Ann Inst Stat Math (2012) 64:167–192
DOI 10.1007/s10463-010-0313-6

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

Tine Buch-Kromann · Jens Perch Nielsen

Received: 4 June 2009 / Revised: 4 December 2009 / Published online: 6 October 2010

Abstract 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.

Keywords Censoring · Champernowne · Counting process theory · Multiplicative correction · Nonparametric estimation · Truncation

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

T. Buch-Kromann (✉)
Department of Mathematical Sciences, University of Copenhagen,
Universitetsparken 5, 2100 Copenhagen Ø, Denmark
e-mail: tbl@math.ku.dk

J. P. Nielsen
Cass Business School, City University London, 106 Bunhill Row,
London EC1Y 8TZ, UK
e-mail: Jens.Nielsen.1@city.ac.uk

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. See also Buch-Kromann (2009) for a comparison of the transformation kernel density estimator and classical EVT and Bagkavos (2008) for an application of transformation in the context of hazard rate estimation.

In this paper we generalise the structured density model of Buch-Kromann et al. (2009) to the filtered data case. This extension however, makes it necessary to use a new mathematical and technical set-up compared with Buch-Kromann et al. (2009). 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. Bouaziz and Lopez (2010) is another recent paper on a general approach to multivariate density estimation based on censored data.

When dealing with filtered data, extensive use of counting process theory, see, i.e. 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 necessary. 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 with the local constant boundary bias of order $O(b)$,where b is the bandwidth, and further studied in Bagkavos (2009). 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 estimators of the location function of the heteroscedastic regression model are studied in Heuchenne and Van Keilegom (2007a) and a parametric version is studied in Heuchenne and Van 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, e.g. 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 ourselves 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 and Bagkavos and Patil (2008) focusing on plug-in bandwidth selectors by applying local linear fitting). 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.

Recently, Linton et al. (2010) introduced and analysed another version of our multiplicate model. Their approach is based on multiplicative hazard estimation based on standard smoothing without any tail correction. However, Linton et al. (2010) was concerned about the body of the distribution not the tail. In iid one-dimensional density estimation, we know from the study of Bolancé et al. (2003) that the standard kernel smoother breaks down when estimating the tail and standard bias correction cannot remedy this situation. Even though the multiplicative model only serves as a pilot guidance of our final fully nonparametric estimator, we do know from other studies, see for example Buch-Kromann (2009) that a good tail behaviour of the pilot estimator is essential for the performance of the final nonparametric estimator. This is our reason to construct an alternative density-based estimator incorporating a correction for heavy tails.

The paper is organized as follows: In Sect. 2, we define the general model and in Sect. 3, we define the estimators of the conditional density. In Sect. 4, the asymptotic properties of the estimators are presented, and Sect. 5 contains an application and a Monte Carlo study which compares the performance of the conditional density estimators. Section 6 is the conclusion.

2 The model

We would like to analyse (X, Y), but Y is not always observed. What we do observe is $(X, \tilde{Y}, D, T)$, where X is a one-dimensional covariate, $\tilde{Y} = Y \wedge C$ 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 $\tilde{Y}$ is only observed when $\tilde{Y} \geq T$. Suppose that Y and C are conditionally independent given X. Let $N(s) = I(\tilde{Y} \leq s, D = 1)$ 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 $R(s) = I(T < s < \tilde{Y})$ 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 variables $(X_1, \tilde{Y}_1, D_1, T_1), \ldots, (X_n, \tilde{Y}_n, D_n, T_n)$. The resulting counting processes $N_1, \ldots, N_n$ 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.

3 Estimating the conditional density

In this section we introduce estimators for the conditional density of filtered data. We first introduce a non-parametric filtered data density estimator taking filtering into consideration by means of counting process theory. This estimator is the fundamental estimator on which all the following density estimators are built, even though its usefulness 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 considerably 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.

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 $\hat{\alpha}_x^{(b)}(t) = \frac{O_t}{E_t}$, 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; $R_i(s) = I(T_i < s < \tilde{Y}_i)$. This gives an obvious candidate for our smoothed conditional survival function $\hat{S}_x^{(b)}(s) = \exp\left\{-\int_0^s \hat{\alpha}_x^{(b)}(u)\,\mathrm{d}u\right\}$.

