Large loss models for general insurance

Tine Buch-Kromann

May 29, 2009

Preface

This thesis has been prepared in partial fulfillment of the requirements for the Ph.D degree at the Department of Mathematical Sciences, University of Copenhagen, Denmark. The work has been carried out in the period from December 2005 to April 2009 under the supervision of Professor Jens Perch Nielsen from Cass Business School and Festina Lente and Professor Thomas Mikosch from the University of Copenhagen.

Each chapter in this thesis corresponds to an article. This means that the chapters are self-contained and can be read independently of the other chapters, even though they are connected, and the complexity of the models increases. This also means that there are notational discrepancies among the chapters, and that repetitions of arguments and theory sometimes occur. The introduction surveys the theory and gives an overview of the papers. Published articles are unchanged except for the references, which have been merged and placed at the end of the thesis. This has the consequence that bibliographical repetitions are avoided, and that the thesis bibliography includes the published form of some references which were referred to as unpublished, submitted or forthcoming in the original article. The thesis also uses a uniform style for the enumeration and the graphical layout of tables and figures. The articles are placed in their logical order rather than in the order they were written or published.

Acknowledgements

First, I would like to thank Professor Jens Perch Nielsen for his encouragement and supervision during the last three years. I also wish to thank Professor Thomas Mikosch for his support and comments on early versions of several papers in the thesis. Also thanks to Codan Insurance for their financial support to write this thesis. Special thanks go to Per Jensen (head of Commercial Pricing, Codan), Bjørn Lunding Sandqvist (former head of Commercial Business Intelligence, Codan) and all my colleagues in Commercial Pricing for their help and support especially concerning data.

At a personal level, I would like to thank my husband Matthias Buch-Kromann for valuable academic discussions, critical questions to my work, and programming support.

Copenhagen, May 2009
Tine Buch-Kromann

Summary

This thesis concerns estimation of the large loss risk of an insurance company. Traditional large loss models in insurance companies are based on extreme value theory and the generalized Pareto distribution (GPD), which is the limit distribution of excesses above a high threshold. In this thesis, we investigate an alternative approach to the problem based on nonparametric statistics.

In the first part of the thesis, we introduce the basic nonparametric kernel density estimator in one dimension, which we improve significantly for heavy-tailed data by applying a Champernowne transformation to the data set. The model is applicable to heavy-tailed insurance losses, but the method is also a reasonable choice in the related field of operational risk estimation. We provide an adaptation of the method to operational risk which takes into account underreporting (i.e., the fact that not all claims are reported), and demonstrate the stabilizing effect of the method compared to pure parametric models.

Comparison of the Champernowne transformed kernel density estimator to similar nonparametric methods demonstrates that the method has superior tail performance especially for heavy-tailed distributions. However, when focusing solely on the tail and comparing with an estimated GPD, improvements can be obtained by estimating the underlying Champernowne distribution in a way that emphasizes the tail. We propose such a method and show that it obtains comparable or superior tail performance when compared to the GPD and g-and-h distributions.

In the last two chapters, the focus is on multivariate large loss estimation. The

Champernowne transformed kernel density estimator is extended to a multivariate version, and multiplicative bias correction is introduced in order to include prior knowledge which imposes some structure on the nonparametric estimation problem; this is desirable for tail estimation in higher dimensions. The improvements gained by introducing structure by means of multiplicative bias correction are demonstrated in two bivariate simulation studies, one study which exclusively concerns the pure multiplicative bias correction, and one which additionally includes the tail-flattening transformation approach.

In the last chapter, the multivariate density estimation is extended to handle truncated and censored data. The general idea of using nonparametric statistics is retained, but the model is now expressed in terms of survival analysis due to the ability of the theory to account for exposure. As in the previous chapter, we modify the model by introducing multiplicative bias correction and tail-flattening transformation, but now within the framework of survival analysis.

Resumé

Denne afhandling omhandler estimation af storskaderisiko i et forsikringsselskab. Traditionelle storskademodeller i forsikringsselskaber er baseret på ekstremværditeori og den generaliserede Pareto fordeling (GPD), som er grænsefordelingen for overskridelser over en høj tærskelværdi. I denne afhandling undersøger vi en alternativ tilgang til problemet, som er funderet i ikke-parametrisk statistik.

I første del af afhandlingen introduceres den basale ikke-parametriske kernetæthedsestimator i én dimension, som signifikant forbedres, specielt for tunghalede data, ved at Champernowne-transformere datasættet. Modellen er anvendelig for tunghalede forsikringsskader, men også i det beslægtede område, operationel risiko, er modellen rimelig. En version af metoden tilpasset til operationel risiko, som specielt tager underrapportering i betragtning (dvs. den kendsgerning, at ikke alle skader rapporteres i forbindelse med operationel risiko), optilles, og metodens stabiliserende effekt sammenlignet med rene parametriske modeller illustreres.

Sammenligning af den Champernowne-transformerede kernetæthedsestimator med tilsvarende ikke-parametriske metoder viser at metoden fitter halen bedre specielt for tunghalede fordelinger. Hvis man imidlertid udelukkende fokuserer på halen og sammenligner med en estimeret GPD, viser der sig forbedringsmuligheder, hvis den underliggende Champernowne fordeling estimeres med en metode, som lægger særlig vægt på haleestimationen. En sådan metode demonstres, og vi viser at den giver et sammenligneligt eller bedre halefit end GPD og g-og-h fordelingerne.

I de to sidste kapitler ligger fokus på flerdimensional storskadeestimation. Den

Champernowne-transformerede kernetæthedsestimator udvides til en flerdimensionel version, og multiplikativ biaskorrektion introduceres for at kunne inkludere en apriori tæthed og derved introducere struktur til det ikke-parametriske estimationsproblem, hvilket er hensigtsmæssigt når fokus ligger på haleestimation i flere dimensioner. Forbedringerne som opnås ved multiplikativ biaskorrectkon er demonstreret i to todimensionale simulationsstudier, ét studie som udelukkende drejer sig om den rene multiplikative biaskorrektion, og ét som også indeholder den haleudglattende transformationsmetode.

I sidste kapitel udvides den flerdimensionale tæthedsestimator til at kunne håndtere trunkerede og censorerede data. Hovedidéen om at anvende ikke-parametrisk statistik bibeholdes, men modellen baserer sig nu på overlevelsesanalyse, på grund at denne teoris evne til at tage hensyn til eksponering. Som i det foregående kapitel modificerer vi modellen ved introduktion af multiplikativ biaskorrektion og haleudglattende transformation for tæthedsestimation baseret på overlevelsesanalyse, men nu inden for rammerne af overlevelsesanalyse.

Contents

Chapter 1

Introduction

This thesis is concerned with the estimation of an insurance company's exposure to large loss risk. The insurance company's earnings and profits are directly determined by the company's ability to determine premium rates correctly: too low premiums are unprofitable, and too high premiums result in the loss of potentially profitable customers. Large loss risk is a substantial component of the premium rates. Although large losses are infrequent, they typically constitute more than half of the total claims expenses in a portfolio, so their impact on the company's performance is substantial. The problem of determining the large loss risk is complicated by the fact that large losses are difficult to predict due to sparse data.

1.1 Estimation of large losses

There are several ways to estimate large losses. Classical extreme value theory studies the asymptotic behaviour of maxima and excesses above a high threshold. These results are approximatively correct for high quantiles under certain conditions on the tail of the underlying distribution. An alternative approach is provided by nonparametric methods which are the statistical foundation of this thesis. Nonparametric

methods produce density estimators on the entire axis, and can be constructed with different degrees of emphasis on the tail. In the following, we give an introduction to the two main approaches to large loss estimation in sections 1.1.1 and 1.1.2.

1.1.1 Extreme value theory

Extreme value theory (EVT) considers the distributional properties of maxima and exceesses above a high threshold. The theory is described in a vast amount of books and papers; see e.g. Leadbetter et al. (1983), Embrechts et al. (1997) and Kotz and Nadarajah (2000). In the following, we give a short and non-exhaustive introduction to the theory, which is primarily based on Embrechts et al. (1997).

The fundamental Fisher-Tippett theorem describes the limit laws for maxima of iid stochastic variables $X_i, i = 1, 2, ...$, and says, that if there exist constants $c_n > 0$ and $d_n \in \mathbb{R}$ for a sequence of iid random variables (X_n) such that

$$c_n^{-1}(M_n - d_n) \xrightarrow{d} H, \qquad n \to \infty, \tag{1.1}$$

for some non-degenerate distribution H, where $M_n = \max(X_1, ..., X_n)$, then H must be one of the three extreme value distributions:

$$\begin{aligned}
&\text{Fréchet :} && \Phi_\alpha(x) = \begin{cases} 0, & x \le 0, \\ \exp\{-x^{-\alpha}\}, & x > 0, \end{cases} && \alpha > 0. \\
&\text{Weibull :} && \Psi_\alpha(x) = \begin{cases} \exp\{-(-x)^{-\alpha}\}, & x \le 0, \\ 1, & x > 0, \end{cases} && \alpha > 0. \\
&\text{Gumbel :} && \Lambda(x) = \exp\{\exp(-x)\}, && x \in \mathbb{R}
\end{aligned}$$

The distribution F of X_i is said to be in the maximum domain of attraction of H, written as $F \in \text{MDA}(H)$, if (1.1) holds. This property of the sequence of partial maxima corresponds to the central limit theorem for properly normalized sums which converge to a normal distribution or infinite variance stable distribution. Notice that the Fréchet distribution is much more heavy-tailed than the Gumbel distribution.

By introducing a shape parameter ξ, the type of the three standard extreme value distributions can be represented by the standard generalized extreme value distribution (GEV) defined by

$$H_\xi = \begin{cases} \exp\{-(1+\xi x)^{-1/\xi}\}, & \xi \neq 0, \\ \exp\{-\exp\{-x\}\}, & \xi = 0, \end{cases} \tag{1.2}$$

with the condition that $1 + \xi x > 0$, which corresponds to $x > \xi^{-1}$ when $\xi > 0$, to $x < \xi^{-1}$ when $\xi < 0$, and to $x \in \mathbb{R}$ when $\xi = 0$. By introducing a location parameter $\mu \in \mathbb{R}$ and a scale parameter $\psi > 0$, a three-parameter family is obtained by defining $H_{\xi,\mu,\psi}(x) = H_\xi((x-\mu)/\psi)$, with a corresponding adjustment of the support.

The limit distribution of excesses above a high threshold, called the generalized Pareto distribution (GPD), is closely related to the limit distribution of maxima. The generalized Pareto distribution is defined as

$$G_\xi(x) = \begin{cases} 1 - (1+\xi x)^{-1/\xi}, & \xi \neq 0, \\ 1 - \exp\{-x\}, & \xi = 0, \end{cases}$$

where $x \geq 0$ for $\xi \geq 0$, and $0 \leq x \leq -1/\xi$ for $\xi < 0$. As for the GEV, location and scale parameters can be introduced.

When using EVT in practice, it is common to assume that the data are iid with a cdf belonging to the maximum domain of attraction of an extreme value distribution, i.e. $F \in \mathrm{MDA}(H_\xi)$. There exist various estimators of the shape parameter, including the Pickands, the Hill and the Deckers-Einmahl-de Haan estimators; see Embrechts et al. (1997), which all depend on graphically based decisions about the choice of threshold. The Peaks Over Threshold (POT) method deals with the problem of estimating a GPD for excesses above a sufficiently high threshold by estimating the parameters, e.g. by means of maximum likelihood estimation or probability-weighted moments. However, in all these procedures, it is crucial to determine a suitable threshold, i.e. to determine from which point in the tail the limit assumption is reasonable. This problem is a classical bias-variance trade-off: choosing the threshold too low

means that the assumption about the tail is inappropriate, whereas choosing the threshold too high means that we have too few data points to reasonably estimate the parameters of the distribution. This problem is often solved by graphical methods, e.g. by choosing a threshold from which on the mean excess function is approximately linear, but automatic approaches also exist, e.g. Dupuis (1999); Cebrián et al. (2003).

EVT is widely used in the literature on insurance loss estimation; see McNeil (1997), McNeil and Saladin (1997), Cebrián et al. (2003), and Sanders (2005) for applications of GPD models of excesses; Chavez-Demoulin and Embrechts (2004) and Chavez-Demoulin and Davison (2005) for time dependent GPD models of excessess by use of smoothing methods; and Corradin (2002) for applications of GPD models in the context of reinsurance. A number of papers recommend modelling the full data set rather than only the tail by using mixture models which combine the GPD distribution with more light-tailed distributions, e.g. Weibull or lognormal distributions; see Frigessi et al. (2002), Knecht and Küttel (2003), and Cooray and Ananda (2005).

1.1.2 Kernel smoothing

Classical kernel smoothing is a simple and intuitive method which produces a density estimator of a data set without any parametric assumptions; see e.g. Silverman (1986), Wand and Jones (1995), and Härdle et al. (2004) for comprehensive introductions. For a data set $X_1, ..., X_n$, the univariate kernel density estimator has the form

$$\widehat{f}_b(x) = \frac{1}{n} \sum_{i=1}^{n} K_b\left(x - X_i\right) \tag{1.3}$$

where $K_b(u) = b^{-1}K(u/b)$ is a nonnegative kernel function which satisfies $\int K(x)\,\mathrm{d}x = 1$ and is symmetric about the origin with finite fourth moment, and where $b > 0$ is a bandwidth which determines the degree of smoothing. The bias and variance are

given by

$$\mathbb{E}\{\widehat{f}_b(x)\} - f(x) \simeq \frac{1}{2}b^2\mu_2(K)f''(x), \tag{1.4}$$

$$\mathbb{V}\{\widehat{f}_b(x)\} \simeq \frac{1}{nb}||K||_2^2 f(x), \tag{1.5}$$

where $\mu_2(K) = \int s^2 K(s)\,\mathrm{d}s$ is the second moment, and $||K||_2^2 = \int K^2(s)\,\mathrm{d}s$ is the squared L_2 norm of K. Notice that the bias is small when b is small, whereas the variance is small when b is large.

An optimal bandwidth can be obtained by optimizing the mean integrated squared error, $\mathrm{MISE}(\widehat{f}_b) = \int \mathbb{E}\left[\{\widehat{f}_b(x) - f(x)\}^2\right]\,\mathrm{d}x$, with the approximate formula

$$\mathrm{AMISE}\left(\widehat{f}_b\right) = \frac{1}{nb}||K||_2^2 + \frac{1}{4}b^4\left\{\mu_2(K)\right\}^2||f''||_2^2, \tag{1.6}$$

which ignores higher order terms. The resulting optimal bandwidth is given by

$$b_{\mathrm{opt}} = \left(\frac{||K||_2^2}{||f''||_2^2\{\mu_2(K)\}^2 n}\right)^{1/5} \sim n^{-1/5}. \tag{1.7}$$

By inserting the optimal bandwidth in (1.6), the optimal rate of convergence can be obtained by

$$\mathrm{AMISE}\left(\widehat{f}_{b_{\mathrm{opt}}}\right) = \frac{5}{4}\left(||K||_2^2\right)^{4/5}\left(\mu_2(K)||f''||_2\right)^{2/5} n^{-4/5} \sim n^{-4/5}.$$

However, the optimal bandwidth (1.7) cannot be calculated directly because $f(x)$ is unknown. Various methods of bandwidth selection have been proposed in the literature. In "Rule-of-Thumb" methods, $f(x)$ in (1.7) is replaced by a parametric density estimator, e.g. the normal distribution in Silverman (1986). "Least squares cross-validation" methods minimize an approximation to the MISE, defined before, by use of a "leave-one-out" density estimator; see e.g. Wand and Jones (1995). More sophisticated bandwidth selection methods, known as "second generation" methods in Jones et al. (1996), seem to be superior with respect to their theoretical and

practical performance, but are computationally more demanding. One example of a "second generation" method is the Sheather and Jones bandwidth selection method; see Sheather and Jones (1991) and Wand and Jones (1995), which is based on a kenel estimator of $||f''||_2^2$. This estimator needs a bandwidth as well, and the procedure can be repeated. A pilot bandwidth is however needed, e.g. estimated by use of a "Rule-of-Thumb" method.

There are various ways in which the basic kernel density estimator (1.3) can be improved. Wand et al. (1991) proposed to transform the data set with a shifted power transformation and estimate the kernel density estimator on the transformed data set. The density of the original data set is then obtained by back-transformation. Similarly, Bolancé et al. (2003) used the shifted power transformation family in an insurance context, but with an alternative parameter estimation method that minimized an approximation of the mean integrated squared error. Other work in this area includes Yang and Marron (1999), who proposed the Johnson Family Transformation, and Clements et al. (2003), who recommeded a Möbius-like mapping.

The transformed kernel density estimator has the form

$$\widetilde{f}_b(x) = \frac{1}{n\, k_{T(x)}} \sum_{i=1}^{n} K_b \left\{ T(x) - T(X_i) \right\} T'(x), \tag{1.8}$$

where $T(x)$ is the transformation function and k_u is a boundary correction which is required if the transformed data belong to a compact interval. The transformed kernel density estimator resembles a classical kernel density estimator with variable bandwidth, since a constant bandwidth on the transformed axis corresponds to an increasing bandwidth on the original axis when the transformation function has compact support. This is always the case if the transformation function is a cdf; see Bolancé et al. (2003). If T belongs to a parametric class with a square-root-n consistent estimator, then the following bias and variance expressions appear; see

Buch-Larsen et al. (2005):

$$\mathbb{E}\{\tilde{f}_b(x)\} - f(x) \simeq \frac{1}{2}b^2\mu_2(K)\left[\left\{\frac{f(x)}{T'(x)}\right\}'\frac{1}{T'(x)}\right]', \tag{1.9}$$

$$\mathbb{V}\{\tilde{f}_b(x)\} \simeq \frac{1}{nb}||K||_2^2 T'(x)f(x). \tag{1.10}$$

By comparing the bias and variance terms of the univariate kernel density estimator and the transformation kernel density estimator, i.e. by comparing (1.4) with (1.9) and (1.5) with (1.10), we note that the bias term $f''(x)$ in (1.4) is replaced by $\left[\left\{\frac{f(x)}{T(x)}\right\}'\frac{1}{T'(x)}\right]'$ in (1.9). This means that the transformation kernel density estimator (1.8) has a superior bias compared to the univariate kernel density estimator (1.3) if the transformation function is close to the true cdf. We also note that the variance in (1.10) is multiplied by $T'(x)$, which means that the variance is superior to (1.5) when $T'(x) < 1$, which is likely to occur in the tail provided the transformation is well chosen.

As mentioned above, boundary correction is needed to ensure a consistent kernel density estimator if the support of the data set is compact. Simple boundary corrections — e.g. renormalization of each kernel such that it integrates to 1, or reflection, which reinstates the "missing mass" by reflecting the estimate in the boundary — ensure a consistent estimator. But the bias is of order $O(b)$ near the boundary, which means that the rate of convergence is $n^{-2/3}$ at the boundary compared to $n^{-4/5}$ elsewhere; see Jones (1993). Jones (1993) also describes more sophisticated boundary corrections with boundary bias of order $O(b^2)$. All of these methods are based on linear combinations of two kernel functions. Of particular interest is the local linear kernel function, which is a linear combination of $K(x)$ and $xK(x)$, and which corresponds to local linear fitting in nonparametric regression. One disadvantage of the $O(b^2)$ boundary corrections mentioned in Jones (1993) is the propensity for taking negative values near the boundary. In Jones and Foster (1996), this drawback is removed by introducing an estimator based on a linear combination of a renormalized kernel density and a local linear kernel density. This modification cre-

ates a non-negative estimator with a performance that matches the performance of the much more complicated boundary correction proposed in Marron and Ruppert (1994). Marron and Ruppert introduced a boundary correction based on a transformation which ensures that the transformed density has first derivative equal to 0 at the boundary, combined with a reflection boundary correction. Unfortunately, the estimator in Jones and Foster (1996) does not necessarily integrate to 1. Zhang et al. (1999) followed the idea in Marron and Ruppert (1994) and proposed a generalized reflection technique based on a transformation that depends on a pilot estimator for the logarithmic derivative of the density at the boundary. This idea was further improved in recent work of Karunamuni and Zhang (2008) by introducing a different (smaller) bandwidth of the transformation. Asymmetric kernel functions are another way to address boundary bias due to compact support of the density function. Chen (1999) proposed beta kernels which produce non-negative estimates free of boundary bias. The support of the beta kernels can be matched to the compact support of the density function, and therefore the estimator has smaller finite-sample variance. In Chen (2000), the idea is extended to situations, where data are bounded only at one end. Here, a gamma kernel is proposed to obtain a non-negative estimator free of boundary bias. Moreover, Scaillet (2004) addresses the problem of estimating densities on the non-negative real line based on inverse Gaussian and reciprocal inverse Gaussian kernel functions.

Various authors have worked on improving bias without aggravating variance, thereby improving the rate of convergence. One method is to introduce higher-order kernels; see Wand and Jones (1995), by relaxing the restriction that the kernel function has to be a density function. By constructing a kernel function with second moment equal to zero, $\mu_2(K) = 0$, the bias is reduced to order $o(h^4)$, and the optimal rate of convergence is then of order $n^{-8/9}$. As mentioned in Jones et al. (1995), higher-order kernels can be interpreted as an additive bias reduction. However, in practice the improvements that one obtains by using higher-order kernels only appear in very large data sets. Moreover, since the support of a higher-order kernel function includes negative values, the resulting density estimate is not necessarily non-negative

and therefore not a density itself. Multiplicative bias correction is an alternative way to obtain an improved rate of convergence. A multiplicative bias corrected kernel density estimator has the form

$$\bar{f}(x) = g(x)\frac{1}{n}\sum_{i=1}^{n} g(X_i)^{-1}K_b(x - X_i). \tag{1.11}$$

The fundamental idea is to let the prior density $g(x)$ capture some of the shape of the true density and then estimate a nonparametric correction which will be smoother than the original density, provided $g(x)$ is not too far away from the true density. The asymptotic properties of the multiplicative bias correction are (see Hjort and Glad (1995))

$$\mathbb{E}\{\bar{f}_b(x)\} - f(x) \simeq \frac{1}{2}b^2\mu_2(K)g(x)r''(x), \tag{1.12}$$

$$\mathbb{V}\{\bar{f}_b(x)\} \simeq \frac{1}{nb}||K||_2^2 f(x), \tag{1.13}$$

where $r(x) = \frac{f(x)}{g(x)}$. By comparing the bias and variance terms of the univariate kernel density estimator and the multiplicative bias corrected kernel density estimator, i.e. by comparing (1.4) with (1.12) and (1.5) with (1.13), we note that, wheareas the variances are identical, the bias depends on the curvature of $r(x)$ in (1.12), which is small when $g(x)$ is close to $f(x)$. Bias improvements are therefore obtained when the prior knowledge density is close to the true density. Hjort and Glad (1995) proposed a multiplicative bias corrected estimator based on a parametric start, and Jones et al. (1995) proposed a multiplicative bias corrected estimator based on a purely non-parametric start. In Jones et al. (1999), the two methods are combined. Hagmann and Scaillet (2007) investigate an estimator based on local multiplicative bias correction and propose an asymmetric gamma kernel in order to address the bounded support and improve the estimation performance. The approach is extended in Gustafsson et al. (2009) by transforming the data to $[0, 1]$ in order to address the heavy-tailedness of the data.

Survival analysis is a topic of interest in the broad literature of nonparametric statis-

tics and kernel smoothing, but is at a first glance only superficially related to large loss modelling. However, as demonstrated in the last chapter in this thesis, survival analysis has turned out to be useful when dealing with truncated and censored data. The following is a short and informal introduction to survival analysis following Martinussen and Scheike (2006); see Andersen et al. (1993) and Martinussen and Scheike (2006) for comprehensive introductions.

Let T^* be a survival time and C be a censoring time independent of T^*. Then the observed survival time is $T = T^* \wedge C$. The variable $D = I(T^* \leq C)$ indicates whether censoring has occurred. We define the "at-risk" indicator as $Y(t) = I(t \leq T)$, which is 1 at time t if neither the event nor the censoring have occured at time t. Let $N(t) = I(T \leq t)$ be a counting process which is 0 until t passes T and 1 thereafter. The quantity $N(t)$ can be decomposed into a compensator (model part) and a martingale (random noise),

$$N(t) = \Lambda(t) + M(t).$$

Assume that T^* has a density f, and let $S(t) = P(T^* > t)$ be the survival function. Then the hazard function is defined as

$$\alpha(t) = \frac{f(t)}{S(t)} = \lim_{h \downarrow 0} \frac{1}{h} P(t \leq T^* < t + h | T^* \geq t).$$

The hazard function defines the distribution uniquely as

$$S(t) = \exp\left\{ - \int_0^t \alpha(s) \, \mathrm{d}s \right\} = \exp\left\{ -A(t) \right\}.$$

We assume the compensator can be written in the form

$$\Lambda(t) = \int_0^t \lambda(s) \, \mathrm{d}s,$$

where $\lambda(s)$ is the intensity process. Let $(T_i, D_i), i = 1, ..., n$ be n independent observations, and $N_i(t) = I(T_i \leq t, D_i = 1)$ and $Y_i(t) = I(t \leq T_i)$ correspond to

the i'th observation. Moreover, define $N(t) = \sum_{i=1}^{n} N_i(t), Y(t) = \sum_{i=1}^{n} Y_i(t)$ and $M(t) = N(t) - \int_0^t Y(s)\alpha(s)\,\mathrm{d}s$. In a nonparametric setup, the cumulative hazard function $A(t)$ can be estimated by means of the Nelson-Aalen estimator

$$\widehat{A}(t) = \int_0^t \frac{J(s)}{Y(s)}\,\mathrm{d}s,$$

where $J(s) = I(Y(s) > 0)$, and the survival function $S(t)$ can be estimated by means of the Kaplan-Meier estimator

$$\widehat{S}(t) = \prod_{s \le t} \left(1 - \Delta\widehat{A}(s)\right) = \prod_{s \le t} \left(1 - \frac{\Delta N(s)}{Y(s)}\right).$$

Estimation of the hazard function in a nonparametric setting has been worked out both in internal and external versions. Beran (1981) and Dabrowska (1987) studied the situation in which there is a time independent covariate. McKeague and Utikal (1990) extended this approach to a model with time dependent covariates, a model which was further developed by Van Keilegom and Veraverbeke (2001). These hazard estimators are so-called internal local constant estimators. In a situation where there are no covariates, they have the form (Ramlau-Hansen (1983))

$$\widehat{\alpha} = \sum_{i=1}^{n} \int K_b(t-s)\frac{1}{Y(s)}\,\mathrm{d}N_i(s).$$

Alternatively, Nielsen and Linton (1995) proposed an external local constant estimator

$$\widetilde{\alpha} = \frac{\sum_{i=1}^{n} \int K_b(t-s)\,\mathrm{d}N_i(s)}{\sum_{i=1}^{n} \int K_b(t-s)Y(s)\,\mathrm{d}s}.$$

In Li and Doss (1995) and Nielsen (1998), the local constant estimators are extended to local linear estimators with superior boundary bias, and in Nielsen and Tanggaard (2001) further extended with a weight function, which enables the authors to identify the Ramlau-Hansen estimator as an instance of one particular weighting scheme; however, the authors argue for an alternative weighting scheme which is less sensitive

to volatile exposure patterns. Moreover, Nielsen and Tanggaard (2001) consider multiplicative as well as additive bias correction and argue for additive bias correction due to its superior theoretical bias and simulation results. In Nielsen et al. (2009), the corresponding local constant and local linear density estimators for filtered data are derived. The local constant density estimator with the so-called "unit" or "natural" weighting, which is the recommended weighting in Nielsen et al. (2009), is similar to the local constant hazard estimator, but with the multiplication of the survival function in the numerator:

$$\widehat{f}(t) = \frac{\sum_{i=1}^{n} \int_0^\infty K_b(t-s) Y_i(s) \hat{S}(s) dN_i(s)}{\int_0^\infty K_b(t-s) Y^{(n)}(s) ds}.$$

Nielsen et al. (2009) studied multiplicative and additive bias correction in the context of filtered data density estimators and compared them in an extensive simulation study. They showed that the multiplicative bias correction is superior to the additive one, in contrast with the results for hazard estimation in Nielsen and Tanggaard (2001).

When the explanatory variables are multidimensional, the purely nonparametric models suffer from a poor rate of convergence, and therefore it is preferable to assume some stucture, e.g. additive or multiplicative regression models; see Hastie and Tibshirani (1990). In the two dimensional case, the models have the following form:

$$f(x,z) = f_1(x) + f_2(z) \qquad f(x,z) = f_1(x) f_2(z).$$

Linton and Nielsen (1995) proposed an alternative kernel procedure that can be used for estimation in both additive and multiplicative regression; the procedure is based on marginal integration and is extended into higher dimensions in Linton et al. (2003).

1.2 Overview and contributions of the thesis

The aim of the thesis is to develop large loss models within the framework of nonparametric statistics. In chapter 2, we present a one-dimensional density estimator based on a parametric distribution and a nonparametric correction, where the parametric distribution is estimated by maximum likelihood. This estimator is the starting point for all the subsequent models presented in the thesis. In chapter 3, the one-dimensional density estimator is adapted to operational risk and the special problem of underreporting (i.e., the problem that not all claims are reported). In chapter 4, we present a case study that suggests that the maximum likelihood estimation procedure for the parametric start might be inappropriate if we are mostly interested in the tail. In chapter 5, we extend this idea and propose an alternative parameter estimation procedure which emphasizes the tail. The performance of our corrected parametric estimator with maximum likelihood and tail emphasizing parameters is compared with the generalized Pareto distribution and the g-and-h distribution, which has received special attention in operational risk in recent years; see e.g. Dutta and Perry (2006). The last part of the thesis concerns multivariate estimators. In chapter 6, we describe how to extend the one-dimensional method to two dimensions and introduce bias correction based on prior knowledge. In chapter 7, the multivariate model is extended to truncated and censored data.

The following is a more detailed overview of each paper and its contributions.

Chapter 2: Kernel density estimation for heavy-tailed distributions using the Champernowne transformation
This chapter is identical to the paper Buch-Larsen et al. (2005). In the paper, we introduce the Champernowne transformed kernel density estimator

$$\widehat{f}(x) = \frac{1}{N\, k_{T(x)}} \sum_{i=1}^{N} K_b(T(x) - T(X_i))T'(x), \tag{1.14}$$

where $K_b(u) = K(u/b)/b$ is a kernel function, b is a bandwidth, k_u is a boundary

correction, and $T(x)$ is a three parameter modified Champernowne distribution with cdf

$$T(x) = \frac{(x+c)^\alpha - c^\alpha}{(x+c)^\alpha + (M+c)^\alpha - 2c^\alpha} \qquad \forall x \in \mathbb{R}_+,$$

given by parameters $\alpha > 0$, $M > 0$, and $c \geq 0$. We investigate the properties of the Champernowne distribution, most importantly the fact that the Champernowne distribution converges to a heavy-tailed Pareto distribution in the tail, and describe how to estimate the parameters by maximum likelihood estimation. We also derive the asymptotic behaviour of (1.14). The paper includes a simulation study which compares the Champernowne transformed kernel density estimator to the kernel density estimators proposed in Clements et al. (2003) and Bolancé et al. (2003). The comparison shows that the Champernowne transformed kernel density estimator has desirable performance particularly for heavy-tailed data. Data studies of automobile claims and employer's liability demonstrate the method's usefulness on insurance data. The overall conclusion is that the method provides an estimator on the entire axis with desirable tail performance compared to similar methods, and that the method is therefore useful with respect to density estimation of heavy-tailed data. The results in this paper are based on Buch-Larsen (2003).

Chapter 3: Nonparametric Estimation of Operational Risk Losses Adjusted for Underreporting

This chapter is identical to the paper Buch-Kromann et al. (2007). In this paper, the Champernowne transformed kernel density estimator is adapted to operational risk and the special problem of underreporting by means of expert judgements. Underreporting is a major challenge in operational risk, which can be modelled by means of an underreporting function. The underreporting function encodes the likelihood that a loss of a particular size is being reported. This likelihood converges to 1 as the claim size approaches infinity, which means that the reported operational risk data set appears to be more heavy-tailed than it really is. The model we set up is the following:

Let $(X_i)_{1 \leq i \leq M}$ with density g be $M \sim \text{Poisson}(\lambda)$ iid operational risk claims which

have occurred, and let $(I(i))_{1\leq i\leq M}$ be an indicator function which encodes whether the claim is reported or not. Then we let $N = \sum_{i=1}^{M} I(i)$ denote the number of reported claims, and we let $(Y_j)_{1\leq j\leq N}$ denote the reported claims from the operational risk data set and assume that they are iid with density f. The probability of observing an operational risk claim is

$$P_{u,g} = \int_0^\infty g(w)u(w)\,\mathrm{d}w,$$

where $u(w) = P(I(1) = 1|X_1 = w)$ is the underreporting function under the assumption that the likelihood of reporting a claim only depends on the value of the claim. Under this model, $N \sim \text{Poisson}(\lambda P_{u,g})$, and the relationship between the density of reported operational risk claims and all operational risk claims, is

$$f(y) = \frac{g(y)u(y)}{P_{u,g}}.$$

We wish to estimate g. Unfortunately, g is the density of all operational risk claims including the unobserved claims, but we can express g as a function of f and u:

$$g(x) = \frac{f(x)\{u(x)\}^{-1}}{\int_0^\infty f(w)\{u(w)\}^{-1}\,\mathrm{d}w}.$$

The function f is the density of the observed claims, so it is possible to estimate f by means of Champernowne transformed kernel density estimation (1.14). The resulting estimator is denoted by $\widehat{f}$. Based on $\widehat{f}$, an obvious estimator of g is:

$$\widehat{g}(x) = \frac{\widehat{f}(x)\{u(x)\}^{-1}}{\int_0^\infty \widehat{f}(w)\{u(w)\}^{-1}\,\mathrm{d}w}.$$

In the paper, we derive the asymptotic behaviour of $\widehat{f}$ and $\widehat{g}$ and apply the method in an operational risk data study with six major business lines based on data from financial institutions. In the data study, we compare the performance of several parametric distributions and the effect of underreporting and nonparametric kernel

smoothing. The conclusion is that it is essential to take underreporting into consideration to obtain reliable estimates. Moreover, the study shows that the choice of parametric distribution is of crucial importance if one takes a purely parametric approach. However, a kernel smoothing correction on top of the parametric distribution has a stabilising effect and makes the choice of underlying parametric model less crucial.

Chapter 4: Estimation of Large Insurance Losses: A Case Study
This chapter is identical to the paper Buch-Kromann (2006). The paper uses the Champernowne transformed kernel density estimator described in chapter 2, but proposes a parameter estimation method of the Champernowne distribution which maximizes tail fit instead of likelihood. The intuition is that the nonparametric correction estimator is able to correct the center of the distribution where the data are dense, but not the tail where the data are sparse, and it is therefore essential to obtain a reliable tail fit in the parametric distribution. The proposed parameter estimation method is a heuristic method which selects the tail parameter α such that the 95% quantile of the empirical distribution and the estimated modified Champernowne distribution are equal. The parameter c is then chosen such that the mean of the estimated modified Champernowne distribution is as close as possible to the empirical mean. This parameter estimation method is called the quantile-mean method (QM). In a data study of employer's liability losses, we compare the performances of our Champernowne estimators in four settings: with ML or QM parameter estimates, and with and without kernel smoothing. Kolmogorov-Smirnov test statistics of the four methods are computed. The test accepts the Champernowne distribution based on ML parameter estimation with and without nonparametric correction, as the ML parameters compute the best overall fit. The test rejects the pure parametric Champernowne distribution with QM parameters, but after the nonparametric correction the estimator is accepted as well. This indicates that by using QM parameters, we get a suboptimal fit of the center of the distribution, even though we obtain a Champernowne estimator with superior tail fit; but the suboptimal center estimation is improved by the nonparametric correction. The methods are moreover

compared with respect to conditional means and the number of claims beyond a given threshold. The comparison shows that the Champernowne distribution seems to underestimate the tail with ML parameters, even when a nonparamtric correction is used, whereas the QM parameters result in a much better tail fit. Finally, we compare the Champernowne transformed kernel density with QM parameters to generalized Pareto distribution (GPD). The comparison shows that our QM method provides a tail fit which is almost comparable to the GPD tail fit, but that the QM method has some additional advantages: firstly, whereas the GPD estimator is only defined in the tail, our estimator is defined on the entire axis; secondly, our method is an automatic procedure, whereas the GPD estimator needs a threshold; and thirdly, the GPD estimator works well for heavy-tailed data, but often results in estimators with finite support when estimating moderately light tails, which never happens with our method.

Chapter 5: Comparison of tail performance of the Champernowne transformed kernel density estimator, the generalized Pareto distribution and the g-and-h distribution
This chapter is identical to the paper Buch-Kromann (2009). The paper is based on the idea in Buch-Kromann (2006): maximum likelihood may not be an optimal parameter estimation criterion when we are mostly interested in the tail of the distribution. In the paper, we introduce a two-stage conditional maximum likelihood parameter estimation method for the three Champernowne parameters, which ensures a superior tail fit coupled with a reasonable fit in the center. The procedure is the following. In the first step, set $c_1 = 0$ and choose (α_1, M_1) by maximizing the conditional log-likelihood function for all data above a threshold t. As the Champernowne distribution converges to a Pareto distribution, this gives a tail approximation of the estimated Champernowne distribution of $\tau x^{-\alpha_1+1}$, where $\tau = \alpha_1 M_1^{\alpha_1}$ is called the tail constant. In the second step, we fix $\widehat{\alpha} = \alpha_1$ and keep the tail constant τ unchanged, but allow c to be non-zero and find $\widehat{c}$ by maximizing an ordinary log-likelihood function. M can then be computed uniquely from $\widehat{\alpha}$, $\widehat{c}$ and τ. The method is called the conditional maximum likelihood (CML) method. In a data driven Monte Carlo simulation study, we compare the Champernowne transformed kernel density

estimator with CML parameters to the corresponding parametric (non-corrected) Champernowne distributions and two other distributions, a generalized Pareto distribution (GPD) and a g-and-h distribution with and without nonparametric correction. The comparison shows that the corrected CML Champernowne transformed kernel density estimator outperforms all the benchmark estimators except for the g-and-h distribution with nonparametric correction. This estimator appears to be superior to the corrected CML Champernowne estimator for heavy-tailed data sets, provided the data sets are large; but for small data sets as well as lighter-tailed data sets, the corrected CML Champernowne estimator is superior.

In the CML estimated Champernowne method as well as the GPD method we need to choose a threshold, and in the evaluation mentioned above, we used optimal thresholds for both methods. However, optimal thresholds are obviously not available in practice, and therefore the method's sensitivity to the choice of threshold is important. A study of the sensitivity shows that the CML Champernowne transformed kernel density estimator is substantially less sensitive to suboptimal choices of threshold than the GPD estimator, which is an important advantage of the CML method. The conclusion in this paper is therefore that the CML Champernowne transformed kernel density estimator is a method which in general has a comparable or superior tail performance compared to the benchmark estimators, while at the same time providing an acceptable fit of the center, unlike the GPD distribution.

Chapter 6: Multivariate density estimation using dimension reducing information and tail flattening transformations

This chapter is idential to the paper Buch-Kromann et al. (2009), which concerns multivariate estimation. The paper presents the transformation approach and multiplicative bias correction in a multivariate setting, and proposes an estimator which combines these two improvements. The multivariate kernel density estimator in the most general form is defined by

$$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_H(x - X_i), \tag{1.15}$$

where K is a multivariate kernel function and H is a $d \times d$ bandwidth matrix so that $K_H(x) = K(H^{-1}x)/\det(H)$. For simplicity, we consider the special case where $H = hI_d$.

For a given auxiliary model of X with density g, suppose $\widehat{g}$ is an estimator of g. A multiplicative corrected estimator of $f(x)$ can be defined by

$$\widetilde{f}(x) = \widehat{g}(x)\frac{1}{n}\sum_{i=1}^{n}\frac{K_H(x - X_i)}{\widehat{g}(X_i)}. \tag{1.16}$$

The transformation kernel density estimator is extended to a multivariate setting. Let $u = T(x, \lambda)$ be a transformation function which only depends on parameters $\lambda \in \Lambda \subseteq \mathbb{R}^p$. Since marginal transformations are convenient, we let $u_j = T_j(x_j, \lambda_j), j = 1, ..., d$. The multivariate transformation kernel density estimator then reduces to

$$\widetilde{f}_T(x) = \widehat{J}(x)\frac{1}{nh^d}\sum_{i=1}^{n}K\frac{\widehat{u}(x) - \widehat{U}_i}{h}, \tag{1.17}$$

where $\widehat{U}_i = T_i(X_i, \widehat{\lambda})$ is the transformed data set, and $\widehat{J} = \prod_{j=1}^{d}\left|\frac{\partial T_j(x_j, \widehat{\lambda})}{\partial x_j}\right|$ is the Jacobian of the empirical transformation.

We combine the multiplicative correction and the transformation approach and obtain the estimator

$$\widetilde{f}_C(x) = \widehat{J}(x)\widetilde{f}_U\{u(x)\}, \tag{1.18}$$

where $\widetilde{f}_U(x) = \widehat{g}_U(u)\frac{1}{nh^d}\sum_{i=1}^{n}\frac{K\left\{(u-\widehat{U}_i)/h\right\}}{\widehat{g}_U(\widehat{U}_i)}$ is the multiplicative corrected density estimator of U, i.e. the multiplicative corrected density estimator on the transformed axis, and $\widehat{g}_U(u) = \widehat{g}\{T^{-1}(u, \widehat{\lambda})\}\{T^{-1}(u, \widehat{\lambda})\}'$ is the auxiliary model density on the transformed axis.

In the paper, we derive the asymptotic theory of the estimators. In a simulation study and a data study, we compare the effects of the proposed improvements with the baseline. In the simulation study, the performance of the pure multiplicative bias correction without tail flattening transformation is investigated both in additive and

multiplicative designs. The additive design is the "home turf" case, which means that the auxiliary model in the multiplicative bias correction is specified correctly, whereas the multiplicative case is specified incorrectly. The performance of the pure nonparametric kernel density estimator and the multiplicatively corrected estimator are measured by means of integrated squared error. The comparison between the two estimators shows that improvements by using multiplicative correction are obtained even in very small sample sizes, in both additive and multiplicative designs.

The data study is based on a heavy-tailed commercial fire insurance data set, where we apply the proposed transformation technique. We use the transformation kernel density estimator with and without multiplicative correction, and compare their performance in a data-driven simulation study, which confirms the conclusion from the simulation study of the pure multiplicative correction. The data study shows that the multiplicative correction of the transformed kernel density estimator improves the estimation performance significantly compared to the transformation kernel density estimator without multiplicative correction. Compared to the auxiliary model, improvements of the multiplicative corrected estimator are obtained when the auxiliary model is incorrectly specified, without aggrevating the performance of the auxiliary model when the auxiliary model is correctly specified.

Chapter 7: Multivariate density estimation using dimension reducing information and tail flattening transformations for truncated or censored data
This chapter is identical to the paper Buch-Kromann and Nielsen (2009). The paper extends the multivariate estimators introduced in the previous chapter to the situation where the data are truncated and censored – in the following referred to "filtering". The extension uses survival analysis to control the exposure. Our data set is $(X_i, \widetilde{Y}_i, D_i, T_i)_{i=1,\dots,n}$ where X_i is a covariate, $\widetilde{Y}_i = Y_i \wedge C_i$ is a claim subjected to censoring C_i, $D_i = I(Y_i \leq C_i)$ indicates whether right censoring has occurred, and T_i is the left-tuncation time, which means that $\widetilde{Y}_i$ is only observed when $\widetilde{Y}_i \geq T_i$.

The non-parametric filtered data density estimator has the form

$$\widehat{f}_x^{(d,b)}(t) = \frac{\sum_{i=1}^n \int K_{d_1}(t-s)K_{d_2}(x-X_i)\widehat{S}^{(b)}_{X_i,(i)}(s)\,\mathrm{d}N_i(s)}{\sum_{i=1}^n \int K_{d_1}(t-s)K_{d_2}(x-X_i)R_i(s)\mathrm{d}s} \tag{1.19}$$

where $\widehat{S}^{(b)}_{X_i,(i)}(s) = \exp\left\{-\int_0^s \widehat{\alpha}^{(b)}_{X_i,(i)}(u)\,\mathrm{d}u\right\}$ is the leave-one-out estimator of the survival function, and

$$\widehat{\alpha}^{(b)}_{X_i,(i)}(t) = \frac{\sum_{j\neq i} \int K_{b_1}(t-s)K_{b_2}(x-X_j)\,\mathrm{d}N_j(s)}{\sum_{j\neq i} \int K_{b_1}(t-s)K_{b_2}(x-X_j)R_j(s)\,\mathrm{d}s}$$

is the leave-one-out hazard estimator. The density estimator (1.19) is a fully nonparametric multivariate density estimator taking filtering into account, which corresponds to (1.15) in the non-filtering case.

The transformation approach is based on the same underlying idea as in the previous papers. Let $\Psi : [0,\infty) \to [0,1)$ be a tail-flattening transformation function, compute the transformed data by means of this transformation, compute the non-parametric filtered data density estimator (1.19) on the transformed data, and back-transform to obtain an estimator on the original axis. The resulting estimator is

$$\widehat{f}^{(d,b)}_{\Psi,x}(t) = \psi(t)\cdot \widehat{k}^{(d,b)}_{\Psi,x}\{\Psi(t)\}, \tag{1.20}$$

where

$$\widehat{k}^{(d,b)}_{\Psi,x}(v) = \frac{\sum_{i=1}^n \int_0^1 K_{d_1}(v-s)K_{d_2}(x-X_i)\widehat{S}^{(b)}_{\Psi,X_i,(i)}(s)\,\mathrm{d}\widetilde{N}_i(s)}{\sum_{i=1}^n \int_0^1 K_{d_1}(v-s)K_{d_2}(x-X_i)R_i\{\Psi^{-1}(s)\}\,\mathrm{d}s}$$

is the non-parametric estimator (1.19) on the transformed axis,

$$\widehat{S}^{(b)}_{\Psi,X_i,(i)}(s) = \exp\left\{-\int_0^s \widehat{\alpha}^{(b)}_{\Psi,X_i,(i)}(u)\,\mathrm{d}u\right\}$$

is the estimator of the survival function on the transformed axis, $R_i\{\Psi^{-1}(s)\}$ is the

"at-risk" indicator on the transformed axis, and

$$\widehat{\alpha}^{(b)}_{\Psi,X_i,(i)}(t) = \frac{\sum_{j\neq i}\int_0^1 K_{b_1}(t-s)K_{b_2}(x-X_j)\,\mathrm{d}\widetilde{N}_j(s)}{\sum_{j\neq i}\int_0^1 K_{b_1}(u-s)K_{b_2}(x-X_j)R_j\{\Psi^{-1}(s)\}\,\mathrm{d}s}$$

is the hazard estimator on the transformed axis. The transformation estimator (1.20) with filtering corresponds to (1.17) in the non-filtering case.

Multiplicative bias correction is introduced in a similarly way as well. Let h_x be a prior knowledge density corresponding to what we previously called the auxiliary density. The multiplicative bias corrected density estimator is

$$\widehat{g}^{(d,b)}_x(t) = h_x(t)\widehat{c}^{(d,b)}_x(t), \tag{1.21}$$

where

$$\widehat{c}^{(d,b)}_x(t) = \frac{\sum_{i=1}^n \int K_{d_1}(t-s)K_{d_2}(x-X_i)\widehat{S}^{(b)}_{X_i,(i)}(s)\left\{h_{X_i}(s)\right\}^{-1}\,\mathrm{d}N_i(s)}{\sum_{i=1}^n \int K_{d_1}(t-s)K_{d_2}(x-X_i)R_i(s)\,\mathrm{d}s}$$

is the multiplicative bias correction. The multiplicative bias corrected estimator (1.21) corresponds to (1.16) in the non-filtering case.

We combine the two techniques to obtain an estimator which benefits particularly from transformation when the data are heavy-tailed, and which incorporates prior knowledge by means of multiplicative correction. This estimator has the form

$$\widetilde{f}^{(d,b)}_{\Psi,x}(t) = \psi(t)\widetilde{k}^{(d,b)}_{\Psi,x}\{\Psi(t)\}, \tag{1.22}$$

where $\widetilde{k}^{(d,b)}_{\Psi,x}(v) = h_x\{\Psi^{-1}(v)\}\widetilde{c}^{(d,b)}_{\Psi,x}(v)$ is the density estimator on the transformed axis, and

$$\widetilde{c}^{(d,b)}_{\Psi,x}(v) = \frac{\sum_{i=1}^n \int_0^1 K_{d_1}(v-s)K_{d_2}(x-X_i)\widehat{S}^{(b)}_{\Psi,X_i,(i)}(s)[h_{X_i}\{\Psi^{-1}(s)\}]^{-1}\,\mathrm{d}\widetilde{N}_i(s)}{\sum_{i=1}^n \int_0^1 K_{d_1}(v-s)K_{d_2}(x-X_i)R_i\{\Psi^{-1}(s)\}\,\mathrm{d}s}$$

is the multiplicative bias correction on the transformed axis. The combined estimator

(1.22) corresponds to (1.18) in the no-filtering case.

We derive the asymptotic properties of the estimators. In a simulation study, we compare the performance of the estimators for different amounts of filtering. The results resemble the results in the previous paper. The multiplicative bias correction improves the nonparametric estimation significantly, and it also improves the prior knowledge estimator when the density is not correctly specified, without worsening it if the prior knowledge is correctly specified. Moreover, the estimators are compared to the estimators in the previous paper. When filtering is not present, the simple estimators in the previous paper outperform the filtering estimators. However, with just small amounts of filtering, the filtering estimators have a significant advantage.

Chapter 2

Kernel density estimation for heavy-tailed distributions using the Champernowne transformation

This chapter is an adapted version of Buch-Larsen et al. (2005).

When estimating loss distributions in insurance, large and small losses are usually split because it is difficult to find a simple parametric model that fits all claim sizes. This approach involves determining the threshold level between large and small losses. In this article a unified approach to the estimation of loss distributions is presented. We propose an estimator obtained by transforming the data set with a modification of the Champernowne cdf and then estimating the density of the transformed data by use of the classical kernel density estimator. We investigate the asymptotic bias and variance of the proposed estimator. In a simulation study, the proposed method shows a good performance. We also present two applications dealing with claims costs in insurance.

2.1 Introduction

In finance and nonlife insurance, estimation of loss distributions is a fundamental part of the business. In most situations, losses are small, and extreme losses are rarely observed, but the number and the size of extreme losses can have a substantial influence on the profit of the company. Standard statistical methodology, such as integrated error and likelihood, does not weigh small and big losses differently in the evaluation of an estimator. These evaluation methods do not, therefore, emphasize an important part of the error: the error in the tail.

Practitioners often decide to analyze large and small losses separately, because no single, classical parametric model fits all claim sizes. This approach leaves some important challenges: choosing the appropriate parametric model, identifying the best way of estimating the parameters and determining the threshold level between large and small losses.

This work presents a systematic approach to the estimation of loss distributions which is suitable for heavy tailed situations. The proposed estimator is obtained by transforming the data set with a parametric estimator and afterwards estimating the density of the transformed data set using the classical kernel density estimator, see Wand and Jones (1995); Silverman (1986)

$$\widehat{f}(y) = \frac{1}{Nb} \sum_{i=1}^{N} K\left(\frac{y - Y_i}{b}\right),$$

where K is the kernel function, b is the bandwidth and Y_i, $i = \{1, ..., N\}$ is the transformed data set. The estimator of the original density is obtained by back-transformation of $\widehat{f}(y)$. We will call this method a *semiparametric estimation procedure* because a parametrized transformation family is used. We propose to use a transformation based on the little-known Champernowne cdf, because it produces good results in all the studied situations and it is straightforward to apply.

