instance, a distribution function or some other monotonic increasing function.

Denote $X$ as the untransformed data and let $\tilde{Y} = \tilde{g}_\theta(X)$. The transformed random variable $Y$ is given by

$$Y = \frac{\sigma_X}{\sigma_{\tilde{Y}}} \tilde{Y} = g_\theta(X) \tag{B.1}$$

where $\sigma_X^2$ and $\sigma_{\tilde{Y}}^2$ are the variances of $X$ and $\tilde{Y}$, respectively. This definition ensures that the ultimate estimator is scale preserving.

The transformed density function can be written as

$$f_Y(y; \theta) = f_X\{g_\theta^{-1}(y)\}(g_\theta^{-1})'(y).$$

Thus, the kernel density estimator for the transformed density $f_Y$ is given by

$$\hat{f_Y}(y; \theta; \lambda) = N^{-1} \sum_{i=1}^{N} K_\lambda(y - Y_i). \tag{B.2}$$

Finally, the ultimate transformed kernel density estimator for the original data can be found by back-transformation

$$\hat{f}_X(x;\theta;\lambda) = N^{-1}\sum_{i=1}^{N} g_\theta^{'}(x)K_\lambda\{g_\theta(x) - g_\theta(X_i)\}. \tag{B.3}$$

Wand, Marron and Ruppert (1991) suggested the shifted power function as the transformation function $\tilde{g}_\theta(.)$. The shifted power function is written as

$$\tilde{g}_\theta(x) = \begin{cases} (x+\theta_1)^{\theta_2}; & \lambda_2 \neq 0 \\ \ln(x+\lambda_1); & \lambda_2 = 0 \end{cases}$$

where $\theta = (\theta_1, \theta_2)$ with $\theta_1 > -\min(X_1, ..., X_N)$ and $\theta_2 < 1$.

Clements, Hurn and Lindsay (2003) developed semiparametric transformed kernel estimators based on the Möibus-like transformation as an alternative to the power shifted transformation. Buch-Larsen *et al.* (2005) and Buch-Kromann (2006) extended their approach by using the modified Champernowne distribution function as the transformation function. It can be shown that the Möibus-like transformation is a special case of the Champernowne distribution.

The distribution function of the modified Champernowne distribution $T_{\alpha,M,c}$ is given by

$$T_{\alpha,M,c}(x) = \frac{(x+c)^\alpha - c^\alpha}{(x+c)^\alpha + (M+c)^\alpha - 2c^\alpha} \tag{B.4}$$

where $\alpha > 0$, $M > 0$ and $c \geq 0$.

To implement the transformed kernel density estimator using the modified Champernowne distribution as the parametric transformation function, we have the following steps:

**i**) Estimate the parameters $\theta = (\alpha, M, c)$ of the modified Champernowne distribution through the *quantile-mean* method, which select parameters in a way that emphasizes the goodness of fit in the right tail. The method assumes that $M$ is equal to the sample (empirical) median and $\alpha$ is selected so that the 95 quantile point of the empirical *cdf* and the estimated distribution are equal. The parameter $c$ is then chosen so that the mean of the estimated distribution is as close as possible to the empirical mean.

**ii**) Transform the data set $X_1, ..., X_N$ into $\tilde{Y}_1, ..., \tilde{Y}_N$ using $\tilde{g}_\theta = \hat{T}_{\hat{\alpha},\hat{M},\hat{c}}$. Calculate the transformed variable $Y$ using the rescaling constant $\frac{\sigma_X}{\sigma_{\tilde{Y}}}$.

**iii**) Assuming that the unknown density $f_X$ is close to a modified Champernowne distribution, the transformed data set will be approximately uniformly distributed. Under the assumption that the transformed distribution is close to a Uniform in the interval $(0, 1)$, we can apply the boundary corrected kernel density estimator to the transformed data $Y_1, ..., Y_N$

$$\hat{f}_Y(y; \theta; \lambda) = \frac{1}{N k_\lambda(y)} \sum_{i=1}^{N} K_\lambda(y - Y_i)$$