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 $\hat{S}_{X_i}(s)\mathrm{d}N_i(s)$ in the kernel hazard estimator above. However, instead of $\hat{S}_{X_i}(s)\mathrm{d}N_i(s)$, we replace $\mathrm{d}N_i(s)$ with $\hat{S}_{X_i,(i)}(s)\mathrm{d}N_i(s)$, where $\hat{S}_{X_i,(i)}(s)$ is a leave-one-out estimator:

$$\hat{S}^{(b)}_{X_i,(i)}(s) = \exp\left\{-\int_0^s \hat{\alpha}^{(b)}_{X_i,(i)}(u)\,\mathrm{d}u\right\} \tag{1}$$

where $\hat{\alpha}^{(b)}_{X_i,(i)}(t) = \frac{\sum_{j\neq i}\int K_{b_1}(t-s)K_{b_2}(x-X_j)\,\mathrm{d}N_j(s)}{\sum_{j\neq i}\int K_{b_1}(t-s)K_{b_2}(x-X_j)R_j(s)\,\mathrm{d}s}$. This trick from Mammen and Nielsen (2007) simplifies the predicability issues in the proofs of the asymptotic results (see Appendix A); moreover, it often improves the performance of the estimators.

Under the assumption that not all data are removed in the filtration process, we arrive at the *non-parametric fitered data density estimator:*

$$\hat{f}_x^{(d,b)}(t) = \frac{\sum_{i=1}^n \int K_{d_1}(t-s)K_{d_2}(x-X_i)\hat{S}^{(b)}_{X_i,(i)}(s)\,\mathrm{d}N_i(s)}{\sum_{i=1}^n \int K_{d_1}(t-s)K_{d_2}(x-X_i)R_i(s)\mathrm{d}s}. \tag{2}$$

The bandwidths $b = (b_1, b_2)$ and $d = (d_1, d_2)$ in (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 $\hat{S}^{(b)}_{X_i,(i)}(s)$ would disturb the results. In the practical application in Sect. 5, where this estimator is applied, we have chosen $(b_1, b_2) = (d_1/2, d_2/2)$.

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 estimation accuracy; 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. (2008), Guillen et al. (2007), Gustafsson (2006), Gustafsson and Nielsen (2008), Gustafsson et al. (2006a,b). Other transformations include 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 $\tilde{N}_i = N_i \circ \Psi^{-1}$, where $(f \circ g)(x) = f\{g(x)\}$. Note that our transformed counting process is defined on $[0, 1]$. Now we calculate the non-parametric filtered data density estimator

(2) on the transformed data set and obtain what we will call the *transformed filtered data density estimator on the* Ψ*-transformed axis*:

$$\hat{k}_{\Psi,x}^{(d,b)}(v) = \frac{\sum_{i=1}^{n} \int_0^1 K_{d_1}(v-s) K_{d_2}(x-X_i) \hat{S}_{\Psi,X_i,(i)}^{(b)}(s)\, \mathrm{d}\tilde{N}_i(s)}{\sum_{i=1}^{n} \int_0^1 K_{d_1}(v-s) K_{d_2}(x-X_i) R_i\{\Psi^{-1}(s)\}\, \mathrm{d}s} \tag{3}$$

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

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

We backtransform (3) to obtain an estimator of $f_x(s)$, called the *transformed filtered data density estimator on the original axis*

$$\hat{f}_{\Psi,x}^{(d,b)}(t) = \psi(t) \times \hat{k}_{\Psi,x}^{(d,b)}\{\Psi(t)\}. \tag{4}$$

$\hat{k}_{\Psi,x}^{(d,b)}$ can be interpreted as an estimator of a correction to ψ, the density corresponding to the transformation function, Ψ.