The semiparametric estimator with shifted power transformation was introduced in

Wand et al. (1991). They showed that the classical kernel density estimator was improved substantially by applying a transformation and suggested the shifted power transformation family. Bolancé et al. (2003) improved the shifted power transformation for highly skewed data by proposing an alternative parameter selection algorithm. The semiparametric estimator with the Johnson family transformation function was studied by Yang and Marron (1999). Hjort and Glad (1995) advocated a semiparametric estimator with a parametric start, which is closely related to the bias reduction method described in Jones et al. (1995). The Möbius-like transformation was introduced in Clements et al. (2003). In contrast to the shifted power transformation, which transforms $(0, \infty)$ into $(-\infty, \infty)$, the Möbius-like transformation transforms $(0, \infty)$ into $(-1, 1)$ and the parameter estimation method is designed to avoid boundary problems. Scaillet (2004) has recently studied nonparametric estimators for probability density function which have support on the non-negative real line using alternative kernels.

The original Champernowne distribution has density, Johnson et al. (1994)

$$f(x) = \frac{c}{x\left(\frac{1}{2}\left(\frac{x}{M}\right)^{-\alpha} + \lambda + \frac{1}{2}\left(\frac{x}{M}\right)^{\alpha}\right)} \qquad x \geq 0, \tag{2.1}$$

where c is a normalizing constant and α, λ and M are parameters. The distribution was mentioned for the first time in 1936 by D.G. Champernowne when he spoke on "The Theory of Income Distribution" at the Oxford Meeting of the Econometric Society Brown (1937). Later, he gave more details about the distribution and its application to economics, Champernowne (1952). When λ equals 1 and the normalizing constant c equals $\frac{1}{2}\alpha$, the density of the original distribution is simply called the *Champernowne distribution*

$$f(x) = \frac{\alpha M^{\alpha} x^{\alpha-1}}{(x^{\alpha} + M^{\alpha})^2}$$

with cdf

$$F(x) = \frac{x^\alpha}{x^\alpha + M^\alpha}. \tag{2.2}$$

The Champernowne distribution converges to a Pareto distribution in the tail, while looking more like a lognormal distribution near 0 when $\alpha > 1$. Its density is either 0 or infinity at 0 (unless $\alpha = 1$).

In the transformation kernel density estimation method, if we transform the data with the Champernowne cdf, the inflexible shape near 0 results in boundary problems. We argue that a modification of the Champernowne with an additional parameter can solve this inconvenience.

We did not choose to work with classical extensions of the Pareto distribution such as the generalised Pareto distribution (GPD), see i.e. Coles (2001). The reason for doing so is that the GPD often estimates distributions of infinite support to have finite support and hence it cannot be used as a transformation. We carried out a small simulation study of a standard log normal distribution; more than half the time the GPD suggested a distribution with finite support. Furthermore, the GPD needs a (hard to pick) threshold from where the distribution starts; such that the transformation methodology meets problems also in the beginning of the distribution.

In this paper we study the transformation kernel density estimation method. The conclusion of the simulation study is that the new approach based on the modified Champernowne distribution is the preferable method, because it is the only estimator which has a good performance in most of the investigated situations. Section 2.2 describes the transformation family and explains the parameter estimation procedure. Section 2.3 presents the semiparametric kernel density estimator and its properties. In section 2.4, the simulation study is presented and section 2.5 shows two applications. Finally, the section 2.6 outlines the main conclusions.

2.2 The modified Champernowne distribution function

We generalize the Champernowne distribution with a new parameter c. This parameter ensures the possibility of a positive finite value of the density at 0 for all α.

Definition 2.1. *The modified Champernowne cdf is defined for $x \geq 0$ and has the form*

$$T_{\alpha,M,c}(x) = \frac{(x+c)^\alpha - c^\alpha}{(x+c)^\alpha + (M+c)^\alpha - 2c^\alpha} \qquad \forall x \in \mathbb{R}_+ \tag{2.3}$$

with parameters $\alpha > 0$, $M > 0$ and $c \geq 0$ and density

$$t_{\alpha,M,c}(x) = \frac{\alpha(x+c)^{\alpha-1}((M+c)^\alpha - c^\alpha)}{((x+c)^\alpha + (M+c)^\alpha - 2c^\alpha)^2} \qquad \forall x \in \mathbb{R}_+.$$

Corresponding to the Champernowne distribution, the modified Champernowne distribution converges to a Pareto distribution in the tail:

$$t_{\alpha,M,c}(x) \to \frac{\alpha\left(((M+c)^\alpha - c^\alpha)^{\frac{1}{\alpha}}\right)^\alpha}{x^{\alpha+1}} \text{ as } x \to \infty.$$

The effect of the additional parameter c is different for $\alpha > 1$ and for $\alpha < 1$. The parameter c has some "scale parameter properties": when $\alpha < 1$, the derivative of the cdf becomes larger for increasing c, and conversely, when $\alpha > 1$, the derivative of the cdf becomes smaller for increasing c. When $\alpha \neq 1$, the choice of c affects the density in three ways. First, c changes the density in the tail. When $\alpha < 1$, positive cs result in lighter tails, and the opposite when $\alpha > 1$. Secondly, c changes the density in 0. A positive c provides a positive finite density in 0:

$$0 < t_{\alpha,M,c}(0) = \frac{\alpha c^{\alpha-1}}{(M+c)^\alpha - c^\alpha} < \infty \qquad \text{when } c > 0.$$

Thirdly, c moves the mode. When $\alpha > 1$, the density has a mode, and positive cs shift the mode to the left. We therefore see that the parameter c also has a shift parameter effect. When $\alpha = 1$, the choice of c has no effect.

Figure 2.1 illustrates the role of c: the two graphs on the top show the cdfs and the densities for the modified Champernowne distribution for fixed $\alpha < 1$ and $M = 3$. In the cdf plot, we see that increasing c results in lower values of the cdf in the interval $[0, M)$ and higher values of the cdf in the interval $[M, \infty)$. In the density plot, we see that increasing c results in a lighter tail and a finite density at 0. In the two graphs in the middle, we have fixed $\alpha = 1$ and $M = 3$. We see that changing c has no effect. The two graphs at the bottom illustrate the effect of increasing c when $\alpha > 1$, for $M = 3$. Notice that the values of the cdf become higher in the interval $[0, M)$ and lower in the interval $[M, \infty)$. The density plot shows that positive cs move the mode to the left and produce a heavier tail.

From a computational point of view, it is simpler to estimate M and then proceed to the other parameters. In the Champernowne distribution, we notice that $T_{\alpha,M,0}(M) = 0.5$. The same holds for the modified Champernowne distribution: $T_{\alpha,M,c}(M) = 0.5$. This suggests that M can be estimated as the empirical median of the data set. The empirical median is a robust estimator, especially for heavy-tailed distributions, as shown in Lehmann (1991). He studied the properties of the median and the mean as an estimator of location for the normal distribution and the Cauchy distribution, and showed that whereas the mean works well as an estimator of location for the normal distribution, it works poorly for the Cauchy distribution due to its heavy tail. Tukey (1960) reached the same conclusion when he studied the efficiency of the median and the mean. He showed that the median efficiency increases as the tail becomes heavier. Corresponding models have also been studied for heavy-tailed distributions, see Stigler (1973); Newcomb (1882, 1886). A similar type of discussion for the variance estimation was done by Huber (1981). As we are especially concerned about heavy tails, we consider the robustness of the median to be important.

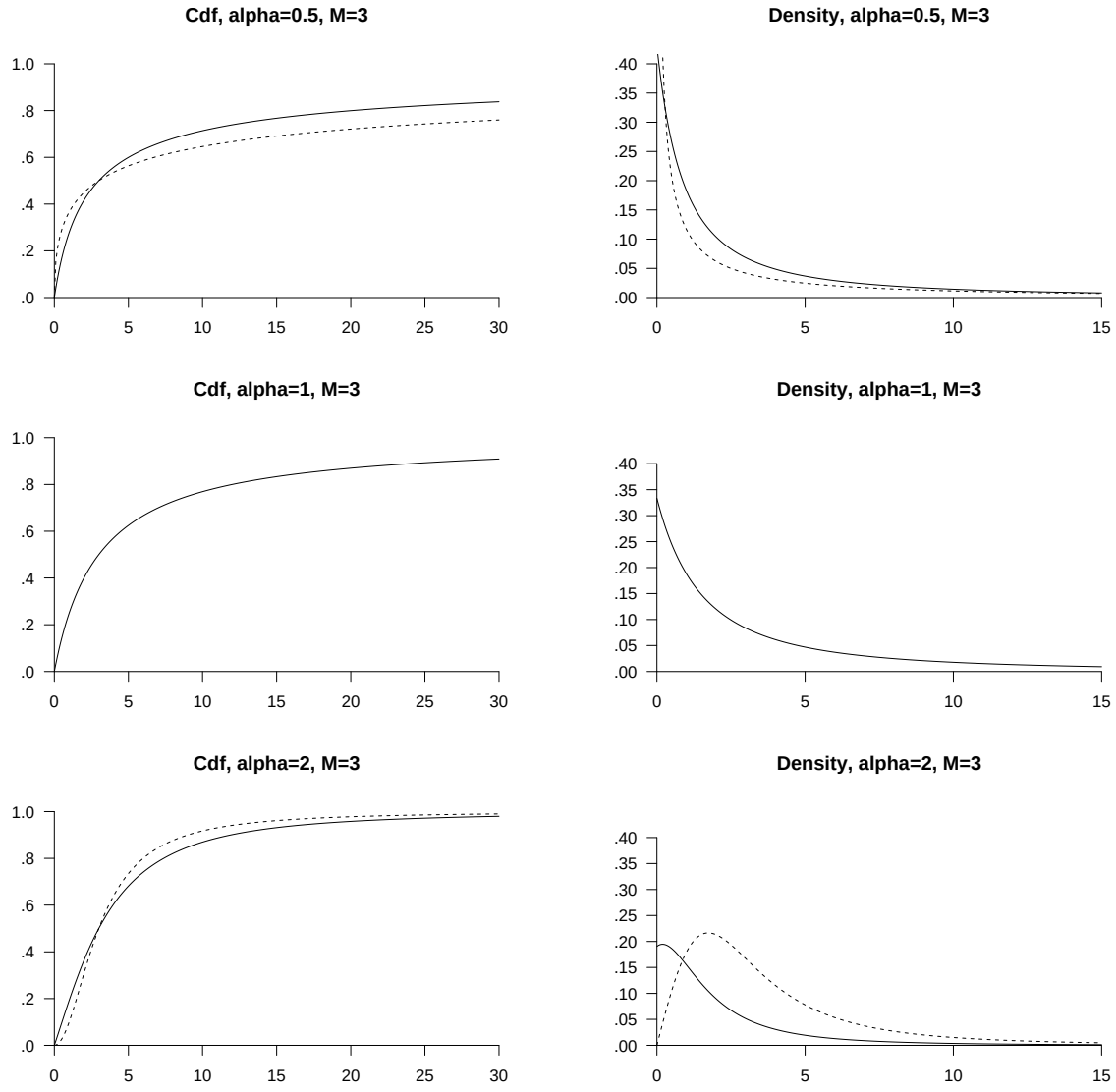

Figure 2.1: Different shapes of the modified Champernowne distribution with different choices of α, as well as the effect of the parameter c. In all plots $c = 0$ dashed line and $c = 2$ solid line.

After parameter M has been estimated as described above, the next step is to estimate the pair (α, c) which maximizes the log likelihood function:

$$\begin{aligned} l(\alpha, c) &= N \log \alpha + N \log\left((M+c)^{\alpha} - c^{\alpha}\right) + (\alpha - 1) \sum_{i=1}^{N} \log(X_i + c) \\ &\quad -2 \sum_{i=1}^{N} \log\left((X_i + c)^{\alpha} + (M+c)^{\alpha} - 2c^{\alpha}\right). \end{aligned} \tag{2.4}$$

For a fixed M, this likelihood function is concave and has a maximum.

2.3 The semiparametric transformation kernel density estimator

In this section we will make a detailed derivation of the estimator based on the modified Champernowne distribution, which we will call KMCE. The resulting estimator is obtained by computing a nonparametric classical kernel density estimator for the transformed data set and, finally, the result is back-transformed.

2.3.1 Transformation with the modified Champernowne distributions

Let X_i, $i = 1, ..., N$, be positive stochastic variables with an unknown cdf F and density f. The following describes in detail the transformation kernel density estimator of f, and Figure 2.2 illustrates the four steps of the estimation procedure for a data set with 1000 observations generated from a Weibull distribution. The resulting transformation kernel density estimator of f based on the Champernowne distribution is denoted by KMCE.

(i) Calculate the parameters $\left(\widehat{\alpha}, \widehat{M}, \widehat{c}\right)$ of the modified Champernowne distribution

as described in section 2.2 to obtain the transformation function. In the first plot in Figure 2.2, we see the estimated transformation function and the true Weibull distribution. Notice that the modified Chapernowne density has a larger mode and that the tail is too heavy.

(ii) Transform the data set X_i, $i = 1, ..., N$, with the transformation function, T:

$$Y_i = T_{\widehat{\alpha},\widehat{M},\widehat{c}}(X_i), \qquad i = 1, ..., N.$$

The transformation function transforms data into the interval $(0, 1)$, and the parameter estimation is designed to make the transformed data as close to a uniform distribution as possible. The transformed data are illustrated in the second plot in Figure 2.2.

(iii) Calculate the classical kernel density estimator on the transformed data, $Y_i, i = 1, ..., N$:

$$\widehat{f}_{\text{trans}}(y) = \frac{1}{N\, k_y} \sum_{i=1}^{N} K_b(y - Y_i),$$

where $K_b(\cdot) = (1/b)K(\cdot)$ and $K(\cdot)$ is the kernel function. The boundary correction, k_y, is required because the Y_i are in the interval $(0, 1)$ so that we need to divide by the integral of the part of the kernel function that lies in this interval. The boundary correction k_y is defined as

$$k_y = \int_{\max(-1,-y/b)}^{\min(1,(1-y)/b)} K(u)\, \mathrm{d}u.$$

The classical kernel density estimator of the transformed data set is illustrated in the third plot in Figure 2.2.

(iv) The classical kernel density estimator of the transformed data set results in the KMCE estimator on the transformed scale. Therefore the estimator of the

density of the original data set, X_i, $i = 1, ..., N$ is :

$$\widehat{f}(x) = \frac{\widehat{f}_{\text{trans}}\left(T_{\widehat{\alpha},\widehat{M},\widehat{c}}(x)\right)}{\left|\left(T^{-1}_{\widehat{\alpha},\widehat{M},\widehat{c}}\right)'\left(T_{\widehat{\alpha},\widehat{M},\widehat{c}}(x)\right)\right|}.$$

The KMCE results for the Weibull data set is seen in the last plot in Figure 2.2.

The expression of the KMCE is:

$$\widehat{f}(x) = \frac{1}{N\ k_{T_{\widehat{\alpha},\widehat{M},\widehat{c}}(x)}} \sum_{i=1}^{N} K_b(T_{\widehat{\alpha},\widehat{M},\widehat{c}}(x) - T_{\widehat{\alpha},\widehat{M},\widehat{c}}(X_i))T'_{\widehat{\alpha},\widehat{M},\widehat{c}}(x). \tag{2.5}$$

2.3.2 Asymptotic theory for the transformation kernel density estimator

In this section, we investigate the asymptotic theory of the transformation kernel density estimator in general. We derive its asymptotic bias and variance.

Theorem 2.1. *Let $X_1, ..., X_N$ be independent identically distributed variables with density f. Let $\widehat{f}(x)$ be the transformation kernel density estimator of $f(x)$*

$$\widehat{f}(x) = \frac{1}{N} \sum_{i=1}^{N} K_b(T(x) - T(X_i))T'(x),$$

where $T(\cdot)$ is the transformation function.

Then the bias and the variance of $\widehat{f}(x)$ *are given by*

$$\begin{aligned}\mathbb{E}\left[\widehat{f}(x)\right] - f(x) &= \frac{1}{2}\mu_2(K)b^2\left(\left(\frac{f(x)}{T'(x)}\right)'\frac{1}{T'(x)}\right)' + o(b^2),\\ \mathbb{V}\left[\widehat{f}(x)\right] &= \frac{1}{Nb}R(K)T'(x)f(x) + o\left(\frac{1}{Nb}\right)\end{aligned}$$

as $N \to \infty$, *where* $\mu_2(K) = \int u^2K(u)\,\mathrm{d}u$ *and* $R(K) = \int K^2(u)\,\mathrm{d}u$.

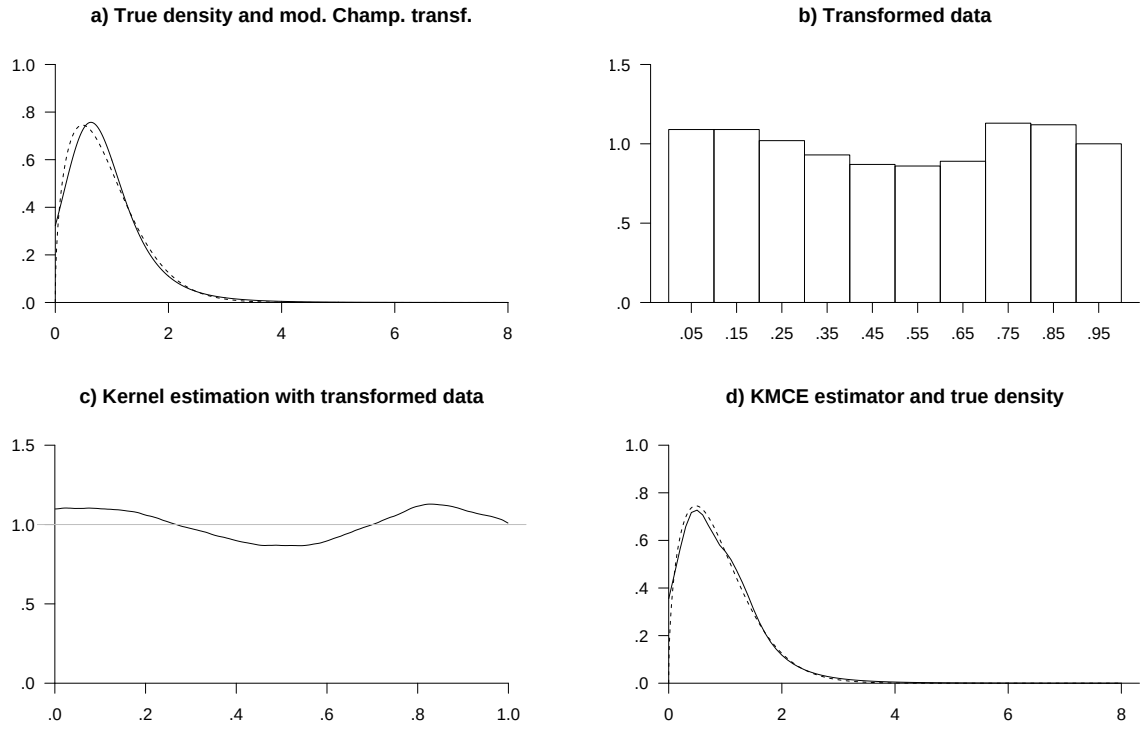

Figure 2.2: Four steps of the KMCE estimator. a) The estimated transformation function (solid line) and the true density (dashed line). b) The histogram of the transformed data set. c) The estimated classical kernel density estimator of the transformed data set. d) The final KMCE estimator (solid line) and the true density (dashed line).

Proof. We assume that $X_1, ..., X_N$ are independent identically distributed variables with density f. Let $\widehat{f}(x)$ be the transformation kernel density estimator of $f(x)$:

$$\widehat{f}(x) = \frac{1}{N}\sum_{i=1}^{N} K_b(T(x) - T(X_i))T'(x),$$

where $T(\cdot)$ is the transformation function. Let the transformed variable have distribution g:

$$Y_i = T(X_i) \sim g(y) = \frac{f(T^{-1}(y))}{T'(T^{-1}(y))}$$

and let $\widehat{g}(y)$ be the classical kernel density estimator of $g(y)$:

$$\widehat{g}(y) = \frac{1}{N}\sum_{i=1}^{N} K_b(y - Y_i).$$

The mean and variance of the classical kernel density estimator is:

$$\mathbb{E}\left[\widehat{g}(y)\right] = g(y) + \frac{1}{2}b^2\mu_2(K)g''(y) + o(b^2), \tag{2.6}$$

$$\mathbb{V}\left[\widehat{g}(y)\right] = \frac{1}{Nb}R(K)g(y) + o\left(\frac{1}{Nb}\right). \tag{2.7}$$

The transformation kernel density estimator can be expressed by the standard kernel density estimator:

$$\widehat{f}(x) = T'(x)\widehat{g}(T(x))$$

implying

$$\begin{aligned}\mathbb{E}\left[\widehat{f}(x)\right] &= T'(x)\mathbb{E}\left[\widehat{g}(T(x))\right] \\ &= T'(x)\left(g(T(x)) + \frac{1}{2}b^2\mu_2(K)\frac{\partial^2 g(T(x))}{\partial(T(x))^2} + o(b^2)\right).\end{aligned} \tag{2.8}$$

Note that

$$g(T(x)) = \frac{f(x)}{T'(x)}, \quad \frac{\partial g(T(x))}{\partial T(x)} = \left(\frac{f(x)}{T'(x)}\right)' \frac{1}{T'(x)},$$

and

$$\frac{\partial^2 g(T(x))}{\partial (T(x))^2} = \left(\left(\frac{f(x)}{T'(x)}\right)' \frac{1}{T'(x)}\right)' \frac{1}{T'(x)}$$

which are used to find the mean of the transformation kernel density estimator

$$\mathbb{E}\left[\widehat{f}(x)\right] = f(x) + \frac{1}{2} b^2 \mu_2(K) \left(\left(\frac{f(x)}{T'(x)}\right)' \frac{1}{T'(x)}\right)' + o(b^2). \tag{2.9}$$

The variance is calculated in a similar way

$$\begin{aligned} \mathbb{V}\left[\widehat{f}(x)\right] &= (T'(x))^2 \mathbb{V}\left[\widehat{g}(T(x))\right] \\ &= (T'(x))^2 \left(\frac{1}{Nb} R(K) g(T(x)) + o\left(\frac{1}{Nb}\right)\right) \\ &= \frac{1}{Nb} R(K) T'(x) f(x) + o\left(\frac{1}{Nb}\right). \end{aligned} \tag{2.10}$$

It is known, Yang (2000) that the classical kernel density estimator follows a nornal distribution asymptotically:

$$\sqrt{Nb}\left(\widehat{g}(y) - \mathbb{E}\left[\widehat{g}(y)\right]\right) \sim \mathbf{N}\left(0, \frac{1}{Nb} R(K) g(y)\right).$$

Then, since $\widehat{f}(x) = T'(x)\widehat{g}(y)$ with $y = T(x)$, then

$$\sqrt{Nb}\left(\widehat{f}(x) - \mathbb{E}\left[\widehat{f}(x)\right]\right) \sim \mathbf{N}\left(0, \frac{1}{Nb} R(K) T'(x) f(x)\right).$$

For a parametric transformation $T(x) = T_\theta(x)$, if we assume that $\hat{\theta}$ is a square-root-n consistent estimator of θ, then it follows that the asymptotic distribution of $\widehat{f}(x)$ with parametric estimated transformation $T_{\hat{\theta}}(x)$, equals the asymptotic distribution

of $\widehat{f}(x)$ with parametric transformation $T_\theta(x)$. □

2.4 Simulation study

This section presents a comparison of our semiparametric method based on the modified Champernowne distributions with two benchmark estimators. We simulate data from four distributions with different tails and different shapes near 0. We measure the error between the estimated density and the true density by using four different error measures. In subsection 2.4.3, we evaluate the performance of the KMCE estimators compared to the estimator described in Clements et al. (2003), in the following called CHL, and the estimator described in Bolancé et al. (2003), in the following called BGN.

2.4.1 The distributions

We have simulated four distributions with different characteristics: *lognormal, lognormal-Pareto, Weibull* and *truncated logistic*. The lognormal distribution has a moderately light tail, and when we mix the lognormal distribution with the Pareto distribution, which is a heavy-tailed distribution, the resulting distribution is also heavy-tailed. The Weibull distribution is a light-tailed distribution that starts at 0 and has a mode. The truncated logistic is a light-tailed distribution that has a positive finite density at 0. The distributions and the chosen parameters are listed in Table 2.1 and Figure 2.3 plots the densities to show the diversity of shapes.

2.4.2 Measuring the error

We measure the performance of the estimators by the error measures L_1, L_2, WISE and E. Let $\widehat{f}(x)$ be the estimated density and $f(x)$ be the true density. The L_1 norm measures the distance between the estimated density and the true density on the

Distribution	**Density for $x > 0$**	**Parameters**
Lognormal(μ, σ^2)	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}x} e^{-\frac{(\log x - \mu)^2}{2\sigma^2}}$	$(\mu, \sigma^2) = (0, .5)$
Mixture of p lognormal(μ, σ) $(1-p)$ Pareto(λ, ρ, c)	$f(x) = p\frac{1}{\sqrt{2\pi\sigma^2}x} e^{-\frac{(\log x - \mu)^2}{2\sigma^2}} +$ $(1-p)(x-c)^{-(\rho+1)}\rho\lambda^\rho$	$(p, \mu, \sigma, \lambda, \rho, c)$ $= (.7, 0, 1, 1, 1, -1)$ $= (.3, 0, 1, 1, 1, -1)$
Weibull(γ)	$f(x) = \gamma x^{(\gamma-1)} e^{-x^\gamma}$	$\gamma = 1.5$
Truncated logistic(s)	$f(x) = \frac{2}{s} e^{\frac{x}{s}} \left(1 + e^{\frac{x}{s}}\right)^{-2}$	$s = 1$

Table 2.1: Distributions used in the simulation study.

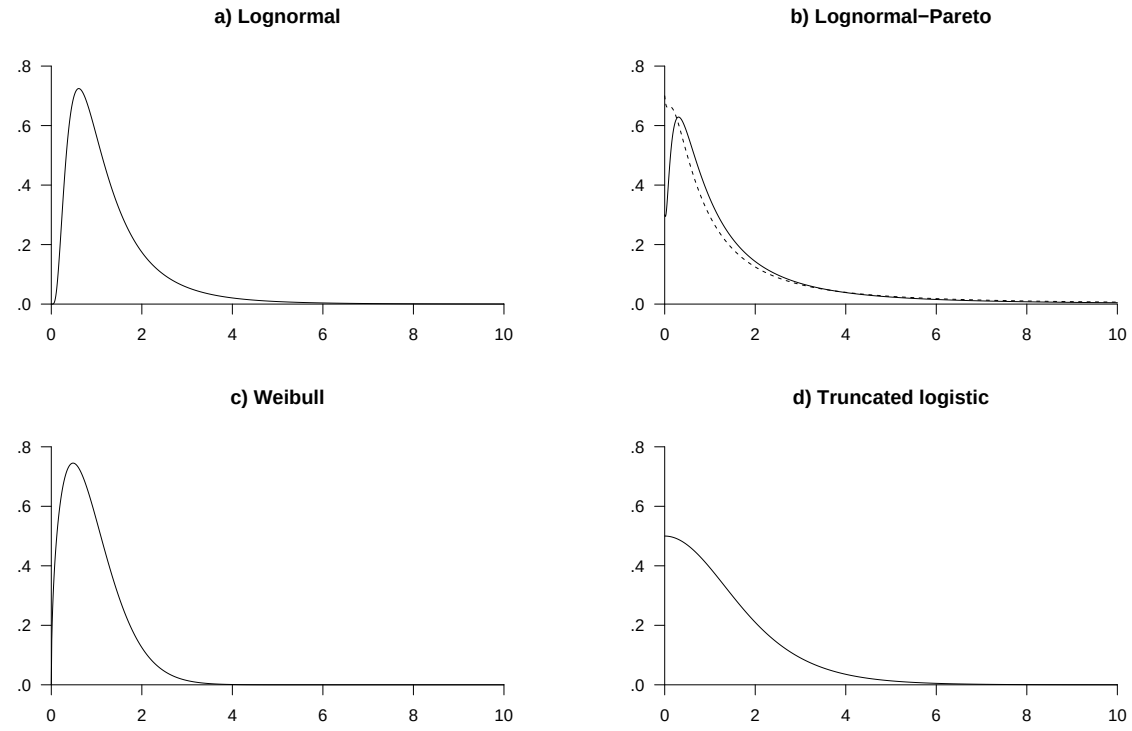

Figure 2.3: Shape of the different distributions used in the simulation study. In plot b) 30% Pareto (solid line) and 70% Pareto (dashed line).

whole support.

$$\mathrm{L}_1 = \int_0^\infty \left| \widehat{f}(x) - f(x) \right| \, dx.$$

We also calculate the L_2 norm between the two distributions.

$$\mathrm{L}_2 = \left(\int_0^\infty \left(\widehat{f}(x) - f(x) \right)^2 \, dx \right)^{1/2}.$$

Both L_1 and L_2 weigh errors of the estimator near 0 and in the tail equally, although the consequences for some real-world situations of a poor estimation in the tail are much more critical than the consequences of a poor estimation near 0.

WISE weighs the distance between the estimated and the true distribution with the squared value of x. This results in an error measure that emphasizes the tail of the distribution, which is very relevant in practice when dealing with income or cost data.

$$\mathrm{WISE} = \left(\int_0^\infty \left(\widehat{f}(x) - f(x) \right)^2 x^2 \, dx \right)^{1/2}.$$

The last error measure, E, calculates the distance between the estimated mean excess function and the true mean excess function. It emphasizes the error in the tail as well.

$$\mathrm{E} = \left(\int_0^\infty \left(\widehat{e}(x) - e(x) \right)^2 f(x) \, \mathrm{d}x \right)^{1/2} \tag{2.11}$$

$$= \left(\int_0^\infty \left(\int_x^\infty u \left(f(u) - \widehat{f}(u) \right) \mathrm{d}u \right)^2 f(x) \, \mathrm{d}x \right)^{1/2}, \tag{2.12}$$

To calulate the error measures, we used the change of variable $y = (x - M)/(x + M)$ proposed by Clements et al. (2003).

2.4.3 Comparison of the estimation methods

We compare the performance of the KMCE, the CHL and the BGN estimators. The comparison is based on data simulated from the four distributions described in Table 2.1, and four sample sizes: $N = 50$, $N = 100$, $N = 500$ and $N = 1000$. Each combination of distribution and sample size is replicated 2000 times. In Table 2.2 and Table 2.3 we show the means of the error measures for the 2000 samples. We show the results obtained when using the rule of thumb method, Silverman (1986) for bandwidth selection. We also investigated bandwidth selection method described in Sheather and Jones (1991) and the conclusions do not change.

For the moderately light-tailed lognormal, all three estimators exhibit good performance in general. The KMCE and CHL estimators show the best performance. The CHL estimator outperforms the KMCE estimator for all N, but it seems to outperform the KMCE estimator near 0 rather than in the tail, as seen by the fact that the improvement obtained on L_1 and L_2 is greater than the improvement obtained on WISE and E. The BGN estimator also performs well in this case. The performance of this estimator is only 3-4% worse than the performance of the KMCE estimator.

For the heavy-tailed distributions, the KMCE estimator shows a significantly better performance than the CHL and the BGN estimators. The performance of the CHL estimator is poor compared to the KMCE estimator. For the 70% lognormal-30% Pareto, the KMCE estimator outperforms the CHL estimator by about 15-20%, and the performance gap seems to become larger when N increases. The largest performance gap occurs with WISE and E, which indicates that the performance gap is mainly in the tail. The BGN estimator is also outperformed by the KMCE estimator: the error measures are about 10% better for the KMCE estimator than the BGN estimator for the 70% lognormal-30% Pareto, but the improvement seems to go down when N increases.

The results for the 30% lognormal-70% Pareto are similar to the previous ones. The KMCE estimator still outperforms the CHL and BGN estimators. This indicates that when the tail becomes heavier, the benefits of using the KMCE estimator instead

			Log-normal	Log-Pareto		Weibull	Tr. Logist.
				p=.7	p=.3		
$N = 50$	L_1	KMCE	.1821	.1713	.1664	.1855	.1732
		CHL	.1811	.1815	.1860	.1955	.2009
		BGN	.1927	.1860	.1895	.1852	.1892
	L_2	KMCE	.1402	.1130	.1099	.1420	.1065
		CHL	.1326	.1311	.1617	.1338	.1183
		BGN	.1456	.1151	.1267	.1391	.1257
	WISE	KMCE	.1391	.1139	.1299	.1178	.1281
		CHL	.1316	.1385	.1693	.1288	.1546
		BGN	.1431	.1347	.1588	.1217	.1403
	E	KMCE	.0373	.0760	.1474	.0313	.0480
		CHL	.0388	.1109	.2295	.0359	.0574
		BGN	.0402	.0974	.2024	.0338	.0523
$N = 100$	L_1	KMCE	.1363	.1287	.1236	.1393	.1294
		CHL	.1381	.1412	.1423	.1533	.1468
		BGN	.1451	.1383	.1413	.1426	.1578
	L_2	KMCE	.1047	.0862	.0837	.1084	.0786
		CHL	.1018	.1021	.1244	.1044	.1000
		BGN	.1100	.0855	.0972	.1079	.0924
	WISE	KMCE	.1039	.0859	.0958	.0886	.0977
		CHL	.1018	.1068	.1276	.1029	.1078
		BGN	.1093	.1004	.1191	.0939	.1241
	E	KMCE	.0268	.0572	.1073	.0224	.0344
		CHL	.0289	.0816	.1677	.0277	.0391
		BGN	.0297	.0716	.1572	.0255	.0443

Table 2.2: The estimated error measures for sample size 50 and 100 based on 2000 repetitions.

of the CHL and the BGN estimators become greater. For the lognormal-Pareto distribution, the parameter c in the KMCE estimator tends to 0 when N increases. Comparing the estimated αs for the KMCE estimator, we observe that the αs are around 1.4-1.8 for the 70% lognormal-30% Pareto distribution, whereas they are around 1.2-1.3 for the 30% lognormal-70% Pareto distribution. This is due to the fact that the 30% lognormal-70% Pareto distribution has a heavier tail than the 70% lognormal-30% Pareto distribution.

For the light-tailed Weibull distribution, we can see that the KMCE, the CHL and the

			Log-normal	Log-Pareto		Weibull	Tr. Logist.
				p=.7	p=.3		
$N = 500$	L_1	KMCE	.0786	.0676	.0646	.0831	.0745
		CHL	.070	.0786	.0786	.0869	.0915
		BGN	.0761	.0703	.0761	.0763	.0791
	L_2	KMCE	.0585	.0480	.0470	.0676	.0437
		CHL	.0560	.0581	.0685	.0594	.0591
		BGN	.0590	.0454	.0601	.0588	.0574
	WISE	KMCE	.0585	.0471	.0517	.0530	.0598
		CHL	.0555	.0587	.0685	.0604	.0718
		BGN	.0579	.0507	.0686	.0510	.0585
	E	KMCE	.0125	.0306	.0591	.0111	.0171
		CHL	.0143	.0405	.0808	.0149	.0235
		BGN	.0145	.0342	.0974	.0129	.0120
$N = 1000$	L_1	KMCE	.0659	.0530	.0507	.0700	.0598
		CHL	.0572	.0606	.0609	.0688	.0730
		BGN	.0684	.0541	.0584	.0583	.0587
	L_2	KMCE	.0481	.0389	.0393	.0582	.0339
		CHL	.0435	.0450	.0528	.0476	.0521
		BGN	.0454	.0360	.0509	.0453	.0434
	WISE	KMCE	.0481	.0384	.0417	.0450	.0501
		CHL	.0434	.0453	.0524	.0478	.0561
		BGN	.0448	.0390	.0539	.0394	.0437
	E	KMCE	.0094	.0251	.0492	.0084	.0126
		CHL	.0107	.0295	.0588	.0113	.0178
		BGN	.0108	.0255	.0780	.0096	.0143

Table 2.3: The estimated error measures for sample size 500 and 1000 based on 2000 repetitions.

BGN estimators show good performance. The KMCE estimator is 5-20% worse on L_2 compared to the CHL estimator, and about 10% better with respect to WISE. This means that the KMCE estimator near 0 is worse than the CHL estimator, whereas the KMCE estimator is better than the CHL estimator in the tail. As compared to the BGN estimator, the KMCE estimator also gives a similar performance.

For the truncated logistic distribution, the KMCE and the BGN estimators show good performance. The bad performance of the CHL estimator is due to the fact that the transformation functions in this case always starts at 0 when $\alpha > 1$. The

estimator therefore transforms the true distribution, which has a positive value at 0, with a function that is 0 at 0, and this gives a positive value divided by 0, which results in a bad fit near 0. We see that the CHL estimator underestimates the true distribution around 0 for all values of N. The KMCE estimator also underestimates the true density around 0, but when N increases, the error around 0 decreases. We have also seen that the KMCE estimator overestimates the tail, which is because the transformation function has a heavy Pareto tail.

The main conclusion of our simulation study is that the KMCE estimator is recommended for heavy tailed situations.

We have designed the simulation study to be comparable to the simulation study in Clements et al. (2003). They compared their estimator to the transformation kernel density estimator with the modified-power transformation function porposed by Wand et al. (1991), which we call WMR, and the transformation kernel density estimator with the iterated transformation function suggested by Yang and Marron (1999), which we called YM. We can therefore use their simulation study to compare our estimator and the BGN estimator with the WMR and the YM estimators, even though the CHL simulation study only compares the L_1 and the L_2 error measures, and not error measures that emphasize the tail.

The CHL simulation study in Clements et al. (2003) shows that the CHL estimator performs well in general compared to WMR and YM. But for some distributions, the CHL estimator is outperformed by one of the other estimators. For the heavy-tailed 70% lognormal-30% Pareto distribution, the CHL estimator is outperformed with respect to L_2 by the YM estimator, but the performance of the KMCE estimator in our simulation study is even better than that of the YM estimator in the heavy-tail situation. For the Weibull distribution, the WMR estimator still gives very good performance compared to both the CHL and KMCE estimators. On the other hand, we are also able to make a comparison between the BGN estimator and the WMR and the YM estimators: the BGN estimator outperforms the WMR and the YM estimator in all situations investigated in the CHL simulation study.

2.5 Data study

In this section, we will apply our semiparametric estimation method to two data sets. The first data set contains automobile claims from a Spanish insurance company, and the second data set is about employer's liability from an Irish insurance company. The first data set was analyzed in detail in Bolancé et al. (2003). It is a typical insurance claims amount data set: it contains a lot of observations and it seems to be heavy-tailed. Unlike the automobile insurance, the liability data set from Ireland is rather light-tailed. The reason is that claims are undeveloped, i.e., large claims are underrepresented in this data set because they take longer to process.

2.5.1 Automobile claims

We study bodily injury payments from automobile accidents occuring in Spain in 1997. The data are divided into two age groups: claims from policyholders who are less than 30 years old, and claims from policyholders who are 30 years old or older. The first group of the data set consists of 1061 observations in the interval $[1; 126000]$ with mean value 402.7. The second group consists of 4061 observations in the interval $[1; 17000]$ with mean value 243.1. Estimation of the parameters in the modified Champernowne distribution function is, for young drivers $\widehat{\alpha}_1 = 1.116$, $\widehat{M}_1 = 66$, $\widehat{c}_1 = 0.000$, and for older drivers $\widehat{\alpha}_2 = 1.145$, $\widehat{M}_2 = 68$, $\widehat{c}_2 = 0.000$, respectively. The bandwidths are $b_1 = 0.172$ and $b_2 = 0.134$. Figure 2.4 presents the classical kernel density estimator of the transformed data separated in the two age groups. We notice that $\alpha_1 < \alpha_2$, which indicates that the data set for young drivers has a heavier tail than the data set for older drivers.

Figure 2.5 shows the resulting KMCE estimator for the two groups of policyholders. The claims have been split into three categories: *Small claims* in the interval $(0; 2000)$, *moderately sized claims* in the interval $[2000; 14000)$, and *extreme claims* in the interval $[14000; \infty)$. The figure illustrates that the tail in the estimated density of young policyholders is heavier than the tail of the estimated density of older

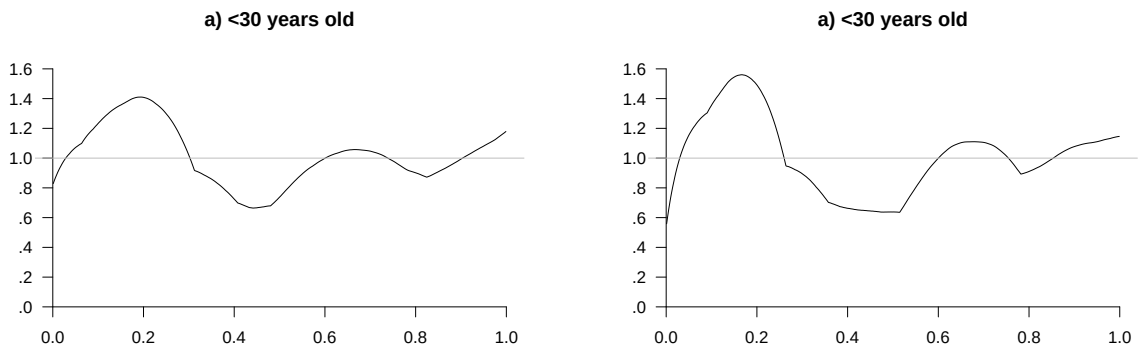

Figure 2.4: Classical kernel density estimator of the transformed automobile claims separated into policyholders < 30 years old and > 30 years old from an insurance company.

policyholders. This can be taken as evidence that young drivers are more likely to claim a large amount so that they should pay a higher premium than older drivers. Therefore the method is useful to identify high risk groups, i.e. those having more extreme claims. The usefulness of the methodology is specially interesting in this point. It alows to plot the estimated density in regions where data are scarce. If risk groups (such as young drivers or type of vehicles) are plotted separately, the density estimates inform about the risk orderings (i.e., which type of customers are likely to claim an extreme cost).

2.5.2 Employer's liability

In this section, we will apply our semiparametric estimation method to the costs of employer's liability from an Irish insurance company. The data set consists of 2522 claims. Here we want to see the effect of not including the additional c parameter in the transformation. The estimation of the parameters in the modified Champernowne distribution is $\widehat{\alpha} = 1.955$, $\widehat{M} = 32379.307$, $\widehat{c} = 64758.614$ and bandwidth $b = 0.147$. When c is assumed equal to 0, then $\widehat{\alpha} = 0.954$ while $\widehat{M}$ is the same because it corresponds to the sample median. Figure 2.6 presents the classical kernel density estimator of the transformed data, using two different values of c.

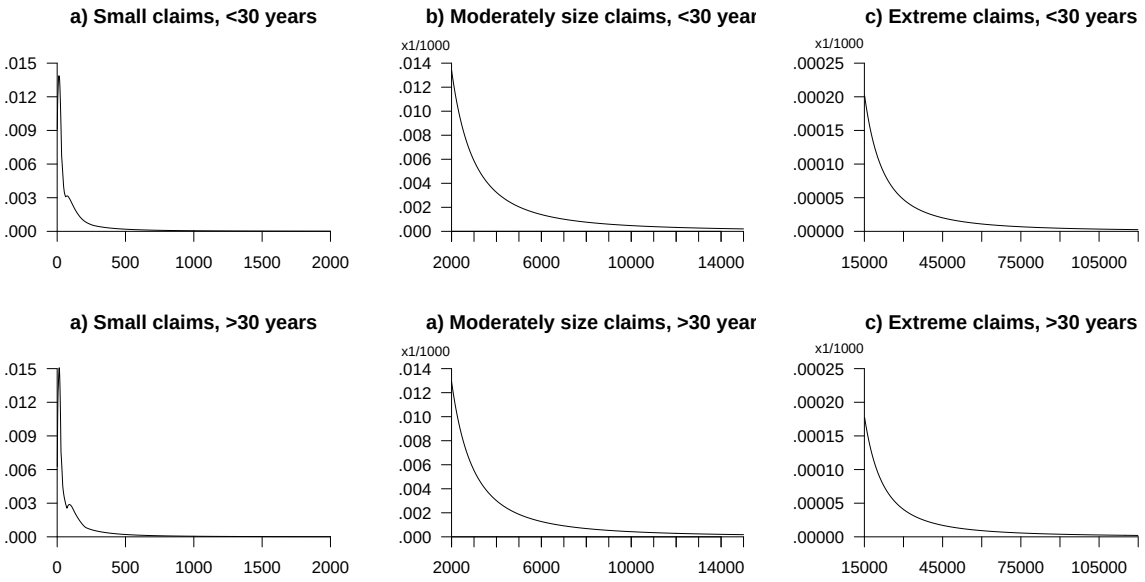

Figure 2.5: KMCE estimator of automobile claims from an insurance company, claims are separated into policyholders < 30 years old and > 30 years old, split into three groups.

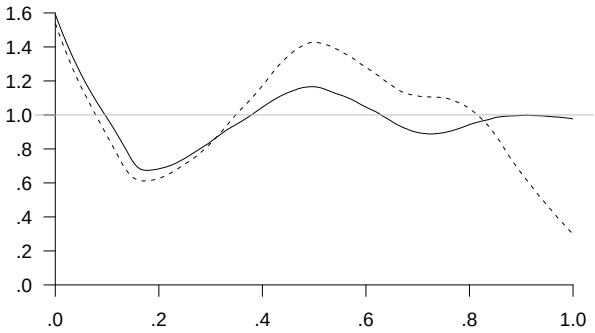

Figure 2.6: Classical kernel density estimation of the EL data transformed with estimated Champernowne ($c = 0$) distribution (dashed line) and modified Champernowne ($c > 0$) distribution (solid line).

In Figure 2.7, we plot the estimators on the original scale. The estimators are nearly identical for small and moderate claims (low costs), whereas the KMCE with

$c = 0$ overestimated the tail. This shows the importance of considering the modified Champernowne distribution.

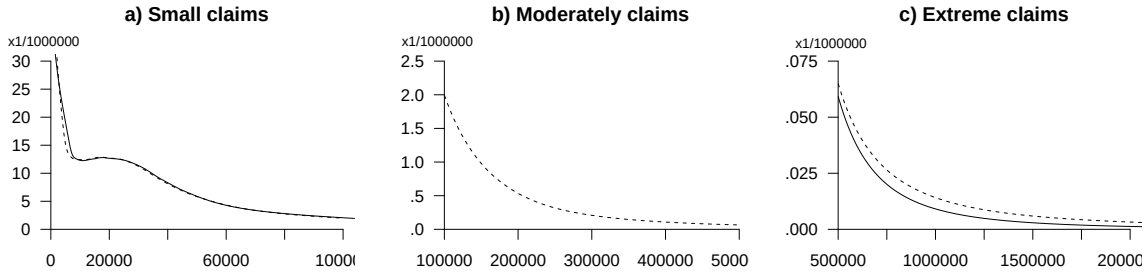

Figure 2.7: KMCE with $c = 0$ (dashed line) and KMCE with $c > 0$ (solid line) estimation of the EL data, separated into three disjoint intervals.

2.6 Conclusion

In this work, we have introduced an alternative method for estimating loss distributions. The method, which we have called a semiparametric transformation kernel density estimator, is based on a parametric estimator that is subsequently corrected with a nonparametric estimator. When we have a lot of information, the estimator is close to a nonparametric estimator, whereas it is close to a parametric estimator when we have little information.

The Champernowne distribution has an inflexible shape near 0, and we have generalised the distribution to the modified Champernowne distribution, which is heavy-tailed as well. We have used this modification for the transformation kernel density estimator.

The KMCE estimator turns out to perform very well compared to existing transformation kernel density estimators. The estimators were compared on simulated data. The KMCE estimator is the only estimator that performed well for all distributions. Therefore, the KMCE estimator is a basis for a unified approach that can be used for all kinds of data.

In insurance companies today, many analyses of loss distributions are based on parametric estimation. Our results show that our proposed method can overcome many disadvantages: in parametric estimation, the analyst must decide on a parametric model and a parameter estimation method. Insurance data sets are often large and the true distribution of real data rarely follows a simple known parametric distribution. We claim that there is no need to separate small and large claims. We believe that the use of our unified method results in an estimation of loss distributions that is very straightforward and can be usefull in practice.

Chapter 3

Nonparametric estimation of operational risk losses adjusted for underreporting

This chapter is an adapted version of Buch-Kromann et al. (2007).[1]

Not all claims are reported when a data base for financial operational risk is created. The probability of reporting increases with the size of the operational risk loss and converges towards one for big losses. Losses in operational risk have different causes and usually follow a wide variety of distributional shapes. Therefore, a method for modelling operational risk based on one or two parametric models is deemed to fail. In this paper we introduce a semiparametric method for modelling operational risk that is capable of taking underreporting into account and being guided by prior knowledge of the distributional shape.

[1] Winning paper of the Operational Risk & Compliance Achievement Award 2007.

3.1 Introduction

This paper contributes to Basel II and to the advanced measurement approach that seems to be part of Solvency II[2]. In Solvency II, insurance companies are allowed to lower their solvency requirement if they can provide internal high-quality models that convincingly show that the standardized methods are overly prudent. Solvency II is based on three pillars that correspond to the pillars in Basel II. In pillar I, concerning capital requirements, the Minimum Capital Requirements (MCR) and Solvency Capital Requirements (SCR)[3] are based on the total risks in the insurance company, which consists of insurance risk, credit risk, market risk and operational risk. While Solvency II specifies standadized methods for the two capital requirements, insurance companies are allowed to reserve capital below the SCR if internal statistical models show that this is prudent. This provides an incentive for insurance companies to spend money and energy on internal models. In this paper we combine two recent developments in operational risk, namely a statistical analysis of the quantitative impact of the failure to report all operational risk claims, and the recent development of smoothing methods that are capable of estimating nonparametric distributions with heavy tails. These two devolopments are described in details below.

When estimating operational risk losses it is a major obstracle that not all losses are observed. To estimate such an underreporting function from the data itself is an incredibly complicated mathematical deconvolution problem, and the rate of convergence of the deconvoluted estimators is often very poor. On top of that, the deconvoluted estimators often rely too heavily on the underlying assumptions about the underreporting function. Inspired by Freedman (1991), we therefore decided to "put our shoes on" (Freedman's phrase) and go out in the world and collect the crucial

[2]Solvency II is the European regulation structure that specifies how insurance companies must manage their risks and reserve prudent solvency margins.

[3]SCR is the capital requirement of the insurance company. FSA increases its control when capital gets below SCR. When capital gets below the MCR, FSA can take over the management of the insurance company and perhaps even stop for new underwriting of business.

data that we need in order to improve our estimation. After extensive interviews with experts in Royal&SunAlliance, a major non-life insurance company, and after a qualitative decision process, we deduced our best guess of an underreporting function. We will show in this paper that an underreporting function created in this way simplifies the theoretical problems and yields a solution that is closely related to what we would have gotten if we had observed all the losses without underreporting. This empirically based function of the expected severity of underreporting for each loss level was introduced by Guillen et al. (2007), but in the purely parametric case, without any non-parametric smoothing. In order to couple the method proposed by Guillen et al. (2007), we note that Buch-Larsen et al. (2005) provided a new approach to nonparametric smoothing that was particularly designed to estimate distributions with heavy tails close to the Pareto type. According to the advanced measurement approach, institutions must have sound estimates of all quantiles up to 99.9%. The institution must furthermore maintain rigorous procedures for developing operational risk models and validating the model. Basel II specifies guidelines and recommendations for the use of external loss data, which underline the importance of using relevant external data, especially when there is reason to believe that the institution is exposed to infrequent, yet potentially severe, losses. The guidelines are first and foremost concerned with documentation of the events that led to the losses. For example, Basel II requires information on actual loss amounts, on the scale of business operations where the event occurred, on the causes and circumstances of the loss events, and other information that could help in assessing the relevance of the loss event for other institutions; similar demands are made for the treatment of internal loss data. Another demand in Basel II is that it must be easy to place the internal data set in the relevant supervisory categories, and that the data set must be provided to the supervisors upon request.