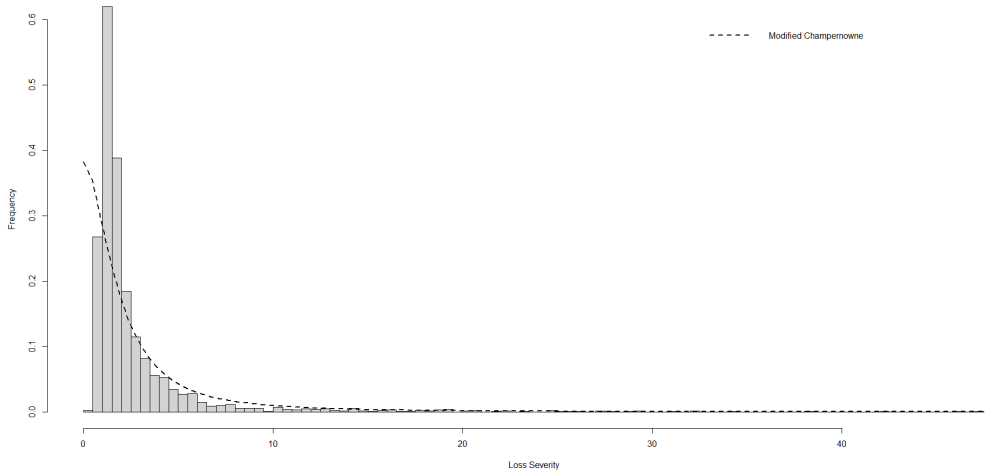
where $k_\lambda(.)$ is the boundary correction, which is needed because $Y$ is constrained to the interval $(0, 1)$

$$k_\lambda(y) = \int_{\max(-1, -\frac{y}{\lambda})}^{\min(1, \frac{1-y}{\lambda})} K(z) dz$$

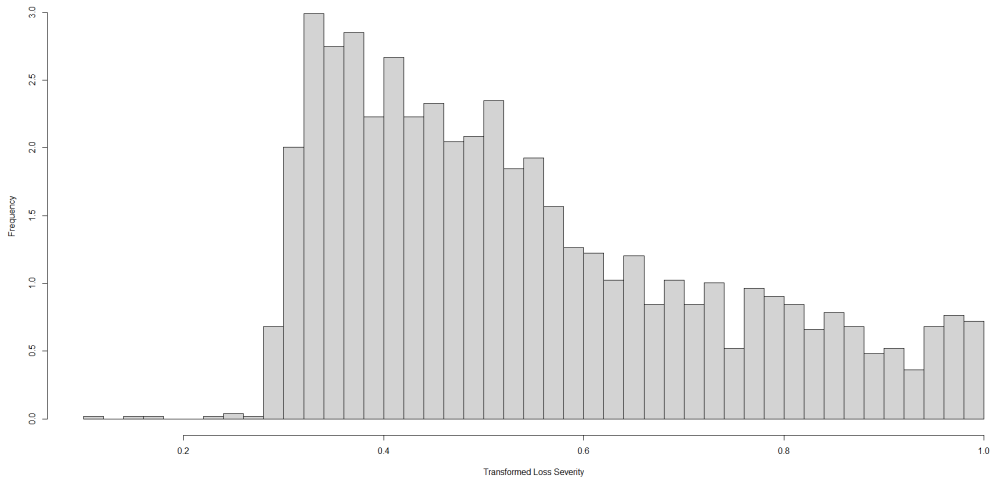
**iv**) The final transformed kernel density estimator for the original data set, $X_1, ..., X_N$, is obtained by a backward transformation such that

$$\hat{f}_X(x; \theta; \lambda) = \hat{f}_Y(y; \theta; \lambda) \hat{T}'_{(\hat{\alpha},\hat{M},\hat{c})}(x)$$