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 (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). The multiplicative bias correction based on h_x is

$$\hat{c}_x^{(d,b)}(t) = \frac{\sum_{i=1}^{n} \int K_{d_1}(t-s) K_{d_2}(x-X_i) \hat{S}_{X_i,(i)}^{(b)}(s) \left\{h_{X_i}(s)\right\}^{-1} \mathrm{d}N_i(s)}{\sum_{i=1}^{n} \int K_{d_1}(t-s) K_{d_2}(x-X_i) R_i(s)\, \mathrm{d}s} \tag{5}$$

and the final multiplicatively bias corrected estimator of $f_x(s)$, called the *filtered data density estmator guided by prior knowledge*, is

$$\hat{g}_x^{(d,b)}(t) = h_x(t) \hat{c}_x^{(d,b)}(t). \tag{6}$$

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 (4) with the multiplicative bias correction from (6) to obtain a multiplicative corrected transformation estimator. On the Ψ-transformed axis the multiplicative bias corrections based on h_x is

$$\tilde{c}^{(d,b)}_{\Psi,x}(v) = \frac{\sum_{i=1}^{n} \int_0^1 K_{d_1}(v-s) K_{d_2}(x-X_i) \hat{S}^{(b)}_{\Psi,X_i,(i)}(s) [\tilde{h}_{X_i}\{\Psi^{-1}(s)\}]^{-1} \, \mathrm{d}\tilde{N}_i(s)}{\sum_{i=1}^{n} \int_0^1 K_{d_1}(v-s) K_{d_2}(x-X_i) R_i\{\Psi^{-1}(s)\} \, \mathrm{d}s} \tag{7}$$

where $\tilde{h}_x(s) = h_x(s)/\psi(s)$ and the density estimator on the Ψ-transformed axis is therefore

$$\tilde{k}^{(d,b)}_{\Psi,x}(v) = \tilde{h}_x\{\Psi^{-1}(v)\} \tilde{c}^{(d,b)}_{\Psi,x}(v) \tag{8}$$

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

$$\tilde{f}^{(d,b)}_{\Psi,x}(t) = \psi(t) \tilde{k}^{(d,b)}_{\Psi,x}\{\Psi(t)\} \tag{9}$$

called the *transformed filtered data density estimator guided by prior knowledge on original axis.*

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 theory 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 replacing $\mathrm{d}N(s)$ with $\hat{S}_X(s)\mathrm{d}N(s)$. This implies that the results obtained about hazard estimation from smoothing $\mathrm{d}N(s)$ can be transferred to density estimation by smoothing $\hat{S}_X(s)\mathrm{d}N(s)$ 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 $Z(x, s) = Pr(X \le x \mid R(s) = 1)$ be the differentiable conditional distribution of the covariate X given that the counting process 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 Sect. 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

$$\sum_{0 \le s \le t+\delta} \left| \frac{R(s)}{n} - \zeta(s) \right| \to o_P(1)$$

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 (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 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 holds:*

$$\sqrt{nd}\left\{\hat{f}_x^{(d,b)}(t) - f_x(t) - d^2\beta_1(x,t)\right\} \Longrightarrow N\{0, \gamma(x,t)\},$$

where

$$\begin{aligned} \beta_1(x,t) &= \mu_2(K)\{\mathcal{B}_1(f,\phi)(x,t) + \mathcal{B}_2(f,\phi)(x,t)\} \\ \gamma(x,t) &= \left\{||K||_2^2\right\}^2 \frac{f_x(t)S_x(t)}{\phi_x(t)}. \end{aligned} \tag{10}$$

The two functionals in (10), $\mathcal{B}_1$ and $\mathcal{B}_2$, both mappings from $\mathbb{R}_+ \times \mathbb{R}_+$ into $\mathbb{R}_+$, are defined by

$$\mathcal{B}_1(f,\phi)(x,t) = \frac{(\partial f_x(t)/\partial t)(\partial \phi_x(t)/\partial t)}{\phi_x(t)} + \frac{\partial^2 f_x(t)/\partial t^2}{2}, \tag{11}$$

$$\mathcal{B}_2(f,\phi)(x,t) = \frac{(\partial f_x(t)/\partial x)(\partial \phi_x(t)/\partial x)}{\phi_x(t)} + \frac{\partial^2 f_x(t)/\partial x^2}{2}. \tag{12}$$

Proof See Appendix A. □

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 (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 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*

$$\sqrt{nd}\left[\hat{g}_x^{(d,b)}(t) - f_x(t) - d^2\beta_2(x, t)\right] \Longrightarrow N\{0, \gamma(x, t)\}$$

where

$$\beta_2(x, t) = h_x(t)\mu_2(K)\{\mathcal{B}_1(c, \phi)(x, t) + \mathcal{B}_2(c, \phi)(x, t)\} \tag{13}$$

and $\gamma(x, t)$, $\mu_2(K)$, $\mathcal{B}_1$ *and* $\mathcal{B}_2$ *are defined in Theorem 1.*