In this paper we use a nonparametric smoothing technique to estimate the distribution of operational losses when underreporting is taken into account. The method is applied to a data base of operational risk from financial institutions with six major business lines. In the database, there is sufficient data in each business line to avoid

the credibility approach of Gustafsson et al. (2006a,b). However, it is possible to combine their credibility technique with our nonparametric smoothing method in order to extend the method to more sparse data sets. We apply our proposed method to the six lines of operational risk claims and combine an internally estimated frequency of expected reported claims with an externally estimated distribution of operational risk losses. The externally available database of operational risk is from financial institution because there are not yet any reliable data on operational risk in the insurance industry. Insurance companies are therefore forced to use operational risk data from other financial institutions. The transformation approach to nonparametric smoothing, originally proposed by Wand et al. (1991), has recently received a fair amount of attention in the particular case of estimating actuarial or financial loss distributions, see Bolancé et al. (2003); Clements et al. (2003); Buch-Larsen et al. (2005); Buch-Kromann (2006); Hagmann et al. (2005); Hagmann and Scaillet (2007).

3.2 Setting up a model for the sampling of operational risk claims with underreporting

Underreporting means that not all operational risk claims in the company are reported. An underreporting function encodes the likelihood that a loss of a particular size is being reported. Because the probability of reporting increases with the size of the operational risk claim, the density of the observed losses in the reported data set is more heavy-tailed compared to the density of all operational risk claims. See Guillen et al. (2007) for further details. In the following we set up a model that first defines all the operational risk claims that have occurred - even though not all of them have been reported - and then models the statistical relationship between the actually reported claims and the total number of claims.

Assume that M independent identically distributed (iid) operational risk claims, $(X_i)_{1\leq i\leq M}$, with density g have occurred where M is a stochastic Poisson(λ)-distributed variable. Since we do not observe all these M claims, let $I(i)$ be an indicator function

taking the value 1 if the i'th claim is observed and 0 otherwise, and let $(I(i))_{1\leq i\leq M}$ be *iid* random variables indicating reported and unreported claims. The random variable $N = \sum_{i=1}^{M} I(i)$ is therefore the reported number of claims. Let $(Y_j)_{1\leq j\leq N}$ be the N reported claims from the operational risk data set and assume that these N claims given $N = n$ are *iid* with density f. We assume furthermore that the underreporting function u only depends on the value of the claim

$$u(x) = P(I(1) = 1|X_1 = x)$$

Under this model the probability of observing an operational risk claim can be written as

$$P_{u,g} = \int_0^\infty g(w)u(w)\,\mathrm{d}w$$

As a result, the random variable N is Poisson distributed with mean $\lambda P_{u,g}$. The relationship between the density of the reported operational risk claims and the density of all operational risk claims, is

$$f(y) = \frac{g(y)u(y)}{P_{u,g}}. \tag{3.1}$$

We model (3.1) with a parametric g as an *apriori* model to obtain a model corrected in a nonparametric way. The non-parametric correction is obtained for $N = n$ by using the fact that if we transform the reported data set $(Y_j)_{1\leq j\leq n}$ with the cumulative distribution function $F(y) = \int_0^y f(w)\,\mathrm{d}w$ we obtain a data set, $Z_j = F(Y_j)$, $j = 1, ..., n$ with density h, where h is a uniform density.

3.3 A transformation approach to tail flattening accounting for underreporting

We wish to find an appropriate nonparametric smoothing estimator for the density g of our operational risk claims. By doing so, we will be able to adjust appropriately for underreporting by means of nonparametric smoothing. Adjusting is always nontrivial in nonparametric smoothing. In Jones et al. (1994) there is an extensive discussion of the difference between adjusting for the design internally (inside the integral) or externally (outside the integral) in the standard nonparametric estimation problem. There seems to be different methods to adjust for an underreporting function in a nonparametric way. We have picked the simplest possible method in terms of implementation and analysis. However, further research might lead to an improved method for the nonparametric correction for underreporting.

We have observations with density f. Based on (3.1) above we express g as a function of f and u in the following way:

$$g(x) = \frac{f(x)\left\{u(x)\right\}^{-1}}{\int_0^\infty f(w)\left\{u(w)\right\}^{-1}\,\mathrm{d}w}.$$

We know from Wand et al. (1991) and Buch-Larsen et al. (2005) that f can be estimated by

$$\widehat{f}(x) = \frac{1}{N}\sum_{j=1}^{N} K_{b,T(x)}\left(T(x) - T(Y_j)\right) T'(y),$$

where $K_{b,T(x)}$ is a kernel function with bandwidth b and boundary correction according to the point of estimation $T(x)$. Buch-Larsen et al. (2005) describes the details behind a standard correction based on the simple boundary correction procedure of density estimation, see Silverman (1986) or Wand and Jones (1995). An obvious estimator of g is therefore

$$\widehat{g}(x) = \frac{\widehat{f}(x)\left\{u(x)\right\}^{-1}}{\int_0^\infty \widehat{f}(w)\left\{u(w)\right\}^{-1}\,\mathrm{d}w}.$$

When we consider the asymptotic properties of this estimator, we notice that the asymptotic distribution of $\widehat{f}$ is well known. From Bolancé et al. (2003) and Buch-Larsen et al. (2005), we have:

Theorem 3.1. *Let the transformation function T be a two times differentiable known function. Assume that f is also two times differentiable. Then the bias of $\widehat{f}$ is given by*

$$\mathbb{E}\widehat{f}(x) - f(x) = \mu_2(K) B_x b^2 + o(b^2)$$

with $B_x = \left[\left\{ \frac{f(x)}{T'(x)} \right\}' \frac{1}{T'(x)} \right]'$ *and the variance is given by*

$$\mathbb{V}\left\{ \widehat{f}(x) \right\} = (nb)^{-1} R(K) T'(x) f(x) + o\left(\frac{1}{nb} \right).$$

where the asymptotics is given for $m \to \infty$ and $n = P_{u,g} \cdot m$ and $\mu_2(K) = 2^{-1} \int w^2 K(w)\, \mathrm{d}w$, $R(K) = \int K^2(w)\, \mathrm{d}w$.

We can derive the asymptotic properties of $\widehat{g}$ from the asymptotic properties of $\widehat{f}$. Let

$$A = f(x) \left\{ u(x) \right\}^{-1},$$
$$\widehat{A} = \widehat{f}(x) \left\{ u(x) \right\}^{-1},$$

$$B = \int_0^\infty f(w) \left\{ u(w) \right\}^{-1} \mathrm{d}w,$$

and

$$\widehat{B} = \int_0^\infty \widehat{f}(w) \left\{ u(w) \right\}^{-1} \mathrm{d}w.$$

Then

$$\widehat{g}(x) - g(x) = \widehat{A}\widehat{B}^{-1} - AB^{-1} = \widehat{B}^{-1} \left(\widehat{A} - A \right) - A\widehat{B}^{-1} B^{-1} \left(\widehat{B} - B \right).$$

Therefore $\widehat{g}(x) - g(x)$ is equivalent from an asymptotic point of view to

$$B^{-1}\left(\widehat{A} - A\right) - AB^{-2}\left(\widehat{B} - B\right).$$

Based on this quick ordering of terms, we can write up the asymptotic theory of $\widehat{g}(x)$. We omit the proof, which is based on the above theorem and the fact that the variance of $\widehat{B}$ is of lower order of magnitude due to the integration, while the integrated bias of $\widehat{B}$ still is of the original order b^2. From this theoretical result, we get two main conclusions for our adjustment for undersmoothing. Firstly, the bias is affected by the way the adjustment is carried out. Secondly, the standard deviation of the final estimator is increased by a local element, the square-root of the underreporting function and by a global element, the square-root of an average of the underreporting function. One can also verify that in the situation where the underreporting function is incorrect, the adjustment method simply results in a biased estimator, where g is different from the function we would like to obtain. However, the theoretical results below are still valid in that situation.

Theorem 3.2. *Let the transformation function T and the underreporting function u be two times differentiable known functions. Assume that g is also two times continuously differentiable. Then the bias of $\widehat{g}$ is*

$$\mathbb{E}\widehat{g}(x) - g(x) = \mu_2(K)b^2\left[B^{-1}B_x\left\{u(x)\right\}^{-1} - AB^{-2}\int B_w\left\{u(w)\right\}^{-1}\,\mathrm{d}w\right] + o(b^2)$$

and the variance is given by

$$\mathbb{V}\left\{\widehat{g}(x)\right\} = \left\{u(x)B\right\}^{-1}(nb)^{-1}R(K)T'(x)g(x) + o\left\{(nb)^{-1}\right\}.$$

3.3.1 The data set

We use a publicly available database with more than 5,000 financial operational risk events from a range of global organizations. For each operational risk event we have information on date, location, loss category and a description of the event. The

Risk category	Number of losses	Maximum loss (£M)	Sample median (£M)	Sample mean (£M)	Standard deviation	Annual frequency
1	1247	6683.8	1.82	32.24	269.43	10
2	538	910.6	2.14	15.60	69.68	20
3	721	221.9	1.98	7.84	20.04	28
4	45	117.6	5.88	22.46	33.25	11
5	2395	39546.4	2.35	74.91	1192.55	3
6	75	104.6	1.56	7.39	17.72	52

Table 3.1: Number of reported operational risk losses in our external data base.

reported operational risk events are categorized into six different event risk categories:

1. Internal fraud.
2. External fraud.
3. Employment practices and workplace safety.
4. Business disruption.
5. Damage to physical assets.
6. Execution, delivery and process management.

As seen in Table 3.1 the number and the severity of the losses differ considerably in the six categories: the number of losses range from 45 events to 2,395, and the loss amounts range from just over £100 million to almost £40 billion. As with most operational risk data sets, the mean is significantly larger than the median which indicates that the distribution of operational risk events is right skewed. Since the external database lacks a reliable estimate of the annual frequency of each event risk category, these are estimated using scenario analysis and are presented in the 7^{th} column in Table 3.1. The scenario analysis is based on the presumed risk of Royal&SunAlliance.

3.4 Estimating the tail flattening transformation and the underreporting function

We use the six underreporting functions proposed by Guillen et al. (2007) based on expert judgements, see Figure 3.1. These underreporting functions are based on parametric modelling and a lot of aggregated experience, and they can therefore be seen as known functions with respect to our asymptotic theory in the previous section. In the same way, we assume our parametrically defined transformation function is known. This type of argument is well known in semiparametric density estimators like ours, see Hjort and Glad (1995); Hjort and Jones (1996); Buch-Larsen et al. (2005). We consider the same three parametric models as Guillen et al. (2007) and estimate them by maximum likelihood. The transformations we use are the parametric cdfs produced by integrating the parametric densities, shown below, see Buch-Larsen et al. (2005) for more details.

- The Champernowne distribution, see Brown (1937) and Champernowne (1952), was generalised in Buch-Larsen et al. (2005). The latter has density

$$f_{\theta_i}(x) = \frac{\alpha_i \left(x + c_i\right)^{\alpha_i - 1} \left((M_i + c_i)^{\alpha_i} - c_i^{\alpha_i}\right)}{\left((x + c_i)^{\alpha_i} + (M_i + c_i)^{\alpha_i} - 2c_i^{\alpha_i}\right)^2} \tag{3.2}$$

  with $\theta_i = \{\alpha_i, M_i, c_i\}$.

- The lognormal distribution

$$f_{\eta_i}(x) = \frac{e^{-\frac{1}{2}\left(\frac{\log x - \mu_i}{\sigma_i}\right)^2}}{x\sigma_i\sqrt{2\pi}} \tag{3.3}$$

  with $\eta_i = \{\mu_i, \sigma_i\}$.

- The Weibull distribution

$$f_{\varsigma_i}(x) = \frac{\gamma_i}{\beta_i}\left(\frac{x}{\beta_i}\right)^{\gamma_i - 1} e^{-\left(\frac{x}{\beta_i}\right)^{\gamma_i}} \tag{3.4}$$

$\varsigma_i = \{\gamma_i, \beta_i\}$.

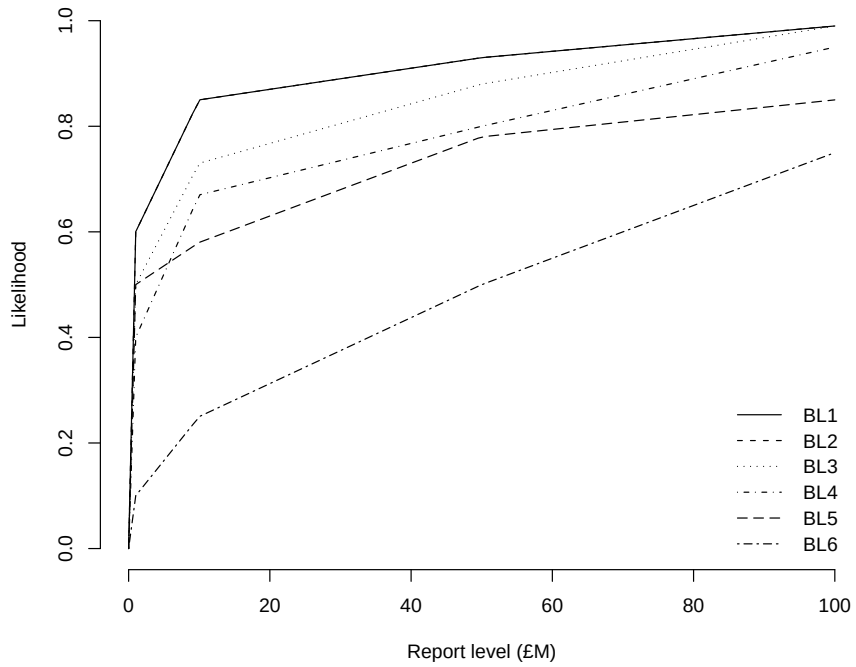

Figure 3.1: The estimated underreporting functions for each of the six business lines. The underreporting functions give the probability of a loss to be reported as a function of the size of the loss.

3.4.1 Aggregated analysis incorporating all six business lines

We use Monte Carlo-simulation to calculate the 99.5% Value-at-Risk (VaR) and Tail-Value-at-Risk (TVaR) for our various versions of estimated distributions. The VaR-measure gives us insight into expected maximal losses for risk tolerance $\alpha = 0.995$, and can be defined by:

$$\mathrm{VaR}_\alpha(S) = \sup\{s \in \mathbb{R} \mid \mathbb{P}(S \leq s) \leq \alpha\}.$$

The TVaR gives us the expectation of the area above the risk tolerance level α and is defined by:

$$\text{TVaR}_\alpha(S) = \mathbb{E}[S \mid S \geq VaR_\alpha(S)].$$

The VaR measure is a common risk measure. However, in contrast to TVaR, VaR is not a coherent risk measure (see Artzner et al. (1999) and Artzner (1999) for a definition of coherent risk measures and a detailed study of VaR and TVaR). That means that VaR does not always fulfill the important property of subadditivity, which loosely speaking means that one will never benefit from splitting up a risk.

Our chosen values of α are inspired by Basel II, which specifies standards within the advanced measurement approach. See Wirch (1999) on VaR and other risk measures. When simulating, we draw $10,000$ operational claims numbers for each risk category using the frequencies from our scenario analysis as our Poisson parameters, see Table 3.1.

$$r_{i,j} \sim Po(\lambda_i), \quad i = 1, 2, ..., 6 \;\; j = 1, 2, ..., 10,000.$$

Then for each of the $60,000$ simulated number of operational risk claims, the $r_{i,j}$´s, we draw $r_{i,j}$ independent identically distributed random variables by means of our nonparametric estimators of the distributions of operational risk claims. First we sample $r_{i,j}$ uniform distributions:

$$v_{i,j,k} \sim U(0,1), \quad k = 1, 2, ..., r_{i,j}.$$

We then calculate the simulated aggregated claim amount for the i^{th} risk category and j^{th} simulation:

$$x_{i,j} = \sum_{k=1}^{r_{i,j}} \widehat{F}_i^{-1}(v_{i,j,k})$$

where $\widehat{F}_i^{-1}$ is the inverse of the nonparametrically estimated cumulative distribution function $\widehat{F}_i$ for the i^{th} risk category. Our value of risk is then based on the 10,000

values of

$$x_{i,\cdot} = \sum_{j=1}^{6} x_{i,j}.$$

3.5 Results

The differences between the distribution assumption with and without kernel smoothing and with and without underreporting is illustrated in Figure 3.2, where we present results for the fourth risk category.

The left-hand graph present the four estimators with the generalized Champernowne distribution as parametric model. By adding the underreporting correction (dashed line) one obtain a much heavier tail compared to the pure parametric approach (solid line). A kernel adjustment (dotted line) on the parametric start have more or less the same appearance in the beginning of the body but demonstrate a lighter tail then the pure parametric density. The fourth estimator (dot-dash line) incorporate both the underreporting effect and kernel adjustment on the parametric start. This estimator present a higher probability that a small loss will occur for this event risk category then the pure parametric model, and consequently becomes more light tailed. The middle and the right-hand graph present the same four estimators but with lognormal- and Weibull distribution as parametric models. These graphs are interpreted analogously as the left-hand graph.

Table 3.2 presents the corrected frequencies for each risk category. The annual frequencies for Royal&SunAlliance are presented in the first row of the table. The six following rows give the corrected frequencies, based on different distribution assumptions with adjustment for the underreporting effect. The abbreviation Ch.UR is the generalised Champerknowne distribution adjusted for underreporting, Ch.UR.KS is the generalised Champerknowne distribution adjusted for underreporting and kernel smoothing.

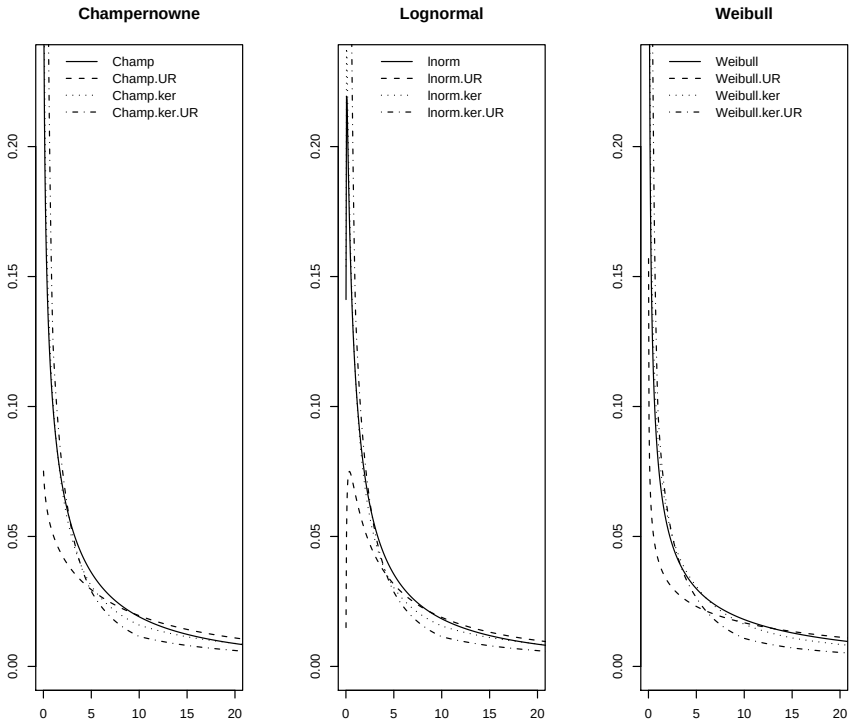

Figure 3.2: The four different estimated models on event risk category 4 with three different distribution assumptions. The solid line represent the pure parametric density, the dashed line correspond to the estimator including the underreporting effect, the dotted line is the semiparametric estimator and the dotdash line incorporates both the underreporting correction and the kernel adjustment.

Table 3.3 presents the total operational loss of the institution; the mean, median, standard deviation; and the 99.5% VaR and TVaR based on different underlying distributions. We consider the three parametric models with and without correction for underreporting and with and without the nonparametric correction based on kernel smoothing. We normalise all results by the results one obtains by using the parametric Weibull distribution without any kind of correction, which seems to be the most popular model among practitioners at the moment.

From Table 3.3 we see that incorporating underreporting clearly increases both the

	Risk Category					
	1	2	3	4	5	6
Unadjusted	10	20	28	11	3	52
Ch.UR	14.4	28.2	46.8	17.9	5.7	290.6
Ch.UR.KS	13.7	27.6	44.9	18.2	5.4	270.4
Ln.UR	14.1	28.1	45.3	17.6	5.6	251.7
Ln.UR.KS	14.1	29.1	48.1	22.7	5.3	391.3
We.UR	13.6	28.6	47.9	22.4	5.3	386.6
We.UR.KS	14.2	33.7	49.3	24.7	5.6	400.3

Table 3.2: The reported frequency for each risk category and the adjusted risk frequencies after adjusting for underreporting with and without the nonparametric correction.

	Mean	Sd	Median	VaR-99.5%	TVaR-99.5%
We	1	1	1	1	1
We.KS	1.71	2.09	1.71	1.88	1.96
We.UR	2.30	1.56	2.34	1.93	1.90
We.UR.KS	1.92	2.21	1.94	2.02	2.11
Ln	0.98	1.54	0.90	1.35	1.45
Ln.KS	1.64	2.26	1.63	1.83	1.84
Ln.UR	2.28	3.02	2.18	2.84	2.90
Ln.UR.KS	1.87	2.43	1.81	1.99	2.05
Ch	1.12	3.11	0.86	2.51	2.81
Ch.KS	1.56	2.14	1.54	1.75	2.14
Ch.UR	2.50	4.65	2.21	3.85	4.10
Ch.UR.KS	2.12	2.74	2.01	2.11	2.36

Table 3.3: The statistical data for the total loss amount. Normalised by the unadjusted Weibull distribution.

mean value, the median value, the standard deviation, the 99.5% quantile and the 99.5% TVaR. This is because more claims are being incorporated into our model when underreporting is taken into account. However, most of these claims are small, therefore the 99.5% quantile is less affected by taking underreporting into account than the other measures in Table 3. When underreporting is not accounted for, kernel smoothing has a tendency to correct the tail into a heavier tail which increases most of the considered measures of risk, while kernel smoothing has the opposite effect when we account for underreporting. It seems that when underreporting is present a major correction is necessary in order to have a sufficient small claim mass in the distribution. This correction takes mass from the tail of the distribution and moves it to the smaller values of the distribution. It is also clear from Table 3.3 that while different parametric models give very different answers, our kernel-smoothed correction has a stabilising effect; it is clear that this stabilising effect affects the quantile estimation as well as the tail value at risk estimation. It does not really matter very much which of the three parametric models we use for our pilot study when a stabilising kernel smoothed correction is performed. However, the choice of parametric model seems to be crucially important, if one decides to stick to a purely parametric approach. One can for example conclude that one gets estimates of the exposure to operational risk that are too optemistic if one uses the widely used parametric Weibull distribution without correcting for underreporting and without a nonparametric correction based on kernel smoothing. We therefore recommend that regulators and practitioners start looking for other approaches with more realistic estimates of the tail behavior of actuarial loss distributions.

Chapter 4

Estimation of large insurance losses: A case study

This chapter is an adapted version of Buch-Kromann (2006).

This paper demonstrates an approach to analyzing liability data recently developed by a Danish insurance company. The approach is based on a Champernowne distribution, which is corrected with a nonparametric estimator. The correction estimator is obtained by transforming the data set with the estimated modified Champernowne cdf and then estimating the density of the transformed data set by using the classical kernel density estimator. Our approach is illustrated by applying it to an actual data set.

4.1 Introduction

This paper demonstrates a unified approach to large loss estimation recently developed in a Danish insurance company. A unified approach was needed because

actuaries and statisticians were spending too much time trying to develop parametric models of losses. Thus they often decided to estimate small and large losses separately because no single parametric model seemed to fit both small and large losses. Apart from the usual challenges such as choosing the appropriate parametric model and identifying the best way of estimating the parameters, a big problem was in determining the threshold between small and large losses, if they are to be estimated separately. Clearly the solution to this problem is fundamentally important to the quality of the estimation.

One approach is to use extreme value theory and generalized Pareto distributions, as described in Embrechts et al. (1997) and Cebrián et al. (2003), to analyze the loss data. As this approach, however, is mainly concerned with the estimation of large losses, it maintains the necessity to determine the threshold between small and large losses.

The approach adopted by the Danish insurance company is based on the work of Buch-Larsen et al. (2005) who developed a unified method based on a semiparametric estimator, i.e., a parametric estimator corrected with a nonparametric correction estimator.[1] The semiparametric estimator is obtained by transforming the data set with the transform function, $T(x)$, which is the cdf of a modified Champernowne distribution. If $X_1, ..., X_N$ represent the data set then the transformed data set is $Z_1, ..., Z_N$ where $Z_i = T(X_i)$, for $i = 1, ..., N$. The density of the transformed data set is estimated by means of a classical kernel density estimator (Wand and Jones, 1995, p. 11):

$$\widehat{g}(z) = \frac{1}{Nb} \sum_{i=1}^{N} K\left(\frac{z - Z_i}{b}\right) \tag{4.1}$$

[1]Semiparametric estimators were introduced in the statistics literature by Wand et al. (1991) who demonstrated that the classical kernel density estimator could be improved by transforming the data set with a shifted power transformation. Since then semiparametric estimators have been used by other authors including Hjort and Glad (1995); Jones et al. (1995); Yang and Marron (1999) and Bolancé et al. (2003). Clements et al. (2003) have developed semiparametric estimators based on a Möbius-like transformation, which is a special case of the Champernowne distribution. This method was further developed by Buch-Larsen et al. (2005) using a modified Champernowne distribution for greater flexibility.

where K is the kernel function, b is the bandwidth. The estimator for the original data set is obtained by an inverse transformation of $\widehat{g}(z)$. This results in an estimator that is close to a parametric estimator for small values of N and "more" nonparametric as N increases. The estimator $\widehat{g}(z)$ is flexible in that it provides good estimates for many different shapes of loss distributions.

In this paper we will provide a detailed outline of the Buch-Larsen et al. (2005) method, which we have called the corrected modified Champernowne method. In addition, we will introduce an alternative parameter estimation method, called the QM method, which provides better estimates of conditional right-tail expected losses compared to those based on maximum likelihood parameter estimation. Moreover, we compare the corrected modified Champernowne method to the generalized Pareto distribution method of Cebrián et al. (2003)

4.2 Estimation of parameters

The modified Champernowne distribution is a generalization of the Champernowne distribution (Brown (1937); Champernowne (1952)) with an extra parameter c to ensure that the pdf of the modified Champernowne distribution is positive at 0 for all α when $c > 0$ and is zero when $c = 0$. The modified Champernowne distribution is defined as:

$$T_{\alpha,M,c}(x) = \frac{(x+c)^{\alpha} - c^{\alpha}}{(x+c)^{\alpha} + (M+c)^{\alpha} - 2c^{\alpha}} \tag{4.2}$$

for $x \geq 0$, with parameters $\alpha > 0, M > 0$ and $c = 0$ and density

$$\frac{\mathrm{d}T_{\alpha,M,c}(x)}{\mathrm{d}x} = \frac{\alpha(x+c)^{\alpha-1}((M+c)^{\alpha} - c^{\alpha})}{((x+c)^{\alpha} + (M+c)^{\alpha} - 2c^{\alpha})^2} \tag{4.3}$$

The inverse cdf is

$$T^{-1}_{\alpha,M,c}(x) = \left[\frac{z(M+c)^{\alpha} - (2z-1)c^{\alpha}}{1-z}\right]^{1/\alpha} - c \tag{4.4}$$

Buch-Larsen et al. (2005) have shown that the modified Champernowne distribution is a heavy-tailed distribution that converges to a Pareto distribution in the tail.

Two estimation methods are used for the parameters a, M, and c of the modified Champernowne distribution: the well-known maximum likelihood method and the quantile-mean method, which selects parameters in a way that emphasizes the goodness-of-fit in the right tail. As $T_{\alpha,M,c}(M) \equiv 0.5$ for all c and α, M is assumed to be equal to the empirical (sample) median in both of these methods. Although this gives a sub-optimal estimate of M, Clements et al. (2003) have argued that it is reasonable to assume that the empirical median is close to the maximum likelihood estimate of M. The empirical median has a further advantage: it is a robust estimator, especially for heavy tailed distributions Lehmann (1991). After the parameter M has been estimated, the estimate of (a, c) is found by each of the methods. The **maximum likelihood estimate** (MLE) is found by maximizing the log likelihood function:

$$\begin{aligned} l(\alpha, c) &= N \log \alpha + N \log((M+c)^\alpha - c^a) + (\alpha - 1) \sum_{i=1}^{N} \log(X_i + c) \\ &\quad -2 \sum_{i=1}^{N} \log(X_i + c)^\alpha + (M+c)^\alpha - 2c^a \end{aligned}$$

The properties of the MLE are well known: it is efficient and ensures the best fit over the entire range of the distribution.

Because the risk of large losses lies in the tail of the loss distribution, we have also tested the **quantile-mean method**, which is a heuristic parameter estimation method. In this method we first select the parameter α so that the 95 quantile point of the empirical or sample cdf and of the estimated modified Champernowne distribution are equal. The parameter c is then chosen so that the mean of the estimated modified Champernowne distribution is as close as possible to the empirical mean.

Though there may be better ways of choosing α and c, it is important to choose

parameters that result in accurate estimates of the number of large losses and the mean because these statistics are important in determining premiums.

4.3 An illustration of density estimation

The data are losses (claims) from employer's liability line of business at Royal & SunAlliance, a British company. The data consist of 34,493 losses ranging from £0 to £4,213,057 without truncations or censoring, i.e., before deductibles and policy limits are applied. The use of untruncated and uncensored loss data is critical to the application of the proposed method.[2] The average loss size is £26,597. The employers are subdivided into 13 trade groups as shown in Table 4.1. For each trade group, the problem is to calculate the expected loss size for a deductible of d (left truncation) and a policy limit (or retention limit) of u (right censoring) where $d < u$.

The employer's liability data set is heavy-tailed, which can be seen by the upward tendency of the empirical mean excess function in Figure 4.1 (left) and the concave departure of the exponential QQ-plot in Figure 4.1 (right).

Table 4.1 shows the MLE and QM estimates of the parameters for the liability data set for each trade group. The M-parameters for MLE and QM are equal because they are estimated in the same way. For the α parameters, no clear tendency is seen, whereas the c-parameters seem to be larger with the QM-method than with the MLE-method.

The estimation method proposed by Buch-Larsen et al. (2005), called the corrected modified Champernowne (CMC) method, is demonstrated by applying it to the data set. The CMC method is essentially a semiparametric transformation kernel density estimator, which is computed by transforming the data set with a modified Champernowne distribution, and applying a nonparametric classical kernel density estimator to the transformed data set. The kernel smoothing function is a correction

[2]For an analysis of losses with truncation and censoring see, for example, Cebrián et al. (2003); Denuit et al. (2006).

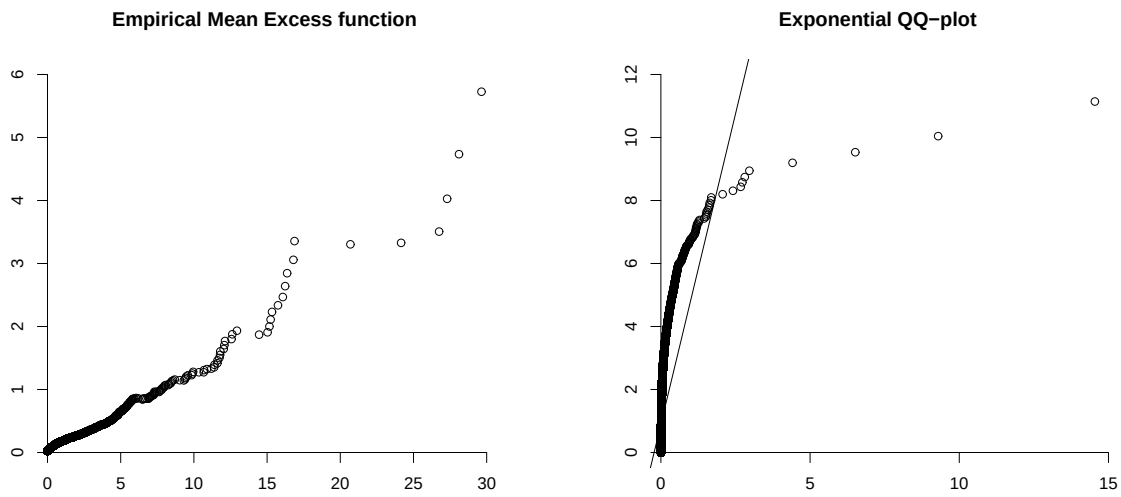

Figure 4.1: Empirical mean excess (left) and exponential QQ-plot (right).

to the parametric modified Champernowne transformation function. Because of the properties of kernel smoothing, the correction will be weak if there are few data points and becomes more pronounced as the sample size increases. This means that the transformed kernel density estimator resembles a parametric estimator for small sample sizes and a nonparametric estimator for larger sample sizes.

Let $X_1^i, ..., X_{N_i}^i$ be the data set with sample size N_i for trade group i with an unknown cdf $F_i(x)$ and density $f_i(x)$. We will use a detailed numerical illustration for trade group 1 only, where $N_1 = 1668$. Figure 4.2 illustrates the four steps of the CMC estimation with QM parameters of f_1.These steps are described in general as follows:

Step 1: Estimate the parameters (α, M, c) of the modified Champernowne distribution as described in Section 4.2 by using either the MLE or QM method. These estimates are displayed in Table 4.1. Figure 4.2(1) shows a histogram for the raw data for trade group 1 and the estimated modified Champernowne distribution with QM-parameters (dotted line).

Step 2: Transform the data set $X_1^i, ..., X_{N_i}^i$ into $Z_1^i, ..., Z_{N_i}^i$ using $Z_j^i = \widehat{T}_i(X_j^i)$ for

Trade	Sample	MLE Estimates			QM Estimates		
Group t	size N_i	$\widehat{\alpha}_{\text{MLE}}$	$\widehat{M}_{\text{MLE}}$	$\widehat{c}_{\text{MLE}}$	$\widehat{\alpha}_{\text{QM}}$	$\widehat{M}_{\text{QM}}$	$\widehat{c}_{\text{QM}}$
1	1,668	1.610	13,616	6,808	1.400	13,616	27,232
2	597	1.401	12,437	0	1.653	12,437	24,874
3	2,112	1.563	8,532	0	1.563	8,532	4,266
4	537	1.563	8,867	0	1.808	8,867	17,733
5	1,083	1.726	9,596	0	1.774	9,596	4,798
6	2,054	1.888	8,777	4,388	1.913	8,777	17,554
7	707	1.458	9,744	0	1.455	9,744	19,487
8	3,620	2.108	8,858	4,429	1.967	8,858	13,287
9	931	1.481	9,423	0	1.629	9,423	14,135
10	6,297	1.935	9,268	4,634	1.950	9,268	13,902
11	1,022	1.656	11,041	0	1.562	11,041	0
12	5,668	1.865	10,629	5,315	1.934	10,629	21,259
13	8,197	1.574	10,790	5,395	1.493	10,790	21,581

Table 4.1: Estimated modified Champernowne parameters for each trade group

$j = 1, ..., N_i$ where $T_{\widehat{\alpha}_i, \widehat{M}_i, \widehat{c}_i}(x) \equiv \widehat{T}_i(x)$ is given in equation (4.2). Figure 4.2(2) shows the histogram for the transformed trade group 1 data.

Step 3: If the unknown distribution $F_i(x)$ is a modified Champernowne distribution, the transformed data set will be uniformly distributed.[3] Even if $F_i(x)$ is not a modified Champernowne distribution, however, the transformed data set is usually close to a uniform distribution because the modified Champernowne distribution is fitted to the data set. Under the assumption that the transformed distribution is close to a uniform distribution on $(0, 1)$, we can use a constant bandwidth when computing the correction estimator by means of a classical kernel density estimator for $Z_1^i, ..., Z_{N_i}^i$:

$$\widehat{g}_i(z) = \frac{1}{N_i \cdot k_{b_i}(z)} \sum_{j=1}^{N_i} K_{b_i}(z - Z_j^i) \tag{4.5}$$

where $K_{b_i}(\cdot)$ is the Epanechnikov kernel function defined in equation (4.8). $k_{b_i}(z)$ is the boundary correction, which is needed because the Z_j^i's are con-

[3]Uniformity can be tested with a chi-square test or Kolmogorov-Smirnov test.

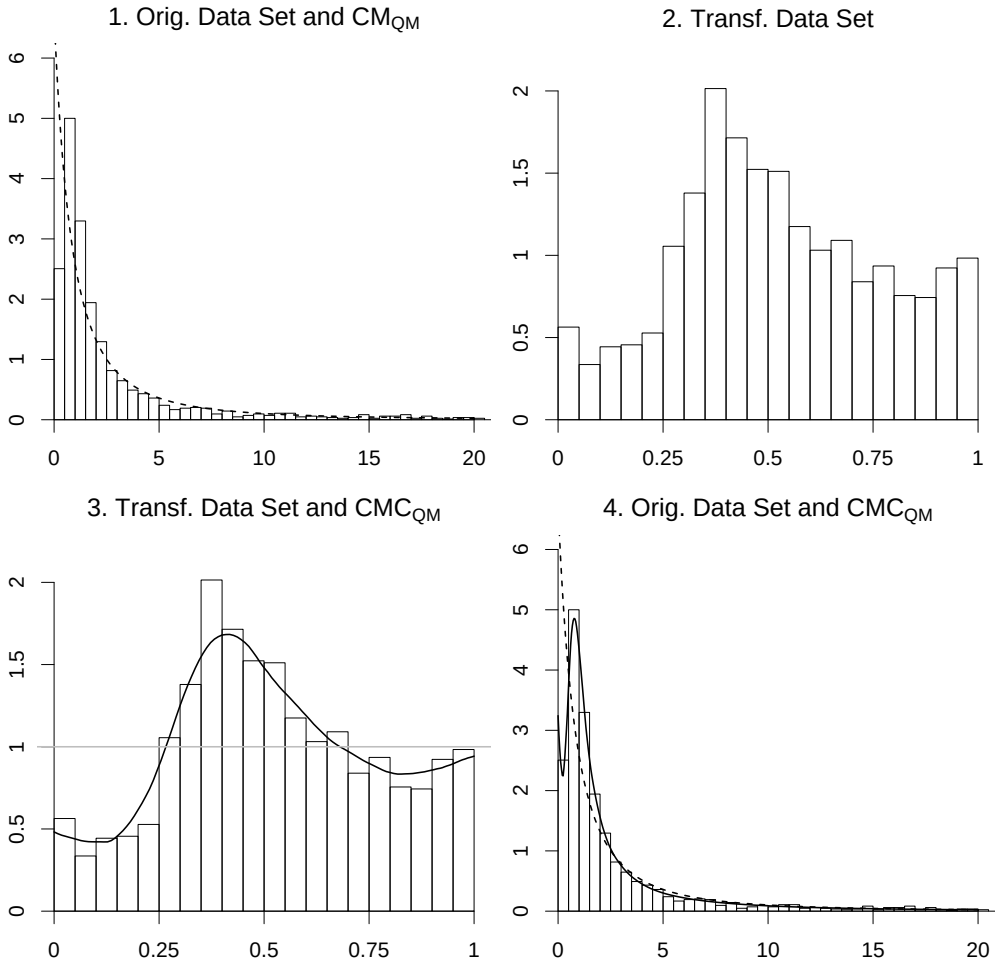

Figure 4.2: Steps in density estimation using the CMC transformation with QM parameter estimates for trade group 1. Dashed lines corresponds to the modified Champernowne distribution with QM parameters, $\mathrm{CM}_{\mathrm{QM}}$, and black lines corresponds to the corrected modified Champernowne distribution with QM parameters, $\mathrm{CMC}_{\mathrm{QM}}$.

strained on the interval $(0, 1)$. The boundary correction $k_{b_i}(z)$ is defined as

$$k_{b_1}(z) = \int_{\max\left(-1,-\frac{z}{b_1}\right)}^{\min\left(1,\frac{1-z}{b_1}\right)} K(u)\,du,$$

The kernel estimator is illustrated in Figure 4.2(3). Notice that near 0, the kernel estimator is below 1, which means that the resulting estimator for f_1 is lower than the density of the estimated modified Champernowne distribution from Step 1. In the interval from 0.25 to 0.6, the kernel estimator is above 1, which means that the kernel estimator has raised the modified Champernowne distribution

Step 4: The kernel estimator, $\widehat{g}_i$, can be interpreted as the final estimator on the transformed axis. The estimated density for the original data set $X_1^i, ..., X_{N_i}^i$ is obtained by an inverse transform such that

$$\widehat{f}_i(x) = \frac{\widehat{g}_i\left(\widehat{T}_i(x)\right)}{\left|\frac{\mathrm{d}\widehat{T}_i^{-1}}{\mathrm{d}z}\Big|_{z=\widehat{T}_i(x)}\right|} \tag{4.6}$$

The resulting estimator for the data from trade group 1 is shown in Figure 4.2(4). The corrected modified Champernowne estimator (solid line) seems to provide a better estimate for the data set than the uncorrected modified Champernowne distribution (dotted line) from Step 2.

These steps can be summarized into the following expression for the final estimator for f_i:

$$\widehat{f}_i(x) = \frac{1}{N \cdot k_b(\widehat{T}(x))} \sum_{i=1}^{N} K_b(\widehat{T}(x) - \widehat{T}(X_i))\widehat{T}'(x) \tag{4.7}$$

As mentioned in Step 3, the Epanechnikov kernel function is used in the kernel estimator. This kernel function is the optimal kernel with respect to efficiency (Wand and Jones, 1995, page 31), i.e., for a fixed number of observations, the Epanechnikov kernel function leads to a better kernel estimator than any other kernel function. The Epanechnikov kernel function has the form

$$K(x) = \begin{cases} \frac{3}{4}(1 - x^2) & \text{if } -1 < x < 1 \\ 0 & \text{otherwise} \end{cases} \tag{4.8}$$

and for bandwidth b

$$K_b(x) = \frac{1}{b} K\left(\frac{x}{b}\right)$$

The choice of bandwidth determines the smoothness of the estimator. The simple normal scale bandwidth selection is used (Wand and Jones, 1995, page 60)

$$b = \left(\frac{40\pi^{\frac{1}{2}}}{N}\right)^{\frac{1}{5}} \widehat{\sigma},$$

where N is the number of observations and $\widehat{\sigma}$ is the standard deviation; this is optimal when g is a normal distribution. For fixed $\widehat{\sigma}$, the bandwidth is decreasing when N increases, and vice versa. Thus, a small data set results in a large bandwidth and a great amount of smoothing in the kernel estimator, and hence a small correction. This ensures that the final estimator $\widehat{f}(x)$ is close to the modified Champernowne distribution from step 1. A large data set results in a small bandwidth, and hence a potentially stronger correction by the kernel estimator to the modified Champernowne distribution from step 1. The asymptotic behavior of the transformation kernel density estimator is described in Buch-Larsen et al. (2005).

Table 4.2 shows the values of the Kolmogorov-Smirnov tests for the modified Champernowne distributions $\mathrm{MC}_{\mathrm{MLE}}$ and $\mathrm{MC}_{\mathrm{QM}}$ from step 1 and the corresponding CMC distributions $\mathrm{CMC}_{\mathrm{MLE}}$ and $\mathrm{CMC}_{\mathrm{QM}}$ are stated for each trade group. In almost all trade groups, the test does not reject the modified Champernowne distribution from step 1 with MLE parameters, whereas the QM parameters result in a rejection in more than half of the trade groups, using 0.05 as the rejection threshold. This confirms the well-known result that MLE produces the best overall fit. However, the test neigher rejects the kernel-smoothed $\mathrm{CMC}_{\mathrm{MLE}}$ estimates with MLE parameters, nor the $\mathrm{CMC}_{\mathrm{QM}}$ estimates with QM parameters.

Next we demonstrate the calculation of conditional means. To avoid numerical problems,[4] all calculations are performed on the transformed axis. We first estimate the

[4]Problems often arise in numerical integration over the interval 0 to ∞ (we assume the integral is convergent). Some (but not all) of these problems can eliminated by an appropriate transformation

Trade Group i	MC_{MLE}	MC_{QM}	CMC_{MLE}	CMC_{QM}
1	0.005	0.009	0.481	0.550
2	0.248	0.010	0.620	0.336
3	0.417	0.065	0.535	0.531
4	0.484	0.159	0.559	0.487
5	0.519	0.176	0.408	0.582
6	0.085	0.018	0.597	0.516
7	0.279	0.090	0.354	0.413
8	0.087	0.038	0.519	0.495
9	0.619	0.184	0.600	0.475
10	0.073	0.000	0.437	0.430
11	0.403	0.253	0.526	0.592
12	0.103	0.013	0.383	0.632
13	0.066	0.002	0.548	0.599

Table 4.2: Kolmogorov-Smirnov tests for corrected (CMC) and uncorrected modified Champernowne (MC)

conditional densities of losses from group i given that they are larger than the deductible. Let $F_j(x|X_j^i > d) = \mathbb{P}\left[X_j^i \leq x|X_j^i > d\right]$. It follows that

$$\widehat{F}_j(x|X_j^i > d) = \frac{\int_d^x \widehat{f}_i(y)\,\mathrm{d}y}{\int_d^\infty \widehat{f}_i(y)\,\mathrm{d}y} = \frac{\int_{\widehat{T}_i(d)}^{\widehat{T}_i(x)} \widehat{g}_i(z)\,\mathrm{d}z}{\int_{\widehat{T}_i(d)}^1 \widehat{g}_i(z)\,\mathrm{d}z} \tag{4.9}$$

where $\widehat{g}_i(z)$ is the classical kernel density estimator given in equation (4.5) and $\widehat{f}_i(x)$ is defined in equation (4.6). Let $X_j^i(d,u)$ denote the insurer's actual loss paid by the insurer that results from the loss X_j^i given a deductible d and a policy limit u, then

$$E\left[X_j^i(d,u)\right] = \frac{\int_d^u (x-d)\widehat{f}_i(x)\,\mathrm{d}x + (u-d)\int_u^\infty \widehat{f}_i(x)\,\mathrm{d}x}{\int_u^\infty \widehat{f}_i(x)\,\mathrm{d}x} \tag{4.10}$$

$$= \frac{\int_{\widehat{T}(d)}^{\widehat{T}(u)} \widehat{T}^{-1}(z)\widehat{g}_i(z)\,\mathrm{d}z + u\int_{\widehat{T}(u)}^1 \widehat{g}_i(z)\,\mathrm{d}z}{\int_{\widehat{T}(d)}^1 \widehat{g}_i(z)\,\mathrm{d}z} - d \tag{4.11}$$

In order to test the goodness-of-fit, we will now compute $R_i(d,u)$ and $S_i(d,u)$, which

so that the integration is done over the interval 0 to 1. For more on numerical integration see, for example, (Ralston and Rabinowitz, 1978, Chapter 4).

Tr.grp	0	25,000	50,000	100,000	250,000	500,000	1,000,000	2,500,000
1	46,395	103,932	158,935	247,935	476,618	783,787	1,207,821	1,519,513
2	32,272	69,969	109,668	175,914	357,761	619,579	1,013,470	1,399,530
3	20,165	59,234	97,517	170,610	370,772	651,681	1,062,555	1,435,935
4	19,717	55,965	87,462	143,640	302,572	539,661	913,017	1,331,167
5	18,350	44,742	73,808	132,056	298,340	542,768	924,039	1,342,388
6	18,469	53,439	79,196	128,825	274,763	496,533	855,227	1,288,418
7	27,659	82,559	132,448	217,471	439,452	737,475	1,157,059	1,490,929
8	17,954	44,303	69,050	117,155	257,922	472,998	825,105	1,266,293
9	21,805	62,074	101,939	169,801	357,251	623,971	1,023,078	1,408,028
10	18,882	47,763	72,662	120,355	262,505	479,694	833,971	1,273,022
11	22,930	49,061	88,242	163,350	365,036	647,448	1,060,089	1,435,471
12	23,759	54,219	81,856	130,384	273,758	492,010	846,952	1,281,096
13	32,430	88,206	138,624	216,229	425,908	714,913	1,128,817	1,473,290

Table 4.3: Conditional expected claims sizes estimated with CMC_{QM} for each trade group and deductible

are ratios of estimated and observed expected conditionals for each trade group, i.e.,

$$R_t(d) = \frac{\mathbb{E}\left[X_j^i(d,u)\right]}{\overline{X}_j^i(d,u)}, \qquad S_t(d) = \frac{\mathbb{E}\left[N_j^i(d)\right]}{\overline{N}_j^i(d)} \tag{4.12}$$

where, for trade group i with deductible d and policy limit u, $\overline{X}_j^i(d,u)$ is the observed conditional expected loss, $N_j^i(d)$ is the number of losses in excess of d, and $\overline{N}_j^i(d)$ is the observed number of losses in excess of d. Figure 4.3 shows plots of $R_1(d,u)$ and $S_1(d)$ for various values of d and $u = 5,000,000$. The parameters are estimated by means of the MLE-method in the two upper plots and by means of the QM-method in the two lower plots.

The plots of $S_t(d)$ show that both the MLE and QM parameters result in reasonable estimates of the number of observations. However, the plots of $R_t(d)$ show that the MLE parameters lead to underestimation of the expected loss in all trade groups, whereas the QM parameters are slightly better in this respect. This may be because MLE estimation assigns equal weight to small and large losses, whereas QM estimation places more emphasis on the tail, which has the biggest effect on the estimated loss. Thus, insurers would be wise to choose estimation methods that put greater

Tr.grp	0	25,000	50,000	100,000	250,000	500,000	1,000,000	2,500,000
1	44,435	99,421	150,588	208,369	364,572	660,494	744,944	242,939
2	35,084	80,771	124,326	207,293	279,611	359,043	180,221	0
3	21,469	66,863	102,769	147,010	168,918	267,415	0	0
4	20,515	62,918	79,133	116,311	89,598	0	0	0
5	20,145	55,599	91,734	114,229	124,775	358,410	0	0
6	21,268	73,225	103,454	150,448	196,683	198,835	0	0
7	28,320	86,489	148,584	172,529	193,729	191,193	0	0
8	19,554	54,378	88,113	107,760	152,140	154,640	33,850	0
9	26,281	92,743	153,164	213,622	224,949	351,758	533,632	0
10	20,813	59,815	94,689	156,765	190,388	209,242	200,246	0
11	32,765	97,685	202,911	389,410	1,699,379	2,124,883	3,022,845	6,792,342
12	24,865	60,025	92,774	133,077	209,587	850,056	803,610	464,448
13	34,128	97,010	152,635	220,197	441,802	835,375	1,592,551	4,550,394

Table 4.4: Observed conditional expected claim sizes for each trade group and deductible

emphasis on the tail losses. Notice that the bottom half of Figure 4.3 shows that the underestimation of the conditional mean is less distinct for the CMC_{QM}. The CMC_{QM} estimators are therefore used to estimate the conditional expected loss for each trade group and for various deductibles; they are shown in Table 4.3 while the actual observed average losses are in Table 4.4. For a general insurance company, these statistics can be used to estimate the rates within each trade group. To continue this illustration, let us compare the corrected modified Champernowne estimation procedure with the generalized Pareto distribution approach (GPD) as exemplified by Cebrián et al. (2003). A loss from trade group i is said to follow a generalized Pareto distribution if its cdf is given by

$$F_i(x) = \begin{cases} 1 - (1 + \xi_i x)^{-\frac{1}{\xi_i}} & \text{if } \xi_i \neq 0 \\ 1 - e^{-x} & \text{if } \xi_i = 0 \end{cases} \tag{4.13}$$

for $\xi_i, x > 0$.