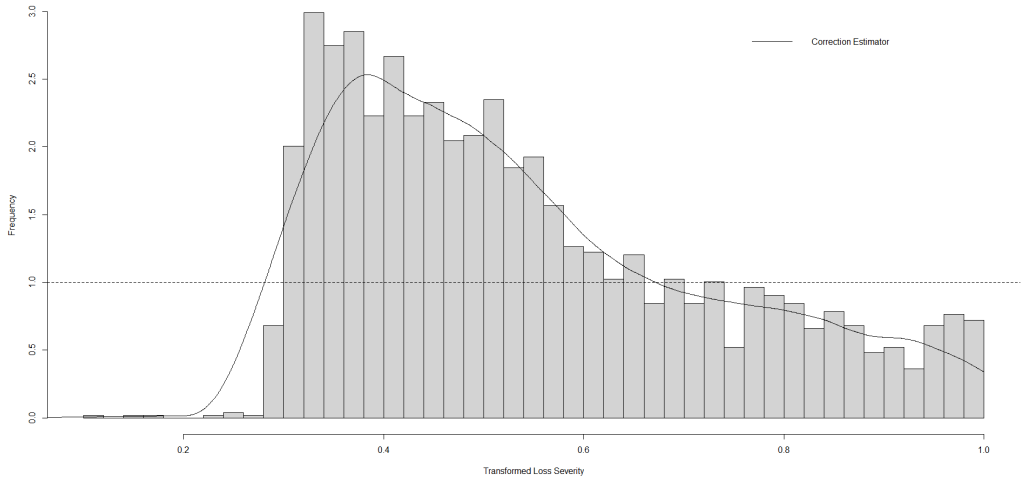
The figure below illustrates these four steps for the Danish Fire Insurance data. The Epanechnikov kernel function was used in the kernel density estimator and the bandwidth parameter $\lambda$ was chosen by the simple normal scale bandwidth selection method.



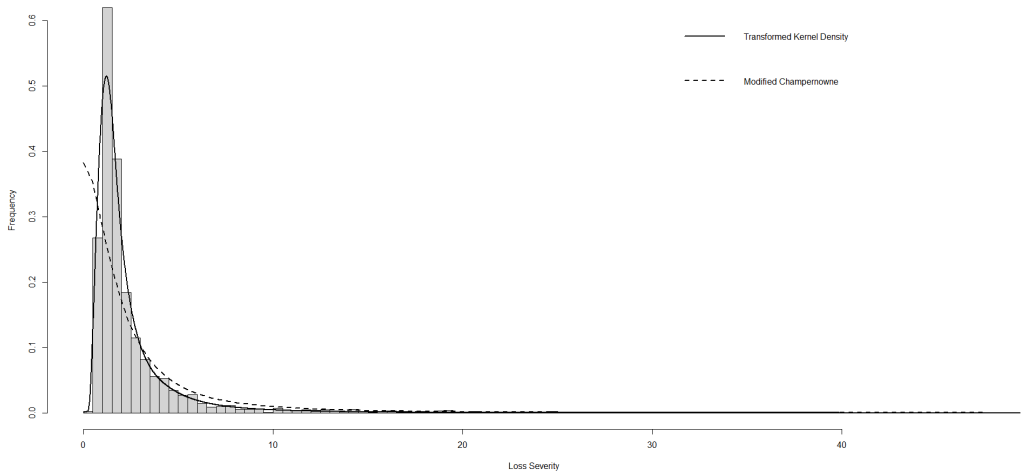

**(a)** Modified Champernowne




**(b)** Transformed Losses




**(c)** Correction Estimator




**(d)** Transformed Kernel Density

**Figure 13.** Transformed Kernel Density. In panel (a) we have the modified Champernowne density superimposed to the empirical histogram. Panel (b) shows the histogram of the transformed data and panel (c) gives the nonparametric correction estimator. Panel (d) illustrates the ultimate transformed kernel density estimator.

As presented in Figure 13, panel (a) shows the histogram for the Danish Fire Insurance data and the estimated modified Champernowne distribution with the quantile-mean (QM) parameters (dashed line). Panel (b) shows the histogram for the transformed data using the estimated modified Champernowne distribution function $\hat{T}_{(\hat{\alpha},\hat{M},\hat{c})}$ as the parametric transformation function.

In panel (c), the nonparametric kernel correction estimator is illustrated. When the correction estimator is below 1 (horizontal dashed line), the final transformed kernel density for $f_X$ will be lower at that point than the density of the estimated modified Champernowne distribution. On the other hand, when the correction estimator is above 1, the final transformed kernel density estimator will have more density at that point than the estimated modified Champernowne distribution.

Finally, in panel (d) we show the transformed kernel density estimator (solid line) for the Danish Fire Insurance data set. The final semiparametric estimator seems to provide a significant superior fit to the data set than the uncorrected modified Champernowne distribution (dashed line).