Proof See Appendix B. □

Now we state the asymptotic theory of the density estimator when a transformation approach is used in our estimation process, i.e. (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 $f_x\left\{\Psi^{-1}(t)\right\}\left[\psi\left\{\Psi^{-1}(t)\right\}\right]^{-1}$. The map, $f\psi^{-1} \circ \Psi^{-1}$ is the conditional density of the dependent 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 3 (Transformed filtered data density estimator) *Suppose that Assumption A is satisfied and suppose that the functional* Ψ *is two times continuously differentiable; then*

$$\sqrt{nd}\left[\hat{f}_{\Psi,x}^{(d,b)}(t) - f_x(t) - d^2\beta_3(x, t)\right] \Longrightarrow N\{0, \psi(t)\gamma(x, t)\}$$

where

$$\begin{aligned}\beta_3(x, t) = \psi(t)\mu_2(K)\Big[&\mathcal{B}_1(f\psi^{-1} \circ \Psi^{-1}, \phi \circ \Psi^{-1})\{x, \Psi(t)\}\\ &+ \mathcal{B}_2(f\psi^{-1} \circ \Psi^{-1}, \phi \circ \Psi^{-1})\{x, \Psi(t)\}\Big]\end{aligned}$$

and $\gamma(x, t)$, $\mu_2(K)$, $\mathcal{B}_1$ *and* $\mathcal{B}_2$ *defined in Theorem 1.*

Proof See Appendix C. □

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 (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 4 (Transformed filtered data density estimator guided by prior knowledge) *Suppose that Assumption A is satisfied and suppose that the functional h is two times continuously differentiable, and the functional Ψ is three 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

$$\begin{aligned}\beta_4(x,t) &= h_x(t)\mu_2(K)\\ &\quad\times\left[\mathcal{B}_1(\tilde{c}\psi^{-1}\circ\Psi^{-1},\phi)\left\{x,\Psi(t)\right\} + \mathcal{B}_2(\tilde{c}\psi^{-1}\circ\Psi^{-1},\phi)\left\{x,\Psi(t)\right\}\right]\end{aligned}$$

and $\tilde{c}_x(t) = f_x(t)/\tilde{h}_x(t)$. $\gamma(x,t)$, $\mu_2(K)$, $\mathcal{B}_1$ and $\mathcal{B}_2$ are defined in Theorem 1.

Proof See Appendix D. □

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 (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

$$T(x) = \frac{(x+c)^{\alpha} - c^{\alpha}}{(x+c)^{\alpha} + (M+c)^{\alpha} - 2c^{\alpha}} \tag{14}$$

with parameters (α, M, c) estimated by a maximum likelihood procedure taking filtering into account (see Appendix E).

First define the transformed filtered data density estimator (3), where the transformation function Ψ is the Champernowne cdf (14) with maximal likelihood parameters as described above. The choice of the Champernowne cdf as transformation function is due to its 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 (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 $\hat{k}_1$ and is defined from (3)

$$\bar{k}_1(v) = \frac{\sum_{i=1}^{n} \int_0^1 K_d(v-s) K_d(x-X_i) \hat{S}^{(b)}_{T,X_i,(i)}(s)\, \mathrm{d}\tilde{N}_i(s)}{\sum_{i=1}^{n} \int_0^1 K_d(v-s) K_d(x-X_i) R_i\{T^{-1}(s)\}\, \mathrm{d}s} \tag{15}$$

where $\tilde{N}_i = N_i \circ T$, and $K_d(u) = d^{-1}K(u/d)$ where K is the Epanechnikov kernel function. 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 Sect. 3. As mentioned, T is the Champernowne cdf, defined in (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. (2010) given by

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

where ε and X is independent, and where the estimator of m is based on the density estimator (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 (4), which takes the corresponding filtering on ε into account. The filtering on ε follows directly from the filtering on Y, i.e. if (Y, X, T, C) is a claim Y, with corresponding EML X, truncation T and censoring C, then $(\tilde{T}, \tilde{C})$, where $\tilde{T} = T/m(X)$ and $\tilde{C} = C/m(X)$ 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 $\hat{h}_x(y)$ 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

$$\bar{k}_2(v) = \frac{\hat{h}_x(T^{-1}(v))}{T'(T^{-1}(v))} \tag{16}$$

At last, we define the transformed filtered data density estimator on transformed axes guided by the prior knowledge $\hat{h}_x$ defined above. The resulting estimator corresponds to (8) based on the Champernowne transformation.