According to Cebrián et al. (2003), we must find the threshold u separating small and large losses by means of one or more graphical tools: (i) an empirical mean excess function plot, (ii) a GPD index plot, or (iii) a Gertensgarbe plot. In the empirical mean excess function plot, the empirical mean excess function is approximately linear

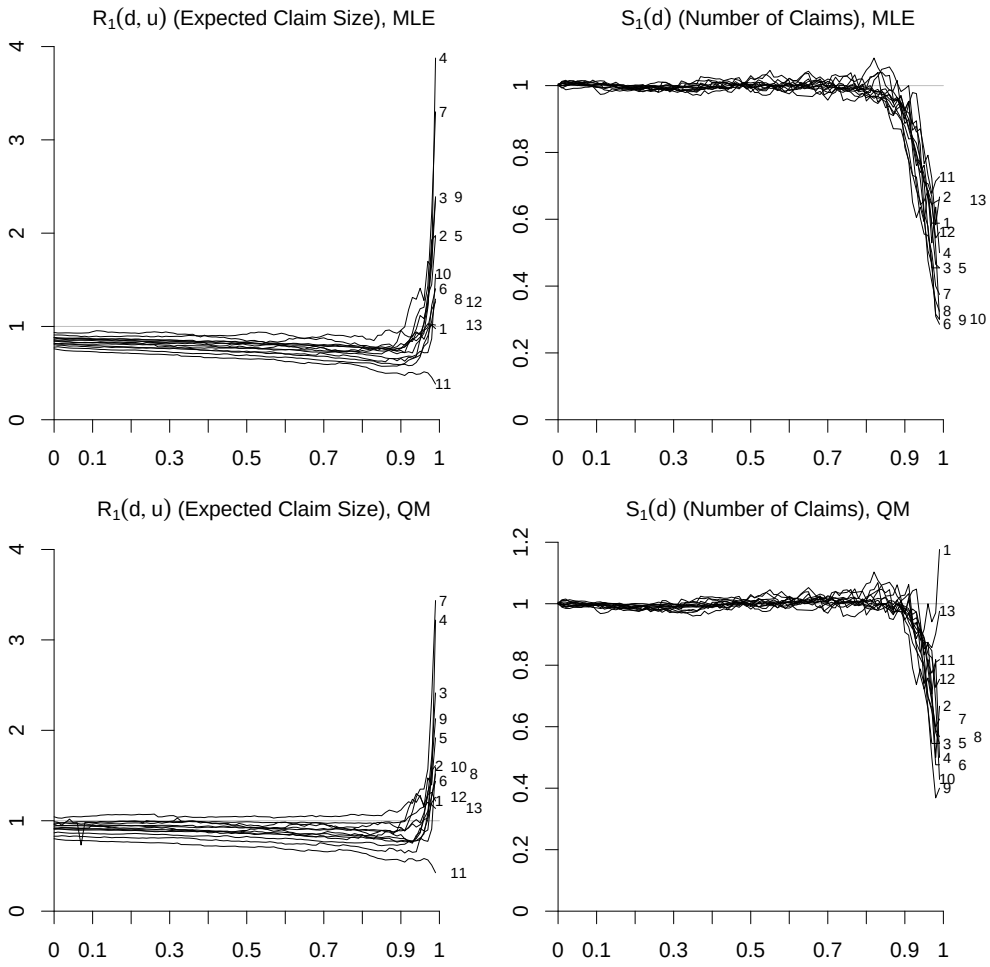

Figure 4.3: Comparing ratios $R_1(d, u)$ (left plots) and $S_1(d)$ (right plots) using MLE and QM methods vs. quantiles.

for $x \geq u$. In the GPD index plot, we compute the maximum likelihood estimator for increasing thresholds, and identify u as the point from which the MLE estimator becomes approximately constant. The Gertensgarbe plot is based on the assumption that the extreme threshold can be found as a change point in the ordered series of claims, and that the change point can be identified by means of a sequential

version of the Mann-Kendall test as the intersection point between a normalized progressive and retrograde rank statistics. The progressive and retrograde curve in the Gertensgarbe plot, however, do not in all cases produce an intersection point: in particular, our data set did not lead to an intersection point, and our choice of thresholds is therefore based on the first two methods. Figure 4.4 shows the GPD index plot and the empirical mean excess plot for trade group 1. In the GPD index plot the chosen threshold corresponds to the 85% quantile where there are 251 observations exceeding the threshold. In the empirical mean excess plot the chosen threshold is 53,571. Table 4.5 shows the chosen thresholds in quantile terms (u_{quan}), in absolute terms (u_{value}), and in number of observations exceeding the threshold (u_{exc}), as well as the estimated GPD parameters, and the Kolmogorov-Smirnov test probabilities. Table 4.5 shows that the estimated GPD's are not rejected by the Kolmogorov-Smirnov tests in any trade group.

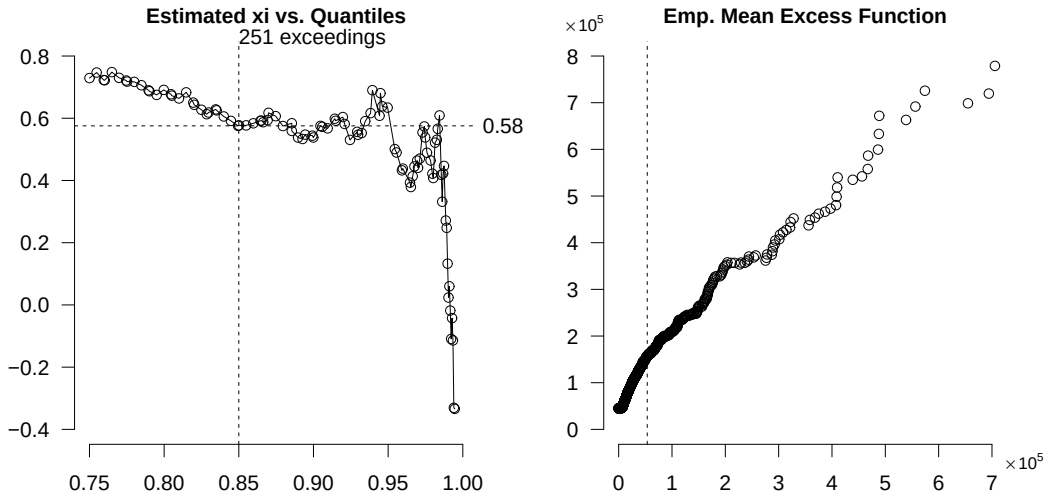

Figure 4.4: GPD index plot (left) and empirical mean excess plot (right) for Trade Group 1.

Table 4.6 displays the conditional means for various deductibles using the estimated GPD parameters. If we compare the conditional expected losses estimated by means of GPD and CMC_{QM} in Tables 4.6 and 4.3, respectively, with the observed conditional expected losses in Table 4.4, we notice that the GPD estimates are closer to the

Trade	Threshold			Parameters		K-S
Group i	u_{quan}	u_{value}	u_{exc}	$\widehat{\xi}$	$\widehat{\beta}$	Test
1	85.0%	53,571	251	0.576	72,494	0,696
2	56.0%	14,621	263	0.876	15,348	0,629
3	90.5%	39,040	201	0.537	48,625	0,769
4	88.0%	28,840	65	0.309	50,974	0,810
5	95.3%	68,107	51	0.149	91,930	0,760
6	90.5%	38,897	196	0.525	49,691	0,570
7	91.0%	48,315	64	0.318	102,541	0,642
8	94.0%	54,866	218	0.257	68,954	0,567
9	95.5%	96,062	42	0.21	164,404	0,770
10	88.0%	31,888	755	0.612	32,577	0,434
11	84.0%	28,339	164	0.787	22,821	0,645
12	95.0%	87,678	284	0.372	75,536	0,490
13	90.0%	57,966	820	0.538	73,313	0,612

Table 4.5: The thresholds, the estimated parameters and the Kolmogorov-Smirnov tests for GPD.

observed means in approximately half of the trade groups, the CMC_{QM} estimates are closer in three others, and the GPD and CMC_{QM} estimates are similar in the others. GPD estimation, however, has some obvious disadvantages:

- It cannot be used to estimate conditional means when the deductible is smaller than the threshold. In such cases the distribution for small losses must be estimated separately;
- No automatic procedure exists for finding the optimal threshold; and
- The GPD only works for heavy-tailed data. For moderately light tails (like the lognormal distribution), GPD estimation will often result in an estimator with finite support Buch-Larsen et al. (2005).

The final phase of the illustration is the validation phase. Whereas a goodness-of-fit test measures how well the estimation fits claims in the data set, a validation study measures how well the method predicts future claims. Therefore, to get a better comparison of the CMC and GPD methods, the data set is randomly partitioned

Tr.grp	0	25,000	50,000	100,000	250,000	500,000	1,000,000	2,500,000
1	$< u_1$	$< u_1$	$< u_1$	275,744	435,556	668,774	1,027,518	1,386,478
2	$< u_2$	147,621	217,969	342,505	642,293	1,005,860	1,453,782	1,654,455
3	$< u_3$	$< u_3$	155,549	207,875	356,409	577,695	929,360	1,326,089
4	$< u_4$	$< u_4$	96,106	118,417	185,071	294,617	502,810	920,995
5	$< u_5$	$< u_5$	$< u_5$	125,509	151,751	195,476	282,747	526,584
6	$< u_6$	$< u_6$	153,549	203,988	347,741	563,296	909,159	1,310,492
7	$< u_7$	$< u_7$	172,940	195,924	264,350	375,953	584,624	977,379
8	$< u_8$	$< u_8$	$< u_8$	127,276	178,946	264,538	431,202	808,021
9	$< u_9$	$< u_9$	$< u_9$	234,495	274,149	339,849	468,285	768,983
10	$< u_{10}$	$< u_{10}$	149,165	215,580	398,335	658,551	1,047,326	1,415,323
11	$< u_{11}$	$< u_{11}$	197,056	299,797	559,587	892,057	1,327,369	1,586,321
12	$< u_{12}$	$< u_{12}$	$< u_{12}$	178,230	264,357	402,829	653,364	1,072,024
13	$< u_{13}$	$< u_{13}$	$< u_{13}$	255,703	401,255	617,947	961,323	1,338,851

Table 4.6: Conditional expected losses for GPD with polity limit $u = 5,000,000$ and various deductibles. Notes: $< u_i$ denotes the deductible is smaller than the threshold.

into two parts: one for estimating model parameters and the other for validation. In other words, the first data set is used to estimate the CMC_{QM} and GPD parameters.

These estimated parameters are then used to calculate conditional expected losses under the CMC_{QM} and GPD methods, which are then compared to the observed conditional expected losses contained in the second data set. The validation study shows that in terms of prediction, which is essential for a general insurance company, the CMC_{QM} performs as well as the GPD method.

4.4 Summary and closing comments

When dealing with heavy-tailed loss distribution data, maximum likelihood estimation of parameters tends to lead to an underestimation of conditional expected losses. For this reason, an alternative, called the quantile-mean method (QM) of parameter estimation, was introduced. Buch-Larsen et al. (2005) corrected modified Champernowne method (CMC) is combined with the QM method to produce decent results. Comparing the CMC method with the generalized Pareto distribution (GPD) method shows that the GPD performs better than the CMC in terms of goodness-

of-fit, whereas our validation study shows that the two methods are comparable in terms of predicting future claims. The CMC method also has some advantages that makes it an attractive alternative compared to GPD: The CMC method estimates the density of the whole range of losses, whereas in GPD estimation, we need to estimate small and large losses separately, which involves finding a threshold from where the data set is GPD. This is normally done by graphical methods, which are difficult to automatize. Finally, the GPD can only be used for heavy-tailed distributions, whereas the CMC also works for lighter-tailed distributions because it always has infinite support. One area for further research is in improving the parameter estimation method and including more sophisticated boundary correction methods. For example, one can combine our work with the methods proposed by Chen (1999, 2000) and Scaillet (2004). We also hope to integrate insights from recent developments in density estimation, such as Hagmann and Scaillet (2007), and to extend our estimation method to handle covariates.

Chapter 5

Comparison of tail performance of the Champernowne transformed kernel density estimator, the generalized Pareto distribution and the g-and-h distribution

This chapter is an adapted version of Buch-Kromann (2009).

Several papers have recommended the Champernowne distribution to describe operational risk losses. This paper compares the tail performance of the Champernowne transformed kernel density estimator, the generalized Pareto distribution (gpd) and the g-and-h distribution. We introduce a new tail-dependent parameter estimation method for the Champernowne distribution, computed by conditional maximum likelihood, and show that, by using this new method, we obtain an estimator that in general outperforms the benchmark estimators with respect to tail performance. At

the same time the new estimator provides a density estimate on the entire axis superior to the g-and-h distribution, and unlike the gpd estimator, which provides a density estimate only above the threshold. The estimators performance are investigated in a Monte Carlo simulation study, and their application to operational risk is illustrated.

5.1 Introduction

Large loss estimation is a central problem in general insurance and occurs both in ordinary insurance portfolios as well as in operational risk estimation.

Regarding operational risk, Basel II provides three schemes for calculating reserves for operational risk ranging: the Basic Approach, the Standardized Approach and the Advanced Measurement Approach. The Advanced Measurement Approach provides the opportunity to use internal statistical models. An important model in this concern is the Loss Distribution Approach (LDA), see McNeil et al. (2005). The model describes the aggregated loss distribution and consists of a severity and a frequency distribution. Various papers deals with this model. In Moscadelli (2004) the underlying severity estimation is based on extreme value theory, whereas Dutta and Perry (2006) recommend the flexible parametric g-and-h distribtuion, which is further studied in Degen et al. (2007) and Degen and Embrechts (2008). Peters and Sisson (2006) takes its starting point in estimating operational risk with LDA as well, but in a Baysian approach, where the severity distribution is assumed to be either the g-and-h distribution or the GB2 distribution. The papers Gustafsson (2006), Gustafsson et al. (2006a), Gustafsson et al. (2006b), Buch-Kromann et al. (2007), Guillen et al. (2007) and Gustafsson and Nielsen (2008) all deal with estimating operational risk by use of the Champernowne distribution with a nonparametric correction.

The classical approach of large loss estimation is extreme value theory, as described in, for example, Embrechts et al. (1997). This theory is based on the fundamental

Fisher-Tippett theorem for maxima, which corresponds to the central limit theorem for sums. The Fisher-Tippett theorem shows that if there exists a limit distribution of maxima of random variables, suitably centered and normalized, then the limit distribution belongs to one of the three types of extreme value distributions. There is a close connection between the extreme value distributions and the generalized Pareto distribution (gpd), which describes the limit distribution of excesses over a high threshold: gpd estimation is the classical way of estimating large losses and the theory is widely used in insurance, see, for example, McNeil and Saladin (1997) and Cebrián et al. (2003).

In spite of the beautiful theoretical properties of extreme value theory, some problems appear, when using the theory in practical large loss estimation. The gpd is the limit distribution of excesses over a high threshold. Therefore, we have to choose from which threshold this assumption is reasonable. This is often done by graphical methods, which are inappropriate in some situations. Moreover, the gpd only describes the distribution of excesses over the threshold, and therefore it does not provide a distribution estimate of losses on the entire axis.

Buch-Larsen et al. (2005) introduced an alternative large loss estimation approach based on nonparametric statistics. They recommended an estimator based on the classical kernel density estimator

$$\bar{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_b(x - X_i)$$

where $X_1, ..., X_n$ is the data set, whose density we want to estimate, and $K_b(u) = K(u/b)/b$ is a kernel function with bandwidth b. They showed that, when introducing a tail flattening transformation, inspired by the work of Wand et al. (1991), for example, the Champernowne cdf with maximum likelihood estimated parameters (ml), this estimator has promising tail performance at the same time as being an estimator on the entire axis. When the transformation function is an estimated cumulative distribution function, this estimator corresponds to a purly parametric estimated distribution with a non-parametric correction, as described in Buch-Larsen

et al. (2005).

However, when focusing only on tail performance and large loss estimation, the estimator presented in Buch-Larsen et al. (2005) is not always as good as the gpd. Buch-Kromann (2006) illustrated some of the problems of the Champernowne transformed kernel density estimator with maximum likelihood parameters in a case study of an insurance liability data set, and the paper also introduced a heuristic parameter estimation method of the Champernowne distribution, which indicated that improvements in the tail fitting seem to be obtainable, when using a parameter estimation method, which emphasizes the tail.

In this paper we use the Champernowne transformed kernel density estimator introduced in Buch-Larsen et al. (2005) and introduce a more formalized parameter estimation method of the Champernowne distribution than the one introduced in Buch-Kromann (2006). We show that the tail performance of the Champernowne transformed kernel density estimator with conditional maximum likelihood (cml) parameters, in general outperforms the tail performance of the gpd estimator in the presented simulation study, whereas the Champernowne transformed kernel density estimator with maximum likelihood estimator does not. This means that the Champernowne transformed kernel density estimator with cml estimator seems to be a better estimator for large loss modeling than the gpd estimator and at the same time provides an estimator on the entire axis.

We benchmark our new Champernowne transformed kernel density estimator against the corresponding parametric (non-corrected) Champernowne distribution and against the estimated g-and-h distribution which has become popular in operational risk as a very flexible distribution with the ability to fit both the center and the tail of the distribution, see, for example, Dutta and Perry (2006). Moreover, we combine the transformation kernel density approach with the g-and-h distribution and compute the g-and-h transformed kernel density estimator. This estimator appears to be superior to the new Champernowne transformed kernel density estimator for large sample sizes, whereas the new Champernowne transformed kernel density estimator is supe-

rior for small sample sizes and for lighter tailed data sets (also large sample sizes). When focusing on the tail very far from the center, the Champernowne transformed kernel density estimator however seems to be superior in almost all situations.

In the simulation study we furthermore compare the estimators performance on the entire axis, called the from-ground-up (FGU) performance. The results show that the Champernowne transformed kernel density estimator with cml parameters has a superior FGU performance compared to the g-and-h distribution both with and without non-parametric correction. The estimator with the best FGU performance is the Champernowne transformed kernel density estimator with maximum likelihood parameters, but this estimator was the estimator with the worst tail performance, which makes this estimator less usable for operational risk, where the focus is on tail estimation. This means, that you pay a price on the FGU performance when using the cml parameters compared to the maximum likelihood parameters due to its special focus on tail performance. In return you get an estimator with a considerably better tail performance.

In the last part of the paper we illustrate the applications of the proposed estimators in a study of aggregated operational risk losses (LDA) and compare their estimated Value-at-risk (VaR) and Tail-value-at-risk (TVaR). We observe that VaR and TVaR of the estimated g-and-h distribution is considerably smaller than the VaR and TVaR of both the Champernowne transformed kernel density estimators and the gpd estimator. Comparison with the results from the simulation study, which showed that the tail performance very far from the center of the Champernowne transformed kernel density estimator with cml parameters was superior to the tail performance of the g-and-h distribution, causes us to doubt that the g-and-h distribution in this case gives us a prudent estimate of the operational risk.

The paper is organized as follows. The first three sections describes how to estimate large losses. In section 5.2, we introduce the transformation kernel density estimator. In section 5.3 we recall the properties of the Champernowne distribution, present the maximum likelihood parameter estimation method from Buch-Larsen et al. (2005)

and introduce the new approach of parameter estimation, the conditional maximum likelihood parameter estimation method, which specially focuses on better tail estimation. In the end of the section, we briefly mention the g-and-h distribution. In section 5.4, we recall the generalized Pareto distribution and some of its properties. Section 5.5 presents a simulation study based on a fire insurance data set which illustrates the performance of the proposed estimators, and section 5.6 is an application til operational risk comparing the proposed estimators. Section 5.7 is the conclusion.

5.2 Transformation kernel density estimators

Inspired by Wand et al. (1991), Buch-Larsen et al. (2005) showed that the tail performance of the kernel density estimator could be significantly improved by using a tail flattening transformation. The resulting transformation kernel density estimator has the form:

$$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_b\left\{T(x) - T(X_i)\right\} T'(x)$$

where $X_1, ..., X_n$ is the data set, whose density we want to estimate, $K_b(u) = K(u/b)/b$ is a kernel function with bandwidth b, for example, the Epanechnikov kernel function as we have used in the simulation study in section 5.5, and $T(x)$ is the transformation function.

When the transformation function returns values on a compact interval, for example, if this is a cdf, it is neccesary to have a boundary correction to ensure that the transformation kernel density estimator is a consistent estimator at the boundary. In this paper we use a simple *renormalization* method, as described in Jones (1993) which ensures that each kernel function integrates to 1. With the notation from Chen (1999) the transformation kernel density estimator with the renormalizing boundary correction is:

$$\widehat{f}(x) = \frac{1}{n\, a_{01}\{T(x), b\}}\sum_{i=1}^{n} K_b\{T(x) - T(X_i)\}T'(x) \tag{5.1}$$

where:

$$a_{sm}(x,b) = \begin{cases} \int_{-1}^{x/b} t^s K^m(t)\,dt, & x \in [0, 1-b] \\ \int_{-(1-x)/b}^{1} t^s K^m(t)\,dt, & x \in [1-b, 1]. \end{cases} \tag{5.2}$$

When the transformation function $T(x)$ is a cdf of a parametric distribution estimated to the data set under investigation, then the kernel density approach can be interpreted as a non-parametric correction to this estimated parametric distribution. Which parametric distribution to use, and how to estimate it, is of crucial importance and depends on the kind of estimation problem you have. In the following, some appropriate parametric distributions will be introduced.

5.3 Parametric distributions

This section introduces the Champernowne distribution which is a heavy-tailed, quite flexible three-parameter distribution. We outline both the maximum likelihood parameter estimation method, as studied in Buch-Larsen et al. (2005) and a new tail-dependent method, called the conditional maximum likelihood parameter estimation method. In the second part, we briefly introduce the g-and-h distribution which belongs to the class of extremely flexible parametric distribution. The g-and-h distribution is a four-parameter distribution and appears as a transformation of the standard normal distribution, see Dutta and Babbel (2002) and Dutta and Perry (2006).

5.3.1 The Champernowne distribution

In Buch-Larsen et al. (2005) the Champernowne distribution is proposed as transformation function. The Champernowne cdf has the form:

$$F_{\text{Champ}}(x) = \frac{(x+c)^\alpha - c^\alpha}{(x+c)^\alpha + (M+c)^\alpha - 2c^\alpha}, \qquad x \geq 0$$

with parameters (α, M, c), and the density function is:

$$f_{\text{Champ}}(x) = \frac{\alpha(x+c)^{\alpha-1}\{(M+c)^{\alpha} - c^{\alpha}\}}{\{(x+c)^{\alpha} + (M+c)^{\alpha} - 2c^{\alpha}\}^2}$$

The Champernowne distribution is a heavy-tailed distribution converging to a Pareto distribution:

$$\frac{f_{\text{Champ}}(x)}{f_{\text{Par}}(x)} \to 1 \qquad \text{as} \qquad x \to \infty \tag{5.3}$$

where $f_{\text{Par}}(x) = \frac{kx_0^k}{(x-\zeta)^{k+1}}$, $x \geq x_0$ with parameters $k = \alpha$, $x_0 = \{(M+c)^{\alpha} - c^{\alpha}\}^{1/\alpha}$ and $\zeta = 0$.

A crucial step when using the Champernowne distribution is the choice of parameter estimators. As described in Buch-Larsen et al. (2005), a natural way is to recognize that $F_{\text{Champ}}(M) = 0.5$ and therefore estimate the parameter M as the empirical median, and then estimate (α, c) by maximizing the loglikelihood function. The choice of M as the empirical median gives a stable estimator, especially for heavy-tailed distributions, and the maximum likelihood estimate of (α, c) ensures the best overall fit of the distribution. As mentioned in Clements et al. (2003), choosing M as the empirical median instead of choosing M by maximum likelihood estimation only makes a marginal difference and simplifies the parameter estimation procedure significantly. We will call the parameters obtained by this estimation method the *maximum likelihood parameters*, even though the method is only an approximation to the maximum likelihood procedure due to the way of estimating M. The Champernowne distribution with maximum likelihood parameters is a purely parametric estimator, and in the following it is called $\widetilde{h}_0(x)$. When the modifed Champernowne distribution with maximum likelihood parameters is used as transformation function in (5.1), or in other words when we estimate a non-parametric correction to $\widetilde{h}_0(x)$, we will denote the resulting estimator by $\widehat{h}_0(x)$.

However, parameters estimated by the maximum likelihood method might not be optimal, especially when one is interested in the estimation of the tail. Maximum likelihood parameters give the best overall fit to the data set. However, in the

tail, which we focus on, there are few data and therefore the maximum likelihood estimated Champernowne distribution might differ significantly from the true distribution. Even though the resulting transformation kernel density estimator corrects some of the deviations between the estimated Champernowne density and the true density, the kernel density estimator does not sufficiently correct the tail estimate, due to sparse data in the tail. Therefore it is crucial to choose a Champernowne distribution with a well fitted tail, even though the fit in the center of the distribution is poor compared to the maximum likelihood estimated Champernowne distribution. In this paper, we use a *cml estimation* method. The procedure is the following: First, set $c_1 = 0$ and choose (α_1, M_1) by maximizing the conditional loglikelihood function,

$$\begin{aligned} \log L_t(\theta) &= n_t \log \alpha + (\alpha - 1) \sum_{j=1}^{n_t} \log(\widetilde{x}_j + c) \\ &\quad -2 \sum_{j=1}^{n_t} \log \left[(\widetilde{x}_j + c)^\alpha + (M + c)^\alpha - 2c^\alpha \right] \\ &\quad +n_t \log \left[(t + c)^\alpha + (M + c)^\alpha - 2c^\alpha \right] \end{aligned} \tag{5.4}$$

for a given threshold t, $c = c_1$, and where $\widetilde{x}_1, ..., \widetilde{x}_{n_t}$ are the n_t largest observations above the threshold t. From (5.3) we know that the Champernowne density converges to a Pareto density. This means that the Champernowne density with parameters $(\alpha_1, M_1, 0)$ is approximated by $\tau x^{-(\alpha_1+1)}$, where $\tau = \alpha_1 M_1^{\alpha_1}$, as x tends to infinity. In the following, we will call τ the *tail constant*. Now, we keep $\widehat{\alpha} = \alpha_1$, and we also keep the tail constant τ unchanged, but now we allow c to be different from 0 and choose $\widehat{c}$ by maximizing the one-dimensional global loglikelihood function, corresponding to (5.4) with threshold $t = 0$, $\alpha = \widehat{\alpha}$ and $M = \left\{ \tau / \widehat{\alpha} + c^{\widehat{\alpha}} \right\}^{1/\widehat{\alpha}} - c$, which ensures an unchanged tail constant. At last, we determine $\widehat{M} = \left\{ \tau / \widehat{\alpha} + \widehat{c}^{\widehat{\alpha}} \right\}^{1/\widehat{\alpha}} - \widehat{c}$.

The intuition for this estimation procedure is that we obtain a triple of parameters, $(\widehat{\alpha}, \widehat{M}, \widehat{c})$, where $\widehat{\alpha}$ and $\widehat{M}$ ensure a good tail fit and $\widehat{c}$ afterwards ensures a good fit in the center of the distribution, but without destroying the tail estimate ensured by unchanged α and tail contant. This estimation procedure provides an estimate on

the entire axis, but unfortunately needs a threshold, t, just like the gpd estimator, see section 5.4. But as we will see in the simulation study in section 5.5 the choice of threshold is less important than the threshold for the gpd estimator. The parametric Champernowne distribution with cml parameters and threshold t, is denoted by $\widetilde{h}_t(x)$, and with a non-parametric correction, it is denonted by $\widehat{h}_t(x)$. Notice, that the conditional maximum likelihood estimator with threshold $t = 0$ corresponds to the ordinary maximum likelihood estimator.

5.3.2 The g-and-h distribution

In recent years, two four-parameter distributions, the Generalized Beta Distribution of Second Kind (GB2) and the g-and-h distribution have become very popular in the area of operational risk estimation, see Degen et al. (2007); Dutta and Perry (2006). In this paper we solely focus on the g-and-h distribution as the GB2 is extremely sensitive to small changes in the parameters as mentioned in Dutta and Babbel (2002) and this property makes the distribution less attractive when it comes to tail estimation. Moreover, as showed in both Dutta and Babbel (2002) and Dutta and Perry (2006) the g-and-h distribution provided superior performance compared to the GB2.

The g-and-h distribution is a very flexible distribution with a wide variety of tail behavior, and it covers a lot of known distributions including the normal and the lognormal distribution, see Martinez and Iglewicz (1984) and Dutta and Babbel (2002). As illustrated in Dutta and Perry (2006) the g-and-h distribution spans a much wider area in skewness-kurtosis than many well-known distributions including the GB2, and it seems to be a reasonable model as a single distribution which is able to fith both the center and the tail of an operational risk severity distribution. As showed in Degen and Embrechts (2008) and Degen et al. (2007), the g-and-h distribution converges extremely slowly to the EVT, which make results of EVT methods inaccurate for g-and-h distribution-like data.

The g-and-h distribution has four parameters and is a strictly increasing transformation of the standard normal distribution:

$$Y = a + b\left\{\exp(gX) - 1\right\}\frac{\exp(hX^2/2)}{g} \qquad x \in \mathbb{R} \tag{5.5}$$

The parameters g and h control the skewness and the kurtosis, respectively, whereas the parameters a and b are location and scale parameters, respectively.

The parameters can be estimated in various ways including the maximum likelihood method and the moment method. In this paper we follow the quantile-based approach described in Dutta and Babbel (2002) and Dutta and Perry (2006) with constant g and h parameters. We denote the g-and-h density estimator by $\widetilde{h}_{\mathrm{gh}}(x)$.

Correspondingly to the Champernowne distribution mentioned above, we can use the cumulative distribution function corresponding to the estimated g-and-h distribution as transformation function in (5.1) to obtain a non-parametricly corrected g-and-h estimator, denoted by $\widehat{h}_{\mathrm{gh}}(x)$.

5.4 Generalized Pareto distributions

From the fundamental Fisher-Tippett theorem in classical extreme value theory, (see e.g. Embrechts et al. (1997)), we know that, if there exist centering and normalizing constants $c_n > 0$ and $d_n \in \mathbb{R}$ so that $c_n^{-1}(M_n - d_n) \to H$ as $n \to \infty$, where $M_n = \max(X_1, ..., X_n)$ for some non-degenerate distribution H, then H must be of one of the three types of extreme value distributions: Fréchet, Weibull or Gumbel distribution.

The extreme value distributions are closely related to the generalized Pareto distribution, which describe the limit distribution of excesses over a high threshold. The

generalized Pareto cumulative distribution function (gpd) has the form

$$F_{\text{gpd}}(x) = \begin{cases} 1 - \left(1 + \xi\frac{x}{\beta}\right)^{-1/\xi}, & \xi \neq 0 \\ 1 - \exp\left(-\frac{x}{\beta}\right), & \xi = 0 \end{cases}$$

with parameters (ξ, β), where $x \geq 0$ if $\xi \geq 0$ and $0 \geq x \geq -1/\xi$ if $\xi < 0$, see (Embrechts et al., 1997, Definition 3.4.9).

Correspondingly to the Champernowne distribution, the gpd is connected to the Pareto distribution, as the gpd for $\xi \geq 0$ can be rewritten as a Pareto cdf with $k = 1/\xi$, $x_0 = \beta/\xi$ and $\zeta = -\beta/\xi$

$$\begin{aligned} F_{\text{gpd}}(x) &= 1 - \left(1 + \xi\frac{x}{\beta}\right)^{-1/\xi} \\ &= 1 - \left(\frac{\beta/\xi}{x - (-\beta/\xi)}\right)^{1/\xi}. \end{aligned}$$

When using the gpd in practice for large loss modeling, one crucial step is the choice of threshold from where the data set is assumed to follow a gpd. The choice of threshold is a classical bias–variance trade-off: choosing the threshold too low, means that assuming the limiting gpd is not appropriate, whereas choosing the threshold to high, means that we have too few data points to estimate the gpd parameters.

Often graphical methods are used in the choice of threshold. As described in Embrechts et al. (1997), one way is to look at the empirical mean excess function and choose a threshold v, such that the empirical mean excess function is approximately linear for $x \geq v$. Another way is to look at a plot of the estimated gpd shape parameter as a function of v. Then choose v so that the estimated gpd shape parameter is approximately constant for $x \geq v$.

One approach when using the gpd distribution to fit the tail, is the Peaks Over Threshold (POT) method, see (Embrechts et al., 1997, section 6.5). When a threshold, v is chosen, the estimated gpd parameters $(\widehat{\xi}, \widehat{\beta})$ are found by the maximum

likelihood method based on data above v. This provide an estimator of the conditional distribution of exceeding above v,

$$\widehat{F}_{|v}(y) = \widehat{F}_{\text{gpd}}(y), \qquad y > 0 \tag{5.6}$$

The unconditional distribution is obtained by a three-parameter gpd with corrected parameters:

$$\widehat{F}(x) = 1 - \left(1 + \widehat{\xi}\,\frac{x - v - \widehat{\mu}}{\widehat{\beta}'}\right)^{-1/\widehat{\xi}}, \qquad x > v \tag{5.7}$$

where $\widehat{\mu} = \widehat{\beta}/\widehat{\xi}\left((n_v/n)^{\widehat{\xi}} - 1\right)$ and $\widehat{\beta}' = \widehat{\beta}\,(n_v/n)^{\widehat{\xi}}$. We denote the resulting density estimator by $\widehat{g}_v(x)$, defined for $x > v$.

5.5 Monte Carlo simulation study

In the following section we illustrate and compare the presented estimators in a Monte Carlo simulation study. The simulation study is based on a typical heavy-tailed loss data set. Operational risk data sets are usually left-truncated and disturbed by under-reporting, Buch-Kromann et al. (2007). To avoid these additional challenges we have used a heavy-tailed fire insurance data set. The characteristics of this data set are similar to a typical operational risk data set, but this data set does not suffer from truncation and under-reporting. Moreover, the data quality is higher compared to most operational risk data sets.

The fire insurance data set originates from the Danish general insurance company, Codan Insurance. The data consist of 2,810 commercial fire claims reported from 1995-2004. The data set is heavy-tailed, with claim sizes ranging from 19 to almost 6 million Dkr. and with an average claim size at 56,220 Dkr. Further details about the data set can be found in Buch-Kromann et al. (2009).

The true distribution of the data is obviously unknown. However, to get simulated test data with realistic claim sizes and known distribution, we estimate a Weibull

and a lognormal distribution by use of the maximum likelihood estimation method to the fire insurance data set:

$$\begin{aligned} f_\theta(x) &= \left(\frac{\alpha}{\beta}\right)\left(\frac{x}{\beta}\right)^{(\alpha-1)} \exp\left\{-\left(\frac{x}{\beta}\right)^{\alpha}\right\} \\ f_\xi(x) &= \frac{1}{\sqrt{2\pi}\sigma x} \exp\left\{-\frac{(\log x-\mu)^2}{2\sigma^2}.\right\} \end{aligned}$$

The estimated parameters for the Weibull distribution are $\theta = (\alpha, \beta) = (0.507, /, 20, 382)$ and for the lognormal distribution $\xi = (\mu, \sigma) = (9.049, 1.83)$.

Moreover, we generate a more heavy-tailed distribution by considering a mixture of the lognormal with a Pareto distribution:

$$f_\psi(x) = p\frac{1}{\sqrt{2\pi}\sigma x} \exp\left\{-\frac{(\log x-\mu)^2}{2\sigma^2}\right\} + (1-p)k\frac{x_0^k}{(x-\zeta)^{k+1}}$$

with parameters $\psi = (p,\, \mu,\, \sigma,\, x_0,\, k,\, \zeta) = (0.7,\, 9.049,\, 1.83,\, 5,000,\, 1,\, -5000)$. At last, we estimate a g-and-h distribution as defined in (5.5) by use of the quantile based estimation method and obtain the parameters $(a,\, b,\, g,\, h) = (6,527,\, 12,264,\, 2.5,\, 0.07)$. Notice that the g-and-h distribution can take negative values even though fire claims are never negative. We follow Dutta and Perry (2006) and use a rejection sampling method to avoid this problem. Based on the four test distributions we simulate $S = 100$ repetitions and measure the error in density estimation by:

$$\text{WISE}_q^\delta(\widehat{f}) = \int_{x_q}^{\infty} \left\{f(x) - \widehat{f}(x)\right\}^2 x^\delta \, dx \tag{5.8}$$

where $f(x)$ is the true density at x and $\widehat{f}(x)$ is the density estimator under investigation. The lower limit, x_q, in the integral in (5.8) is the claim size value corresponding to the q's empirical quantile. In the paper, we use $q = 0.8$, ie. the 80% quantile, when measuring tail performance of an estimator, and $q = 0$ when measuring FGU performance. The parameter $\delta = \{0, 1, 2\}$ in (5.8) decides the weight in the error measure: the bigger δ, the more weight is put into the tail deviation between the true

and the estimated density. The performance of each density estimator is measured by the average WISE obtained from the $S = 100$ repetitions and is called $\text{AWISE}_q^\delta(\widehat{f})$.

The purpose of the simulation study is to compare the performance – primarily the tail performance – of the gpd estimator, $\widehat{g}_v(x)$ and the Champernowne transformed kernel density estimator with cml parameters, $\widehat{h}_t(x)$ and with maximum likelihood parameters, $\widehat{h}_0(x)$. We benchmark the performance against the parametric Champernowne distribution with cml parameters, $\widetilde{h}_t(x)$, the Champernowne distribution with maximum likelihood parameters, $\widetilde{h}_0(x)$ and against the estimated g-and-h distribution, both with non-parametric correction, $\widehat{h}_{\text{gh}}(x)$, and without non-parametric correction, $\widetilde{h}_{\text{gh}}(x)$.

To calculate $\widehat{g}_v(x)$ and $\widehat{h}_t(x)$ we need to choose thresholds. This is not possible by using graphical methods due to the large number of data sets. Instead we calculate the *global optimal threshold*

$$u_{\text{opt}}^\delta \quad = \quad \arg\min_u \text{AWISE}_{0.8}^\delta(\widehat{f}_u), \tag{5.9}$$

where $\widehat{f}_u$ is either $\widehat{g}_v(x)$ or $\widehat{h}_t(x)$. The global optimal gpd and cml thresholds, v_{opt}^δ and t_{opt}^δ, are listed in the last two columns in Tables 5.1-5.3. Comparing them, we see that both v_{opt}^δ and t_{opt}^δ fluctuates and are far from constant. Moreover, there seems to be a tendency of higher thresholds for larger δ which is resonable because AWISE with higher δ gives more weight to the tail. Figures 5.1-5.3 show the thresholds influence on the tail performance for $\widehat{g}_v(x)$ and $\widehat{h}_t(x)$ for selected values of n, and these plots give a more in-depth understanding of the global thresholds. By studying the figures we recognize, that the curves for tail performance of $\widehat{h}_t(x)$, generally speaking, are flatter around the global minimum point than the curves for $\widehat{g}_v(x)$. That means, that the choice of threshold is less crucial for tail performance of $\widehat{h}_t(x)$ than for $\widehat{g}_v(x)$, which is an advantage for $\widehat{h}_t(x)$. This characteristic is caused by the fact that $\widehat{g}_v(x)$ is based exclusively on the observations above the threshold, whereas $\widehat{h}_t(x)$ is based on all observations including the observations below the threshold.

	$\widehat{h}^0_{\text{opt}}$	$\widehat{h}_0$	$\widehat{h}_{\text{gh}}$	$\widehat{g}^0_{\text{opt}}$	$\widetilde{h}^0_{\text{opt}}$	$\widetilde{h}_0$	$\widetilde{h}_{\text{gh}}$	v^0_{opt}	t^0_{opt}
Weibull									
$n = 50$	1.37e-08	2.20e-08	2.66e-08	1.76e-08	1.23e-08	2.63e-08	4.40e-08	0.66	0.43
$n = 100$	6.80e-09	1.23e-08	9.69e-09	8.63e-09	7.39e-09	1.61e-08	1.89e-08	0.36	0.45
$n = 500$	2.77e-09	7.26e-09	4.04e-09	2.87e-09	3.73e-09	1.39e-08	8.31e-09	0.69	0.77
$n = 1000$	1.38e-09	3.94e-09	1.93e-09	1.57e-09	2.93e-09	1.18e-08	3.11e-09	0.74	0.56
$n = 2000$	7.60e-10	2.95e-09	1.40e-09	7.90e-10	2.51e-09	1.22e-08	5.21e-09	0.76	0.67
$n = 5000$	4.63e-10	1.57e-09	7.72e-10	5.36e-10	2.40e-09	1.17e-08	3.90e-09	0.77	0.77
lognormal									
$n = 50$	1.20e-08	1.54e-08	2.47e-08	1.68e-08	2.70e-08	1.93e-08	3.92e-08	0.23	0.22
$n = 100$	5.73e-09	6.07e-09	7.51e-09	7.16e-09	2.74e-08	7.32e-09	2.06e-08	0.30	0.30
$n = 500$	2.14e-09	2.56e-09	2.30e-09	2.56e-09	2.19e-08	4.48e-09	7.28e-09	0.39	0.36
$n = 1000$	9.54e-10	1.07e-09	8.02e-10	1.24e-09	2.18e-08	2.85e-09	2.01e-09	0.38	0.32
$n = 2000$	4.82e-10	7.52e-10	3.91e-10	7.06e-10	2.13e-08	2.92e-09	1.68e-09	0.74	0.38
$n = 5000$	3.00e-10	4.53e-10	2.13e-10	4.46e-10	2.12e-08	2.50e-09	1.21e-09	0.76	0.32
lognormal-Pareto									
$n = 50$	1.39e-08	1.74e-08	2.42e-08	1.70e-08	3.57e-08	2.09e-08	3.62e-08	0.15	0.20
$n = 100$	6.65e-09	6.58e-09	8.56e-09	7.51e-09	3.67e-08	7.45e-09	1.47e-08	0.15	0.27
$n = 500$	2.55e-09	2.61e-09	2.34e-09	2.86e-09	3.37e-08	3.76e-09	5.16e-09	0.30	0.33
$n = 1000$	1.16e-09	1.04e-09	9.76e-10	1.32e-09	3.42e-08	2.27e-09	5.11e-09	0.41	0.29
$n = 2000$	5.99e-10	6.64e-10	4.13e-10	7.74e-10	3.37e-08	2.01e-09	2.14e-09	0.72	0.33
$n = 5000$	3.53e-10	4.08e-10	2.43e-10	5.42e-10	3.32e-08	1.76e-09	1.44e-09	0.75	0.29
g-and-h									
$n = 50$	7.51e-09	9.90e-09	1.31e-08	1.09e-08	4.77e-08	1.42e-08	2.07e-08	0.11	0.10
$n = 100$	4.03e-09	3.80e-09	3.84e-09	4.52e-09	4.27e-08	5.24e-09	8.15e-09	0.44	0.10
$n = 500$	1.38e-09	1.51e-09	1.15e-09	1.71e-09	3.75e-08	4.19e-09	1.89e-09	0.48	0.10
$n = 1000$	6.81e-10	7.23e-10	4.45e-10	8.68e-10	3.30e-08	3.36e-09	4.78e-10	0.49	0.10
$n = 2000$	3.33e-10	4.88e-10	1.98e-10	5.04e-10	3.36e-08	3.55e-09	2.77e-10	0.76	0.10
$n = 5000$	2.01e-10	3.14e-10	1.17e-10	3.56e-10	3.34e-08	3.31e-09	1.45e-10	0.77	0.10

Table 5.1: Tail performance. AWISE with $\delta = 0$ corresponding to the Champernowne transformed kernel density estimator with cml parameters $\widehat{h}^0_{\text{opt}}$, with maximum likelihood parameters $\widehat{h}_0$, the g-and-h transformed kernel density estimator $\widehat{h}_{\text{gh}}$, the corresponding parametric distributions $\widetilde{h}^0_{\text{opt}}$, $\widetilde{h}_0$ and $\widetilde{h}_{\text{gh}}$ together with the gpd estimator $\widehat{g}^0_{\text{opt}}$ for the four test distributions. The last two colums are the optimal thresholds for the gpd, (v^0_{opt}) and the cml (t^0_{opt}) estimators.

The threshold for $\widehat{h}_t(x)$ only determines the point from where the tail estimation is made. Special focus on Figures 5.3 shows that $\widehat{g}_v(x)$ has a very poor estimation of the far tail of the g-and-h distributed data set, whereas $\widehat{h}_t(x)$ has a considerably better tail performance for this type of data. This consideration corresponds to the

	$\widehat{h}^1_{\text{opt}}$	$\widehat{h}_0$	$\widehat{h}_{\text{gh}}$	$\widehat{g}^1_{\text{opt}}$	$\widetilde{h}^1_{\text{opt}}$	$\widetilde{h}_0$	$\widetilde{h}_{\text{gh}}$	v^1_{opt}	t^1_{opt}
Weibull									
$n = 50$	1.33e-03	2.22e-03	2.48e-03	1.80e-03	1.35e-03	2.674e-03	3.329e-03	0.68	0.43
$n = 100$	6.95e-04	1.44e-03	1.15e-03	8.55e-04	8.33e-04	1.946e-03	1.760e-03	0.36	0.46
$n = 500$	2.96e-04	9.54e-04	5.35e-04	3.10e-04	4.31e-04	1.806e-03	6.727e-04	0.69	0.78
$n = 1000$	1.55e-04	6.51e-04	3.34e-04	1.64e-04	3.68e-04	1.704e-03	3.310e-04	0.74	0.76
$n = 2000$	9.18e-05	5.08e-04	2.59e-04	8.18e-05	3.05e-04	1.690e-03	4.183e-04	0.76	0.76
$n = 5000$	6.00e-05	3.11e-04	1.68e-04	5.54e-05	2.92e-04	1.675e-03	2.910e-04	0.77	0.87
lognormal									
$n = 50$	8.93e-04	1.09e-03	1.70e-03	1.15e-03	1.99e-03	1.28e-03	2.236e-03	0.13	0.21
$n = 100$	4.26e-04	4.77e-04	6.73e-04	5.32e-04	1.82e-03	5.66e-04	1.339e-03	0.30	0.30
$n = 500$	1.74e-04	2.26e-04	2.04e-04	2.21e-04	1.31e-03	3.61e-04	4.493e-04	0.45	0.28
$n = 1000$	7.98e-05	1.15e-04	7.74e-05	1.10e-04	1.33e-03	2.43e-04	1.442e-04	0.49	0.32
$n = 2000$	4.40e-05	8.83e-05	3.65e-05	5.71e-05	1.25e-03	2.42e-04	1.126e-04	0.76	0.37
$n = 5000$	2.87e-05	6.23e-05	1.85e-05	3.62e-05	1.24e-03	2.11e-04	7.618e-05	0.77	0.39
lognormal-Pareto									
$n = 50$	8.83e-04	1.04e-03	1.45e-03	9.92e-04	2.22e-03	1.18e-03	1.803e-03	0.11	0.20
$n = 100$	4.09e-04	4.30e-04	6.00e-04	4.71e-04	2.04e-03	4.83e-04	8.123e-04	0.15	0.25
$n = 500$	1.71e-04	1.95e-04	1.73e-04	2.08e-04	1.68e-03	2.69e-04	2.929e-04	0.42	0.27
$n = 1000$	7.53e-05	8.77e-05	7.89e-05	9.61e-05	1.74e-03	1.68e-04	2.763e-04	0.59	0.29
$n = 2000$	4.04e-05	6.22e-05	2.98e-05	5.29e-05	1.66e-03	1.52e-04	1.123e-04	0.74	0.29
$n = 5000$	2.46e-05	4.28e-05	1.67e-05	3.48e-05	1.64e-03	1.35e-04	7.688e-05	0.76	0.30
g-and-h									
$n = 50$	7.05e-04	9.22e-04	1.07e-03	8.75e-04	2.91e-03	1.25e-03	1.334e-03	0.23	0.10
$n = 100$	3.47e-04	4.01e-04	4.14e-04	4.02e-04	2.86e-03	6.12e-04	6.826e-04	0.44	0.10
$n = 500$	1.39e-04	2.00e-04	1.31e-04	1.87e-04	2.38e-03	5.58e-04	1.675e-04	0.69	0.10
$n = 1000$	6.75e-05	1.03e-04	5.17e-05	1.02e-04	2.21e-03	4.75e-04	4.983e-05	0.58	0.10
$n = 2000$	3.51e-05	8.04e-05	2.19e-05	5.27e-05	2.18e-03	5.05e-04	2.786e-05	0.76	0.10
$n = 5000$	2.19e-05	5.44e-05	1.19e-05	3.79e-05	2.19e-03	4.81e-04	1.430e-05	0.77	0.61

Table 5.2: Tail performance. AWISE with $\delta = 1$ corresponding to the Champernowne transformed kernel density estimator with cml parameters $\widehat{h}^1_{\text{opt}}$, with maximum likelihood parameters $\widehat{h}_0$, the g-and-h transformed kernel density estimator $\widehat{h}_{\text{gh}}$, the corresponding parametric distributions $\widetilde{h}^1_{\text{opt}}$, $\widetilde{h}_0$ and $\widetilde{h}_{\text{gh}}$ together with the gpd estimator $\widehat{g}^1_{\text{opt}}$ for the four test distributions. The last two colums are the optimal thresholds for the gpd, (v^1_{opt}) and the cml (t^1_{opt}) estimators.

conclusion in Degen et al. (2007) which is that the gpd estimator has a poor tail performance especially for g-and-h parameter values with a large ratio g/h. This is exactly the characteristic for this data set ($g = 2.5, h = 0.07$).

The tail performance of $\widehat{g}_v(x)$ and $\widehat{h}_t(x)$ with optimal thresholds, v^δ_{opt} and t^δ_{opt} (called

	$\widehat{h}^2_{\rm opt}$	$\widehat{h}_0$	$\widehat{h}_{\rm gh}$	$\widehat{g}^2_{\rm opt}$	$\widetilde{h}^2_{\rm opt}$	$\widetilde{h}_0$	$\widetilde{h}_{\rm gh}$	$v^2_{\rm opt}$	$t^2_{\rm opt}$
Weibull									
$n = 50$	269.60	1624.00	1190.00	279.40	484.300	2285.00	863.30	0.37	0.63
$n = 100$	128.90	1213.00	528.40	130.80	311.400	2043.00	490.90	0.71	0.85
$n = 500$	56.43	710.70	231.00	49.10	187.000	1679.00	149.00	0.69	0.82
$n = 1000$	30.58	588.30	194.60	24.17	192.200	1740.00	117.30	0.74	0.91
$n = 2000$	22.00	439.70	155.30	12.84	169.700	1630.00	76.07	0.76	0.90
$n = 5000$	15.65	283.70	112.50	8.51	164.800	1670.00	52.81	0.79	0.92
lognormal									
$n = 50$	189.80	339.00	703.90	248.70	362.600	312.70	481.00	0.39	0.46
$n = 100$	86.95	184.30	261.60	111.30	242.900	183.10	223.90	0.43	0.41
$n = 500$	33.60	90.10	53.59	37.34	129.200	96.78	76.08	0.68	0.70
$n = 1000$	17.44	81.18	26.01	17.81	132.100	86.18	34.02	0.67	0.84
$n = 2000$	10.29	67.52	12.54	9.71	113.300	77.10	20.41	0.76	0.93
$n = 5000$	6.18	60.36	5.11	6.11	109.800	74.86	10.76	0.77	0.91
lognormal-Pareto									
$n = 50$	159.90	278.10	876.40	231.100	323.00	253.00	276.80	0.47	0.51
$n = 100$	68.78	135.80	312.80	97.360	214.40	132.80	187.90	0.46	0.35
$n = 500$	25.52	57.40	64.52	33.320	124.20	59.98	53.07	0.64	0.33
$n = 1000$	13.98	47.46	36.43	14.360	128.00	48.47	35.61	0.69	0.39
$n = 2000$	7.99	37.56	9.48	8.047	112.60	41.62	12.80	0.77	0.87
$n = 5000$	4.62	33.71	2.72	4.748	109.70	39.64	8.16	0.79	0.88
g-and-h									
$n = 50$	410.60	2021.00	2173.00	725.20	672.00	1458.00	1642.00	0.80	0.19
$n = 100$	156.80	1827.00	670.50	496.80	377.20	1629.00	426.60	0.71	0.29
$n = 500$	73.65	427.90	114.70	124.10	228.10	363.20	120.40	0.74	0.21
$n = 1000$	43.70	370.50	51.12	53.86	218.80	299.50	55.88	0.79	0.29
$n = 2000$	29.96	311.40	24.47	32.11	195.70	278.00	29.52	0.80	0.25
$n = 5000$	18.95	266.50	9.68	25.86	193.90	249.50	12.81	0.78	0.95

Table 5.3: Tail performance. AWISE with $\delta = 2$ corresponding to the Champernowne transformed kernel density estimator with cml parameters $\widehat{h}^2_{\rm opt}$, with maximum likelihood parameters $\widehat{h}_0$, the g-and-h transformed kernel density estimator $\widehat{h}_{\rm gh}$, the corresponding parametric distributions $\widetilde{h}^2_{\rm opt}$, $\widetilde{h}_0$ and $\widetilde{h}_{\rm gh}$ together with the gpd estimator $\widehat{g}^2_{\rm opt}$ for the four test distributions. The last two colums are the optimal thresholds for the gpd, ($v^2_{\rm opt}$) and the cml ($t^2_{\rm opt}$) estimators.

$\widehat{g}^\delta_{\rm opt}$ and $\widehat{h}^\delta_{\rm opt}$, respectively) is listed in Tables 5.1-5.3 together with the tail performance of $\widehat{h}_0(x)$ and $\widehat{h}_{\rm gh}$ and the parametric estimators, $\widetilde{h}^\delta_{\rm opt}(x)$, $\widetilde{h}_0(x)$ and $\widetilde{h}_{\rm gh}(x)$. Comparing $\widehat{g}^\delta_{\rm opt}$ and $\widehat{h}^\delta_{\rm opt}$, we recognize that $\widehat{h}^\delta_{\rm opt}$ outperforms $\widehat{g}^\delta_{\rm opt}$ by 30% on average for $\delta = 0$ and $\delta = 1$, and by 20% for $\delta = 2$. However, especially for $\delta = 2$ this

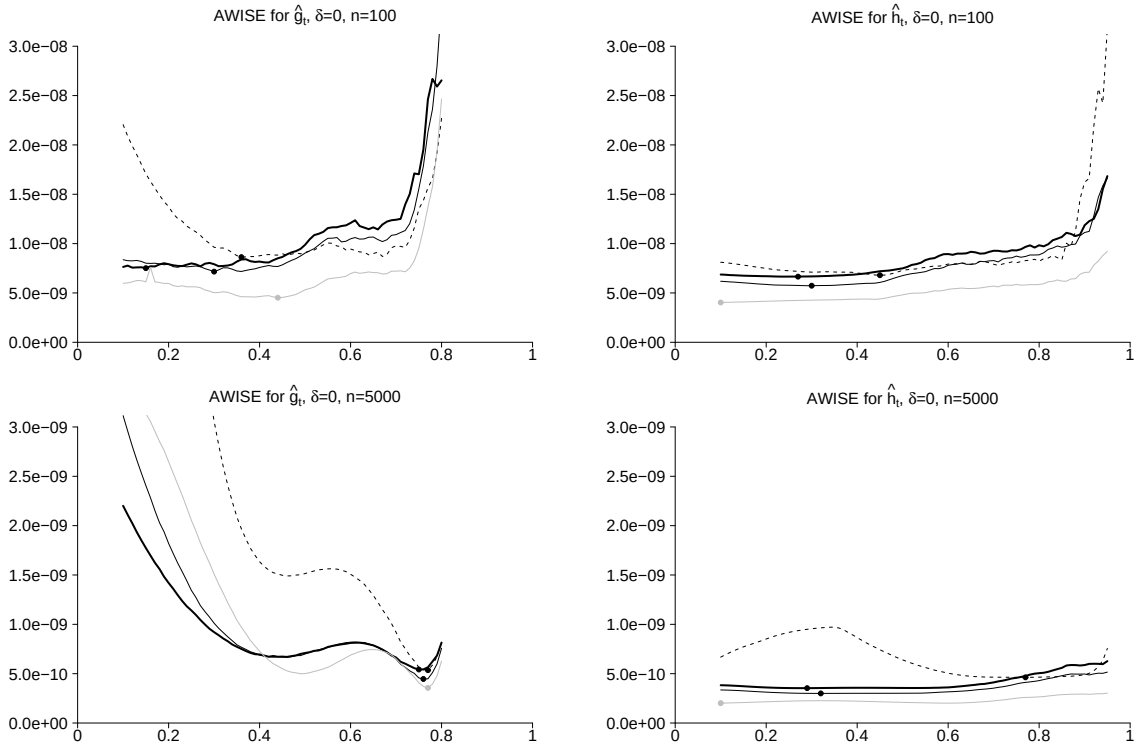

Figure 5.1: The thresholds influence on tail performance with $\delta = 0$ for the gpd estimator, $\widehat{g}_t$ (left plots) and the Champernowne transformed kernel density estimator with cml parameters, $\widehat{h}_t$ (right plots). Black dashed lines correspond to the Weibull data, black solid lines corresponds to lognormal data, black thick lines correspond to lognormal-Pareto data and gray lines corresponds to g-and-h data. The points on the plots corresponds to the optimal thresholds. The upper plots corresponds to a sample size of $n = 100$ and the lower plots corresponds to a sample size of $n = 5000$.

average procentage hides significant differences, where $\widehat{g}^2_{\text{opt}}$ seems to outperform $\widehat{h}^2_{\text{opt}}$ for large- and lighter-tailed data sets, whereas the $\widehat{h}^2_{\text{opt}}$ outperforms $\widehat{g}^2_{\text{opt}}$ on small- and heavier-tailed data sets. Generally holds, for all the selected values of δ that the superior tail performance of $\widehat{h}^\delta_{\text{opt}}$ is most pronounced for heavy-tailed data and particularly the g-and-h distributed data. Comparing $\widehat{h}^\delta_{\text{opt}}$ with the parametric benchmark estimators $\widetilde{h}^\delta_{\text{opt}}$, $\widetilde{h}_0$ and $\widetilde{h}_{\text{gh}}$ shows that $\widehat{h}^\delta_{\text{opt}}$ significantly outperforms the parametric estimators; this effect increases with the size of the data set. The only exception is $\widetilde{h}_{\text{gh}}$ on g-and-h distributed data sets. In this situation $\widehat{h}^\delta_{\text{opt}}$ only outperforms $\widetilde{h}_{\text{gh}}$

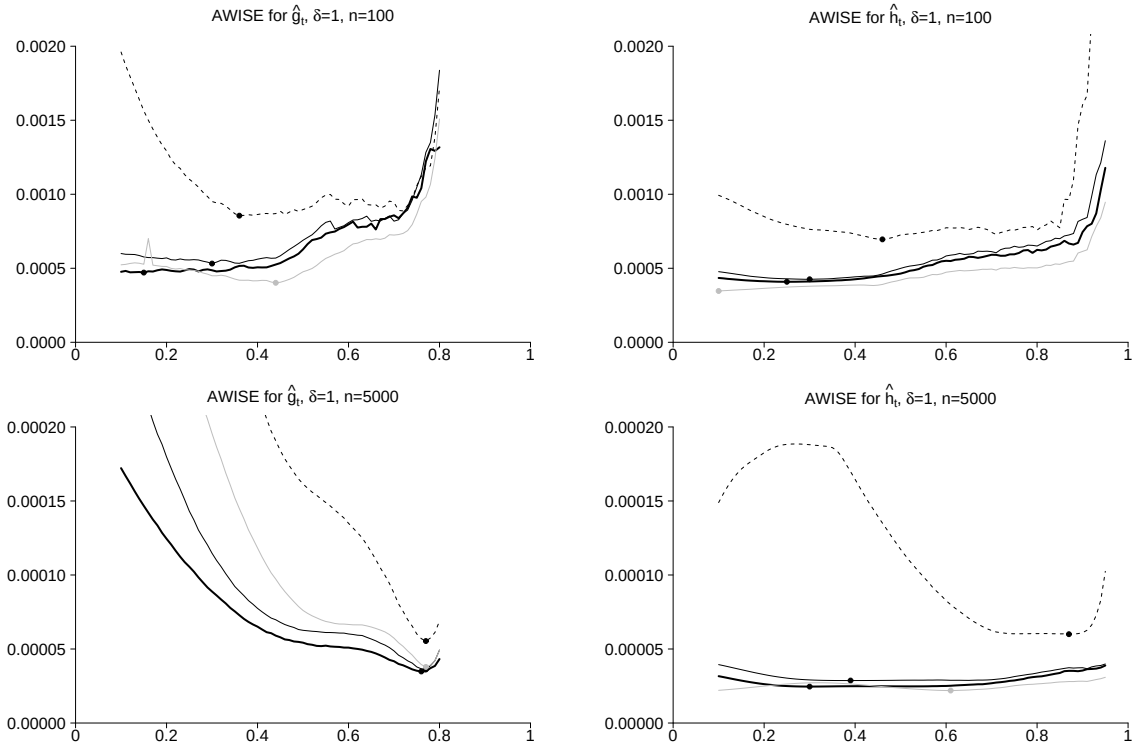

Figure 5.2: The thresholds influence on tail performance with $\delta = 1$ for the gpd estimator, $\widehat{g}_t$ (left plots) and the Champernowne transformed kernel density estimator with cml parameters, $\widehat{h}_t$ (right plots). Black dashed lines correspond to the Weibull data, black solid lines corresponds to lognormal data, black thick lines correspond to lognormal-Pareto data and gray lines corresponds to g-and-h data. The points on the plots corresponds to the optimal thresholds. The upper plots corresponds to a sample size of $n = 100$ and the lower plots corresponds to a sample size of $n = 5000$.

for small sizes of data sets. As expected, the most efficient estimator is the true parametric estimator.

Finally, we compare the tail performance of $\widehat{h}^{\delta}_{\text{opt}}$ and $\widehat{h}_{\text{gh}}$, the non-parametric corrected g-and-h distribution. This estimator appears to be superior to $\widehat{h}^{\delta}_{\text{opt}}$ for large data sets with heavy tails, whereas the tail performance of $\widehat{h}^{\delta}_{\text{opt}}$ is superior for smaller sample sizes and lighter-tailed data sets of all sample sizes. For $\delta = 2$, the tail performance of $\widehat{h}^{2}_{\text{opt}}$ is superior to $\widehat{h}_{\text{gh}}$ in almost all situations.

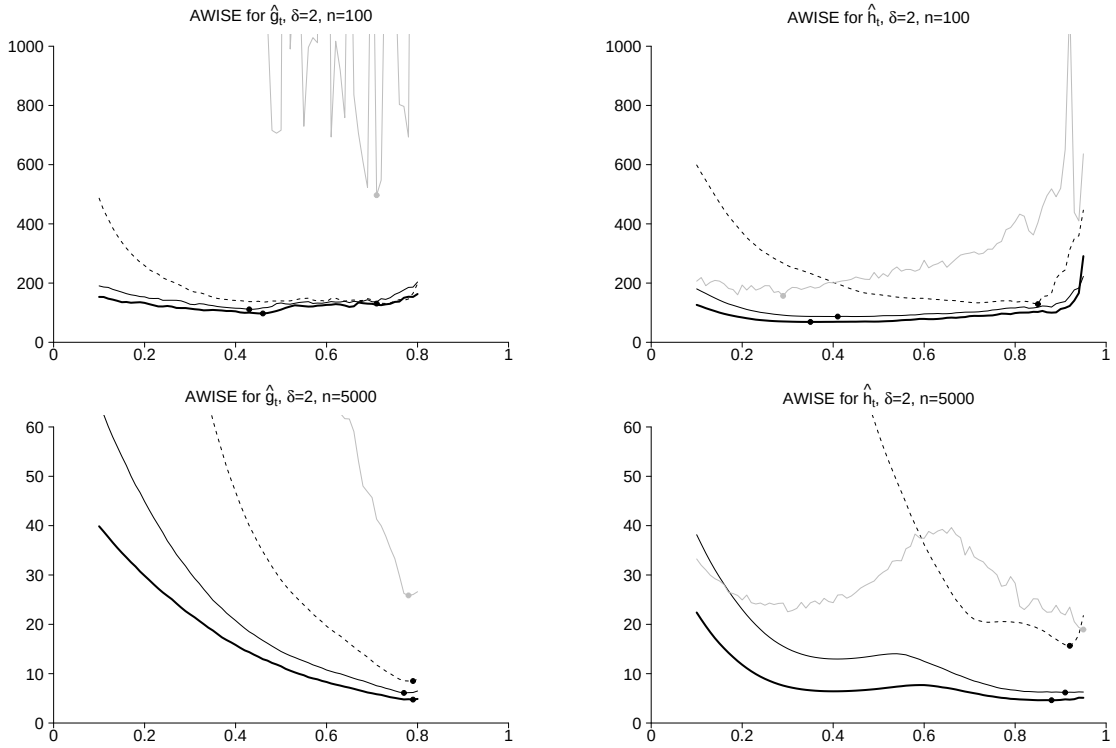

Figure 5.3: The thresholds influence on tail performance with $\delta = 2$ for the gpd estimator, $\widehat{g}_t$ (left plots) and the Champernowne transformed kernel density estimator with cml parameters, $\widehat{h}_t$ (right plots). Black dashed lines correspond to the Weibull data, black solid lines corresponds to lognormal data, black thick lines correspond to lognormal-Pareto data and gray lines corresponds to g-and-h data. The points on the plots corresponds to the optimal thresholds. The upper plots corresponds to a sample size of $n = 100$ and the lower plots corresponds to a sample size of $n = 5000$.

Comparing $\widehat{h}^{\delta}_{\text{opt}}$ and $\widehat{h}_0$ in Tables 5.1-5.3, we see that $\widehat{h}^{\delta}_{\text{opt}}$ has a significantly better tail performance than $\widehat{h}_0$ for all the chosen values of δ, but most pronounced for large values of δ and n. However, when it comes to FGU performance, the picture changes. In Tables 5.4-5.6 we have computed AWISE^{δ}_0 for $\widehat{h}^{\delta}_{\text{opt}}$, $\widehat{h}_0$, $\widehat{h}_{\text{gh}}$ and the parametric estimators $\widetilde{h}^{\delta}_{\text{opt}}$, $\widetilde{h}_0$ and $\widetilde{h}_{\text{gh}}$. Notice, that a comparison with $\widehat{g}^{\delta}_{\text{opt}}$ on the entire axis is not possible, because the gpd estimator is not defined below its threshold. We recognize that $\widehat{h}_0$ has a significantly better FGU performance than $\widehat{h}^{\delta}_{\text{opt}}$ for both $\delta = 0$ and $\delta = 1$ where the FGU performance deviation is the larger the lighter the tail. Only for

	$\widehat{h}^0_{\text{opt}}$	$\widehat{h}_0$	$\widehat{h}_{\text{gh}}$	$\widetilde{h}^0_{\text{opt}}$	$\widetilde{h}_0$	$\widetilde{h}_{\text{gh}}$
Weibull						
$n = 50$	1.91e-05	1.25e-05	2.76e-05	2.75e-05	1.43e-05	4.99e-05
$n = 100$	2.20e-05	1.34e-05	2.79e-05	3.42e-05	1.63e-05	5.46e-05
$n = 500$	2.57e-05	1.84e-05	2.69e-05	3.94e-05	2.58e-05	6.62e-05
$n = 1000$	2.52e-05	1.92e-05	2.57e-05	3.97e-05	2.80e-05	7.54e-05
$n = 2000$	2.41e-05	1.95e-05	2.45e-05	3.97e-05	2.97e-05	8.26e-05
$n = 5000$	2.23e-05	1.80e-05	2.28e-05	3.97e-05	3.02e-05	6.68e-05
lognormal						
$n = 50$	2.31e-06	2.11e-06	9.60e-06	5.21e-06	2.16e-06	1.49e-05
$n = 100$	1.54e-06	1.17e-06	5.18e-06	5.99e-06	1.23e-06	1.11e-05
$n = 500$	7.68e-07	5.22e-07	1.69e-06	5.78e-06	6.95e-07	5.51e-06
$n = 1000$	6.35e-07	4.71e-07	1.16e-06	6.04e-06	6.13e-07	3.73e-06
$n = 2000$	5.23e-07	4.10e-07	7.25e-07	5.87e-06	5.79e-07	2.81e-06
$n = 5000$	4.46e-07	4.08e-07	3.64e-07	5.91e-06	6.10e-07	1.71e-06
lognormal-Pareto						
$n = 50$	2.39e-06	2.395e-06	8.53e-06	5.31e-06	2.46e-06	1.51e-05
$n = 100$	1.35e-06	1.348e-06	5.21e-06	5.96e-06	1.30e-06	1.05e-05
$n = 500$	6.03e-07	4.995e-07	1.80e-06	5.92e-06	5.03e-07	4.80e-06
$n = 1000$	4.74e-07	4.162e-07	1.21e-06	6.22e-06	4.58e-07	4.37e-06
$n = 2000$	3.58e-07	3.158e-07	9.68e-07	6.07e-06	3.79e-07	3.33e-06
$n = 5000$	2.81e-07	2.809e-07	5.68e-07	6.08e-06	3.78e-07	2.10e-06
g-and-h						
$n = 50$	1.06e-05	7.97e-06	1.91e-05	1.72e-05	1.11e-05	3.12e-05
$n = 100$	1.06e-05	7.26e-06	1.51e-05	1.79e-05	1.20e-05	1.94e-05
$n = 500$	1.05e-05	4.69e-06	4.86e-06	1.86e-05	1.63e-05	7.90e-06
$n = 1000$	9.61e-06	3.45e-06	4.07e-06	1.92e-05	1.79e-05	4.64e-06
$n = 2000$	7.75e-06	2.26e-06	2.84e-06	1.95e-05	1.93e-05	4.88e-06
$n = 5000$	5.27e-06	1.36e-06	1.67e-06	1.98e-05	2.00e-05	2.37e-06

Table 5.4: FGU performance. AWISE with $\delta = 0$ corresponding to the Champernowne transformed kernel density estimator with cml parameters $\widehat{h}^0_{\text{opt}}$, with maximum likelihood parameters $\widehat{h}_0$, the g-and-h transformed kernel density estimator $\widehat{h}_{\text{gh}}$ and the corresponding parametric distributions $\widetilde{h}^0_{\text{opt}}$, $\widetilde{h}_0$ and $\widetilde{h}_{\text{gh}}$ for the four test distributions.

$\delta = 2$ is enough weight is put into the tail of the error measure, so that $\widehat{h}^2_{\text{opt}}$ outperforms $\widehat{h}_0$. That means, that we pay a price on FGU performance, when we choose a parameter estimation method of the transformation function, which specially focuses on fitting the tail, but we obtain obviously a superior tail performance. Comparing the parametric distributions $\widetilde{h}^\delta_{\text{opt}}$, $\widetilde{h}_0$ and $\widetilde{h}_{\text{gh}}$ with the corresponding non-parametric corrected estimatores $\widehat{h}^\delta_{\text{opt}}$, $\widehat{h}_0$ and $\widehat{h}_{\text{gh}}$, respectively, shows that the non-parametric

	$\widehat{h}^1_{\text{opt}}$	$\widehat{h}_0$	$\widehat{h}_{\text{gh}}$	$\widetilde{h}^1_{\text{opt}}$	$\widetilde{h}_0$	$\widetilde{h}_{\text{gh}}$
Weibull						
$n = 50$	6.84e-03	6.50e-03	1.68e-02	9.90e-03	8.88e-03	4.17e-02
$n = 100$	4.33e-03	3.79e-03	9.08e-03	9.13e-03	6.20e-03	2.72e-02
$n = 500$	2.48e-03	1.70e-03	4.13e-03	7.80e-03	3.93e-03	1.27e-02
$n = 1000$	1.82e-03	1.12e-03	3.18e-03	7.62e-03	3.47e-03	6.23e-03
$n = 2000$	1.45e-03	7.96e-04	2.69e-03	7.44e-03	3.44e-03	8.92e-03
$n = 5000$	1.08e-03	4.91e-04	2.21e-03	7.49e-03	3.35e-03	6.29e-03
lognormal						
$n = 50$	5.20e-03	5.20e-03	2.04e-02	1.16e-02	5.11e-03	4.05e-02
$n = 100$	2.30e-03	2.63e-03	8.60e-03	1.20e-02	2.77e-03	2.46e-02
$n = 500$	1.24e-03	9.73e-04	1.79e-03	1.07e-02	1.46e-03	8.07e-03
$n = 1000$	7.83e-04	5.85e-04	1.03e-04	1.10e-02	1.03e-03	3.88e-03
$n = 2000$	4.86e-04	3.45e-04	6.21e-04	1.06e-02	9.23e-04	2.76e-03
$n = 5000$	2.91e-04	2.11e-04	2.34e-04	1.07e-02	8.32e-04	1.50e-03
lognormal-Pareto						
$n = 50$	5.02e-03	5.19e-03	1.63e-02	1.17e-02	4.88	3.55e-02
$n = 100$	2.75e-03	2.59e-03	7.29e-03	1.15e-02	2.38e-03	1.81e-02
$n = 500$	1.05e-03	8.86e-04	1.89e-03	1.07e-02	9.49e-04	5.29e-03
$n = 1000$	6.54e-04	5.27e-04	1.09e-03	1.11e-02	6.63e-04	4.68e-03
$n = 2000$	3.84e-04	2.94e-04	6.84e-04	1.08e-02	5.19e-04	2.84e-03
$n = 5000$	2.15e-04	1.70e-04	2.93e-04	1.08e-02	4.53e-04	1.54e-03
g-and-h						
$n = 50$	2.05e-02	1.56e-02	4.68e-02	4.20e-02	2.04e-02	8.01e-02
$n = 100$	1.83e-02	1.23e-02	3.02e-02	4.29e-02	1.95e-02	5.53e-02
$n = 500$	1.35e-02	6.59e-03	9.19e-03	4.15e-02	1.98e-02	1.51e-02
$n = 1000$	1.15e-02	4.76e-03	6.87e-03	4.16e-02	2.00e-02	9.45e-03
$n = 2000$	9.07e-03	3.20e-03	5.24e-03	4.13e-02	2.01e-02	7.52e-03
$n = 5000$	6.31e-03	2.00e-03	3.06e-03	4.16e-02	2.02e-02	4.10e-03

Table 5.5: FGU performance. AWISE with $\delta = 1$ corresponding to the Champernowne transformed kernel density estimator with cml parameters $\widehat{h}^1_{\text{opt}}$, with maximum likelihood parameters $\widehat{h}_0$, the g-and-h transformed kernel density estimator $\widehat{h}_{\text{gh}}$ and the corresponding parametric distributions $\widetilde{h}^1_{\text{opt}}$, $\widetilde{h}_0$ and $\widetilde{h}_{\text{gh}}$ for the four test distributions.

correction in almost all situations improves the estimators significantly. Moreover, we observe that the FGU performance of $\widehat{h}^\delta_{\text{opt}}$ is superior compare to $\widehat{h}_{\text{gh}}$.

From a theoretical point of view, all four test distributions, Weibull, lognormal, lognormal-Pareto and g-and-h, tend to a gpd distribution in the tail, and therefore one might expect, that $\widehat{g}^\delta_{\text{opt}}$ has a better tail estimation than $\widehat{h}^\delta_{\text{opt}}$ above a certain point in the tail. However, when working with large loss estimation in practice, it is

	$\widehat{h}^2_{\text{opt}}$	$\widehat{h}_0$	$\widehat{h}_{\text{gh}}$	$\widetilde{h}^2_{\text{opt}}$	$\widetilde{h}_0$	$\widetilde{h}_{\text{gh}}$
Weibull						
$n = 50$	338.80	1678.00	1159.00	657.20	2365.00	1207.00
$n = 100$	201.10	1243.00	596.70	490.40	2104.00	769.00
$n = 500$	98.54	718.10	243.70	352.70	1713.00	300.10
$n = 1000$	67.27	592.80	203.80	354.80	1773.00	168.00
$n = 2000$	50.57	441.80	161.00	328.20	1661.00	181.80
$n = 5000$	36.26	285.00	116.30	324.20	1701.00	115.40
lognormal						
$n = 50$	230.40	377.20	561.00	477.60	350.50	789.70
$n = 100$	111.20	205.60	298.90	355.10	205.40	482.80
$n = 500$	44.05	97.02	90.81	227.10	105.50	134.00
$n = 1000$	27.22	85.44	50.87	233.30	92.22	55.64
$n = 2000$	16.93	69.54	20.29	209.80	81.96	36.72
$n = 5000$	10.91	61.62	7.47	207.90	79.15	21.59
lognormal-Pareto						
$n = 50$	197.10	311.30	724.50	424.30	284.50	456.80
$n = 100$	89.12	154.10	236.30	304.60	149.60	253.60
$n = 500$	32.45	63.04	88.80	207.40	64.88	78.65
$n = 1000$	18.53	51.05	49.05	214.20	52.13	61.57
$n = 2000$	10.93	39.28	8.99	195.80	44.09	26.27
$n = 5000$	7.78	34.76	5.19	193.50	41.75	16.00
g-and-h						
$n = 50$	546.90	3209.00	3361.00	874.40	1653.00	1821.00
$n = 100$	270.40	3784.00	731.20	643.90	2021.00	658.60
$n = 500$	135.00	615.10	215.10	437.50	434.50	162.30
$n = 1000$	66.29	497.60	77.89	426.00	363.00	64.66
$n = 2000$	43.99	383.40	24.22	406.40	324.20	43.13
$n = 5000$	30.24	296.70	6.83	405.50	290.00	21.16

Table 5.6: FGU performance. AWISE with $\delta = 2$ corresponding to the Champernowne transformed kernel density estimator with cml parameters $\widehat{h}^2_{\text{opt}}$, with maximum likelihood parameters $\widehat{h}_0$, the g-and-h transformed kernel density estimator $\widehat{h}_{\text{gh}}$ and the corresponding parametric distributions $\widetilde{h}^2_{\text{opt}}$, $\widetilde{h}_0$ and $\widetilde{h}_{\text{gh}}$ for the four test distributions.

relevant to know from which point this possibly happens in the tail. Therefore, for given δ, define:

$$\widetilde{x}_i = \inf\{x | \forall y \geq x : |\widehat{g}^{i,\delta}_{\text{opt}} - f(x)| < |\widehat{h}^{i,\delta}_{\text{opt}} - f(x)|\} \tag{5.10}$$

that is the minimum point (if it exists) from which $\widehat{g}^{i,\delta}_{\text{opt}}$ is closer to the true density f than $\widehat{h}^{i,\delta}_{\text{opt}}$ for a given repetition i.

Simulation studies not presented in the paper show that $\widetilde{x}_i$ exists and is around the 99% quantile for almost all repetitions when the distribution is light-tailed and the sample sizes are reasonable large, whereas $\widetilde{x}_i$ rarely exists for heavy-tailed distributions and small sample sizes. That means that $\widehat{g}^{\delta}_{\text{opt}}$ fits the very high quantiles above 99% of the light-tailed Weibull distribution more accurately than $\widehat{h}^{\delta}_{\text{opt}}$ and that $\widehat{g}^{\delta}_{\text{opt}}$ also fits very high quantiles of the lognormal distribution more exactly than $\widehat{h}^{\delta}_{\text{opt}}$ when the sample sizes are large. However, when it comes to the lognormal distribution with small sample sizes, the heavy-tailed lognormal-Pareto distribution and the g-and-h distribution, then $\widehat{h}^{\delta}_{\text{opt}}$ mostly seems to fit the high quantiles better than $\widehat{g}^{\delta}_{\text{opt}}$.

The general conclusion of the Monte Carlo study is that the Champernowne transformed kernel density estimator with cml parameters on the whole seems to be a better tail estimator than the gpd estimator at the same time as being an estimator on the entire axis. The Champernowne transformed kernel density estimator with maximum likelihood parameters has a superior FGU performance, but this estimator has a substantially poorer tail performance which makes it less attractive as an estimator for operational risk. Compared to the parametric g-and-h distribution the Champernowne transformed kernel density estimator with cml parameters has a superior tail as well as FGU performance. The g-and-h distribution is significantly improved by use of non-parametric correction, and this estimator seems to be superior to the Champernowne transformed kernel density estimator with cml parameters in some situations when the sample sizes are very large. For small sample sizes and lighter tails and for the tail performance with $\delta = 2$ the Champernowne transformed kernel density estimator with cml parameters are superior to the g-and-h transformed kernel density estimator, and therefore the Champernowne transformed kernel density estimator with cml parameters is our final recommendation for operational risk estimation.

5.6 An application to operational risk

In this section we demonstrate the proposed methods on an operational risk data set. The data set consists of 5,021 financial operational risk events with the corresponding information of risk event category defined equivalent to the categories in Basel II. Descriptive statistics for the data set can be found in Table 5.7. The data set corresponds to the data set applied in Buch-Kromann et al. (2007) and more information about the data set can be found there.

Risk category	Number of losses	Max loss (£M)	Median (£M)	Mean (£M)	Standard deviation	Annual frequency
Internal fraud	1247	6683.8	1.82	32.24	269.43	10
External fraud	538	910.6	2.14	15.60	69.68	20
Employment practices and workplace safety	721	221.9	1.98	7.84	20.04	28
Business disruption	45	117.6	5.88	22.46	33.25	11
Damage to physical assets	2395	39546.4	2.35	74.91	1192.55	3
Execution, delivery and process management	75	104.6	1.56	7.39	17.72	52

Table 5.7: Descriptive statistics for the operational risk data set.

As mentioned in the introduction, LDA is an important model in operational risk modeling. The LDA consists of a severity and a frequency distribution and describes the aggregated loss distribution,

$$Y_j = \sum_{i=1}^{N_j} X_{ij}.$$

To each of the six operational risk event risk groups, we estimate the Champernowne transformed kernel density estimator with cml parameters. For this data set it is not possible to choose optimal thresholds as the true distribution of the data set is unknown just like it always is in practice. Instead we choose a threshold which

optimize the goodness-of-fit:

$$t_{\text{gof}} = \arg\min_{t} \sum_{i=1}^{n} \{F_{\text{emp}}(X_i) - \widehat{F}_t(X_i)\}^2, \tag{5.11}$$

where F_{emp} is the empirical cdf. To avoid outlying values, particularly when the sample size is small, it seems preferable to choose the goodness-of-fit optimizing threshold between 60% and 80%. This approach provides a threshold based on the *empirical* and not the true distribution. The estimator based on the goodness-of-fit optimized threshold is called $\widehat{h}_{\text{gof}}$.

Furthermore, we estimate a gpd distribution with a fixed threshold at $u_1 = 0.85$ for all the six event risk groups, and combine the gpd estimator with the empirical distribution in such a way that everything below $u_2 = 0.95$ is fitted by the empirical distribution, and everything above $u_2 = 0.95$ is fitted by the gpd distribution. This estimator (estimated to each single operational risk event risk group) is called $\widehat{g}_{0.85}$.

In addition, we estimate the Champernowne transformed kernel density estimator with maximum likelihood parameters, $\widehat{h}_0$, the g-and-h transformed kernel density estimator, $\widehat{h}_{\text{gh}}$ and the parametric estimators, $\widetilde{h}_{\text{gof}}$, $\widetilde{h}_0$ and $\widetilde{h}_{\text{gh}}$ together with the purely empirical distribution, called p_{emp}.

We compute a simulation study to calculate the 99.9% VaR and the 99.9% TVaR for our various versions of estimated severity distributions. The risk tolerance value 99.9% is chosen according to the Basel II standards for operational risk.

The simulation setup follows the setup in Buch-Kromann et al. (2007) and is the following. We draw 10,000 operational risk claims numbers for each risk category using a Poisson distribution (the Poisson parameters are stated in the last column in Table 5.7).

$$r_{ij} \sim \text{Poisson}(\lambda_i), \qquad i = 1, ..., 6, \qquad j = 1, ..., 10,000$$

For each of the simulated number of operational risk claims, we draw r_{ij} iid uniformly distributed random variables

$$u_{ijk} \sim U(0,1), \qquad k = 1, ..., r_{ij}$$

and based on these, we calculate the simulated operational risk claims by means of our estimated severity distributions

$$x_{ijk} = \widehat{F}_i^{-1}(u_{ijk})$$

where $\widehat{F}_i^{-1}(x)$ is the inverse cumulative distribution function corresponding to one of the eight estimated severity distributions $\widehat{h}_{\text{gof}}$, $\widehat{h}_0$, $\widehat{h}_{\text{gh}}$, $\widehat{g}_{0.85}$, $\widetilde{h}_{\text{gof}}$, $\widetilde{h}_0$, $\widetilde{h}_{\text{gh}}$ and p_{emp}.

This simulation study give us the aggregated losses according to each of the eight severity distributions in 10,000 years. According to these data sets we find the mean, standard deviation, median, VaR-99.9% and TVaR-99.9%. This is repeated 200 times and the averages are stated in Table 5.8. The numbers in parentheses and italics in Table 5.8 are the standard deviation of the estimates. We recognize that the mean is significantly larger than the median, which indicates that the distribution of each operational risk events group is right skewed. Focusing on the VaR-99.9% and TVaR-99.9% in Table 5.8 illustrates the significant disparities between the estimators. The empirical distribution estimates significantly lower VaR-99.9% and TVaR-99.9% than the other estimators, but also the g-and-h estimators with and without non-parametric correction are substantially lower than the Champernowne estimators and the gpd estimator. The VaR and TVaR for various values of quantiles are illustrated in Figure 5.4 for some selected estimators. The plots indicate that $\widehat{h}_0$ substantially overestimates the operational risk as regards VaR as well as TVaR, and, compared with the results in the Monte Carlo simulation study, we conclude that this estimator presumably overestimate the risk. On the other hand $\widetilde{h}_{\text{gh}}$ is substatially lower than both $\widehat{h}_{\text{gof}}(x)$ and $\widehat{g}_{0.85}(x)$ and this indicates that this estimator might

produce imprudent estimates for this data set.

	Mean	SD	Median	VaR-99.9%	TVaR-99.9%
$\widehat{h}_{\text{gof}}$	3,744*(6,269)*	131,146*(623,800)*	1,390*(7)*	133,780*(41,932)*	1,666,230*(6,261,244)*
$\widehat{h}_0$	44,171*(80,835)*	3,084,456*(7,758,720)*	1,594*(12)*	1,658,766*(695,492)*	37,990,347*(80,780,170)*
$\widehat{h}_{\text{gh}}$	2,518*(159)*	10,208*(12,263)*	1,655*(11)*	65,253*(15,188)*	198,421 *(151,049)*
$\widehat{g}_{0.85}$	6,082*(9,966)*	304,488*(977,792)*	1,294*(6)*	258,669*(103,758)*	3,865,910*(9,941,709)*
$\widetilde{h}_{\text{gof}}$	7,099*(14,700)*	296,669*(1,463,434)*	2,151*(10)*	264,527*(85,689)*	3,699,343*(14,683,800)*
$\widetilde{h}_0$	40,176*(71,558)*	2,809,963*(6,847,125)*	1,396*(11)*	1,520,906*(634,974)*	34,618,228*(71,517,690)*
$\widetilde{h}_{\text{gh}}$	2,907*(163)*	10,656*(12,435)*	1,911*(13)*	71,220*(16,528)*	207,439*(154,093)*
p_{emp}	1,597*(19)*	1,971*(128)*	1,247*(5)*	33,364*(2,502)*	37,522*(1,524)*

Table 5.8: The performance of the various versions of severity estimators in the operational risk application. The values in parentheses and italics are the standard deviation of the estimates. The severity estimators are the Champernowne transformed kernel density estimator with gof optimized cml parameters, $\widehat{h}_{\text{gof}}$, the Champernowne transformed kernel density estimator with maximum likelihood parameters, $\widehat{h}_0$, the g-and-h transformed kernel density estimator, $\widehat{h}_{\text{gh}}$, the gpd estimator, $\widehat{g}_{0.85}$, the parametric Champernowne distribution with gof optimized cml parameters, $\widetilde{h}_{\text{gof}}$, the Champernowne distribution with maximum likelihood parameters, $\widetilde{h}_0$, the g-and-h distribution, $\widetilde{h}_{\text{gh}}$, and the empirical distribution, p_{emp}.

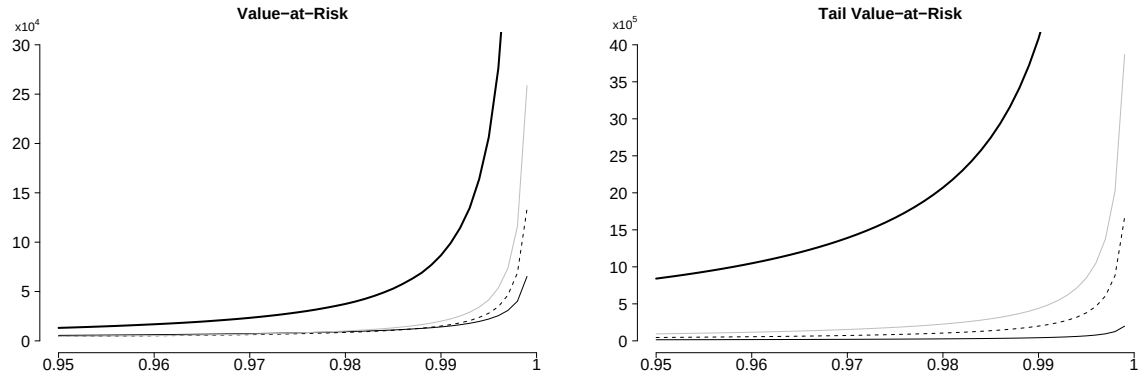

Figure 5.4: VaR (left) and TVaR (right) of the operational risk data set as a function of quantiles for some selected estimators. Black thick lines corresponds to the Champernowne transformed kernel density estimator with maximum likelihood parameters $\widehat{h}_0$, gray lines corresponds to the gpd estimator $\widehat{g}_{0.85}$, black dashed lines corresponds to the Champernowne transformed kernel density estimator with cml parameters $\widehat{h}_{\text{gof}}$, and black solid lines corresponds to the g-and-h distribution, $\widetilde{h}_{\text{gh}}$.

5.7 Conclusion

In this paper we introduce a new tail-dependent parameter estimation method (the cml method) of the Champernowne distribution which substantially improves the performance of the Champernowne transformed kernel density estimator compared to the Champernowne transformed kernel density estimator with maximum likelihood parameters as described in Buch-Larsen et al. (2005). As benchmark estimators we use the generalized Pareto distribution and the g-and-h distribution which has become popular in operational risk as a flexible distribution with the ability to fit both the center and the tail. Comparing these estimator shows that the Champernowne transformed kernel density estimator with cml parameters is superior as regard tail performance and performance on the entire axis.

Furthermore we combine the transformation kernel density approach and the g-and-h distribution and introduce the g-and-h transformed kernel density estimator. This estimator outperforms the Champernowne transformed kernel density estimator with cml parameters for very large data sets with a moderately heavy tail, most likely because the g-and-h distribution has an extra parameter compared to the Champernowne distribution, which yield an additional flexibility which is an advantage for large data sets. Focusing on the very far tail (AWISE with $\delta = 2$) shows that the Champernowne transformed kernel density estimator with cml parameters seems to be superior. Comparing the two estimators' performance as estimators on the entire axis shows that the Champernowne transformed kernel density estimator with cml parameters is superior to the g-and-h transformed kernel density estimator.

As most operational risk data sets in practice are very small, and as the very far tail is of particular interest in operational risk, our overall conclusion is to recommend the Champernowne transformed kernel density estimator with cml parameters.

In the last part of the paper, we illustrate the proposed methods in a study of the aggregated loss distribution. This study indicates that the g-and-h distribution underestimate the very far tail and thereby not produce prudent operational risk

estimates for this data set as required in Basel II.

Chapter 6

Multivariate density estimation using dimension reducing information and tail flattening transformations

This chapter is an adapted version of Buch-Kromann et al. (2007).

We propose a nonparametric multiplicative bias corrected transformation estimator designed for heavy tailed data. The multiplicative correction is based on prior knowledge and has a dimension reducing effect at the same time as the original dimension of the estimation problem is retained. Adding a tail-flattening transformation improves the estimation significantly – particularly in the tail – and provides significant graphical advantages by allowing the density estimation to be visualized in a simple way. The combined method is demonstrated on a fire insurance data set and provides excellent performance in a data-driven simulation study.

6.1 Introduction

We study a two-dimensional nonparametric density estimation problem that arises in the estimation of right-skewed distributions. One particular application that we have in mind is density estimation of insurance claims distributions. In this paper we suggest a nonparametric multidimensional density estimator based on prior knowledge, which has a dimension reducing effect.

The interest on multivariate analysis of risks has recently grown enormously, but most authors work in the framework of parametric models. For instance, Brodin and Rootzén (2009) applied a bivariate distribution to modelling wind storm and hurricane risks. Genest et al. (2009) note that modeling and measurement of multivariate risk in insurance and finance is an extremely challenging and important area of research. Recent contributions based on the parametric approach include the ones by Valdez et al. (2009), Li and Peng (2009), Hashorva (2008), Gebizlioglu and Yagci (2008) or Kallenberg (2008). Simultaneously, only few authors suggest a nonparametric analysis (see, Koekemoer and Swanepoel (2008a), Vilar et al. (2009), Cao et al. (2009) or Bolancé et al. (2008b)).

We have realized that the multidimensional nature of a problem can provide new insights, i.e., using information on risk scores can help to predict insurance claims severity by estimating the conditional density. However, compared to the majority of other authors, we are specially concerned about tail estimation in the challenging situation of right-skewed data. A major issue for multivariate density estimation is the curse of dimensionality whereby the optimal rate of convergence declines rapidly with dimensions, Stone (1980), and in practice this problem is even more pronounced when focusing on tail estimation. Moreover, as shown in Silverman (1986) the finite sample performance of standard nonparametric estimators is poor even when the dimension is quite small. One approach to this problem is working with restricted models that reduce the dimensionality. In many cases though it may be felt too unrealistic to rely so much on a restrictive model. When the model is not true the proposed estimators are not consistent and may be badly biased leading to misleading

inferences. Therefore we shall retain the high dimensional model assumption and seek to improve the performance of standard kernel estimation.

Nonparametric kernel density estimation has received considerable attention in the literature, see Wand and Jones (1995); Scott (1992) for useful introductions, and the standard kernel density estimator has been improved in several ways. One important contribution is Hjort and Glad (1995) which introduced a multiplicative bias correction based on a parametric distribution which improved the performance of the kernel density estimator substantially. In one dimension a parametric start is necessary to obtain an improvement in the rate of convergence, however in many dimensions it is sufficient to have a dimension reducing model. We therefore extend the idea in Hjort and Glad (1995) and propose a method based on prior knowledge. Our method is based on a structured auxiliary model which is multiplicatively bias corrected in a nonparametric way. The multiplicative correction is obtained after a tail flattening transformation, e.g. the Champernowne c.d.f. which has shown desirable properties especially when dealing with heavy-tailed distributions, see Buch-Larsen et al. (2005). That means that the structured auxiliary model and the tail flattening transformation is our prior knowledge which is nonparametrically corrected. We show that this approach improves the performance of the nonparametric kernel density estimator substantially if the prior knowledge is correct or almost correct.

The paper also includes a Monte Carlo study and an application. The Monte Carlo study is designed only to test the performance of the multiplicative bias correction. We have simulated from both an additive and a multiplicative model based on iid uniform variables and standard normal distributed error terms. This study shows that our proposed multiplicative bias correction brings improvements even in very small sample sizes. The application is based on a fire insurance data set, which consists of claim sizes and the corresponding explanatory variable: the estimated maximum loss. This is a heavy tailed data set, which makes tail flattening transformation essential. First, we apply the proposed methods on the fire data set and demonstrate the graphical advantage of working on transformed scales. Thereafter, we perform a simulation study based on the fire insurance data set, which gives us

a kind of bootstrapped evidence of the performance of our proposed density estimators. As this is the first paper focusing on the special problems that arise when focusing on heavy tails in multidimensional kernel density estimation, we will use the experience from corresponding work in one dimension and keep it as simple and clear as possible. Likewise, we stick to a data-driven simulation study in two dimensions corresponding to the dimension of the original data. The simulation study compares the performance of the proposed transformation kernel density estimator with and without multiplicative correction both when the auxiliary model is correct and when it is not correct. We conclude that when the auxiliary model is not correct we obtain a substantial improvement by using the multiplicative bias corrected transformation kernel density estimator compared to both the transformation kernel density estimator without multiplicative bias correction and the auxiliary model. When the auxiliary model is true, the multiplicative corrected estimator performs as well as the auxiliary.

In the simpler one-dimensional case, there has recently been a lot of activity to understand optimal transformation methods and optimal nonparametric smoothing methodology in our context. Some early papers in this direction was Bolancé et al. (2003) and Buch-Larsen et al. (2005) that updated the well known transformation method of Wand et al. (1991) to actuarial loss distributions. Simultaneously and independently, Clements et al. (2003) suggested the Mobius transformation. Buch-Larsen et al. (2005) noticed that the Mobius transformation turns out to be a special case of the well known Champernowne distribution that has a long history as a useful distribution for long tail estimation, see Brown (1937); Champernowne (1952). Based on an extensive simulation study Buch-Larsen et al. (2005) suggests to combine a modified version of the Champernowne distribution with a simple local constant kernel density estimator when estimating heavy tail distributions. This approach proved to be an improvement to the earlier studies quoted above. In the context of operational risk, Buch-Kromann et al. (2007) concluded that the transformation approach worked very well and led to a robustifying property when combined with alternative prior assumptions of parametric distributions. Also, Gustafsson et al.

(2006b), Guillen et al. (2007) and Gustafsson and Nielsen (2008) took advantage of the transformation approach defined in Buch-Larsen et al. (2005). However, Bolancé et al. (2008a) and Gustafsson et al. (2009) replicated the above simulations and showed that the more complicated approach of local beta density estimators and a beta distribution transformation approach proved to be a genuine improvement to the local constant estimator that had proved so hard to beat. However, the improvement was relatively modest compared to the increased level of complexity introduced by this new less known class of density estimators. When generalising the method to multidimensions as we do in this first paper on the topic, we have chosen to stick to the simple local constant kernel estimator that clearly is the benchmark to which all other approaches should be compared to. In Buch-Kromann et al. (2009) the Champernowne transformed kernel density estimator with different parameter estimation methods is compared as regards tail performance as well as performance on the entire axis to the generalized Pareto distribution and the g-and-h distribution which has become popular particularly in operational risk as a very flexible distribution with the ability to fit both the center and the tail of the distribution, see Dutta and Perry (2006), Degen et al. (2007) and Degen and Embrechts (2008). This comparison is generally in favour of the Champernowne transformed kernel density estimation.

In section 6.2 we describe the statistical setup for the multiplicative bias reduction method for joint and conditional densities. In section 6.3 we describe the multivariate transformation approach. In section 6.4 we combine the multiplicative bias reduction from section 6.2 with the transformation approach from section 6.3 to obtain a final estimator which benefits from both approaches. In section 6.5 we present the asymptotic theory of the described estimators. Section 6.6 presents a Monte Carlo study of the pure multiplicative bias correction and section 6.7 is an application on a heavy tailed fire insurance data set, where we demonstrate the performance of the proposed density estimator including tail flattening transformation. Section 6.7 also includes a data-driven simulation study of the presented estimators. Section 6.8 is the conclusion.

6.2 Multiplicative bias correction by structured nonparametric model

Suppose that $X = (X^1, \ldots, X^d)$ is a d-dimensional absolutely continuous random vector and we are interested in estimating the density function f based on a random sample of vectors $X_1, \ldots, X_n$. We are primarily concerned with the case where the support is unbounded in at least some directions. We are also interested in the conditional density of X^1 given $X^2, \ldots, X^d$, and functionals thereof like the conditional expectation and conditional median. A commonly employed estimator of $f(x)$ is the kernel estimator

$$\widehat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K_H \left(x - X_i \right), \tag{6.1}$$

where K is a multivariate kernel function and H is a $d \times d$ bandwidth matrix so that $K_H(x) = K(H^{-1}x)/\det(H)$. For pedagogic simplicity we consider the special case where $H = hI_d$. In practice one needs to make scale adjustments when the components of X have different marginals but we shall avoid this for notational simplicity. The conditional density $f(x^1|x^2, \ldots, x^d)$ can be estimated by $\widehat{f}(x^1|x^2, \ldots, x^d) = \widehat{f}(x)/\widehat{f}(x^2, \ldots, x^d)$, where $\widehat{f}(x^2, \ldots, x^d)$ is the corresponding estimator of the density of $(X^2, \ldots, X^d)$. See Chen et al. (2001) for a recent discussion of bandwidth issues in conditional density estimation.

Suppose that there is an auxiliary model for X with density denoted by g. The sort of model we have in mind is semiparametric: it depends in a known fashion on parameters $\theta \in \Theta \subseteq \mathbb{R}^p$ and on unknown one-dimensional functions $m_1, \ldots, m_R$ for some R. We can capture this with the general notation $g(.) = G(.; \theta, m_1, \ldots, m_R)$. The function m_j is defined on the domain of a one-dimensional random variable Z_j, where $Z_j = \psi_j(X)$ for some known measurable function ψ_j. It is natural to assume that $R \leq d$ here. For example, g might be the product density or g might be elliptically symmetric. Other examples include partially linear partially additive regression models among components of X. We consider a specific example in the application below.

Suppose that we can estimate the parameter θ and the functions $m_1, \ldots, m_R$ by estimates $\widehat{\theta}$ and $\widehat{m}_1, \ldots, \widehat{m}_R$, so that we can write $\widehat{g}(.) = G(.; \widehat{\theta}, \widehat{m}_1, \ldots, \widehat{m}_R)$, which is our estimate of the function g. We then estimate $f(x)$ by the multiplicative correction estimator

$$\widetilde{f}(x) = \widehat{g}(x) \frac{1}{n} \sum_{i=1}^{n} \frac{K_H(x - X_i)}{\widehat{g}(X_i)}. \tag{6.2}$$

For computation of the estimate $\widetilde{f}(x)$ at point x we need to compute $\widehat{g}(X_i)$ for all X_i in a neighbourhood of x determined by the bandwidth H.