$$\bar{k}_3(v) = \frac{\hat{h}_x\{T^{-1}(s)\} \sum_{i=1}^n \int_0^1 K_d(v-s) K_d(x-X_i) \hat{S}^{(b)}_{T,X_i,(i)}(s)[\hat{h}_{X_i}\{T^{-1}(s)\}]^{-1} \, \mathrm{d}\tilde{N}_i(s)}{\sum_{i=1}^n \int_0^1 K_d(v-s) K_d(x-X_i) R_i\{T^{-1}(s)\} \, \mathrm{d}s} \tag{17}$$

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 1 illustrates how the exposure of the fire claims data set is affected by the two filtering schemes compared with 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 (15). In Fig. 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

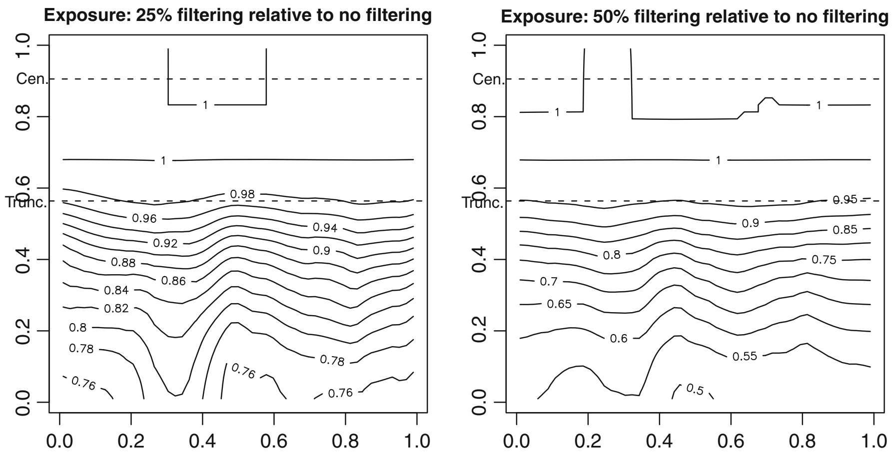

Fig. 1 Smoothed exposure of no filtering scheme relative to smoothed exposure of, respectively, 25% (*left*) and 50% (*right*) filtering scheme

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 and 5 claims, respectively.

5.1 Application

In the application study, we compute the transformed filtered data density estimator both without and with prior knowledge, i.e. (15) and (17), and plot them on the transformed axes together with the prior knowledge density (16) in the three data filtering schemes.

The transformed filtered data density estimator (15) of the fire claims data set is illustrated in Fig. 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 25 and 50% filtering schemes, as illustrated in Fig. 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 (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

$$\begin{aligned}
\theta_{x,0} &= (1.66, 2.56 \times 10^7, 6.06 \times 10^{-5}), & \theta_{y,0} &= (0.82, 7.54 \times 10^3, 3.44 \times 10^3)\\
\theta_{x,25} &= (1.67, 2.60 \times 10^7, 6.22 \times 10^{-5}), & \theta_{y,25} &= (0.82, 7.47 \times 10^3, 2.84 \times 10^3)\\
\theta_{x,50} &= (1.65, 2.78 \times 10^7, 6.80 \times 10^{-5}), & \theta_{y,50} &= (0.82, 7.67 \times 10^3, 2.17 \times 10^3)
\end{aligned}$$

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 (16) in the no-filtering, the 25- and 50% filtering schemes are illustrated in Fig. 3. As in Fig. 2 we recognize that the