In the case of conditional density $f(x^1|x^2, \ldots, x^d) = f(x)/f(x^2, \ldots, x^d)$, it is natural to start with an auxiliary model for the conditional density $g(x^1|x^2, \ldots, x^d)$, and suppose that one has an estimate $\widehat{g}(x^1|x^2, \ldots, x^d)$. Then define

$$\widetilde{f}(x^1|x^2, \ldots, x^d) = \frac{\widehat{g}(x^1|x^2, \ldots, x^d)}{\widehat{f}(x^2, \ldots, x^d)} \frac{1}{nh^d} \sum_{i=1}^{n} \frac{K\left(\frac{x - X_i}{h^d}\right)}{\widehat{g}(X_i^1|X_i^2, \ldots, X_i^d)}, \tag{6.3}$$

where $\widehat{f}(x^2, \ldots, x^d)$ is an estimate of the density $f(x^2, \ldots, x^d)$. This allows one to avoid specifying an auxiliary model for $f(x^2, \ldots, x^d)$. An alternative estimator is $\widetilde{f}(x)/\widetilde{f}(x^2, \ldots, x^d)$, which requires an auxiliary model for the full joint density. Provided the kernels are positive and the auxiliary density estimators $\widehat{g}$ are positive, the resulting estimators (6.2) and (6.3) are positive. One may make further adjustments to $\widetilde{f}(x)$ and $\widetilde{f}(x^1|x^2, \ldots, x^d)$ in order to make them integrate to one, see Glad et al. (2003).

This is a generalization of the principles involved in Hjort and Glad (1995) where the model g is fully parametric. In this paper g is a structured model containing some parametric components and some one-dimensional nonparametric components. The advantage of this is that such models might be more realistic and closer to the functional form of f, thereby producing better statistical performance. The basic motivation of the estimators (6.2) and (6.3) is that of prewhitening. The advantage of our approach over the parametric pilot approach is in terms of the bias: since our auxiliary model is in some sense larger, we expect to achieve smaller biases.

6.3 The multivariate transformation approach

In this section we propose a multivariate version of the univariate transformation approach from Buch-Larsen et al. (2005). An advantage of the 'transformation' method, evident from the theoretical analysis, is that it works well in the tail by effectively increasing the bandwidth there, and indeed has been interpreted as a form of variable bandwidth method, Yang and Marron (1999) and Bolancé et al. (2003).

Consider the invertible transformations $u = T(x;\lambda)$ depending only on parameters $\lambda \in \Lambda \subseteq \mathbb{R}^p$ and known functions T. Usually it is convenient to take just marginal transformations so that $u_j = T_j(x_j;\lambda_j), j = 1,\ldots,d$. It follows by the transformation theorem that

$$f(x) = J(x) f_U\{u(x)\}, \tag{6.4}$$

where f_U is the density of $U = (U_1,\ldots,U_d) = (T_1(X_1;\lambda_1),\ldots,T_d(X_d;\lambda_d))$, while J is the Jacobian of the transformation, in this case $J(x) = \prod_{j=1}^{d} |\partial T_j(x_j;\lambda_j)/\partial x_j|$. Then suppose that there exists an estimator $\widehat{\lambda}$ of λ computed from the data and let

$$\widetilde{f}_T(x) = \widehat{J}(x)\frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{\widehat{u}(x) - \widehat{U}_i}{h}\right), \tag{6.5}$$

where $\widehat{U}_i = T_i(X_i;\widehat{\lambda})$, while $\widehat{J}$ is the Jacobian of the empirical transformation

$$\widehat{J}(x) = \prod_{j=1}^{d}\left|\frac{\partial T_j(x_j;\widehat{\lambda})}{\partial x_j}\right|.$$

In general the transformed variable U can have the same support as X or different support, for example the unit cube. Examples of transformations include the Champernowne transformation, Buch-Larsen et al. (2005). This has a Pareto-like tail but with some extra flexibility. The heavy tail is useful in some applications. Note that when T_j is actually the marginal c.d.f. of X_j, U_j is uniformly distributed

on $[0,1]$ and the Jacobian is just the product of the densities of X_j. In this case, the most natural way of estimating the parameters λ is by (quasi) maximum likelihood since $T(x;\lambda)$ can be interpreted as a model for the multivariate c.d.f. of the data. The transformation ensures better tail behaviour, as discussed in Buch-Larsen et al. (2005).

6.4 Multiplicatively corrected transformation approach

In the following we propose to combine the transformation approach from section 6.3 with the multiplicative bias reduction method from section 6.2. The advantage of this is that we are doing our density estimation and our primary multiplicative bias correction on a scale of our choosing, for example on the unit cube. If the transformation is particularly good one has almost equally spaced data on the unit cube and bandwidth choice can be less crucial. The full algorithm is given below.

1. Estimate the auxiliary model density $\widehat{g}(x)$ using appropriate techniques.

2. Transform the data to $\widehat{U}_i = T(X_i;\widehat{\lambda})$ and the implied auxiliary model density to $\widehat{g}_U(u)$, where $\widehat{g}_U(u) = \widehat{g}\{T^{-1}(u;\widehat{\lambda})\}\{T^{-1}(u;\widehat{\lambda})\}'$.

3. Estimate the density of U by the multiplicative correction estimator

$$\widetilde{f}_U(u) = \widehat{g}_U(u)\frac{1}{nh^d}\sum_{i=1}^{n}\frac{K\left(\frac{u-\widehat{U}_i}{h}\right)}{\widehat{g}_U(\widehat{U}_i)} \tag{6.6}$$

4. Transform back from U to X to obtain a density estimator of X on the original axis

$$\widetilde{f}_C(x) = \widehat{J}(x)\widetilde{f}_U\{u(x)\}. \tag{6.7}$$

This algorithm benefits from transforming into the unit cube because this means that the issue of different scaling of the variables is handled automatically.

6.5 Distribution theory

6.5.1 Full model case

In this section we derive the distribution theory for our procedures in a general setting. The main insight is that although there are nonparametric components in the auxiliary model g, they are of lower dimension than f and so the estimation error in $\widehat{g}$ is of smaller order and can be ignored under the general model assumption. Define $g(x)$ as the limiting value of $\widehat{g}(x)$. We assume that this is well-defined. It may depend on the method used to compute $\widehat{g}$. Define also

$$\beta(x) = \frac{1}{2}\mu_2(k)g(x)\nabla_2 r(x) \quad ; \quad v(x) = ||K||_2^2 f(x),$$

where $r(x) = f(x)/g(x)$ and $\nabla_2 g(x) = \sum_{j=1}^{d} \partial^2 g(x)/\partial x_j^2$ is the trace Hessian operator, while $\mu_2(k) = \int k(t)t^2 dt$ and $||K||_2^2 = \int K(u)^2 du$. In the appendix we show

THEOREM 1. *Suppose that assumptions A in Appendix A are satisfied. Then as* $n \to \infty$,

$$\sqrt{nh^d}\{\widetilde{f}(x) - f(x) - h^2\beta(x)\} \Longrightarrow N(0, v(x)).$$

PROOF See Appendix B.

When $h = O(n^{-1/(d+4)})$ one obtains convergence in distribution at rate $n^{-2/(d+4)}$, which is the optimal rate under our smoothness conditions. The limiting variance is the same as that of the standard kernel estimator but the bias is different. It depends on the curvature of r rather than f: when $r(x)$ is flat (i.e., g is close to f near x) then the bias is small. In the extreme case when the model is true, i.e., $g(x) = f(x)$, the bias constant is zero, and the rate of convergence can potentially be increased

by taking a larger bandwidth; we discuss this in section 6.5.2.

We next present the distribution theory for the transformation estimator. Define

$$\beta_T(x) = \frac{1}{2}\mu_2(k)J(x)\nabla_2 f_U\{u(x)\} \quad ; \quad v_T(x) = ||K||_2^2 J(x)f(x).$$

THEOREM 2. *Suppose that assumptions A and B in Appendix A are satisfied. Then as* $n \to \infty$,

$$\sqrt{nh^d}\{\widetilde{f}_T(x) - f(x) - h^2\beta_T(x)\} \Longrightarrow N(0, v_T(x)).$$

PROOF See Appendix C.

In this case, both the bias and the variance are different from the standard kernel estimator. The bias depends on the curvature of the density of the transformed variable, so that when U is actually uniform (i.e., the transform $T(x;\lambda_0)$ is the c.d.f. of X) the bias constant is zero. The variance can be smaller in the tails due to the Jacobian term, i.e., when $J(x) < 1$, $v_T(x) < v(x)$; this condition is likely to be met out in the tails provided the transformation is well chosen. For example, when T_j is the c.d.f. of X_j, then $J(x) = \prod_{j=1}^d f_j(x_j) \to 0$ as $x_j \to \infty$. In this case, the transformation estimator has finite relative error in the tail, in the sense that

$$\lim_{x\to\infty} \text{avar}\left\{\sqrt{nh^d}\frac{\widetilde{f}_T(x)}{f(x)}\right\} < \infty. \tag{6.8}$$

By comparison, the relative error of $\widehat{f}(x)$ (and hence of $\widetilde{f}(x)$) in the tail becomes infinite.

Finally, we provide the theory for the combination estimator. Define $r_U(u) = f_U(u)/g_U(u)$ and

$$\beta_C(x) = \frac{1}{2}\mu_2(k)J(x)g_U\{u(x)\}\nabla_2 r_U\{u(x)\} \quad ; \quad v_C(x) = ||K||_2^2 J(x)f(x).$$

THEOREM 3. *Suppose that assumptions A and B in Appendix A are satisfied. Then as* $n \to \infty$,

$$\sqrt{nh^d}\{\widetilde{f}_C(x) - f(x) - h^2\beta_C(x)\} \Longrightarrow N(0, v_C(x)).$$

PROOF See Appendix D.

The variance is the same as the variance of the transformation estimator, while the bias is slightly different. Specifically, the magnitude of the bias depends on the curvature of $r_U(u) = f_U(u)/g_U(u)$ in u-space. If the auxiliary model g is 'good', i.e., g_U is close to uniform, and if the transformation T is 'good', i.e., f_U is close to uniform, then the ratio $f_U(u)/g_U(u)$ is also close to uniform. However, it can be that $f_U(u)/g_U(u)$ is approximately constant even when the two functions $f_U(u)$ and $g_U(u)$ are not constant, and in such cases one will have small bias. Note that when $r(x) = 1$ for all x, then $r_U(u) = 1$ also.

Finally, we present the theory for the conditional density estimator (6.3). This theory parallels the theory for the joint density estimators and so we omit the results for the transformation estimator and the combined estimator. Define:

$$\begin{aligned}
\beta(x^1|x^2,\ldots,x^d) &= \beta_1(x^1|x^2,\ldots,x^d) + \beta_2(x^1|x^2,\ldots,x^d) \\
\beta_1(x^1|x^2,\ldots,x^d) &= \frac{1}{2}\mu_2(k)\frac{g(x^1|x^2,\ldots,x^d)}{f(x^2,\ldots,x^d)}\sum_{j=1}^{d}\frac{\partial^2\{r(x^1|x^2,\ldots,x^d)f(x^2,\ldots,x^d)\}}{\partial(x^j)^2} \\
\beta_2(x^1|x^2,\ldots,x^d) &= -\frac{1}{2}\mu_2(k)\frac{f(x^1|x^2,\ldots,x^d)}{f(x^2,\ldots,x^d)}\sum_{j=2}^{d}\frac{\partial^2 f(x^2,\ldots,x^d)}{\partial(x^j)^2} \\
v(x^1|x^2,\ldots,x^d) &= ||K||_2^2\frac{f(x^1|x^2,\ldots,x^d)}{f(x^2,\ldots,x^d)},
\end{aligned}$$

where $r(x^1|x^2,\ldots,x^d) = f(x^1|x^2,\ldots,x^d)/g(x^1|x^2,\ldots,x^d)$.

THEOREM 4. *Suppose that assumption A in Appendix A are satisfied. Then as*

$n \to \infty$,

$$\sqrt{nh^d}\{\widetilde{f}(x^1|x^2,\ldots,x^d) - f(x^1|x^2,\ldots,x^d) - h^2\beta(x^1|x^2,\ldots,x^d)\} \\ \Longrightarrow N(0, v(x^1|x^2,\ldots,x^d)).$$

PROOF See Appendix E.

The asymptotic variance is the same as that of $\widehat{f}(x^1|x^2,\ldots,x^d)$. The bias depends on the curvature of $r(x^1|x^2,\ldots,x^d)f(x^2,\ldots,x^d)$; for comparison the bias of $\widehat{f}(x^1|x^2,\ldots,x^d)$ depends on the curvature of $f(x)$ and $f(x^2,\ldots,x^d)$, see Chen et al. (2001).

6.5.2 Home turf case

We now consider what happens under the 'home turf' assumption; for brevity we just consider the case of Theorem 1. We suppose therefore that the auxiliary model is true, i.e., $g(x) = f(x)$ for all x. We shall also suppose that $\widehat{g}(\cdot)$ behaves like a one-dimensional smoother in the sense that it has an expansion of the form

$$\widehat{g}(x') - g(x') = h_g^2\beta_g(x') + \frac{1}{\sqrt{nh_g}}\omega^{1/2}(x')Z_n(x') + \mathcal{R}_n(x') \tag{6.9}$$

for x' in a neighborhood of x, where $\beta_g(\cdot)$ and $\omega^{1/2}(\cdot)$ are bounded continuous deterministic functions, $Z_n(\cdot)$ is a sum of independent random variables satisfying $Z_n(x') \Longrightarrow N(0,1)$ for each x', and $\mathcal{R}_n(\cdot)$ is a remainder term that is of smaller order in probability than the first two terms. The quantity h_g is a bandwidth sequence that we shall suppose is of order $n^{-1/5}$.We have the following corollary

COROLLARY 1. *Suppose that assumptions A1,A2, and C1,C2 in Appendix A are satisfied. Then as* $n \to \infty$,

$$\sqrt{nh_g}\{\widetilde{f}(x) - f(x)\} \Longrightarrow N(0, \omega(x)).$$

PROOF See Appendix F.

This shows that under the home turf case one obtains the faster rate of convergence provided bandwidth is chosen correctly. Furthermore, one also achieves a bias correction – the asymptotic distribution is centered at zero. Compare this with the result in (Hjort and Glad, 1995, section 8.3). They establish root-n consistency of their density estimator under a fixed bandwidth assumption but their estimator is not as efficient as the parametric start itself.

6.5.3 Bandwidth choice

The issue of bandwidth choice is very important but notoriously difficult to resolve. Our procedures for $\widetilde{f}$ and $\widetilde{f}_C$ involve smoothing to compute $\widehat{g}$ and then further smoothing to compute $\widetilde{f}$ or $\widetilde{f}_C$. One approach would be to use a least squares cross validation procedure as defined in Wand and Jones (1995) to jointly select the bandwidths but this can be quite computationally demanding. Instead we recommend using a rule of thumb plug-in method based on the asymptotic mean squared error expansions given above. Studies of corresponding estimators in one dimension show that improvements can be obtained by using more sophisticated bandwidth choices, however, the gain is small compared to the computational complexity which are introduced by more complicated procedures, see Buch-Larsen et al. (2005) and Gustafsson (2006) for the one dimensional case. Due to the well-chosen transformation function, the transformed data has a much more homogeneous structure compared to the original data set, and the fact that the bandwidths are used on the transformed data, means that simple bandwidth selection methods work fairly well for our estimators. Regarding the bandwidth for the auxiliary models, we recommend to use procedures developed specially for those models.

6.6 Monte Carlo study

In this Monte Carlo study we investigate the performance of the pure multiplicative bias correction without tail flattening transformation. We investigate the following additive and multiplicative designs:

$$Y_i = X_i + \varepsilon_i \tag{6.10}$$

$$Y_i = X_i \varepsilon_i, \tag{6.11}$$

where in each case ε_i are i.i.d standard normal and X_i are i.i.d. uniform on $[1, 2]$.

We compute the standard kernel density estimates, corresponding to (6.1), $\widehat{f}(y, x)$ and $\widehat{f}(x)$ with bandwidths $2 \times s \times n^{-1/6}$, where s is the standard deviation of the variable, and the estimates

$$\widehat{g}(y, x) = \widehat{f_\varepsilon}(y - x)\widehat{f}(x) \text{ and } \widehat{g}(y|x) = \widehat{f_\varepsilon}(y - x)$$

with bandwidths $2 \times s \times n^{-1/5}$, which assumes the additive model prevails. In each case a Gaussian kernel is used. Actually, since we compute the estimates at the sample points, the following simple matrix definitions are used

$$\widetilde{f} = \widehat{g}. * (W * (1./\widehat{g}) \text{ and } \widetilde{f}_{|x} = (\widehat{g}_{|x}./\widehat{f}_x). * (W * (1./\widehat{g}_{|x}),$$

where W is the $n \times n$ matrix with elements $k((Y_i - Y_j)/h_y)k((X_i - X_j)/h_x)/nh_xh_y$, $\widehat{g}$ is the $n \times 1$ vector with typical element $\widehat{g}(Y_i, X_i)$, $\widehat{g}_{|x}$ is the $n \times 1$ vector with typical element $\widehat{g}(Y_i|X_i)$, and $\widehat{f}_x$ is the $n \times 1$ vector with typical element $\widehat{f}(X_i)$. $\widetilde{f}$ corresponds to a two dimensional multiplicative correction density estimator (6.2) and $\widetilde{f}_{|x}$ is the corresponding conditional estimator.

For an estimator $\widehat{f}$, define the performance measure

$$\text{ISE}(\widehat{f}) = \frac{1}{n} \sum_{i=1}^{n} \left[\widehat{f}(X_i) - f(X_i) \right]^2 \tag{6.12}$$

for each sample. We averaged over 1000 samples to produce the mean integrated squared error (MISE). We compare the performance in Table 6.1 and Table 6.2. The additive design belongs to the "home turf" case, but the multiplicative is not correctly specified. Our method brings improvements even in very small sample sizes and for both unconditional and conditional density estimation and in both additive and multiplicative designs.

n	MISE($\widehat{f}$)	MISE($\widetilde{f}$)	MISE($\widehat{f}_{\vert x}$)	MISE($\widetilde{f}_{\vert x}$)
10	0.01207	0.01007	0.02388	0.02235
20	0.00997	0.00680	0.02131	0.01857
50	0.00744	0.00383	0.01800	0.01484
100	0.00581	0.00235	0.01509	0.01211
200	0.00435	0.00133	0.01253	0.00996

Table 6.1: Comparison of integrated squared error: Additive case.

n	MISE($\widehat{f}$)	MISE($\widetilde{f}$)	MISE($\widehat{f}_{\vert x}$)	MISE($\widetilde{f}_{\vert x}$)
10	0.00712	0.00612	0.01305	0.01247
20	0.00608	0.00460	0.01207	0.01109
50	0.00483	0.00311	0.01049	0.00926
100	0.00381	0.00220	0.00892	0.00783
200	0.00289	0.00150	0.00743	0.00654

Table 6.2: Comparison of integrated squared error: Multiplicative case.

6.7 Application

The application is based on a real commercial fire insurance data set from the Danish general insurance company Codan Insurance that contains claims reported from 1995 to 2004. The data set specifies two characteristics, for each individual claim, the claims amount Y, and the risk score that is provided by the expected maximum loss (EML), here called X. Our application corresponds to the two-dimensional case, $d = 2$. We will estimate the joint density of (Y, X) and the conditional density of $(Y|X)$. In this situation, the tail flattening transformation approach can be used to visualizing the dependence structure between the two variables. Moreover, since

the fire insurance data set is heavy-tailed, the estimation will benefit from the tail properties of the transformation approach, as described in section 6.3. In the last part of the section, a data-driven simulation study based on the same fire insurance data set compares the proposed estimators. The study shows that the transformation approach can be improved by using the multiplicative correction, without losing the visualization properties and the tail estimation properties from the transformation approach.

6.7.1 Analysis of the fire insurance data set

Data are taken from fire policies, which normally consist of three types of coverage: buildings, contents and loss of production. The analysis presented here only covers the fire claims on buildings. The data set consists of 2810 fire claims from a main trade group covering residences. This main trade group constitutes approximately 30% of the total claims cost of the fire building claims in the firm and it is therefore an important group of risk.

The claims are uncensored with claims from 19 Danish Krone (Dkr.) to about 6 million Dkr. and an average claim size at 56,220 Dkr. The claims are right-skewed with a skewness at $2.3 \cdot 10^{17}$, and therefore we use a log-scale plot in the histogram of the claim sizes in Figure 6.1 (left). Even on a log-scale, the claims are right-skewed. For each claim size in the data set, the corresponding EML is observed. The average EML is 47,417,302 Dkr. and EML is right-skewed as well. The histogram for the EML is shown in Figure 6.1 (right) on a log-scale. We expect that large claims sizes arise from policies with a large EML. This is confirmed in Figure 6.2 (left), which shows a plot of $(\log(X), \log(Y))$.

In the first estimation step we estimate the transformation kernel density for (Y, X), $\widehat{f}_T(y, x)$, by use of (6.5) in two dimensions. The bandwidth is chosen different for each component by use of Silverman's normal scale bandwidth, Silverman (1986). We have taken marginal transformations $u_j = T_j(x^j; \lambda_j)$. The univariate transformation

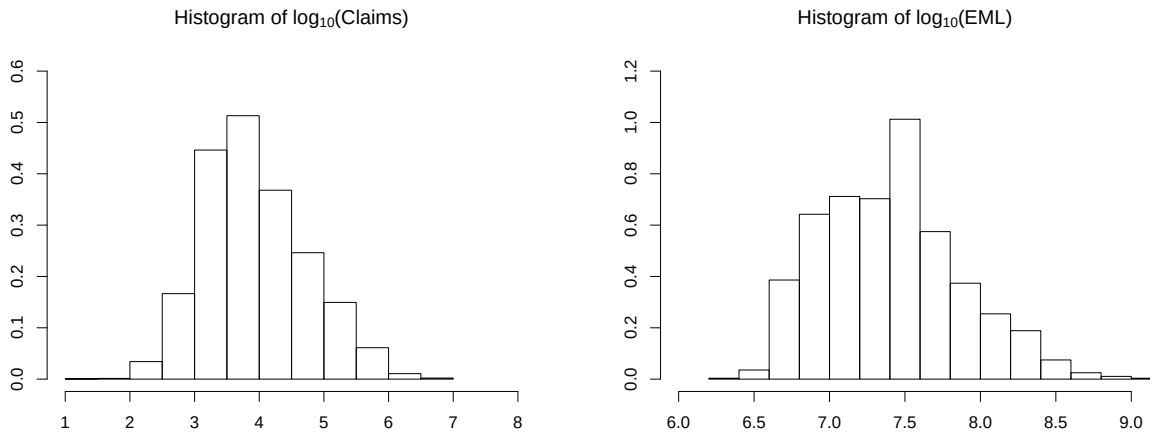

Figure 6.1: Histograms on logarithmic-axes of the claims and the EML's.

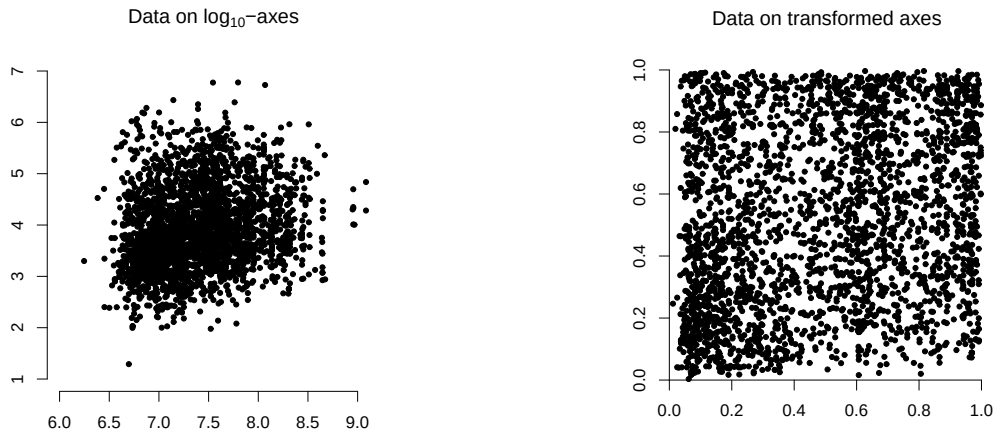

Figure 6.2: The fire claims data set on logarithmic- and Champernowne transformed axes.

function used in this application for each component is the modified Champernowne distribution, see Buch-Larsen et al. (2005), and parameter estimates are obtained by maximum likelihood for each component. The c.d.f. of the modified Champernowne distribution depends on three parameters and is equal to:

$$T_j(x^j;\lambda_j) = \frac{(x^j+c)^\alpha - c^\alpha}{(x^j+c)^\alpha + (M+c)^\alpha - 2c^\alpha} \tag{6.13}$$

The transformed data $(\widehat{U}_i^1, \widehat{U}_i^2)$ are presented in Figure 6.2 (right). We observe that the data are almost uniformly distributed. The conditional density of Y given X is obtained from the joint and the marginal density and is denoted $\widehat{f}_T(y|x)$. In the left plot in Figure 6.3 the estimated joint density of (U^1, U^2), $\widehat{f}_U(u^1, u^2)$ is shown. In the right plot in Figure 6.3 the conditional density of U^1 given U^2, $\widehat{f}_U(u^1|u^2)$, is shown. The hills in the bottom left corner and the upper right corner confirm the expectation that small policies (small EML's) have lower expected claim sizes than large policies (large EML's). The advantage of looking at the density plots on the transformed scale is that we can illustrate the density of the whole domain in one plot. It is a well known empirical fact that density estimates of heavy tailed insurance claims data presented in the original scale consist of an enormous concentration of small costs and several large costs far off in the tail. As a consequence, unless the domain is split in parts, the scale imposed by the size of the tail masks the behaviour near zero. Indeed, it is very difficult to compare two estimated densities in the whole domain when dealing with insurance claims data, because only the tails seem to matter. This is the reason why the results on the transformed scale are so useful, because this is a good way to compare the joint occurrence on the same scale.

The estimation approach described so far has treated the multivariate (bivariate) aspects of the estimation problem and also addressed the skewness by using the transformation approach. However, as argued in section 6.6, we expect that multiplicative correction will improve the results substantially, and therefore we define a median regression model as our auxiliary model. The median regression model is

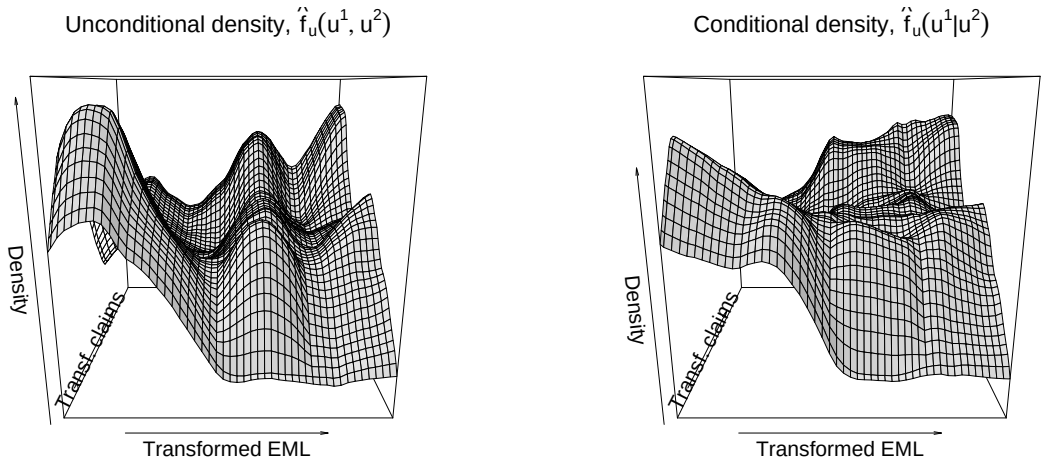

Figure 6.3: The nonparametric unconditional and conditional densities of EML and fire claims on transformed axes.

given by

$$Y = m(X)\varepsilon$$

where ε is lognormal distributed with median one for identification. The model is estimated with local polynomials and bandwidth b_x. To simplify the bandwidth selection problem, we use U^2 instead of X as explanatory variable and use a constant bandwidth of two times the Silverman's normal scale bandwidth which is approximately $b_u = 0.28$ for this data set. We did try more complicated bandwidth selection methods. However, simulations showed that the simple Silverman's normal scale bandwidth did better. The auxiliary model is therefore $Y = m(U^2)\varepsilon_u$.

From the auxiliary model we obtain the conditional density of Y given X, $\widehat{g}(y|x)$, and the corresponding conditional density on the transformed axes, $\widehat{g}_U(u^1|u^2)$. By use of the estimated marginal density of X, we obtain the joint densities of (Y, X) and (U^1, U^2), called $\widehat{g}(y, x)$ and $\widehat{g}_U(u^1, u^2)$, respectively. The densities $\widehat{g}_U(u^1, u^2)$ and $\widehat{g}_U(u^1|u^2)$ are shown in Figure 6.4. The tendencies in the densities correspond to the nonparametric densities in Figure 6.3, however, the auxiliary model introduces structure into the model which results in a more smooth density function.

Finally, the multiplicative corrected estimator of the joint density is obtained on

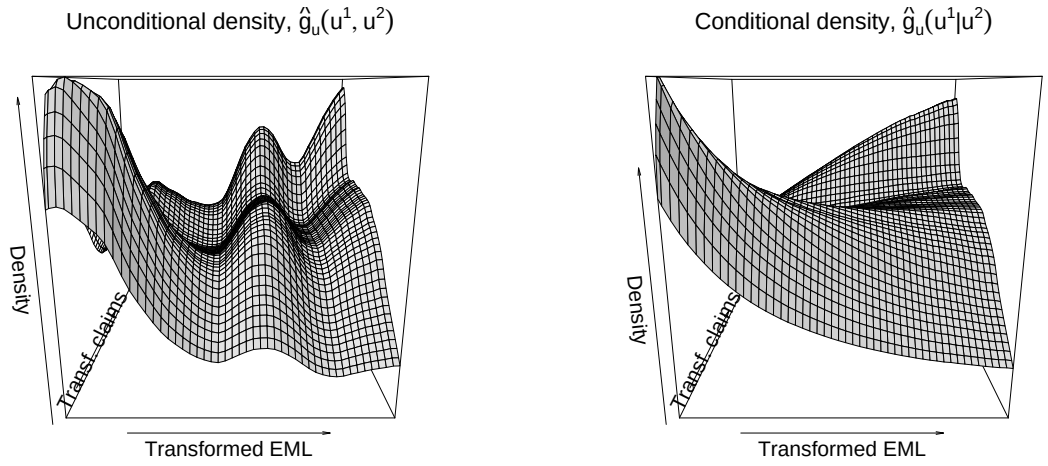

Figure 6.4: The unconditional and conditional densities of EML and fire claims on transformed axes under the auxiliary model.

the transformed axes, $\widetilde{f}_U(u^1, u^2)$, as in (6.6) where the same Silverman's normal scale bandwidth is used for each component. By back transformation (6.7) the multiplicative corrected estimator $\widetilde{f}_C(y, x)$ appears on the original axes.

The conditional densities on the transformed and the original axes, $\widetilde{f}_U(u^1|u^2)$ and $\widetilde{f}_C(y|x)$ are found by means of the marginal densities. In Figure 6.5 the multiplicative corrected estimate of the joint density of (U^1, U^2), $\widetilde{f}_U(u^1, u^2)$, and the conditional density of U^1 given U^2, $\widetilde{f}_U(u^1|u^2)$, are shown. The tendencies of hills in the bottom left and the upper right corner, meaning that large policies generate larger claims on average, is significant for the multiplicative corrected estimator as well. However, the smooth structure which was obtained in the auxiliary model is now corrected nonparametrically.

6.7.2 Data-driven simulation study

In section 6.6 we showed that the pure multiplicative correction seemed to improve estimation significantly. In this section we want to extend the investigation to a situation where a tail flattening transformation is applied. To investigate the per-

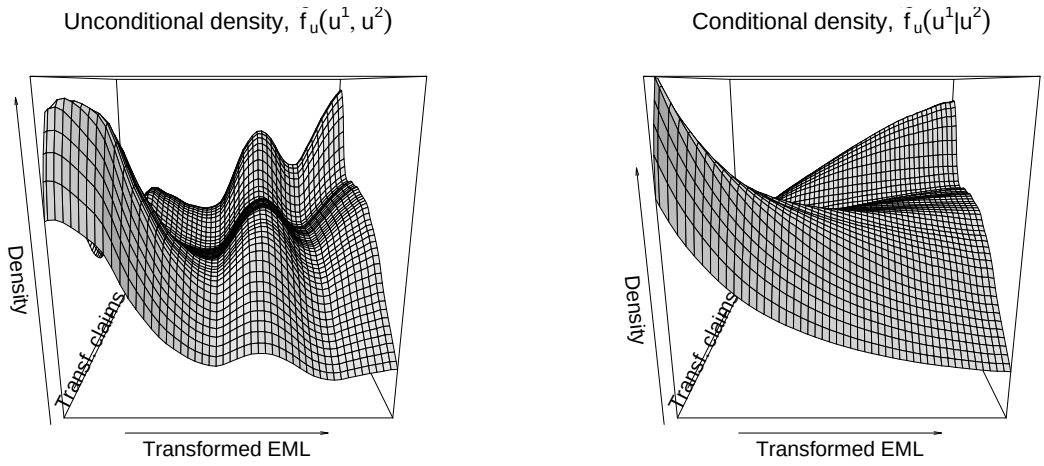

Figure 6.5: The unconditional and conditional densities of EML and fire claims on transformed axes under the multiplicative corrected model.

formance of the multiplicative correction in this situation, we perform a data-driven Monte Carlo experiment. In the design of the Monte Carlo study we want the simulated data to be as close to the "real-world" data as possible, to test the performance of our model in a real-world setup. The simulation study shows that the multiplicative corrected transformation estimator improves the performance compared to the estimator without multiplicative correction, and that the multiplicative corrected transformation estimator also improve the auxiliary model when this model is not quite correct without making it worse when the auxiliary model is correct (the home turf case).

We base the Monte Carlo simulation on the same data set as described in section 6.7.1 and describe the relationship between EML and claims in the data set by two multiplicative models

$$Y = \alpha X^{\beta} \varepsilon_1 \tag{6.14}$$

$$Y = \alpha X^{\beta} \varepsilon_2(X) \tag{6.15}$$

This approach ensures that we have simulated data sets of realistic order with a known distribution function.

We estimate the parameters in (6.14) by the least square method and obtain the parameters: $\hat{\alpha} = 182.37$, $\hat{\beta} = 0.32$, Moreover, we assume the residuals in (6.14) are i.i.d. and lognormal distributed, $\varepsilon_1 \sim \log \mathrm{N}(-1.62; 1.8)$. In (6.15) we let the parameters in the lognormal distribution of the residuals depends on x, $\varepsilon_2 \sim \log \mathrm{N}(\mu_x, \sigma_x)$, with a linear dependence on the Champernowne transformed axis, $\sigma_x = 1.5 + 0.5T(x)$ and $\mu_x = -0.5\sigma_x$. The estimated parameters $\hat{\alpha}$ and $\hat{\beta}$ from (6.14) is maintained.

To simulate data of "real-world" amounts, we first sample n values from X, (the EML's in the data set) and call the data set X^*. Corresponding to X^* we thereafter, simulate n random variables, called Y_1^*, from the estimated multiplicative model (6.14) and n random variables, called Y_2^*, from the estimated multiplicative model (6.15).

Moreover, for each sample we define the performance measure, ISE, for an estimator $\hat{f}$

$$\mathrm{ISE}(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{f}(X_i) - f(X_i)\right\}^2 \tag{6.16}$$

We compute 100 samples for each n, and estimate for each sample, (X^*, Y_1^*), the nonparametric density estimator, $\hat{f}_{1,T}(y|X^* = x)$, the auxiliary density estimator, $\hat{g}_{1,T}(y|X^* = x)$, the multiplicative corrected density estimator, $\tilde{f}_{1,C}(y|X^* = x)$, and the performance measure ISE, for each of the estimators. Thereafter, we average over the ISE for each estimator to produce the mean integrated squared error (MISE) for each estimator. Likewise for each sample, (X^*, Y_2^*), we calculate $\hat{f}_{2,T}(y|X^* = x)$, $\hat{g}_{2,T}(y|X^* = x)$, $\tilde{f}_{2,C}(y|X^* = x)$ and the average performance measure, MISE, for each estimator. The results are collected in Table 6.3.

The performance of $\mathrm{MISE}\{\tilde{f}_{1,C}(y|x)\}$ is comparable to $\mathrm{MISE}\{\hat{g}_{1,T}(y|x)\}$. This means that when the auxiliary model is the correct model of the data set as it is in model (6.14), the multiplicative correction makes as good results as the (correct) auxiliary model. However, $\mathrm{MISE}\{\tilde{f}_{2,C}(y|x)\}$ is smaller than $\mathrm{MISE}\{\hat{g}_{2,T}(y|x)\}$ for all n in the Monte Carlo experiment, which means that, in the much more realistic case, where the auxiliary model is not the correct model of the data set, the multiplicative cor-

rection of the auxiliary model improves the performance of the auxiliary model. The multiplicative corrected transformation estimator is therefore as good as the auxiliary models in the home turf case, whereas it improves the estimation when the auxiliary model is not correct.

	n=100		**n=500**		**n=1000**	
Model, m	m=1	m=2	m=1	m=2	m=1	m=2
MISE$\{\hat{f}_{m,T}(y\|x)\}$	0.06755	0.06071	0.02980	0.02791	0.02064	0.01888
MISE$\{\hat{g}_{m,T}(y\|x)\}$	0.05607	0.06244	0.02057	0.02882	0.01459	0.02300
MISE$\{\tilde{f}_{m,C}(y\|x)\}$	0.05429	0.05774	0.02110	0.02514	0.01493	0.01882

Table 6.3: Monte Carlo experiment with 100 samples and n observations in each sample.

6.8 Conclusion

This paper introduces an estimator which combines a multiplicative bias correction approach with the transformation approach in two dimensions. The estimator both benefits from the multiplicative correction through the dimension reducing information from the auxiliary model, and from the tail flattening transformation approach which moreover improves the visualization when transforming to the unit cube.

In the multivariate case there are many alternative bias reduction methods for density estimation. Our proposal has involved the use of semiparametric and structured nonparametric models that can be quite good approximations to unconstrained densities in certain aspects without being completely correct. These models have been the subject of quite a lot of recent work, and the estimation technology and distribution theory has provided a firm foundation for their use in applications. We also believe that they can greatly assist in the estimation of unconstrained multivariate densities and our theoretical and empirical work supports this.

In the last part of the paper we performed a Monte Carlo study of the pure multiplicative correction without tail flattening transformation, which shows that multiplicative correction seems to improve the results. Moreover, we made an application of the

estimation method to a heavy tailed fire insurance data set and investigated the performance of the transformation estimator with and without multiplicative correction in a data-driven simulation study. This study showed that the multiplicative correction significantly improves the performance of the transformation estimator when the auxiliary model is not perfectly correct, without aggravating the performance of the auxiliary model in the home turf case.

6.9 Appendix

6.9.1 Appendix A

Here we state the regularity conditions.

Assumption A.

1. *Suppose that f is twice continuously differentiable on its support $\mathcal{X} \subset \mathbb{R}^d$, and strictly positive at the interior point x.*

2. *Suppose that $K(u) = \prod_{j=1}^{d} k(u_j)$, where k is a continuous density function symmetric about zero (a second order kernel) with compact support.*

3. *Suppose that $nh^d \to \infty$ and $\limsup_n nh^{d+4} < \infty$.*

4. *The function g is well defined and twice continuously differentiable at x. For some $\epsilon > 0$,*
$$\sup_{|x-x'|\leq\epsilon} |\widehat{g}(x') - g(x')| = o_p(n^{-1/2}h^{-d/2}). \tag{6.17}$$

Assumption A4 is satisfied under a variety of conditions for many estimators in structured nonparametric and semiparametric models, it just requires that $\widehat{g}$ converges to some limit g faster than $\widehat{f}$. Typically, one can obtain one-dimensional uniform

convergence rates for $\widehat{g}(x)$, i.e.,
$\sup_{|x-x'|\leq\epsilon} |\widehat{g}(x') - g(x')| = O_p((\log n/n)^{-2/5})$, which would imply (6.17).

Assumption B.

1. *There exists a* $\lambda_0 \in \Lambda \subseteq \mathbb{R}^p$ *such that* $\sqrt{n}(\widehat{\lambda} - \lambda_0) = O_p(1)$.

2. *The transformation* $T : \mathbb{R}^d \longmapsto \mathbb{R}^d$ *is invertible and twice continuously differentiable in* λ. *Furthermore, there exists a non-negative function* $d(.)$ *with* $Ed(X) < \infty$ *such that for some sequence* $\delta_n \to 0$,

$$\max_{1\leq j\leq d} \sup_{\lambda:\sqrt{n}\|\lambda-\lambda_0\|\leq\delta_n} \left\| \frac{\partial^2 T_j}{\partial\lambda\partial\lambda^\top}(x;\lambda) \right\| \leq d(x).$$

3. *The kernel* k *is twice continuously differentiable.*

Assumption C.

1. *The bandwidth* h *satisfies* $hn^{1/5d} \to \infty$

2. *For some* $\epsilon > 0$, *the expansion (6.9) holds with*

$$\begin{aligned} \sup_{x':|x-x'|\leq\epsilon} |Z_n(x')| &= O_p\left(\sqrt{\log n} n^{-1/2} h_g^{-1/2}\right) \quad ; \\ \sup_{x':|x-x'|\leq\epsilon} |\mathcal{R}_n(x')| &= o_p(n^{-1/2} h_g^{-1/2}). \end{aligned}$$

In the sequel for sequences A_n, B_n, let $A_n \simeq B_n$ mean that $A_n/B_n \to 1$.

6.9.2 Appendix B

Proof of Theorem 1. We show that $\widetilde{f}(x)$ is equivalent to

$$\overline{f}(x) = g(x)\frac{1}{nh^d}\sum_{i=1}^{n} \frac{K\left(\frac{x-X_i}{h}\right)}{g(X_i)}, \tag{6.18}$$

in the sense that $\widetilde{f}(x) - \overline{f}(x) = o_p(n^{-1/2}h^{-d/2})$.

Write $1/\widehat{g}(X_i) - 1/g(X_i) = -\{\widehat{g}(X_i) - g(X_i)\}/\widehat{g}(X_i)g(X_i)$ for all i. Then,

$$\begin{aligned}\widetilde{f}(x) - \overline{f}(x) &= \{\widehat{g}(x) - g(x)\} \frac{1}{nh^d} \sum_{i=1}^{n} \frac{K\left(\frac{x-X_i}{h}\right)}{g(X_i)} \\ &\quad - g(x) \frac{1}{nh^d} \sum_{i=1}^{n} \frac{K\left(\frac{x-X_i}{h}\right)}{g(X_i)} \frac{\{\widehat{g}(X_i) - g(X_i)\}}{\widehat{g}(X_i)} \\ &\quad - \{\widehat{g}(x) - g(x)\} \frac{1}{nh^d} \sum_{i=1}^{n} \frac{K\left(\frac{x-X_i}{h}\right)}{g(X_i)} \frac{\{\widehat{g}(X_i) - g(X_i)\}}{\widehat{g}(X_i)} \\ &\equiv R_1 + R_2 + R_3. \end{aligned} \tag{6.19}$$

Firstly, note that

$$\frac{1}{nh^d} \sum_{i=1}^{n} \frac{K\left(\frac{x-X_i}{h}\right)}{g(X_i)} = O_p(1)$$

by the Markov inequality because by a change of variables and dominated convergence

$$\begin{aligned} E\left\{ \frac{1}{h^d} \frac{|K\left(\frac{x-X_i}{h}\right)|}{g(X_i)} \right\} &= \frac{1}{h^d} \int |K\left(\frac{x-X}{h}\right)| \frac{f(X)}{g(X)} dX \\ &= \int |K(s)| \frac{f(x-sh)}{g(x-sh)} ds \to \frac{f(x)}{g(x)} \int |K(s)| ds < \infty. \end{aligned}$$

Therefore, $R_1 = o_p(n^{-1/2}h^{-d/2})$. For large enough n, $\{x' : K\left(\frac{x-x'}{h}\right) \neq 0\} \subset \{x' : |x - x'| \leq \epsilon\}$, since K has compact support, and so

$$\left| \frac{1}{nh^d} \sum_{i=1}^{n} \frac{K\left(\frac{x-X_i}{h}\right)}{g(X_i)} \frac{\{\widehat{g}(X_i) - g(X_i)\}}{\widehat{g}(X_i)} \right| \leq \frac{1}{nh^d} \sum_{i=1}^{n} \frac{\left|K\left(\frac{x-X_i}{h}\right)\right|}{g(X_i)} \times \frac{\sup_{|x-x'|\leq\epsilon} |\widehat{g}(x') - g(x')|}{\inf_{|x-x'|\leq\epsilon} \widehat{g}(x')}.$$

By the triangle inequality: $\inf_{|x-x'|\leq\epsilon} \widehat{g}(x') \geq \inf_{|x-x'|\leq\epsilon} g(x') - \sup_{|x-x'|\leq\epsilon} |\widehat{g}(x') - g(x')| = \inf_{|x-x'|\leq\epsilon} g(x') - o_p(1)$. By continuity and positivity of g at x, $\inf_{|x-x'|\leq\epsilon} g(x') > 0$.

It follows from Assumption 4 that

$$|R_2| \leq o_p(n^{-1/2}h^{-d/2})\frac{1}{nh^d}\sum_{i=1}^{n}\frac{\left|K\left(\frac{x-X_i}{h}\right)\right|}{g(X_i)} = o_p(n^{-1/2}h^{-d/2}).$$

Finally, $R_3 = o_p(n^{-1/2}h^{-d/2})$ by the same arguments. Therefore, $\widetilde{f}(x) - \overline{f}(x) = o_p(n^{-1/2}h^{-d/2})$.

Furthermore,

$$\sqrt{nh^d}\left\{\overline{f}(x) - f(x) - h^2\beta(x)\right\} \Longrightarrow N(0, v(x)), \tag{6.20}$$

by the CLT for kernel smoothers. Specifically,

$$\begin{aligned} E\{\overline{f}(x)\} &= g(x)\frac{1}{h^d}\int K\left(\frac{x-X}{h}\right)\frac{f(X)}{g(X)}dX \\ &= g(x)\int K(s)\, r(x-sh)ds \\ &= f(x) + h^2\beta(x) + o(h^2) \end{aligned}$$

by a change of variable and Taylor expansion. Furthermore,

$$\begin{aligned} \mathrm{var}\{\overline{f}(x)\} &= g(x)^2\frac{1}{nh^{2d}}\mathrm{var}\left\{\frac{K\left(\frac{x-X}{h}\right)}{g(X)}\right\} \\ &= g(x)^2\frac{1}{nh^{2d}}\left\{\int K^2\left(\frac{x-X}{h}\right)\frac{f(X)}{g^2(X)}dX - \left\{\int K\left(\frac{x-X}{h}\right)\frac{f(X)}{g(X)}dX\right\}^2\right\} \\ &= g^2(x)\frac{1}{nh^{2d}}\left\{h^d\int K^2(s)\,\frac{f(x-sh)}{g^2(x-sh)}ds - \left\{h^d\int K(s)\, r(x-sh)ds\right\}^2\right\} \\ &= g^2(x)\frac{1}{nh^d}\int K^2(s)\,\frac{f(x-sh)}{g^2(x-sh)}ds + O(n^{-1}) \\ &= f(x)||K||_2^2\frac{1}{nh^d} + o(n^{-1}h^{-d}). \end{aligned}$$

The Lindeberg central limit theorem (6.20) follows because the kernel is of bounded support. □

6.9.3 Appendix C

Proof of Theorem 2. Let

$$\widetilde{f}_T(x) = J(x)\frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{u(x)-U_i}{h}\right), \tag{6.21}$$

where $U_i = T(X_i;\lambda_0)$ and $u(x) = T(x;\lambda_0)$. Then

$$\begin{aligned} E\{\widetilde{f}_T(x)\} &= J(x)\frac{1}{h^d}\int K\left(\frac{u(x)-U}{h}\right) f_U(U)dU \\ &= J(x)\int K(t)\, f_U\{u(x)-th\}dt \\ &\simeq J(x)f_U\{u(x)\} + \frac{h^2}{2}J(x)\nabla_2 f_U\{u(x)\}\mu_2(k) \\ &= f(x) + \frac{h^2}{2}J(x)\nabla_2 f_U\{u(x)\}\mu_2(k) \end{aligned}$$

by (6.4), where $\nabla_2 f_U(u) = \sum_{j=1}^{d}\partial^2 f_U(u)/\partial u_j^2$. Furthermore, by change of variable and dominated convergence arguments, and by (6.4)

$$\begin{aligned} \mathrm{var}\{\widetilde{f}_T(x)\} &= J^2(x)\frac{1}{nh^{2d}}\left[E\left\{K\left(\frac{u(x)-U_i}{h}\right)^2\right\} - E^2\left\{K\left(\frac{u(x)-U_i}{h}\right)\right\}\right] \\ &= J^2(x)\frac{1}{nh^d}\frac{1}{h^d}\int K\left(\frac{u(x)-U}{h}\right)^2 f_U(U)dU + O(n^{-1}) \\ &= J^2(x)\frac{1}{nh^d}\int K(t)^2\, f_U\{u(x)-th\}dt + O(n^{-1}) \\ &\simeq \frac{1}{nh^d}J^2(x)f_U\{u(x)\}||K||_2^2 = \frac{1}{nh^d}J(x)f(x)||K||_2^2. \end{aligned}$$

By Taylor series expansion, for $j = 1,\ldots,d$,

$$\widehat{U}_{ji} - U_{ji} = T_j(X_i;\widehat{\lambda}) - T_j(X_i;\lambda_0) = \frac{\partial T_j}{\partial\lambda}(X_i;\lambda_0)(\widehat{\lambda}-\lambda_0) + \frac{1}{2}(\widehat{\lambda}-\lambda_0)^\top\frac{\partial^2 T_j}{\partial\lambda\partial\lambda^\top}(X_i;\overline{\lambda}_j)(\widehat{\lambda}-\lambda_0),$$

where $\overline{\lambda}_j$ is an intermediate point. Therefore,

$$\begin{aligned}
&\frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{u(x)-\widehat{U}_i}{h}\right) - \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{u(x)-U_i}{h}\right) \\
&= \frac{-1}{nh^{d+1}}\sum_{j=1}^{d}\sum_{i=1}^{n}\frac{\partial K}{\partial u_j}\left(\frac{u(x)-U_i}{h}\right)(\widehat{U}_{ji}-U_{ji}) + \\
&+ \frac{1}{2nh^{d+2}}\sum_{j,j'=1}^{d}\sum_{i=1}^{n}\frac{\partial^2 K}{\partial u_j \partial u_{j'}}\left(\frac{u(x)-\overline{U}_i^j}{h}\right)(\widehat{U}_{ji}-U_{ji})(\widehat{U}_{j'i}-U_{j'i}) \\
&\equiv L_n + Q_n,
\end{aligned}$$

where $\overline{U}_i^j$ are intermediate points. Since $\widehat{\lambda}$ is root-n consistent, there exists a sequence $\delta_n \to 0$ such that $\Pr[\sqrt{n}||\widehat{\lambda}-\lambda_0|| \geq \delta_n] \to 0$. Then, for some $O_p(1)$ random variables:

$$|Q_n| \leq \frac{1}{h^2}\left\|\widehat{\lambda}-\lambda_0\right\|^2 \times O_p(1) = O_p(n^{-1}h^{-2}),$$

$$|L_n| \leq \left\|\widehat{\lambda}-\lambda_0\right\| \times O_p(1).$$

It follows that $\widehat{f}_T(x) - \widetilde{f}_T(x) = O_p(n^{-1/2}) + O_p(n^{-1}h^{-2}) = o_p(n^{-1/2}h^{-d/2})$. □

6.9.4 Appendix D

Proof of Theorem 3. Let $\overline{f}_C(x) = J(x)\overline{f}_U\{u(x)\}$, where

$$\overline{f}_U(u) = g_U(u)\frac{1}{nh^d}\sum_{i=1}^{n}\frac{K\left(\frac{u-U_i}{h}\right)}{g_U(U_i)}.$$

We have

$$\begin{aligned} E\{\overline{f}_C(x)\} &= J(x)E[\overline{f}_U\{u(x)\}] \\ &= J(x)g_U\{u(x)\}\frac{1}{h^d}\int K\left(\frac{u(x)-U}{h}\right)\frac{f_U(U)}{g_U(U)}dU \\ &= J(x)g_U\{u(x)\}\int K(t)r_U\{u(x)-th\}dt \\ &\simeq J(x)f_U\{u(x)\} + \frac{h^2}{2}J(x)g_U\{u(x)\}\nabla_2 r_U\{u(x)\}\int k(t)t^2dt. \end{aligned}$$

$$\text{var}\{\overline{f}_C(x)\} \simeq J(x)^2\frac{1}{nh^d}f_U\{u(x)\}||K||_2^2 = \frac{1}{nh^d}J(x)f(x)||K||_2^2.$$

One shows that $\widetilde{f}_C(x) - \overline{f}_C(x) = o_p(n^{-1/2}h^{-d/2})$ by similar arguments to Theorem 1. □

6.9.5 Appendix E

Proof of Theorem 4. Let

$$\overline{f}(x^1|x^2,\ldots,x^d) = \frac{g(x^1|x^2,\ldots,x^d)}{\widehat{f}(x^2,\ldots,x^d)} \frac{1}{nh^d} \sum_{i=1}^{n} \frac{K\left(\frac{x-X_i}{h}\right)}{g(X_i^1|X_i^2,\ldots,X_i^d)}. \tag{6.22}$$

We claim that $\widetilde{f}(x^1|x^2,\ldots,x^d)$ is well approximated by $\overline{f}(x^1|x^2,\ldots,x^d)$. This follows by the same arguments in Theorem 1. Then write

$$\overline{f}(x^1|x^2,\ldots,x^d) = T_{n1} + T_{n2} + R_n,$$

$$\begin{aligned}
T_{n1} &= \frac{g(x^1|x^2,\ldots,x^d)}{f(x^2,\ldots,x^d)} \frac{1}{nh^d} \sum_{i=1}^{n} \frac{K\left(\frac{x-X_i}{h}\right)}{g(X_i^1|X_i^2,\ldots,X_i^d)} \\
T_{n2} &= -\frac{g(x^1|x^2,\ldots,x^d)}{f(x^2,\ldots,x^d)} \frac{1}{nh^d} \sum_{i=1}^{n} \frac{K\left(\frac{x-X_i}{h}\right)}{g(X_i^1|X_i^2,\ldots,X_i^d)} \frac{\widehat{f}(x^2,\ldots,x^d) - f(x^2,\ldots,x^d)}{f(x^2,\ldots,x^d)} \\
R_n &= \frac{g(x^1|x^2,\ldots,x^d)}{2f(x^2,\ldots,x^d)} \frac{1}{nh^d} \sum_{i=1}^{n} \frac{K\left(\frac{x-X_i}{h}\right)}{g(X_i^1|X_i^2,\ldots,X_i^d)} \frac{\left\{\widehat{f}(x^2,\ldots,x^d) - f(x^2,\ldots,x^d)\right\}^2}{f(x^2,\ldots,x^d)\widehat{f}(x^2,\ldots,x^d)}
\end{aligned}$$

We have

$$\begin{aligned}
E[T_{n1}] &= \frac{g(x^1|x^2,\ldots,x^d)}{f(x^2,\ldots,x^d)} \frac{1}{h^d} \int k\left(\frac{x^1 - X^1}{h}\right) \frac{f(X^1|X^2,\ldots,X^d)}{g(X^1|X^2,\ldots,X^d)} dX_1 \prod_{j=2}^{d} k\left(\frac{x^j - X^j}{h}\right) \\
&\quad \times f(X^2,\ldots,X^d) dX^2 \ldots dX^d \\
&\simeq \frac{g(x^1|x^2,\ldots,x^d)}{f(x^2,\ldots,x^d)} \int \left\{ r(x^1|X^2,\ldots,X^d) + \frac{h^2}{2} \frac{\partial^2 r(x^1|X^2,\ldots,X^d)}{\partial (x^1)^2} \int k(t) t^2 dt \right\} \\
&\quad \times \frac{1}{h^{d-1}} \prod_{j=2}^{d} k\left(\frac{x^j - X^j}{h}\right) f(X^2,\ldots,X^d) dX^2 \ldots dX^d
\end{aligned}$$

$$
\begin{aligned}
&\simeq \frac{g(x^1|x^2,\ldots,x^d)}{f(x^2,\ldots,x^d)}\Big\{r(x^1|x^2)f(x^2,\ldots,x^d)\\
&+\frac{h^2}{2}\frac{\partial^2 r(x^1|x^2,\ldots,x^d)}{\partial(x^1)^2}f(x^2,\ldots,x^d)\mu_2(k)\Big\}\\
&+\frac{h^2}{2}\sum_{j=2}^{d}\frac{\partial^2\{r(x^1|x^2,\ldots,x^d)f(x^2,\ldots,x^d)\}}{\partial(x^j)^2}\mu_2(k)\\
&\simeq f(x^1|x^2,\ldots,x^d)+\frac{h^2}{2}\frac{\partial^2 r(x^1|x^2,\ldots,x^d)}{\partial(x^1)^2}g(x^1|x^2,\ldots,x^d)\mu_2(k)\\
&+\frac{h^2}{2}\frac{g(x^1|x^2,\ldots,x^d)}{f(x^2,\ldots,x^d)}\sum_{j=2}^{d}\frac{\partial^2\{r(x^1|x^2,\ldots,x^d)f(x^2,\ldots,x^d)\}}{\partial(x^j)^2}\mu_2(k).
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{var}[T_{n1}] &\simeq \left\{\frac{g(x^1|x^2,\ldots,x^d)}{f(x^2,\ldots,x^d)}\right\}^2\frac{1}{nh^d}\int K(t)^2dt\frac{f(x^1|x^2,\ldots,x^d)}{g(x^1|x^2,\ldots,x^d)^2}f(x^2,\ldots,x^d)\\
&=\frac{1}{nh^d}\int K(t)^2dt\times\frac{f(x^1|x^2,\ldots,x^d)}{f(x^2,\ldots,x^d)}.
\end{aligned}
$$

Furthermore

$$
T_{n2}=-\left\{f(x^1|x^2,\ldots,x^d)+O_p(h^2)+O_p(n^{-1/2}h^{-d/2})\right\}\frac{\widehat{f}(x^2,\ldots,x^d)-f(x^2,\ldots,x^d)}{f(x^2,\ldots,x^d)},
$$

where

$$
\widehat{f}(x^2,\ldots,x^d)-f(x^2,\ldots,x^d)=\frac{h^2}{2}\mu_2(k)\sum_{j=2}^{d}\frac{\partial^2 f(x^2,\ldots,x^d)}{\partial(x^j)^2}+O_p(n^{-1/2}h^{-(d-1)/2})+o_p(h^2).
$$

The result now follows. □

6.9.6 Appendix F

Proof of Corollary 1. When $g(x) = f(x)$, we have

$$E\overline{f}(x) = f(x)\frac{1}{h^d}\int \frac{K\left(\frac{x-X}{h}\right)}{f(X)} f(X)dX = f(x)\frac{1}{h^d}\int K\left(\frac{x-X}{h}\right)dX = f(x).$$

Furthermore, $\text{var}\{\overline{f}(x)\} = f(x)\{||K||_2^2/nh^d - O(1/n)\}$. By C1 it follows that $\overline{f}(x) = f(x) + o_p(n^{-1/2}h_g^{-1/2})$.

Continuing (6.19) we obtain

$$\begin{aligned}\widetilde{f}(x) - \overline{f}(x) &= \{\widehat{g}(x) - g(x)\}\frac{1}{nh^d}\sum_{i=1}^{n}\frac{K\left(\frac{x-X_i}{h}\right)}{g(X_i)} \\ &\quad -g(x)\frac{1}{nh^d}\sum_{i=1}^{n}\frac{K\left(\frac{x-X_i}{h}\right)}{g^2(X_i)}\{\widehat{g}(X_i) - g(X_i)\} + R_4, \qquad (6.23)\end{aligned}$$

where R_4 is of smaller order using C2. Substituting the bias terms of (6.9) into the expansion (6.23) we obtain

$$h_g^2\left\{\beta_g(x) - f(x)\frac{1}{nh^d}\sum_{i=1}^{n}\frac{K\left(\frac{x-X_i}{h}\right)}{f(X_i)^2}\beta_g(X_i)\right\} = o_p(h_g^2), \qquad (6.24)$$

because

$$\begin{aligned}E\left\{\frac{1}{nh^d}\sum_{i=1}^{n}\frac{K\left(\frac{x-X_i}{h}\right)}{f(X_i)^2}\beta_g(X_i)\right\} &= \int \frac{1}{h^d}\frac{K\left(\frac{x-X}{h}\right)}{f(X)}\beta_g(X)dX \\ &= \int K(u)\frac{\beta_g(x+uh)}{f(x+uh)}du \\ &= \frac{\beta_g(x)}{f(x)} + o(1).\end{aligned}$$

The variance of (6.24) is $o(h_g^4)$ using the same arguments. Regarding the stochastic

terms in (6.23). Note that the stochastic part of

$$\frac{1}{nh^d}\sum_{i=1}^{n}\frac{K\left(\frac{x-X_i}{h}\right)}{g^2(X_i)}\left\{\widehat{g}(X_i)-g(X_i)\right\}$$

is an average of the stochastic part of a one-dimensional kernel smoother over a lot of terms (a consequence of the large bandwidth assumption), and therefore is of smaller order. Therefore,

$$\widetilde{f}(x)-f(x)=\frac{1}{\sqrt{nh_g}}\omega^{1/2}(x)Z_n(x)+o_p(n^{-2/5})$$

and the result follows.

□

Chapter 7

Multivariate density estimation using dimension reducing information and tail flattening transformations for truncated or censored data

This chapter is an adapted version of Buch-Kromann and Nielsen (2009).

This paper introduces a multivariate density estimator for truncated and censored data with special emphasis on extreme values based on survival analysis. A local constant density estimator is considered. We extend this estimator by means of tail flattening transformation, dimension reducing prior knowledge and a combination of both. The asymptotic theory is derived for the proposed estimators. It shows that the extensions might improve the performance of the density estimator when the

transformation and the prior knowledge is not too far away from the true distribution. A simulation study shows that the density estimator based on tail flattening transformation and prior knowledge substantially outperforms the one without prior knowledge, and therefore confirms the asymptotic results. The proposed estimators are illustrated and compared in a data study of fire insurance claims.

7.1 Introduction

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be n independent identically distributed stochastic variables. We wish to estimate various functionals of the conditional distribution of Y_1 given X_1. In particular we are concerned about functionals emphasizing the importance of extreme high values of the dependent variable, and we want to profit by some complexity reducing structure or prior knowledge on a useful parametric model.

However, Y is subject to truncation and censoring – in the following filtering is an abbreviation for data that might have been truncated or censored. One prominent example where this statistical problem arises is in general insurance. The censoring applies when there is some upper limit on the insurance policy. This happens either as part of the actual contract or as a consequence of poor data collection where only the actual expense of the company is recorded disregarding amounts paid by the reinsurance company. Typically, an insurance company holds an excess of loss contract where the reinsurance company covers amounts above some threshold value exactly corresponding to the right censoring mechanism described above. Left truncation exactly corresponds to the widely used deductibles. A loss below the deductible value is covered by the individual policy holder without even noticing the insurance company.

Even in the simple one-dimensional case without any filtering our estimation problem is non-trivial and has given rise to an enormous amount of theory on the extreme value behaviour of distributions and its estimation; the so called extreme value theory (EVT), see Embrechts et al. (1997) for a prominent textbook on this. However,

most of this literature is based on the asymptotic behaviour of the right tail of the distribution, and in practise most EVT methods are based on personal judgements. Also, there are surprisingly few simulation studies spelling out the actual benefits of EVT methods. This led Bolancé et al. (2003) and Buch-Larsen et al. (2005) to view this one-dimensional problem as a standard estimation problem attempting to improve estimation considering the classical trade off between variance and bias present in all problems of statistical inference. The extreme tail was accounted for by transformation methods inspired by the pioneering paper of Wand et al. (1991). In the working paper version of Bolancé et al. (2003), a simulation study was carried out where it was shown that classical EVT models did not work very well for any of the distributions considered in the study. Moreover, see Buch-Kromann (2009) for a comparison of the transformation kernel density estimator and classical EVT.

In this paper, we follow Buch-Kromann et al. (2009) and generalize it to the filtered data case. Buch-Kromann et al. (2009) extended the approach of Buch-Larsen et al. (2005) to a multivariate setting where the loss distribution is allowed to depend on covariates. This led to various methods of multivariate density estimation and its adjustment guided by structured models.

When dealing with filtered data, extensive use of counting process theory, see ie. Martinussen and Scheike (2006), and the pioneering work of internal hazard estimators in Beran (1981), Dabrowska (1987), McKeague and Utikal (1990) and Van Keilegom and Veraverbeke (2001) and the alternative external hazard estimator introduced in Nielsen and Linton (1995) are neccesary. All these papers deal with locally constant estimators. They are extended to locally linear versions in Li and Doss (1995) and Nielsen (1998) with superior boundary bias of order $O(b^2)$ compared to the local constant boundary bias of order $O(b)$,where b is the bandwidth. The paper Van Keilegom and Akritas (1999) proposed a new estimator of the conditional cumulative density function based on a fully nonparametric heteroscedastic regression model, which improved the estimator significantly, when the censoring in the tail is "heavy". The conditional density and hazard functions under this model are studied in Van Keilegom and Veraverbeke (2002). Consistent nonparametric estima-

tors of the location function of the heteroscedastic regression model are studied in Heuchenne and Keilegom (2007a) and a parametric version is studied in Heuchenne and Keilegom (2007b). When dealing with multivariate estimation problems, the rate of convergence of the standard estimators is poor, see Stone (1980), and the interpretation might be difficult. One way to solve these problems is to make assumptions about the structure of the problem, eg. additive or multiplicative models, as studied in Hastie and Tibshirani (1990), Linton and Nielsen (1995) and Linton et al. (2003).

In this paper, we restrict ourself to the locally constant estimator for reasons of notation and presentation. The widely available methodology of regression is not appropriate for this type of problems where we need a full model specification and not just mean functions or quantiles. We extend the approach of the study in Buch-Kromann et al. (2009) to the more complicated setting where filtering is present, and we use counting process theory for this task. The authors in Nielsen et al. (2009) note that nonparametric smoothing of densities can be generalised in such a way that in a filtered data context it corresponds to local polynomial hazard estimation weighted with the classical Kaplan-Meier estimator. Without filtering, this locally constant estimator simply collapses to the standard kernel density estimator. It is also noticed in Nielsen et al. (2009) that they do not recommend this estimator in general for filtered data. The reason is what they call exposure robustness indicating that another weighting, the so called natural weighting combined with a smooth version of the Kaplan-Meier estimator, works just as well as standard kernel density estimation when there is no filtering or when filtering is happening in a smooth and nonsurprising way. However, when lack of robustness is present in the exposure pattern, the method with natural weighting and a smoothed Kaplan-Meier estimator significantly outperform the other method. Therefore, Nielsen et al. (2009) suggested always to use the latter approach since there was no pain, only gain, see also Nielsen and Tanggaard (2001) for a study about weighting functions in kernel hazard estimation. We generalise this latter approach to the multivariate setting. First we define a smoothed conditional Kaplan-Meier estimator as a simple functional of the

multivariate kernel hazard estimator of Nielsen and Linton (1995). Then we define our nonparametric conditional density estimator as a weighted version of this very same local constant multivariate kernel hazard estimator, where the weight is the smoothed conditional Kaplan-Meier estimator. Once a conditional density estimator is available, we can approximate this density to our complexity reducing structure. Finally, we apply this structured density to guide a bias correction leading to our final smooth nonparametric density estimator. In this way, we add some structure to our estimation problem caused by the curse of dimensionality as described in Linton and Nielsen (1995) and Linton et al. (2003). However, we allow a nonparametric correction of this structure in the final multiplicative correction step of our procedure.

The paper is organized as follows. In section 7.2, we define the general model and in section 7.3, we define the estimators of the conditional density. In section 7.4, the asymptotic properties of the estimators are presented, and section 7.5 contains an application and a Monte Carlo study which compares the performance of the conditional density estimators. Section 7.6 is the conclusion.

7.2 The model

We would like to analyse (X, Y), but Y is not always observed. What we do observe is $(X, \widetilde{Y}, D, T)$, where X is a one-dimensional covariate, $\widetilde{Y} = Y \wedge C$ is Y subject to right censoring, $D = I(Y \leq C)$ is an indicator of right censoring has occurred and T is the truncation time, which means that $\widetilde{Y}$ is only observed when $\widetilde{Y} \geq T$. Suppose that Y and C are conditionally independent given X. Let $N(s) = I(\widetilde{Y} \leq s, D = 1)$ be a counting process with stochastic intensity λ with respect to its natural filtration $\mathcal{F}_y = \sigma\{X, T, D, N(s), s < y\}$, see Jacobsen (1982), Andersen et al. (1993) and Martinussen and Scheike (2006) for solid introductions to the formulation of this type of models. Hence N has a compensator Λ that equals the integrated stochastic intensity and $M = N - \Lambda$ is a martingale. We assume that the stochastic intensity

function λ can be written as $\lambda(s) = \alpha_X(s)R(s)$, where $\alpha_X(s)$ is the conditional hazard of the distribution of Y given X and $R(s) = I(T < s < \widetilde{Y})$ is the "at-risk" indicator, indicating whether the counting process is able to jump at time s. Then $S_X(s) = \exp\left\{-\int_0^s \alpha_X(u)\,\mathrm{d}u\right\}$ is the conditional survival function and $f_X(s) = \alpha_X(s)S_X(s)$ is the conditional density.

Our final notational definition in this section concerns our actually observed stochastic variables. We assume that we observe independent and identically distributed variabels $(X_1, \widetilde{Y}_1, D_1, T_1), \ldots, (X_n, \widetilde{Y}_n, D_n, T_n)$. The resulting counting processes N_1,...,N_n have stochastic intensities $\lambda_1, \ldots, \lambda_n$ and compensators $\Lambda_1, \ldots, \Lambda_n$ with corresponding martingales $M_1, \ldots, M_n$. Our aim is to estimate the conditional density $f_x(s)$ given $X = x$, possibly guided by prior knowledge and structured models.

7.3 Estimating the conditional density

In this section we introduce estimators for the conditional density of filtered data. We first introduce a non-parametric filetered data density estimator which takes filering into consideration by means of counting proces theory. This estimator is the fundamental estimator on which all the following density estimators are build, even though its usefullness is limited especially for heavy-tailed data. Subsequently, we introduce two extensions of the non-parametric filtered data density estimator, namely tail flattening transformations and multiplicative correction guided by prior knowledge. Tail flattening transformations improve performance of non-parametric estimators considerable and multiplicative correction guided by prior knowledge allow us to "remove" the simple and rough trends in data and thereby improve the non-parametric estimation. At last we combine tail flattening transformations and multiplicative correction in our recommended density estimator for filtered data.

7.3.1 The non-parametric filtered data density estimator

In the simple case where we have a homogeneous Poisson process, it is well known that the maximum likelihood estimator of the hazard of the process equals observed occurrences divided by the total exposure time of the process. Now let us consider a local version of this: we take all observed occurrences localized around some covariate or time and divide by the total exposure in this neighbourhood. This gives us a local hazard estimator which depends on the covariate or time. One can even become slightly more sophisticated and weigh these occurrences or exposure times according to how far away they are from the covariate or time value that we want to know the intensity of. This latter case is exactly the local kernel hazard estimator of Nielsen and Linton (1995) that we will use in the following. Let K be some mean zero probability density with finite variance and finite support and let $K_b(u) = \frac{1}{b}K\left(\frac{u}{b}\right)$, where b is a bandwidth. Moreover, let $\widehat{\alpha}_x^{(b)}(t) = \frac{O_t}{E_t}$, where $O_t = \sum_{i=1}^n \int K_{b_1}(t-s)K_{b_2}(x-X_i)\,\mathrm{d}N_i(s)$, is the total localised and smoothed number of occurrences, and $b = (b_1, b_2)$ are bandwidths corresponding to the time and the covariate X, respectively. $E_t = \sum_{i=1}^n \int K_{b_1}(t-s)K_{b_2}(x-X_i)R_i(s)\,\mathrm{d}s$, is the total localised and smoothed exposure; $R_i(s) = I(T_i < s < \widetilde{Y}_i)$. This gives an obvious candidate for our smoothed conditional survival function $\widehat{S}_x^{(b)}(s) = \exp\left\{-\int_0^s \widehat{\alpha}_x^{(b)}(u)\,\mathrm{d}u\right\}$.

We have two obvious candidates for the conditional density. One follows from the fact that the density is just a function of the hazard and the suvival function, so that one can plug it in to the estimated conditional hazard and survival function. However, we prefer a more direct estimator that is the natural generalisation of the estimator of Nielsen et al. (2009). They show that if the counting process in their case is replaced by the integral of the estimated survival function with respect to the counting process, then local polynomial density estimators can be written as direct minimization of a natural least squared loss criteria. In our case, this corresponds to replacing $\mathrm{d}N_i(s)$ by $\widehat{S}_{X_i}(s)\mathrm{d}N_i(s)$ in the kernel hazard estimator above. However, instead of $\widehat{S}_{X_i}(s)\mathrm{d}N_i(s)$, we replace $\mathrm{d}N_i(s)$ with $\widehat{S}_{X_i,(i)}(s)\mathrm{d}N_i(s)$, where $\widehat{S}_{X_i,(i)}(s)$ is

a leave-one-out estimator:

$$\widehat{S}^{(b)}_{X_i,(i)}(s) = \exp\left\{-\int_0^s \widehat{\alpha}^{(b)}_{X_i,(i)}(u)\,\mathrm{d}u\right\} \tag{7.1}$$

where $\widehat{\alpha}^{(b)}_{X_i,(i)}(t) = \frac{\sum_{j\neq i}\int K_{b_1}(t-s)K_{b_2}(x-X_j)\,\mathrm{d}N_j(s)}{\sum_{j\neq i}\int K_{b_1}(t-s)K_{b_2}(x-X_j)R_j(s)\,\mathrm{d}s}$. This is a well-known trick from Mammen and Nielsen (2007) to simplify the predicability issues in the proofs of the asymptotic results, see appendix 7.7.1, and which moreover often improves the performance of the estimators.

Under the assumption that something is observed after filtration, we arrive at the *non-parametric fitered data density estimator:*

$$\widehat{f}^{(d,b)}_x(t) = \frac{\sum_{i=1}^n \int K_{d_1}(t-s)K_{d_2}(x-X_i)\widehat{S}^{(b)}_{X_i,(i)}(s)\,\mathrm{d}N_i(s)}{\sum_{i=1}^n \int K_{d_1}(t-s)K_{d_2}(x-X_i)R_i(s)\mathrm{d}s}. \tag{7.2}$$

The bandwidths $b = (b_1, b_2)$ and $d = (d_1, d_2)$ in (7.2) allow us to undersmooth the conditional survival function that we use as an auxiliary variable while estimating the conditional density. The consequence of this undersmoothing is that the conditional survival function can be seen as known from the point of view of asymptotic theory. Otherwise, bias from $\widehat{S}^{(b)}_{X_i,(i)}(s)$ would disturb the results.

7.3.2 The transformed filtered data density estimator

When dealing with heavy tail distributions, tail flattening transformations, as introduced in Wand et al. (1991), have shown to improve the estimation results significantly, see Bolancé et al. (2003) and Buch-Larsen et al. (2005) for simulation studies in one dimension and Buch-Kromann et al. (2009) in the multivariate case. Moreover, tail flattening transformations have shown robustifying properties when combined with alternative prior assumptions of parametric distributions, see Buch-Kromann et al. (2007).

Let $\Psi : [0, \infty) \to [0, 1)$ be a candidate of a tail flattening transformation, where Ψ is a

cdf. Let ψ be the density corresponding to Ψ that we assume to be differentiable, and let $\Psi^{-1}(t)$ be the inverse of the cdf $\Psi(t)$. Ψ could be the Champernowne cdf, see Buch-Larsen et al. (2005), as this is a flexible and widely useable transformation function, e.g. in operational risk, see Bolancé et al. (2008a); Guillen et al. (2007); Gustafsson (2006); Gustafsson and Nielsen (2008); Gustafsson et al. (2006a,b). However other transformation functions could be i.e. transformations to normality, see Koekemoer and Swanepoel (2008a,b), the Mobius-like transformation, see Clements et al. (2003) or the Johnson families, see Yang and Marron (1999).

We transform our data with Ψ and obtain the transformed counting process $\widetilde{N}_i = N_i \circ \Psi$. Now we calculate the non-parametric filtered data density estimator (7.2) on the transformed data set and obtain what we will call the *the transformed filtered data density estimator on the Ψ-transformed axis*:

$$\widehat{k}^{(d,b)}_{\Psi,x}(v) = \frac{\sum_{i=1}^{n} \int_0^1 K_{d_1}(v-s) K_{d_2}(x-X_i) \widehat{S}^{(b)}_{\Psi,X_i,(i)}(s)\, \mathrm{d}\widetilde{N}_i(s)}{\sum_{i=1}^{n} \int_0^1 K_{d_1}(v-s) K_{d_2}(x-X_i) R_i\{\Psi^{-1}(s)\}\, \mathrm{d}s} \tag{7.3}$$

where $\widehat{S}^{(b)}_{\Psi,X_i,(i)}(s) = \exp\left\{-\int_0^s \widehat{\alpha}^{(b)}_{\Psi,X_i,(i)}(u)\, \mathrm{d}u\right\}$ is the leave-one-out estimator of the survival function on the Ψ-transformed axis, $R_i\{\Psi^{-1}(s)\}$ is the "at-risk" indicator on the transformed axis, and the leave-one-out hazard estimator on the transformed axis is given by:

$$\widehat{\alpha}^{(b)}_{\Psi,X_i,(i)}(t) = \frac{\sum_{j\neq i} \int_0^1 K_{b_1}(t-s) K_{b_2}(x-X_j)\, \mathrm{d}\widetilde{N}_j(s)}{\sum_{j\neq i} \int_0^1 K_{b_1}(u-s) K_{b_2}(x-X_j) R_j\{\Psi^{-1}(s)\}\, \mathrm{d}s}.$$

We backtransform (7.3) to obtain an estimator of $f_x(s)$, called the *transformed filtered data density estimator on the original axis*

$$\widehat{f}^{(d,b)}_{\Psi,x}(t) = \psi(t) \cdot \widehat{k}^{(d,b)}_{\Psi,x}\{\Psi(t)\}. \tag{7.4}$$

In addition to being a density estimator on the Ψ-transformed axis, $\widehat{k}^{(d,b)}_{\Psi,x}$ can be

interpreted as a correction estimator of ψ, the density corresponding to the transformation function, Ψ.

7.3.3 The filtered data density estimator guided by prior knowledge

Assume we have a prior knowledge indicating that $h_x(s)$ is close to $f_x(s)$. By introducing a multiplicative bias correction (7.5) based on the prior knowledge h_x, where h_x could be some appropriate parametric model, we reduce the complexity of the estimation problem, see Nielsen et al. (2009); Mammen and Nielsen (2007); Nielsen and Tanggaard (2001). Even though Nielsen and Tanggaard (2001) show that additive bias correction often is better than multiplicative correction, we choose multiplicative correction to ensure that the resulting estimator is positive. The multiplicative bias correction based on h_x is

$$\widehat{c}_x^{(d,b)}(t) = \frac{\sum_{i=1}^n \int K_{d_1}(t-s)K_{d_2}(x-X_i)\widehat{S}^{(b)}_{X_i,(i)}(s)\left\{h_{X_i}(s)\right\}^{-1} \mathrm{d}N_i(s)}{\sum_{i=1}^n \int K_{d_1}(t-s)K_{d_2}(x-X_i)R_i(s)\,\mathrm{d}s} \tag{7.5}$$

and the final multiplicatively bias corrected estimator of $f_x(s)$, called the *filtered data density estmator guided by prior knowledge*, is

$$\widehat{g}_x^{(d,b)}(t) = h_x(t)\widehat{c}_x^{(d,b)}(t). \tag{7.6}$$

Notice, that even though h_x is a parametric model then $\widehat{g}_x^{(d,b)}(t)$ is a fully nonparametric estimator of $f_x(s)$.

7.3.4 The transformed filtered data density estimator guided by prior knowledge

Until now we have introduced a transformation approach that improves the performance especially for heavy tailed distributions, and we have also discussed how to incorporate prior knowledge by multiplicative correction. Now we combine the tail flattening transformation approach (7.4) with the multiplicative bias correction from (7.6) to obtain a multiplicative corrected transformation estimator. On the Ψ-transformed axis the multiplicative bias correction based on h_x is

$$\widetilde{c}^{(d,b)}_{\Psi,x}(v) = \frac{\sum_{i=1}^{n} \int_0^1 K_{d_1}(v-s) K_{d_2}(x-X_i) \widehat{S}^{(b)}_{\Psi,X_i,(i)}(s) [h_{X_i}\{\Psi^{-1}(s)\}]^{-1} \, \mathrm{d}\widetilde{N}_i(s)}{\sum_{i=1}^{n} \int_0^1 K_{d_1}(v-s) K_{d_2}(x-X_i) R_i\{\Psi^{-1}(s)\} \, \mathrm{d}s} \tag{7.7}$$

and the density estimator on the Ψ-transformed axis is therefore

$$\widetilde{k}^{(d,b)}_{\Psi,x}(v) = h_x\{\Psi^{-1}(v)\} \widetilde{c}^{(d,b)}_{\Psi,x}(v) \tag{7.8}$$

in the following called the *transformed filtered data density estimator guided by prior knowledge on transformed axis.*

After back transformation we obtain an estimator of $f_x(s)$ on the original axis

$$\widetilde{f}^{(d,b)}_{\Psi,x}(t) = \psi(t) \widetilde{k}^{(d,b)}_{\Psi,x}\{\Psi(t)\} \tag{7.9}$$

called the *transformed filtered data density estimator guided by prior knowledge on original axis.*

7.4 Asymptotic properties

The intuition behind the proof of the asymptotic theory is similar to what is known from the theory of multivariate hazard estimation. In the proof of asymptotic the-

ory of the multivariate hazard estimator in Nielsen and Linton (1995) a counting process is spilt into a martingale and its compensator: $N(s) = M(s) + \Lambda(s)$, giving $\mathrm{d}N(s) = \mathrm{d}M(s) + \mathrm{d}\Lambda(s) = \mathrm{d}M(s) + \lambda(s)\mathrm{d}s = \mathrm{d}M(s) + \alpha(s)R(s)\mathrm{d}s$. In our proof, we replace $\mathrm{d}N(s)$ with $S_X(s)\mathrm{d}N(s) = f_X(s)R(s)\mathrm{d}s + S_X(s)\mathrm{d}M(s)$ and show that this is equivalent to replace $\mathrm{d}N(s)$ with $\widehat{S}_X(s)\mathrm{d}N(s)$. This implies that the results obtained about hazard estimation from smoothing $\mathrm{d}N(s)$ can be transferred to density estimation by smoothing $\widehat{S}_X(s)\mathrm{d}N(s)$ as in Nielsen et al. (2009).

To simplify the notation, we assume in the following that the scale of time t and covariate X is the same, and therefore we let $b = b_1 = b_2$ and $d = d_1 = d_2$.

Let $Z(x, s) = Pr(X \leq x \mid R(s) = 1)$ be the differentiable conditional distribution of the covariate X given that the counting proces can jump at time s and let $z(x, s) = \partial Z(x, s)/\partial s$ be the corresponding density of Z with respect to the two-dimensional Lesbesque-measure. Also let $\phi_x(s) = z(x, s)r(s)$, where $r(s) = \mathbb{E}\{R(s)\}$ as defined in section 7.2. Let f be the functional mapping (x, t) into $f_x(t)$ and let ϕ be the functional mapping (x, t) into $\phi_x(t)$. Both are mappings from $\mathbb{R} \times \mathbb{R}_+$ into $\mathbb{R}_+$.

Assumption A .

1. *Suppose that f is twice continuously differentiable and strictly positive at the interior point* (x, t) of $\mathbb{R} \times \mathbb{R}_+$.

2. *Suppose that the two dimensional functional ϕ is twice continuously differentiable and strictly positive at the interior point* (x, t) of $\mathbb{R} \times \mathbb{R}_+$.

3. *Suppose that $nd^2 \to \infty, d \to 0, b/d \to 0$ and $d^2/b \to 0$.*

4. *Suppose that for a constant $\delta > 0$, it holds that*

$$\sum_{0 \leq s \leq t+\delta} \left| \frac{R(s)}{n} - \zeta(s) \right| \to o_P(1)$$

where $\zeta : [0, t + \delta] \to \mathbb{R}_+$ is a continuous strictly positive function.

Now we are able to write up the asymptotic theory of our non-parametric filtered data density estimator (7.2). From a theoretical point of view, the theory of this estimator is close to the theory of the non-parametric locally constant kernel hazard estimators considered in Nielsen and Linton (1995) and Nielsen (1998).

Theorem 7.1. (Non-parametric filtered data density estimator)
Suppose that assumption A is satisfied. Define the kernel moments $||K||_2^2 = \int K^2(u)\,\mathrm{d}u$ *and* $\mu_2(K) = \int K(u)u^2\,\mathrm{d}u$*, where the kernel function* K *is a density function with finite support, mean zero and finite variance. Then the following hold:*

$$\sqrt{n}d\left\{\widehat{f}_x^{(d,b)}(t) - f_x(t) - d^2\beta_1(x,t)\right\} \Longrightarrow N\left\{0, \gamma(x,t)\right\},$$

where

$$\begin{aligned} \beta_1(x,t) &= \mu_2(K)\{\mathcal{B}_1(f,\phi)(x,t) + \mathcal{B}_2(f,\phi)(x,t)\} \\ \gamma(x,t) &= \{||K||_2^2\}^2 \frac{f_x(t)S_x(t)}{\phi_x(t)}. \end{aligned} \tag{7.10}$$

The two functionals in (7.10), $\mathcal{B}_1$ *and* $\mathcal{B}_2$*, both mappings from* $\mathbb{R}_+ \times \mathbb{R}_+$ *into* $\mathbb{R}_+$*, are defined by*

$$\mathcal{B}_1(f,\phi)(x,t) = \frac{(\partial f_x(t)/\partial t)(\partial \phi_x(t)/\partial t)}{\phi_x(t)} + \frac{\partial^2 f_x(t)/\partial t^2}{2}, \tag{7.11}$$

$$\mathcal{B}_2(f,\phi)(x,t) = \frac{(\partial f_x(t)/\partial x)(\partial \phi_x(t)/\partial x)}{\phi_x(t)} + \frac{\partial^2 f_x(t)/\partial x^2}{2}. \tag{7.12}$$

Proof. See appendix 7.7.1. □

Now we are ready to state the asymptotic theory of the above density estimator when prior knowledge, represented by $h_x(t)$, is used to bias correct the original estimator, i.e. the filtered data density estimator guided by prior knowledge (7.6). The resulting asymptotic theory is very similar to the asymptotic theory without bias correction. However, the bias expression is changed such that it is the curvature of the true

density divided by the prior knowledge that enters our bias expression. Therefore, this approach improves performance when our prior knowledge is sufficiently precise to capture essential properties of the curvature of the problem. If the prior knowledge does not have this quality, it will not be helpful in our estimation process.

Theorem 7.2. (Filtered data density estimator guided by prior knowledge) *Suppose that assumption A is satisfied. Moreover, suppose that the functional* $h : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}^+$ *mapping* (x,t) *into* $h_x(t)$ *is two times continuously differentiable, and that* $c : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}^+$ *maps* (x,t) *into* $c_x(t) = f_x(t) \{h_x(t)\}^{-1}$*, then*

$$\sqrt{n}d \left[\widehat{g}_x^{(d,b)}(t) - f_x(t) - d^2\beta_2(x,t)\right] \Longrightarrow N\{0, \gamma(x,t)\}$$

where

$$\beta_2(x,t) = h_x(t)\mu_2(K) \{\mathcal{B}_1(c,\phi)(x,t) + \mathcal{B}_2(c,\phi)(x,t)\} \tag{7.13}$$

and $\gamma(x,t), \mu_2(K), \mathcal{B}_1$ *and* $\mathcal{B}_2$ *are defined in Theorem 7.1.*

Proof. See appendix 7.7.2. □

Now we state the asymptotic theory of the density estimator when a transformation approach is used in our estimation process, i.e. (7.4). The asymptotic theory is similar to the asymptotic theory with multiplicative bias correction guided by prior knowledge. The bias expression is changed such that it is the curvature of the transformed density that enters our bias expression. Therefore, the transformation approach improves performance when the transformation captures essential properties of the curvature of the problem. In the transformation approach, the variance is also affected since it is multiplied by the density of the transformation. This is because the transformation approach acts similarly to a nearest neighbour type of approach compressing the data through the transformation. The variance is affected in a similar fashion as with nearest neighborhood methods accounting for the changed amount of information present in a bandwidth distance. Let $f\psi^{-1} \circ \Psi^{-1}$ be the functional mapping (x,t) into $f_x\{\Psi^{-1}(t)\}[\psi\{\Psi^{-1}(t)\}]^{-1}$. The map, $f\psi^{-1} \circ \Psi^{-1}$

is the conditional density of the latent variable Y after the transformation has taken place. Since we carry out the nonparametric density estimation on this transformed axis, it is not surprising that the main term in the bias of this approach is the bias of the density estimator on this axis.

Theorem 7.3. (Transformed filtered data density estimator)
Suppose that assumption A is satisfied and suppose that the functional Ψ is two times continuously differentiable, then

$$\sqrt{n}d\left[\widehat{f}^{(d,b)}_{\Psi,x}(t) - f_x(t) - d^2\beta_3(x,t)\right] \Longrightarrow N\left\{0, \psi(t)\gamma(x,t)\right\}$$

where

$$\begin{aligned}\beta_3(x,t) &= \psi(t)\mu_2(K) \\ &\left[\mathcal{B}_1(f\psi^{-1}\circ\Psi^{-1},\phi)\left\{x,\Psi(t)\right\} + \mathcal{B}_2(f\psi^{-1}\circ\Psi^{-1},\phi)\left\{x,\Psi(t)\right\}\right]\end{aligned}$$

and $\gamma(x,t), \mu_2(K), \mathcal{B}_1$ and $\mathcal{B}_2$ defined in Theorem 7.1.

Proof. See appendix 7.7.3. □

Let $c\psi^{-1}\circ\Psi^{-1}$ be the same functional as $f\psi^{-1}\circ\Psi^{-1}$, but with c replacing f. Then we can state the asymptotic theory of the transformed filtered data density estimator guided by prior knowledge (7.9). From this approach we both get the advantage of the nearest neighbour type of quality of the transformation and the bias reducing advantage of our prior knowledge. The practical advantages of this approach are seen in the numerical results in the next section.

Theorem 7.4. (Transformed filtered data density estimator guided by prior knowledge)
Suppose that assumption A is satisfied and suppose that the functionals h and Ψ are

two times continuously differentiable, then

$$\sqrt{n}d\left[\tilde{f}^{(d,b)}_{\Psi,x}(t) - f_x(t) - d^2\beta_4(x,t)\right] \Longrightarrow N\left\{0, \psi(t)\gamma(x,t)\right\}$$

where

$$\begin{aligned}\beta_4(x,t) &= \psi(t)h_x(t)\mu_2(K) \\ &\quad \left[\mathcal{B}_1(c\psi^{-1}\circ\Psi^{-1},\phi)\left\{x,\Psi(t)\right\} + \mathcal{B}_2(c\psi^{-1}\circ\Psi^{-1},\phi)\left\{x,\Psi(t)\right\}\right]\end{aligned}$$

and $\gamma(x,t), \mu_2(K), \mathcal{B}_1$ *and* $\mathcal{B}_2$ *defined in Theorem 7.1.*

Proof. See appendix 7.7.4. □

7.5 Numerical results

In this section, we analyse a data set that originates from the Danish general insurance company, Codan Insurance, and contains commercial fire claims reported from 1995 to 2004. The data set consists of 2810 claims Y, and for each claim the corresponding estimated maximum loss (EML) X, is reported. The data set is heavy-tailed with claim sizes ranging from 19 to almost 6 million DKK. with average claim size at 56,220 DKK.

This section contains an application study and a Monte Carlo simulation study. In the application study, we compute the transformed filtered data density estimator both without and with prior knowledge and illustrate the estimators ability of taking filtering into account. In the Monte Carlo study we compare the performance of the same two estimators and benchmark against the prior knowledge estimator both when the prior knowledge is true and when the prior knowledge is roughly and not exactly true. Moreover, we compare the results with the performance of the standard two-dimensional transformation kernel density estimator studied in Buch-Kromann et al. (2009).

The transformation approach both improves the estimation performance and the visualization properties. When dealing with heavy-tailed data as commercial fire claims, a classical kernel density estimator without transformation and with constant bandwidths as defined in (7.2) has a very bad performance and therefore it is omitted in this study. We transform the claims as well as the EMLs with the three parameter Champernowne cdf

$$T(x) = \frac{(x+c)^\alpha - c^\alpha}{(x+c)^\alpha + (M+c)^\alpha - 2c^\alpha} \tag{7.14}$$

with parameters (α, M, c) estimated by a maximum likelihood procedure taking filtering into account, see appendix 7.7.5.

First, define the transformed filtered data density estimator (7.3), where the transformation function Ψ is the Champernowne cdf (7.14) with maximum likelihood parameters as described above. The choice of the Champernowne cdf as transformation function is due to ist ability to capture different distribution shapes and its special utility for heavy-tailed data, see Buch-Larsen et al. (2005) and Buch-Kromann et al. (2009) for further details about the Champernowne cdf. Notice that the explanatory variable, X in (7.3) is the Champernowne-transformed EMLs, which lie between 0 and 1. The choice of Champernowne transformation of both claims and EMLs ensures that the two variables are of the same scale, and therefore the simplification $d = d_1 = d_2$ and $b = b_1 = b_2$ is reasonable. The estimator is called $\widehat{k}_1$ and is defined from (7.3)

$$\bar{k}_1(v) = \frac{\sum_{i=1}^n \int_0^1 K_d(v-s) K_d(x - X_i) \widehat{S}^{(b)}_{T,X_i,(i)}(s) \, \mathrm{d}\widetilde{N}_i(s)}{\sum_{i=1}^n \int_0^1 K_d(v-s) K_d(x - X_i) R_i\{T^{-1}(s)\} \, \mathrm{d}s} \tag{7.15}$$

where $\widetilde{N}_i = N_i \circ T$. The bandwidth d is a simple Silverman-rule-of-thumb, see Silverman (1986), and $b = d/2$ to ensure the undersmoothing of the conditional survival function as mentioned in section 7.3 at page 158. As mentioned, T is the Champernowne cdf, defined in (7.14).

Thereafter, we define the prior knowledge. For that purpose, we set up a median

regression model corresponding to the model described in Linton et al. (2007) given by

$$Y = m(X)\varepsilon.$$

where ε and X is independent, and where the estimator of m is based on the density estimator (7.15), but with doubled bandwidths to ensure a smooth shape. The choice of this model is motivated by its ability to capture the shape of the distribution in a crude and smooth way. The density estimator of ε is a one-dimensional version of the transformation filtering data density estimator (7.4), which takes the corresponding filtering on ε into account. The filtering on ε follows directly from the filtering on Y, ie. if (Y, X, T, C) is a claim Y, with corresponding EML X, truncation T and censoring C, then $(\widetilde{T}, \widetilde{C})$, where $\widetilde{T} = T/m(X)$ and $\widetilde{C} = C/m(X)$ is the corresponding filtering on ε under the median regression model. However, the estimation procedure in this paper is slightly more complicated, due to the possible filtering on ε that needs to be taken into consideration. Let $\widehat{h}_x(y)$ be the resulting prior knowledge density on original axis estimated as if it was known, see Buch-Kromann et al. (2009), and then let $\bar{k}_2$ be prior knowledge density on the Champernowne-transformed axis, defined as

$$\bar{k}_2(v) = \frac{\widehat{h}_x(T^{-1}(v))}{T'(T^{-1}(v))} \tag{7.16}$$

At last, we define the transformed filtered data density estimator on transformed axes guided by the prior knowledge $\widehat{h}_x$ defined above. The resulting estimator corresponds to (7.8) based on the Champernowne transformation.

$$\bar{k}_3(v) = \frac{\widehat{h}_x\{T^{-1}(s)\} \sum_{i=1}^n \int_0^1 K_d(v-s) K_d(x - X_i) \widehat{S}^{(b)}_{T,X_i,(i)}(s) [\widehat{h}_{X_i}\{T^{-1}(s)\}]^{-1} \, \mathrm{d}\widetilde{N}_i(s)}{\sum_{i=1}^n \int_0^1 K_d(v-s) K_d(x - X_i) R_i\{T^{-1}(s)\} \, \mathrm{d}s} \tag{7.17}$$

where d is equal to a double Silverman-rule-of-thumb bandwidth and $b = d/2$.

To illustrate the estimator's ability to handle filtering data, we set up a filtering scheme. We simulate truncation for 25% randomly chosen claims and choose the

truncation levels for these claims uniformly on 0 to 10,000 DKK., which corresponds to the 0% and 58% empirical quantiles, respectively. Analogously, we simulate censoring for 25% randomly chosen claims and choose the censoring levels uniformly on 100,000 to 6,000,000 DKK., corresponding to the 89% and 100% empirical quantiles. We will refer to this filtering scheme as the *25% filtering scheme*. Analogously, we compute a *50% filtering scheme*, where filtering is simulated on 50% of the claims.

Figure 7.1 illustrates how the exposure of the fire claims data set is affected by the two filtering schemes compared to the no filtering scheme. We plot the smoothed exposures for the two filtering schemes relative to the exposure without filtering. The smoothed exposures correspond to the denominator of (7.15). In Figure 7.1 the truncation can be recognized clearly in both the 25% and the 50% filtering scheme, whereas the censoring is much less clear on the relative exposure plots for both filtering schemes. This is due to the chosen values of truncation and censoring levels, which are based on realistic filtering levels for the underlying commercial fire insurance data set. In the 25% filtering scheme, 283 claims are influenced by left-truncation and only 6 claims are influenced by right-censoring, whereas in the 50% filtering scheme the corresponding claims numbers are 561 claims and 5 claims, respectively.

7.5.1 Application

In the application study, we compute the transformed filtered data density estimator both without and with prior knowledge, i.e. (7.15) and (7.17), and plot them on the transformed axes together with the prior knowledge density (7.16) in the three data filtering schemes.

The transformed filtered data density estimator (7.15) of the fire claims data set is illustrated in Figure 7.2 in the three filtering schemes. The three plots are very similar. This means, that the dependence structure between X and Y is almost identical even though we have made a systematic reduction in the exposure in the

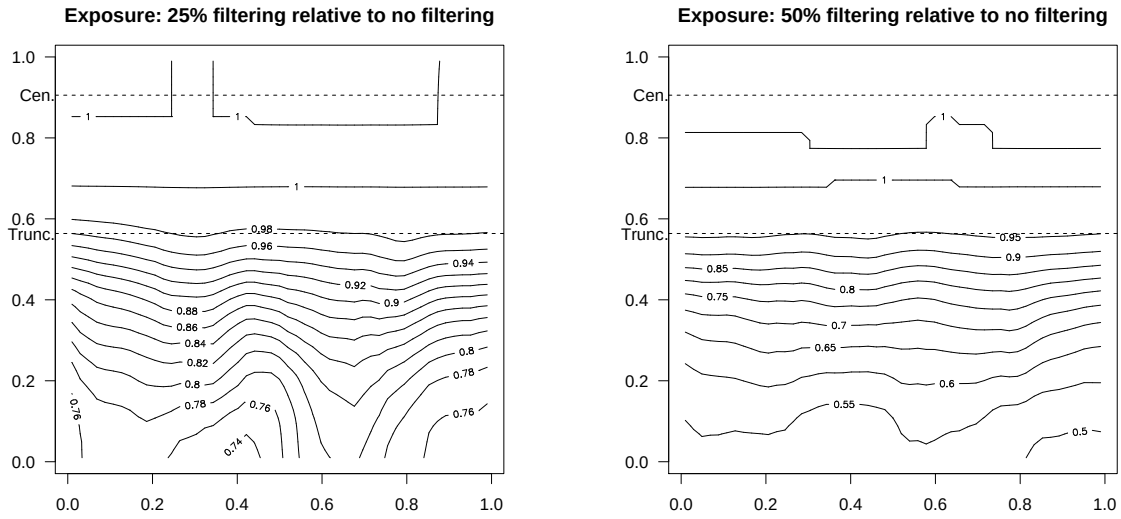

Figure 7.1: Smoothed exposure of no filtering scheme relative to smoothed exposure of respectively 25% (left) and 50% (right) filtering scheme.

25% and the 50% filtering schemes, as illustrated in Figure 7.1. Also the marginal distributions of X and Y are similar due to the maximum likelihood procedure's ability to take filtering into account: The estimated parameters of the Champernowne transformation function (7.14) $\theta_{j,\phi} = (\alpha_{j,\phi}, M_{j,\phi}, c_{j,\phi})$, where $j = \{X, Y\}$ indicates whether the parameters correspond to either X or Y, and $\phi = \{0, 25, 50\}$ indicates the chosen filtering scheme, are

$$\begin{aligned}
\theta_{x,0} &= (1.66, 2.56 \cdot 10^7, 6.06 \cdot 10^{-5}), \qquad \theta_{y,0} = (0.82, 7.54 \cdot 10^3, 3.44 \cdot 10^3) \\
\theta_{x,25} &= (1.67, 2.60 \cdot 10^7, 6.22 \cdot 10^{-5}), \qquad \theta_{y,25} = (0.82, 7.47 \cdot 10^3, 2.84 \cdot 10^3) \\
\theta_{x,50} &= (1.65, 2.78 \cdot 10^7, 6.80 \cdot 10^{-5}), \qquad \theta_{y,50} = (0.82, 7.67 \cdot 10^3, 2.17 \cdot 10^3)
\end{aligned}$$

The fact that both the dependence structure and the marginal distributions seem to be similar, indicates the transformed filtered data density estimator's ability to take filtering into consideration.

The prior knowledge density on transformed axes (7.16) in the no filtering, the 25% and the 50% filtering schemes are illustrated in Figure 7.3. As in Figure 7.2 we

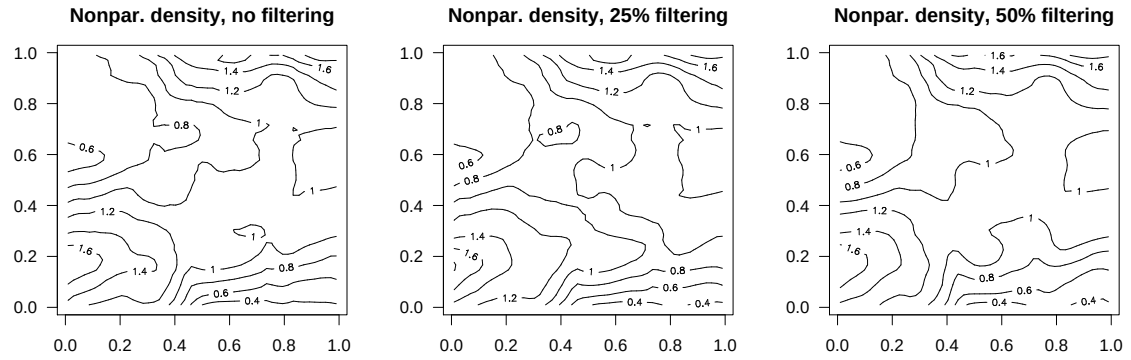

Figure 7.2: The transformed filtered data density estimator (7.15) computed on the fire claim data set without filtering (left), with the 25% (middle) and the 50% (right) filtering scheme. We recognize that the density estimates are almost identical which illustrates the density estimators ability to take filtering into account.

recognize that the shapes in the three plots illustrating the dependence structures are almost identical due to the method's ability to take filtering into account. We mention that the prior knowledge estimator puts perhaps to much structure into the density estimator. However, if it is not too wrong, then the multiplicative bias correction will correct is and benefit from it in the final estimator.