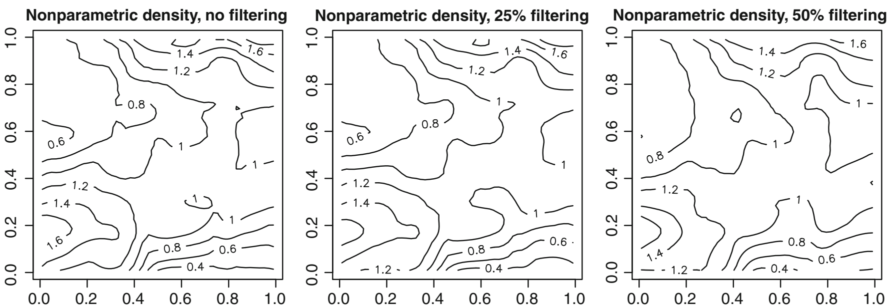

Fig. 2 The transformed filtered data density estimator (15) computed on the fire claim data set without filtering (*left*), with the 25% (*middle*) and the 50% (*right*) filtering scheme. We recognize that the density estimates are almost identical which illustrates the density estimator's ability to take filtering into account

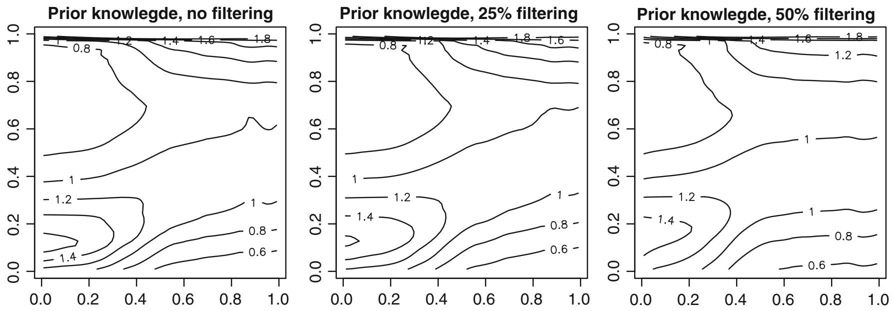

Fig. 3 The prior knowledge density (16) computed on the fire claim data set without filtering (*left*), with the 25% (*middle*) and the 50% (*right*) filtering scheme. We recognize that the density estimates are almost identical which illustrates the density estimator's ability to take filtering into account

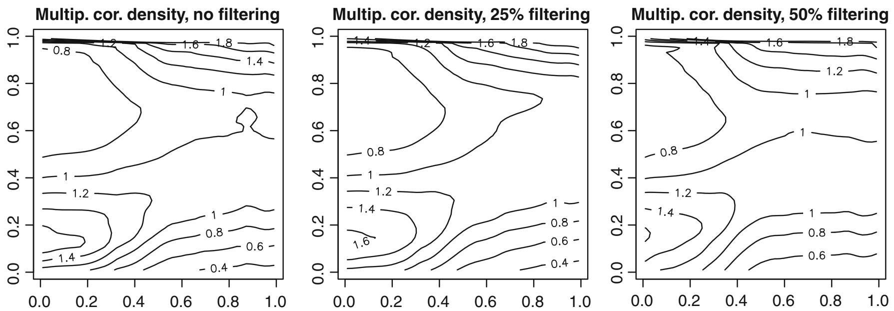

Fig. 4 The transformed filtered data density estimator guided by prior knowledge (17) computed on the fire claim data set without filtering (*left*), with the 25% (*middle*) and the 50% (*right*) filtering scheme. We recognize that the density estimates are almost identical which illustrates the density estimators ability to take filtering into account

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 too much structure into the density estimator. However, if it is not too wrong, then the multiplicative bias correction will correct it and benefit from it in the final estimator.

At last we illustrate the transformed filtered data density estimator guided by prior knowledge (17) in the three filtering schemes on Champernowne transformed axes in Fig. 4. Compared with Fig. 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 Fig. 4.

5.2 Monte Carlo study

In the Monte Carlo study, we want to compare the three performance of the estimators defined in Sect. 5 and illustrated in the application study. The simulation is based on the commercial fire insurance data set we have described 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 (14) with parameters $\theta = (1.66, 2.56 \times 10^7, 6.06 \times 10^{-5})$. 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 $n = \{100, 500, 1{,}000\}$ 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 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 (15), the prior knowledge in the form of the median regression density estimator (16) and the transformed filtered data density estimator guided by prior knowledge (17). We call the density estimators $\bar{k}_{i,\phi}(x)$, where $i = \{1, 2, 3\}$ is the type of estimator defined analogously to Sect. 5, and where $\phi = \{0, 25, 50\}$ is the filtering scheme and compare them with 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.