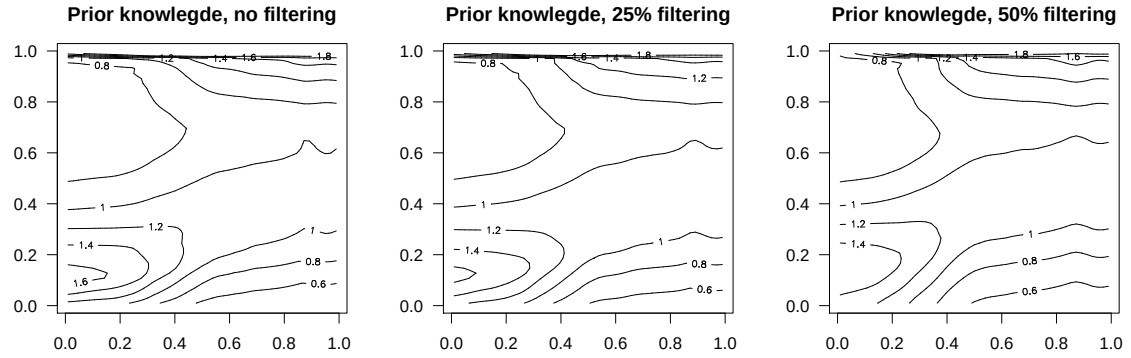

Figure 7.3: The prior knowledge density (7.16) computed on the fire claim data set without filtering (left), with the 25% (middle) and the 50% (right) filtering scheme. We recognize that the density estimates are almost identical which illustrates the density estimators ability to take filtering into account.

At last we illustrate the transformed filtered data density estimator guided by prior

knowledge (7.17) in the three filtering schemes on Champernowne transformed axes in Figure 7.4. Compared to Figure 7.3, some structure from the median regression density estimator (prior knowledge) is inherited. This is because we have a good prior knowledge. However, the multiplicative bias corrected density estimator has the opportunity to correct the density estimator in regions where the prior knowledge seems to be wrong. We also recognize the similarities between the dependence structures of the density estimators in the three filtering schemes in Figure 7.4.

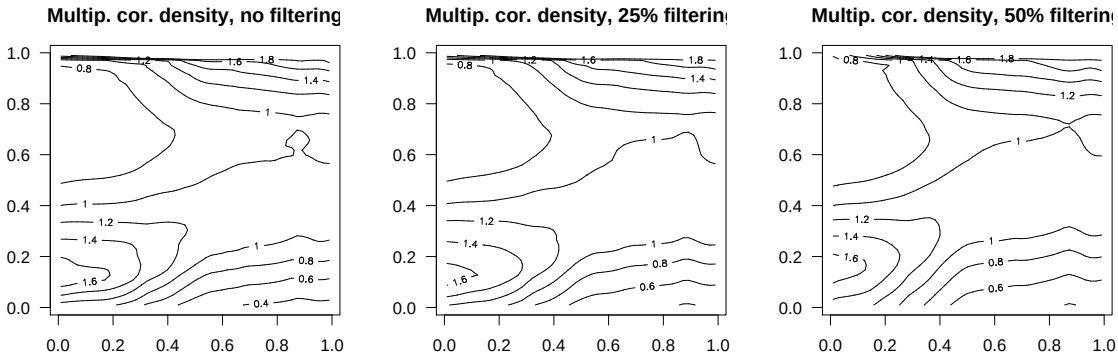

Figure 7.4: The transformed filtered data density estimator guided by prior knowledge (7.17) computed on the fire claim data set without filtering (left), with the 25% (middle) and the 50% (right) filtering scheme. We recognize that the density estimates are almost identical which illustrates the density estimators ability to take filtering into account.

7.5.2 Monte Carlo study

In the Monte Carlo study, we want to compare the performance of the three estimators defined in section 7.5 and illustrated in the application study. The simulation is based on the commercial fire insurance data set we have descibed above. We compute a multiplicative model with iid lognormal residuals independent of X

$$Y = \alpha X^{\beta} \varepsilon_1$$

to the data set and obtain the following estimates

$$\alpha = 182.37 \qquad \beta = 0.32 \qquad \varepsilon_1 \sim \log N(-1.62, 1.8)$$

We will refer to this model as *model 1*.

Thereafter, we define *model 2* based on model 1, but now we assume that the parameters in the lognormal distribution of the residuals depend on x:

$$Y = \alpha X^{\beta} \varepsilon_2(x)$$

where $\varepsilon_2(x) \sim \log N(\mu_x; \sigma_x)$. We choose the residual parameter's dependence of x so that the dependence is linear on the Champernowne transformed axis

$$\sigma_x = 1.5 + 0.5 T_\theta(x) \qquad \mu_x = -0.5\sigma_x$$

where $T_\theta(x)$ is the Champernowne cdf defined in (7.14) with parameters $\theta = (1.66, 2.56 \cdot 10^7, 6.06 \cdot 10^{-5})$. In model 2, we use the same values of α and β as in model 1.

Now, we simulate $S = 100$ data sets with sample size $n = \{100, 500, 1000\}$ from model 1 and model 2 and with the X's bootstrapped from the original EML values in the commercial fire insurance data set. Moreover, we simulate a 25% and a 50% filtering scheme to each data set.

We mention that data simulated from model 1 corresponds to estimation with a true prior knowledge, whereas data simulated from model 2 corresponds to estimation with a roughly and not exactly true prior knowledge.

To each of the simulated data sets, we compute the transformed filtered data density estimator (7.15), the prior knowledge in the form of the median regression density estimator (7.16), and the transformed filtered data density estimator guided by prior knowledge (7.17). We call the density estimators $\bar{k}_{i,\phi}(x)$, where $i = \{1, 2, 3\}$ is the type of estimator defined analogously to section 7.5, and where $\phi = \{0, 25, 50\}$ is the filtering scheme, and compare them to the true density on the Champernowne

transformed axis, called $k(x)$, from either model 1 or model 2, with the following performance measure

$$\mathrm{ISE}(\bar{k}_{i,\phi}) = \frac{1}{n}\sum_{i=1}^{n}\{\bar{k}_{i,\phi}(X_i) - k(X_i)\}^2$$

where $(X_i)_{i=1,\ldots,n}$ are the bootstrapped X's in the sample.

	n=100		**n=500**		**n=1000**	
	model 1	model 2	model 1	model 2	model 1	model 2
MISE($\bar{k}_{1,0}$)	0.08147	0.07545	0.03556	0.03384	0.02516	0.02311
MISE($\bar{k}_{2,0}$)	0.07129	0.06554	0.03475	0.03152	0.02966	0.02520
MISE($\bar{k}_{3,0}$)	0.06515	0.06273	0.03770	0.03184	0.03346	0.02397
MISE($\bar{k}_{1,25}$)	0.09621	0.08348	0.04102	0.03649	0.02827	0.02440
MISE($\bar{k}_{2,25}$)	0.08581	0.07047	0.03702	0.03329	0.02762	0.02632
MISE($\bar{k}_{3,25}$)	0.07943	0.06700	0.03655	0.03169	0.02897	0.02322
MISE($\bar{k}_{1,50}$)	0.14757	0.11873	0.05607	0.04275	0.03981	0.02994
MISE($\bar{k}_{1,50}$)	0.14600	0.10957	0.05813	0.04396	0.04021	0.03509
MISE($\bar{k}_{1,50}$)	0.12855	0.10120	0.04835	0.03844	0.03331	0.02943

Table 7.1: Monte Carlo simulation comparing the performance of the estimators. For $\bar{k}_{i,\phi}$ $i = \{1, 2, 3\}$ corresponds to the type of estimator: $i = 1$ is the transformed filtered data density estimator, $i = 2$ is the prior knowledge density and $i = 3$ is the transformed filtered data density estimator guided by prior knowledge. $\phi = \{0, 25, 50\}$ indicating the filtering scheme.

In Table 7.1 at page 176, the average of the ISE errors are presented for each estimator, each n and each model. First, we notice that $\bar{k}_{3,\phi}$ outperforms $\bar{k}_{1,\phi}$ almost everywhere, even when the prior knowledge $\bar{k}_{2,\phi}$ has a poorer performance than $\bar{k}_{1,\phi}$. It seems that $\bar{k}_{3,\phi}$'s outperformance of $\bar{k}_{1,\phi}$ increases the more filtering we have. Furthermore, we observe that the performance gets worse when we increase the filtering. This is expected since we remove some information. The performance gap between the no filtering scheme and 25% filtering scheme is on average about 5% whereas the performance gap between the no filtering scheme and 50% filtering scheme is on average about 30%. Moreover, we notice that the performance gap between no filtering and filtering seems to decrease when the number of observations increases. Compar-

ing $\bar{k}_{1,\phi}$ and $\bar{k}_{2,\phi}$ we notice that the performance of $\bar{k}_{2,\phi}$ is always better when then the number of observations in the data set is small, whereas $\bar{k}_{1,\phi}$ is more competitive to $\bar{k}_{2,\phi}$, when the number of observations increases, especially when the prior knowledge (model 2) is not true. Comparing $\bar{k}_{3,\phi}$ and $\bar{k}_{2,\phi}$, we recognize that $\bar{k}_{3,\phi}$ almost always improves the performance of the prior knowledge when prior knowledge is not true (model 2), without aggravating the performance when the prior knowledge is true (model 1). Particulaly, when a large amount of filtering is present, $\bar{k}_{3,\phi}$ seems to be a desirable estimator.

At last, we compare the Monte Carlo results with the results from Buch-Kromann et al. (2009). If we compare the results without filtering, we recognize that the method described in this paper is about 20% worse than the corresponding estimates in the study. This means that if we have a data set without filtering, we should always use a standard multidimensional kernel density estimator. However, standard multidimensional kernel density estimators can not take filtering into consideration, and therefore the performance of these methods is very bad for filtered data. A small comparison, not presented in this paper shows that the performance gap between a standard multidimensional kernel density method and the filtering data density estimators described in this paper is about 20% in the 25% filtering scheme, and 115% in the 50% filtering scheme, and that the gap increases with the number of observations. This means that even for small amounts of filtering, it seems neccesary to use a method which is capable of taking filtering into account.

7.6 Conclusion

This paper presents a method for multivariate density estimation of truncated or censored data that pays special attention to extreme values. The estimation is based on a local constant estimator extended with dimension reducing prior knowledge and a tail flattening transformation. The asymptotic theory shows that these extensions will improve the performance of the estimator when the prior knowledge and the

transformation are not too different from the true distribution. A simulation study supports the asymptotic theory and shows substantial improvements in performance when using multiplicative bias correction.

7.7 Appendix

7.7.1 Proof of theorem 7.1

The proof of Theorem 7.1 is divided into two parts: First we analyse $\hat{f}_x^{(d,b)}$, where the leave-one-out estimator $\widehat{S}_{X_i,(i)}^{(b)}$ defined in (7.1) has been replaced by S_{X_i}. In the second part, we show that from an asymptotic point of view, we really can replace $\widehat{S}_{X_i,(i)}^{(b)}$ by S_{X_i}.

When analysing (7.2)

$$\widehat{f}_x^{(d,b)}(t) = \frac{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)\widehat{S}_{X_i,(i)}^{(b)}\,\mathrm{d}N_i(s)}{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)R_i(s)\,\mathrm{d}s}$$

we first notice that $\widehat{f}_x^{(d,b)}(t)$ has the same structure as the local constant hazard estimator

$$\widehat{\alpha}_x^{(d)}(t) = \frac{O_t}{E_t} = \frac{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)\,\mathrm{d}N_i(s)}{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)R_i(s)\,\mathrm{d}s}. \tag{7.18}$$

The only difference is the conditional survival function $\widehat{S}_{X_i,(i)}^{(b)}$ that enters the expression of $\widehat{f}_x^{(d,b)}$ but not $\widehat{\alpha}_x^{(d)}$.

When analysing $\widehat{\alpha}_x^{(d)}$, Nielsen and Linton (1995) divided the error of the hazard estimator into a variable part $V_x(t)$ converging in distribution and describing the asymptotic variance, and a stable part $B_x(t)$ converging in probability and describing the asymptotic bias. We have

$$\widehat{\alpha}_x^{(d)}(t) - \alpha_x(t) = V_x(t) + B_x(t),$$

where

$$V_x(t) = \frac{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)\,\mathrm{d}M_i(s)}{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)R_i(s)\,\mathrm{d}s}$$

and

$$B_x(t) = \frac{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)\{\alpha_{X_i}(s)-\alpha_x(t)\}R_i(s)\,\mathrm{d}s}{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)R_i(s)\,\mathrm{d}s}.$$

Now define

$$\overline{f}_x^{(d)}(t) = \frac{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)S_{X_i}(s)\,\mathrm{d}N_i(s)}{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)R_i(s)\,\mathrm{d}s},$$

where the only difference from $\hat{f}_x^{(d,b)}$ is that we have replaced $\widehat{S}_{X_i,(i)}^{(b)}$ by S_{X_i}.

When analysing $\overline{f}_x^{(d)}(t)$, we divide the error into its variable part $\overline{V}_x(t)$ and its stable part $\overline{B}_x(t)$ similarly to what is done for $\widehat{\alpha}_x^{(d)}(t)$ in Nielsen and Linton (1995):

$$\overline{f}_x^{(d)}(t) - f_x(t) = \overline{V}_x(t) + \overline{B}_x(t).$$

where

$$\overline{V}_x(t) = \frac{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)S_{X_i}(s)\,\mathrm{d}M_i(s)}{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)R_i(s)\,\mathrm{d}s}$$

and

$$\overline{B}_x(t) = \frac{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)\{f_{X_i}(s)-f_x(t)\}R_i(s)\,\mathrm{d}s}{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)R_i(s)\,\mathrm{d}s}$$

We first notice that $\overline{B}_x(t)$ is exactly the same functional of the density $f_x(s)$ as $B_x(t)$ is of the functional $\alpha_x(s)$. Therefore the asymptotic expression of $\overline{B}_x(t)$ is found by taking the asymptotic expression of $B_x(t)$ and then replacing the conditional hazard of this latter expression with our conditional density. From Theorem 1(b) in Nielsen and Linton (1995), we get that

$$d^{-2}\,\overline{B}_x(t) \xrightarrow{P} \mu_2(K)\{\mathcal{B}_1(f,\phi)(x,t) + \mathcal{B}_2(f,\phi)(x,t)\}$$

where $\mathcal{B}_1(f,\phi)(x,t)$ and $\mathcal{B}_2(f,\phi)(x,t)$ is defined in (7.11) and (7.12), respectively.

We can interpret $\overline{V}_x(t)$ by relating it to the corresponding expression for the hazard

$V_x(t)$. The only difference between these two expressions is that $S_{X_i}(s)$ enters in front of $dM_i(s)$ in the marginal integral of $\overline{V}_x(t)$, but not in $V_x(t)$. We therefore see that the asymptotic variance of $\overline{V}_x(t)$ is identical to the asymptotic variance of $V_x(t)$, but with the component $S_x^2(t)$ entering the compensator in the variance calculation, cf. Theorem 1(a) in Nielsen and Linton (1995):

$$\sqrt{n}d\overline{V}_x(t) \Rightarrow N\left\{0, \gamma_1(x,t)\right\}$$

where

$$\begin{aligned}\gamma_1(x,t) &= \{||K||_2^2\}^2 \frac{\alpha_x(t)S_x^2(t)}{\phi_x(t)} \\ &= \{||K||_2^2\}^2 \frac{f_x(t)S_x(t)}{\phi_x(t)}\end{aligned}$$

In the second part of the proof, we show that $\hat{f}_x^{(d,b)}(t)$ and $\overline{f}_x^{(d)}(t)$ are equivalent from an asymptotic point of view. First note that

$$\begin{aligned}&\left|\widehat{f}_x^{(d,b)}(t) - \overline{f}_x^{(d)}(t)\right| \\ &\quad= \left|\frac{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)\left(\widehat{S}_{X_i,(i)}^{(b)}(s) - S_{X_i}(s)\right)\,\mathrm{d}N_i(s)}{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)R_i(s)\,\mathrm{d}s}\right| \\ &\quad= \left|\frac{\sum_{i=1}^n K_d(x-X_i)\int K_d(t-s)\left(\widehat{S}_{X_i,(i)}^{(b)}(s) - S_{X_i}(s)\right)\,\mathrm{d}N_i(s)}{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)R_i(s)\,\mathrm{d}s}\right| \\ &\quad= \left|\frac{\sum_{i=1}^n K_d(x-X_i)h_i}{\sum_{j=1}^n K_d(x-X_j)}\,\frac{\sum_{j=1}^n K_d(x-X_j)}{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)R_i(s)\,\mathrm{d}s}\right| \\ &\quad\leq |\Theta(x)|\left|\sum_{i=1}^n a_i(x)h_i\right|\end{aligned}$$

where $h_i = \int K_d(t-s)\left(\widehat{S}_{X_i,(i)}^{(b)}(s) - S_{X_i}(s)\right)\,\mathrm{d}N_i(s)$,
$\Theta(x) = \left|\frac{n^{-1}\sum_{j=1}^n K_d(x-X_j)}{n^{-1}\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)R_i(s)\,\mathrm{d}s}\right|$ and $a_i(x) = \frac{K_d(x-X_i)}{\sum_{j=1}^n K_d(x-X_j)}$.

The numerator of $\Theta(x)$ is a kernel density estimator, and therefore it converges to a constant. Moreover, from the proof of Theorem 1 in Nielsen and Linton (1995), we know that the denominator of $\Theta(x)$ converges in probability. Therefore $|\Theta(x)| = O_P(1)$ can be neglected.

It now remains to be shown that

$$\begin{aligned}\Xi(x) &= \sum_{i=1}^{n} a_i(x)h_i \\ &= o_P\left(d^2 + n^{-1/2}d^{-1}\right)\end{aligned}$$

We know that $\ddot{S}^{(b)}_{X_i,(i)}(s) = \widehat{S}^{(b)}_{X_i,(i)}(s)$ with probability 1, where

$$\ddot{S}^{(b)}_{X_i,(i)}(s) = \exp\left\{-\int_0^s \ddot{\alpha}^{(b)}_{X_i,(i)}(u)\,\mathrm{d}u\right\},$$

and the hazard estimator, $\ddot{\alpha}^{(b)}_{X_i,(i)}(u) = \frac{\sum_{j\neq i}\int K_b(t-s)K_b(x-X_j)\,\mathrm{d}N_j(s)}{\max\left\{\sum_{j\neq i}\int K_b(t-s)K_b(x-X_j)R_j(s)\,\mathrm{d}s\,,\frac{n\zeta(u)}{2}\right\}}$ is a leave-one-out hazard estimator with smoothed exposure bounded from below, which follows from Assumption A, p. 164. Therefore it is sufficient to show that

$$\begin{aligned}\ddot{\Xi}(x) &= \sum_{i=1}^{n} a_i(x)\ddot{h}_i \\ &= o_P\left(d^2 + n^{-1/2}d^{-1}\right)\end{aligned}$$

where

$$\begin{aligned}\ddot{h}_i &= \int K_d(t-s)\left(\ddot{S}^{(b)}_{X_i,(i)}(s) - S_{X_i}(s)\right)\mathrm{d}N_i(s) \\ &= \int K(u)\left(\ddot{S}^{(b)}_{X_i,(i)}(t-du) - S_{X_i}(t-du)\right)\mathrm{d}N_i(t-du).\end{aligned}$$

The boundedness of the smoothed exposure from below in the hazard estimator $\ddot{\alpha}^{(b)}_{X_i,(i)}(u)$ ensures that the second moment of $\ddot{h}_i$ exists. This is essentially the same trick as used in Mammen and Nielsen (2007), p. 886. From algebra we know that

$(\sum_{i=1}^n a_i(x)\ddot{h}_i)^2 \leq \sum_{i=1}^n a_i(x)\ddot{h}_i^2$ since $\sum_{i=1}^n a_i(x) = 1$. Therefore

$$\ddot{\Xi}^2(x) \leq \sum_{i=1}^n a_i(x)\ddot{h}_i^2.$$

Taking the conditional expectation given X_i, we get

$$\mathbb{E}[\ddot{\Xi}^2(x)|X_i] = \sum_{j=1}^n a_i(x)\mathbb{E}\left[\ddot{h}_i^2|X_i\right].$$

For the survival function estimator with artificial exposure, $\ddot{S}^{(b)}_{(X_i),(i)}(s)$, the proof and result from Theorem 1 in Linton et al. (2003) holds for the second moment and we therefore get

$$\begin{aligned}
\mathbb{E}[\ddot{h}_i^2|X_i] &= \mathbb{E}\left[\left\{\int K(u)\left(\ddot{S}^{(b)}_{X_i,(i)}(t-du) - S_{X_i}(t-du)\right)\,\mathrm{d}N_i(t-du)\right\}^2\middle| X_i\right] \\
&= \mathbb{E}\left[\int K^2(u)\left(\ddot{S}^{(b)}_{X_i,(i)}(t-du) - S_{X_i}(t-du)\right)^2\,\mathrm{d}N_i(t-du)\middle| X_i\right] \\
&= \mathbb{E}\left[\int K^2(u)\left(\ddot{S}^{(b)}_{X_i,(i)}(t-du) - S_{X_i}(t-du)\right)^2\,\mathrm{d}\Lambda_i(t-du)\middle| X_i\right] \\
&= \int K^2(u)\mathbb{E}\left[\left(\ddot{S}^{(b)}_{X_i,(i)}(t-du) - S_{X_i}(t-du)\right)^2\middle| X_i\right]\,\mathrm{d}\Lambda_i(t-du) \\
&= \int K^2(u)\mathbb{E}\left[\left(\ddot{S}^{(b)}_{X_i,(i)}(t-du) - S_{X_i}(t-du)\right)^2\middle| X_i\right]\,\alpha_{X_i}(t-du)R_i(t-du)\,\mathrm{d}u \\
&= g_1(X_i)b^4 + g_2(X_i)n^{-1}b^{-1}
\end{aligned}$$

where the third equality holds because $\ddot{S}^{(b)}_{X_i,(i)}$ is a leave-one-out estimator and hence predictable. Moreover, the main components in $\ddot{S}^{(b)}_{X_i,(i)}(v)$ and $S^{(b)}_{X_i,(i)}(v)$ are $\int_0^v \ddot{\alpha}_y^{(b)}(u)\,\mathrm{d}u$ and $\int_0^v \alpha_y^{(b)}(u)\,\mathrm{d}u$; exactly the marginally integrated hazards considered in Linton et al. (2003). The last equality therefore follows from Theorem 1 in Linton et al. (2003), where the functions g_1 corresponding to the bias and g_2 corresponding to the

variance are continuous functions, and X_i belongs to a bounded interval. Therefore

$$\begin{aligned}\mathbb{E}[\ddot{\Xi}^2(x)|X_i] &\leq \sum_{i=1}^{n} a_i(x)\left(g_1(X_i)b^4 + g_2(X_i)n^{-1}b^{-1}\right) \\ &= O_P\left(b^4 + n^{-1}b^{-1}\right)\end{aligned}$$

which gives

$$\begin{aligned}\mathbb{E}[\ddot{\Xi}^2(x)] &= \mathbb{E}\left[\mathbb{E}[\ddot{\Xi}^2(x)|X_i]\right] \\ &= O_P\left(b^4 + n^{-1}b^{-1}\right)\end{aligned}$$

and hence

$$\begin{aligned}\left|\widehat{f}_x^{(b_1,b_2)}(t) - \overline{f}_x^{(b_1)}(t)\right| &= O_P\left(b^2 + n^{-1/2}b^{-1/2}\right) \\ &= o_P\left(d^2 + n^{-1/2}d^{-1}\right)\end{aligned}$$

where the last equality holds when $d > b$ and $d^2 < b$.

7.7.2 Proof of theorem 7.2

The proof of Theorem 7.2 is analogous to the proof of Theorem 7.1 in appendix 7.7.1. We define

$$\overline{g}_x^{(d)}(t) = h_x(t)\frac{\sum_{i=1}^{n}\int K_d(t-s)K_d(x-X_i)S_{X_i}(s)\{h_{X_i}(s)\}^{-1}\,\mathrm{d}N_i(s)}{\sum_{i=1}^{n}\int K_d(t-s)K_d(x-X_i)R_i(s)\,\mathrm{d}s},$$

and divide the error of $\overline{g}_x^{(d)}(t)$ into its variable part $\widehat{V}(t)$ and its stable part $\widehat{B}(t)$:

$$\overline{g}_x^{(d,b)}(t) - f_x(t) = \widehat{V}(t) + \widehat{B}(t)$$

where

$$\widehat{V}_x(t) = \frac{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)S_{X_i}(s)h_x(t)\{h_{X_i}(s)\}^{-1}\,\mathrm{d}M_i(s)}{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)R_i(s)\,\mathrm{d}s}$$

and

$$\begin{aligned}\widehat{B}_x(t) &= \frac{h_x(t)\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)\left[S_{X_i}(s)\{h_{X_i}(s)\}^{-1}\alpha_{X_i}(s) - \frac{f_x(t)}{h_x(t)}\right]R_i(s)\,\mathrm{d}s}{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)R_i(s)\,\mathrm{d}s} \\ &= h_x(t)B^*(t)\end{aligned}$$

where $B^*(t) = \frac{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)\{c_{X_i}(s)-c_x(t)\}R_i(s)\,\mathrm{d}s}{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)R_i(s)\,\mathrm{d}s}$

The variable part $\widehat{V}_x(t)$ corresponds to $\overline{V}_x(t)$ in the proof of Theorem 7.1 in appendix 7.7.1, but with an extra term $h_x(t)\{h_{X_i}(s)\}^{-1}$ that enters in front of $\mathrm{d}M_i(s)$. But $h_x(t)\{h_{X_i}(s)\}^{-1}$ is asymptotically equivalent to 1, and the asymptotics of the variable part of $\widehat{g}_x^{(d,b)}(t)$ is therefore identical to the asymptotics of the variable part of $\widehat{f}_x^{(d,b)}(t)$.

When it comes to the stable part $\widehat{B}_x(t)$, we note that $B_x^*(t)$ corresponds to $\overline{B}_x(t)$ in the proof of Theorem 7.1 in appendix 7.7.1, but with c instead of f. The final asymptotics of $\widehat{B}_x(t)$ is therefore identical to the asymptotics of $\overline{B}_x(t)$, but with c replacing f. We therefore have:

$$d^{-2}\,\widehat{B}_x(t) \xrightarrow{P} \mu_2(K)h_x(t)\{\mathcal{B}_1(c,\phi)(x,t) + \mathcal{B}_2(c,\phi)(x,t)\}$$

The second part of the proof, where we have to show that $\hat{g}_x^{(d,b)}(t)$ and $\overline{g}_x^{(d)}(t)$ are equvalent from an asymptotic point of view, corresponds to the proof of Theorem 7.1 in appendix 7.7.1.

7.7.3 Proof of theorem 7.3

The proof of Theorem 7.3 is based on a combination of the proof of Theorem 7.1 above and the technique used in the proof of the multivariate transformation approach without filtering in Theorem 2 in Buch-Larsen et al. (2005). Like in this latter paper, we argue that we can simply consider the pointwise asymptotic theory of the kernel density estimator on the transformed axes. That is, we can use the result from Theorem 1 on the transformed axes where the kernel density estimation is carried out. The conditional density on the transformed axes is $f_x\left\{\Psi^{-1}(v)\right\}\left[\psi\left\{\Psi^{-1}(v)\right\}\right]^{-1}$. We get the bias expression of Theorem 7.3 after we have back-transformed and multiplied by $\psi(t)$ as part of this process.

When it comes to the variance, we follow Buch-Larsen et al. (2005) in showing that the variance equals the variance calculated on the transformed axes – where a division of ψ comes from the expression of the density on the transformed axes – and then during the backtransformation we get a multiplication by ψ^2. The final result is that the variance is multiplied by ψ compared to the variance in Theorem 7.3.

7.7.4 Proof of theorem 7.4

The proof of Theorem 7.4 is based on a straight-forward combination of the proof of Theorem 7.2 and the proof of Theorem 7.3 and we leave it out.

7.7.5 Maximum likelihood parameters for the Champernowne distribution

The following describes the procedure for estimating the parameters of the Champernowne distribution (7.14) by a maximum likelihood procedure taking filtering into account.

Let $(\widetilde{Y}_i, X_i, T_i, D_i)_{i=1,\ldots,n}$ be the data set to which we want to estimate a Champer-

nowne distribution, where $\widetilde{Y}_i = Y_i \wedge C_i$ is the Y_i's subjected to right censoring, X_i is the covariate, T_i is the truncation and $D_i = I(Y_i \leq C_i)$ is the "at-risk" indicator. Let $N_i(s) = I(\widetilde{Y}_i, D = 1)$ be the corresponding counting process with intensities $\lambda_i(s)$, and let $R_i(s) = I(T_i < s < \widetilde{Y}_i)$ be the "at-risk" indicator. We can estimate a Champernowne distribution to this data set by assuming the following parametric model

$$\lambda_i(t, \theta) = \alpha(t, \theta) R_i(t)$$

where $\alpha(t, \theta) = \frac{\alpha(t+c)^{\alpha-1}}{(t+c)^{\alpha}+(M+c)^{\alpha}-2c^{\alpha}}$ is the parametric hazard function for the Champernowne distribution and $\theta = (\alpha, M, c)$ is the parameters in the Champernowne distribution.

Then it follows from Andersen et al. (1993) that the likelihood function is

$$L(\theta) = \left(\prod_{0<t\leq\infty} \alpha(t\theta)^{dN_{\cdot}(t)} \right) \exp\left(\int_0^{\tau} \alpha(u, \theta) R_{\cdot}(u) \, \mathrm{d}u \right)$$

where $N_{\cdot}(s) = \sum_{i=1}^{n} N_i(s)$ and $R_{\cdot}(s) = \sum_{i=1}^{n} R_i(s)$.

We therefore determine the parameters of the Champernowne distribution by maximizing the log likelihood function with respect to θ

$$\log L(\theta) = \sum_{i=1}^{n} \log\{\alpha(\widetilde{Y}_i, \theta)\} D_i - \int_0^{\infty} \alpha(u, \theta) R_i(u) \, \mathrm{d}u$$

Bibliography

P. K. Andersen, Ø. Borgan, R. D. Gill, and N. Keiding. *Statistical models based on counting processes*. Springer Series in Statistics. Springer-Verlag, New York, 1993. ISBN 0-387-97872-0.

P. Artzner. Application of coherent risk measures to capital requirements in insurance. *N. Am. Actuar. J.*, 3(2):11–25, 1999. ISSN 1092-0277. SOA Seminar: Integrated Approaches to Risk Measurement in the Financial Services Industry (Atlanta, GA, 1997).

P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent measures of risk. *Math. Finance*, 9(3):203–228, 1999. ISSN 0960-1627.

R. Beran. Nonparametric regression with randomly censored survival data. *Technical report, University of California, Berkeley*, 1981.

C. Bolancé, M. Guillén, and J. P. Nielsen. Kernel density estimation of actuarial loss functions. *Insurance: Mathematics and Economics*, 32:19–36, 2003.

C. Bolancé, M. Guillén, and J. P. Nielsen. Inverse beta transformation in kernel density estimation. *Statistics and Probability Letters*, 78(13):1757–1764, 2008a.

C. Bolancé, M. Guillén, E. Pelican, and R. Vernic. Skewed bivariate models and nonparametric estimation for the cte risk measure. *Insurance: Mathematics and Economics*, 43(3):386–393, 2008b.

E. Brodin and H. Rootzén. Univariate and bivariate gpd methods for predicting extreme wind storm losses. *Insurance: Mathematics and Economics*, 44(3):345–356, 2009.

Brown. Report of the Oxford Meeting, September 25-29, 1936. *Econometrica*, 5(4): 361–383, 1937.

T. Buch-Kromann. Estimation of large insurance losses: A case study. *Journal of Actuarial Practice*, 13:191–211, 2006.

T. Buch-Kromann. Comparison of tail performance of the champernowne transformed kernel density estimator, the generalized pareto distribution and the g-and-h distribution. *Journal of Operational Risk*, 4(2):1–25, 2009.

T. Buch-Kromann and J. P. Nielsen. Multivariate density estimation using dimension reducing information and tail flattening transformations for truncated or censored data, 2009. Submitted.

T. Buch-Kromann, M. Englund, J. Gustafsson, J. P. Nielsen, and F. Thuring. Non-parametric estimation of operational risk losses adjusted for under-reporting. *Scandinavian Actuarial Journal*, 4:293–304, 2007.

T. Buch-Kromann, M. Guillén, O. Linton, and J. P. Nielsen. Multivariate density estimation using dimension reducing information and tail flattening transformations, 2009. Submitted.

T. Buch-Larsen. A unified approach to the estimation of financial and actuarial loss distributions. Master's thesis, University of Copenhagen, Dept. of Mathematics and Statistics, Laboratory of Actuarial Mathematics, 2003.

T. Buch-Larsen, J. P. Nielsen, M. Guillén, and C. Bolancé. Kernel density estimation for heavy-tailed distributions using the Champernowne transformation. *Statistics*, 39(6):503–518, 2005. ISSN 0233-1888.

R. Cao, J. M. Vilar, and A. Devia. Modelling consumer credit risk via survival analysis. *SORT*, 33(1):3–20, 2009. With discussion and a rejoinder by the authors.

A. C. Cebrián, M. Denuit, and P. Lambert. Generalized Pareto fit to the society of actuaries' large claims database. *N. Am. Actuar. J.*, 7(3):18–36, 2003. ISSN 1092-0277.

D. G. Champernowne. The graduation of income distributions. *Econometrica*, 20: 591–615, 1952.

V. Chavez-Demoulin and A. C. Davison. Generalized additive modelling of sample extremes. *J. Roy. Statist. Soc. Ser. C*, 54(1):207–222, 2005. ISSN 0035-9254.

V. Chavez-Demoulin and P. Embrechts. Smooth extremal models in finance and insurance. *The Journal of Risk and Insurance*, 71(2):183–199, 2004.

S. X. Chen. Beta kernel estimators for density functions. *Comput. Statist. Data Anal.*, 31(2):131–145, 1999. ISSN 0167-9473.

S. X. Chen. Probability density function estimation using gamma kernels. *Ann. Inst. Statist. Math.*, 52(3):471–480, 2000. ISSN 0020-3157.

X. Chen, O. Linton, and P. Robinson. The estimation of conditional densities. *The Journal of Statistical Planning and Inference Special Issue in Honor of George Roussas*, pages 71–84, 2001. ISSN 0020-3157.

A. Clements, S. Hurn, and K. Lindsay. Mobius-like mappings and their use in kernel density estimation. *J. Amer. Statist. Assoc.*, 98(464):993–1000, 2003. ISSN 0162-1459.

S. Coles. *An introduction to statistical modeling of extreme values*. Springer Series in Statistics. Springer-Verlag London Ltd., London, 2001. ISBN 1-85233-459-2.

K. Cooray and M. M. A. Ananda. Modeling actuarial data with a composite lognormal-Pareto model. *Scand. Actuar. J.*, (5):321–334, 2005. ISSN 0346-1238.

S. Corradin. Economic risk capital and reinsurance: an extreme value theory's application to fire claims of an insurance company, 2002. Paper based on talks at "Sixth International Congress on Insurance: Mathematics and Economics, Lisbon,

2002", "Second Conference in Actuarial Science an Finance, Samos, 2002" and "III Workshop Finanza Matematica, Verona, 2002" and on a seminar at University of Pavia.

D. M. Dabrowska. Nonparametric regression with censored survival time data. *Scand. J. Statist.*, 14(3):181–197, 1987. ISSN 0303-6898.

M. Degen and P. Embrechts. Evt-based estimation of risk capital and cenvergence of high quantiles. *Applied Probability Trust*, 3:696–715, 2008.

M. Degen, P. Embrechts, and D. D. Lambrigger. The quantitative modeling of operational risk: between g-and-h and evt. *Astin Bulletin*, 2(37):265–291, 2007.

Denuit, Purcaru, and V. Keilegom. Bivariate archimedean copula models for censored data in non-life insurance. *Journal of Actuarial Practice*, 13:5–32, 2006.

D. J. Dupuis. Exceedances over high thresholds: A guide to threshold selection. *Extremes*, 1(3):251–261, 1999.

K. Dutta and J. Perry. A tale of tails: an empirical analysis of loss distribution models for estimating operational risk capital, 2006. Federal Reserve Bank of Boston, Working Paper No. 06-13.

K. K. Dutta and D. F. Babbel. On measuring skewness and kurtosis in short rate distributions: The case of the us dollar london inter bank offer rates, 2002. Wharton Financial Institutions Center Working Paper.

P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling extremal events*, volume 33 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1997. ISBN 3-540-60931-8. For insurance and finance.

D. Freedman. Statistical models and shoe leader. *Social Methodology*, 21:219–313, 1991.

A. Frigessi, O. Haug, and H. Rue. A dynamic mixture model for unsupervised tail estimation without threshold selection. *Extremes*, 5(3):219–235 (2003), 2002. ISSN 1386-1999.

O. L. Gebizlioglu and B. Yagci. Tolerance intervals for quantiles of bivariate risks and risk measurement. *Insurance: Mathematics and Economics*, 42(3):1022–1027, 2008.

C. Genest, H. U. Gerber, M. J. Goovaerts, and R. J. Laeven. Editorial to the special issue on modeling and measurement of multivariate risk in insurance and finance. *Insurance: Mathematics and Economics*, 44(2):143–145, 2009.

I. K. Glad, N. L. Hjort, and N. G. Ushakov. Correction of density estimators that are not densities. *Scand. J. Statist.*, 30(2):415–427, 2003. ISSN 0303-6898.

M. Guillen, J. Gustafsson, J. P. Nielsen, and P. Pritchard. Using external data in operational risk capital. *The Geneva papers*, 32:178–189, 2007.

J. Gustafsson. Modelling operational risk with kernel density estimation using the champernowne transformation. *The ICFAI Journal of Risk & Insurance*, 3(4): 39–75, 2006.

J. Gustafsson and J. P. Nielsen. A mixing model for operational risk. *The Journal of Operational Risk*, 3(3):25–37, 2008.

J. Gustafsson, J. P. Nielsen, P. Pritchard, and D. Roberts. Quantifying operational risk guided by kernel smoothing and continuous credibility. *The ICFAI Journal of Finacial Risk Management*, 3(2):23–47, 2006a.

J. Gustafsson, J. P. Nielsen, P. Pritchard, and D. Roberts. Quantifying operational risk guided by kernel smoothing and continuous credibility: A practioner's view. *The Journal of Operational Risk*, 1(1):43–55, 2006b.

J. Gustafsson, M. Hagmann, J. P. Nielsen, and O. Scaillet. Local transformation kernel density estimation of loss distributions. *Journal of Business and Economic Statistics*, 27(2):2–15, 2009.

M. Hagmann and O. Scaillet. Local multiplicative bias correction for asymmetric kernel density estimators. *J. Econometrics*, 141(1):213–249, 2007. ISSN 0304-4076.

M. Hagmann, O. Renault, and O. Scaillet. Estimation of recovery rate densities: Non-parametric and semi-parametric approaches versus industry practice, in recovery risk: The next challenge in credit risk management. *A., Sironi, A., Altman E., Risk Publications*, pages 323–346, 2005.

W. Härdle, M. Müller, S. Sperlich, and A. Werwatz. *Nonparametric and semiparametric models*. Springer Series in Statistics. Springer-Verlag, New York, 2004. ISBN 3-540-20722-8.

E. Hashorva. Tail asymptotic results for elliptical distributions. *Insurance: Mathematics and Economics*, 43(1):158–164, 2008.

T. J. Hastie and R. J. Tibshirani. *Generalized additive models*, volume 43 of *Monographs on Statistics and Applied Probability*. Chapman and Hall Ltd., London, 1990. ISBN 0-412-34390-8.

C. Heuchenne and I. V. Keilegom. Location estimation in nonparametric regression with censored data. *Journal of Multivariate Analysis*, 98:1558–1582, 2007a.

C. Heuchenne and I. V. Keilegom. Nonlinear regression with censored data. *Technometrics*, 49:34–44, 2007b.

N. L. Hjort and I. K. Glad. Nonparametric density estimation with a parametric start. *Ann. Statist.*, 23(3):882–904, 1995. ISSN 0090-5364.

N. L. Hjort and M. C. Jones. Locally parametric nonparametric density estimation. *Ann. Statist.*, 24(4):1619–1647, 1996. ISSN 0090-5364.

P. J. Huber. *Robust statistics*. John Wiley & Sons Inc., New York, 1981. ISBN 0-471-41805-6. Wiley Series in Probability and Mathematical Statistics.

M. Jacobsen. *Statistical analysis of counting processes*, volume 12 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1982. ISBN 0-387-90769-6.

N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous univariate distributions. Vol. 1*, volume 1 of *Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*, chapter 20, pages xxii+756. John Wiley & Sons Inc., New York, second edition, 1994. ISBN 0-471-58495-9. A Wiley-Interscience Publication.

M. C. Jones. Simple boundary correction for kernel density estimation. *Statistics and Computing*, 3:135–146, 1993.

M. C. Jones and P. J. Foster. A simple nonnegative boundary correction method for kernel density estimation. *Statist. Sinica*, 6(4):1005–1013, 1996. ISSN 1017-0405.

M. C. Jones, S. J. Davies, and B. U. Park. Versions of kernel-type regression estimators. *J. Amer. Statist. Assoc.*, 89(427):825–832, 1994. ISSN 0162-1459.

M. C. Jones, O. Linton, and J. P. Nielsen. A simple bias reduction method for density estimation. *Biometrika*, 82(2):327–338, 1995. ISSN 0006-3444.

M. C. Jones, J. S. Marron, and S. J. Sheather. A brief survey of bandwidth selection for density estimation. *J. Amer. Statist. Assoc.*, 91(433):401–407, 1996. ISSN 0162-1459.

M. C. Jones, D. F. Signorini, and N. L. Hjort. On multiplicative bias correction in kernel density estimation. *Sankhyā Ser. A*, 61(3):422–430, 1999. ISSN 0581-572X.

W. C. M. Kallenberg. Modelling dependence. *Insurance: Mathematics and Economics*, 42(1):127–146, 2008.

R. J. Karunamuni and S. Zhang. Some improvements on a boundary corrected kernel density estimator. *Statist. Probab. Lett.*, 78(5):499–507, 2008. ISSN 0167-7152.

M. Knecht and S. Küttel. The czeledin distribution function. *Astin Bulletin*, 2003.

G. Koekemoer and J. W. H. Swanepoel. A semi-parametric method for transforming data to normality. *Statistics and Computing*, 18(3):241–257, 2008a.

G. Koekemoer and J. W. H. Swanepoel. Transformation kernel density estimation with applications. *Journal of Computational and Graphical Statistics*, 17(3):750–769, 2008b.

S. Kotz and S. Nadarajah. *Extreme value distributions*. Imperial College Press, London, 2000. ISBN 1-86094-224-5. Theory and applications.

M. R. Leadbetter, G. Lindgren, and H. Rootzén. *Extremes and related properties of random sequences and processes*. Springer Series in Statistics. Springer-Verlag, New York, 1983. ISBN 0-387-90731-9.

E. L. Lehmann. *Theory of point estimation*. The Wadsworth & Brooks/Cole Statistics/Probability Series. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1991. ISBN 0-534-15978-8. Reprint of the 1983 original.

D. Li and L. Peng. Goodness-of-fit test for tail copulas modeled by elliptical copulas. *Statistics and Probability Letters*, 79(8):1097–1104, 2009.

G. Li and H. Doss. An approach to nonparametric regression for life history data using local linear fitting. *Ann. Statist.*, 23(3):787–823, 1995. ISSN 0090-5364.

O. Linton and J. P. Nielsen. A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, 82(1):93–100, 1995. ISSN 0006-3444.

O. Linton, E. Mammen, J. P. Nielsen, and I. V. Keilegom. Nonparametric regression with filtered data, 2007. Under revision in Bernoulli.

O. B. Linton, J. P. Nielsen, and S. van de Geer. Estimating multiplicative and additive hazard functions by kernel methods. *Ann. Statist.*, 31(2):464–492, 2003. ISSN 0090-5364. Dedicated to the memory of Herbert E. Robbins.

E. Mammen and J. P. Nielsen. A general approach to predictability issue in survival analysis with applications. *Biometrika*, 94(4):873–892, 2007.

J. S. Marron and D. Ruppert. Transformations to reduce boundary bias in kernel density estimation. *J. Roy. Statist. Soc. Ser. B*, 56(4):653–671, 1994. ISSN 0035-9246.

J. Martinez and B. Iglewicz. Some properties of the tukey g and h family distribution. *Communications in Statistics - Theory Methodology*, 13(3):353–369, 1984.

T. Martinussen and T. H. Scheike. *Dynamic regression models for survival data*. Statistics for Biology and Health. Springer, New York, 2006. ISBN 978-0387-20274-7; 0-387-20274-9.

I. W. McKeague and K. J. Utikal. Inference for a nonlinear counting process regression model. *Ann. Statist.*, 18(3):1172–1187, 1990. ISSN 0090-5364.

A. McNeil. Estimating the tails of loss severity distributions using extreme value theory. *ASTIN Bulletin*, 27:117–137, 1997.

A. McNeil and T. Saladin. The peaks over thresholds method for estimating high quantiles of loss distributions, 1997. Proceedings of 28th International ASTIN Colloquium.

A. J. McNeil, R. Frey, and P. Embrechts. *Quantitative risk management*. Princeton Series in Finance. Princeton University Press, Princeton, NJ, 2005. ISBN 0-691-12255-5. Concepts, techniques and tools.

M. Moscadelli. The modelling of operational risk:experiences with the analysis of the data collected by the basel committee, 2004. Bank of Italy, Working Paper No. 517.

S. Newcomb. Discussion and results of observations on transits of mercury from 1677 to 1881. *Astronomical Papers*, 1:363–487, 1882.

S. Newcomb. A Generalized Theory of the Combination of Observations so as to Obtain the Best Result. *Amer. J. Math.*, 8(4):343–366, 1886. ISSN 0002-9327.

J. P. Nielsen. Marker dependent kernel hazard estimation from local linear estimation. *Scand. Actuar. J.*, (2):113–124, 1998. ISSN 0346-1238.

J. P. Nielsen and O. B. Linton. Kernel estimation in a nonparametric marker dependent hazard model. *Ann. Statist.*, 23(5):1735–1748, 1995. ISSN 0090-5364.

J. P. Nielsen and C. Tanggaard. Boundary and bias correction in kernel hazard estimation. *Scand. J. Statist.*, 28(4):675–698, 2001. ISSN 0303-6898.

J. P. Nielsen, C. Jones, and C. Tanggaard. Local linear density estimation for filtered survival data. *To appear in Statistics*, 2009.

G. W. Peters and S. A. Sisson. Bayesian inference, monte carlo sampling and operational risk. *Journal of Operational Risk*, 1(3):27–50, 2006.

A. Ralston and P. Rabinowitz. *A first course in numerical analysis*. McGraw-Hill Book Co., New York, second edition, 1978. ISBN 0-07-051158-6. International Series in Pure and Applied Mathematics.

H. Ramlau-Hansen. Smoothing counting process intensities by means of kernel functions. *Ann. Statist.*, 11(2):453–466, 1983. ISSN 0090-5364.

D. E. A. Sanders. The modelling of extreme events. *British Actuarial Journal*, 11 (3):519–557, 2005.

O. Scaillet. Density estimation using inverse and reciprocal inverse Gaussian kernels. *J. Nonparametr. Stat.*, 16(1-2):217–226, 2004. ISSN 1048-5252. The International Conference on Recent Trends and Directions in Nonparametric Statistics.

D. W. Scott. *Multivariate density estimation*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York, 1992. ISBN 0-471-54770-0. Theory, practice, and visualization, A Wiley-Interscience Publication.

S. J. Sheather and M. C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc. Ser. B*, 53(3):683–690, 1991. ISSN 0035-9246.

B. W. Silverman. *Density estimation for statistics and data analysis*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1986. ISBN 0-412-24620-1.

S. M. Stigler. Simon Newcomb, Percy Daniell, and the history of robust estimation 1885–1920. *J. Amer. Statist. Assoc.*, 68:872–879, 1973. ISSN 0162-1459.

C. J. Stone. Optimal rates of convergence for nonparametric estimators. *Ann. Statist.*, 8(6):1348–1360, 1980. ISSN 0090-5364.

J. W. Tukey. A survey of sampling from contaminated distributions. In *Contributions to probability and statistics*, pages 448–485. Stanford Univ. Press, Stanford, Calif., 1960.

E. A. Valdez, J. Dhaene, M. Maj, and S. Vanduffel. Bounds and approximations for sums of dependent log-elliptical random variables. *Insurance: Mathematics and Economics*, 44(3):385–397, 2009.

I. Van Keilegom and M. G. Akritas. Transfer of tail information in censored regression models. *Ann. Statist.*, 27(5):1745–1784, 1999. ISSN 0090-5364.

I. Van Keilegom and N. Veraverbeke. Hazard rate estimation in nonparametric regression with censored data. *Ann. Inst. Statist. Math.*, 53(4):730–745, 2001. ISSN 0020-3157.

I. Van Keilegom and N. Veraverbeke. Density and hazard estimation in censored regression models. *Bernoulli*, 8(5):607–625, 2002. ISSN 1350-7265.

J. M. Vilar, R. Cao, C. González-Fragueiro, and M. Ausin. Nonparametric analysis of aggregate loss models. *Journal of Applied Statistics*, 36(2):149–166, 2009.

M. P. Wand and M. C. Jones. *Kernel smoothing*, volume 60 of *Monographs on Statistics and Applied Probability*. Chapman and Hall Ltd., London, 1995. ISBN 0-412-55270-1.

M. P. Wand, J. S. Marron, and D. Ruppert. Transformations in density estimation. *J. Amer. Statist. Assoc.*, 86(414):343–361, 1991. ISSN 0162-1459. With discussion and a rejoinder by the authors.

J. L. Wirch. Raising value at risk. *N. Am. Actuar. J.*, 3(2):106–115, 1999. ISSN 1092-0277. SOA Seminar: Integrated Approaches to Risk Measurement in the Financial Services Industry (Atlanta, GA, 1997).

L. Yang. Root-n convergent transformation-kernel density estimation. *J. Nonparametr. Statist.*, 12(4):447–474, 2000. ISSN 1048-5252.

L. Yang and J. S. Marron. Iterated transformation-kernel density estimation. *J. Amer. Statist. Assoc.*, 94(446):580–589, 1999. ISSN 0162-1459.

S. Zhang, R. J. Karunamuni, and M. C. Jones. An improved estimator of the density function at the boundary. *J. Amer. Statist. Assoc.*, 94(448):1231–1241, 1999. ISSN 0162-1459.