In Table 1, 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

Table 1 Monte Carlo simulation comparing the performance of the estimators

	$n = 100$		$n = 500$		$n = 1{,}000$	
	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

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

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. Comparing $\bar{k}_{1,\phi}$ 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). Particularly, when a large amount of filtering is present, $\bar{k}_{3,\phi}$ seems to be a desirable estimator.

When comparing the Monte Carlo results of our filtering estimation approach with the method of Buch-Kromann et al. (2009), see Table 2, we see that this approach of this latter paper is better at estimating the structured density when no filtering is present. However, when filtering is present the method of Buch-Kromann et al. (2009) breaks down as expected while our method still works well.

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

Table 2 Monte Carlo simulation comparing the performance of the estimators in Buch-Kromann et al. (2009)

	$n = 100$		$n = 500$		$n = 1{,}000$	
	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2
MISE($\bar{h}_{1,0}$)	0.06755	0.06071	0.02980	0.02791	0.02064	0.01888
MISE($\bar{h}_{2,0}$)	0.05607	0.06244	0.02057	0.02882	0.01459	0.02300
MISE($\bar{h}_{3,0}$)	0.05429	0.05774	0.02110	0.02514	0.01493	0.01882
MISE($\bar{h}_{1,25}$)	0.09934	0.08161	0.05174	0.04361	0.03910	0.03298
MISE($\bar{h}_{2,25}$)	0.08787	0.08542	0.04195	0.04488	0.03186	0.03736
MISE($\bar{h}_{3,25}$)	0.08519	0.08139	0.04196	0.04066	0.03237	0.03297
MISE($\bar{k}_{1,50}$)	0.19136	0.15584	0.12603	0.09889	0.10986	0.08855
MISE($\bar{k}_{1,50}$)	0.17858	0.15591	0.11564	0.10045	0.10321	0.09158
MISE($\bar{k}_{1,50}$)	0.17672	0.15031	0.11575	0.09561	0.10404	0.08900

For $\bar{h}_{i,\phi}$ $i = \{1, 2, 3\}$ corresponds to the type of estimator: $i = 1$ is the transformed kernel density estimator, $i = 2$ is the prior knowledge density and $i = 3$ is the multiplicative corrected transformation kernel density estimator. $\phi = \{0, 25, 50\}$ indicating the filtering scheme

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.

Appendix A: Proof of Theorem 1

The Proof of Theorem 1 is divided into two parts: First we analyse $\hat{f}_x^{(d,b)}$, where the leave-one-out estimator $\hat{S}^{(b)}_{X_i,(i)}$ defined in (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 $\hat{S}^{(b)}_{X_i,(i)}$ by S_{X_i}.

When analysing (2)

$$\hat{f}_x^{(d,b)}(t) = \frac{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)\hat{S}^{(b)}_{X_i,(i)}\,\mathrm{d}N_i(s)}{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)R_i(s)\,\mathrm{d}s}$$

we first notice that $\hat{f}_x^{(d,b)}(t)$ has the same structure as the local constant hazard estimator

$$\hat{\alpha}_x^{(d)}(t) = \frac{O_t}{E_t} = \frac{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)\,\mathrm{d}N_i(s)}{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)R_i(s)\,\mathrm{d}s}. \tag{18}$$

The only difference is the conditional survival function $\hat{S}^{(b)}_{X_i,(i)}$ that enters the expression of $\hat{f}_x^{(d,b)}$, but not $\hat{\alpha}_x^{(d)}$.

When analysing $\hat{\alpha}_x^{(d)}$, 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

$$\hat{\alpha}_x^{(d)}(t) - \alpha_x(t) = V_x(t) + B_x(t),$$

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

$$\bar{f}_x^{(d)}(t) = \frac{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)S_{X_i}(s)\,\mathrm{d}N_i(s)}{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)R_i(s)\,\mathrm{d}s},$$

where the only difference from $\hat{f}_x^{(d,b)}$ is that we have replaced $\hat{S}_{X_i,(i)}^{(b)}$ by S_{X_i}.

When analysing $\bar{f}_x^{(d)}(t)$, we divide the error into its variable part $\bar{V}_x(t)$ and its stable part $\bar{B}_x(t)$ similarly to what is done for $\hat{\alpha}_x^{(d)}(t)$ in Nielsen and Linton (1995):

$$\bar{f}_x^{(d)}(t) - f_x(t) = \bar{V}_x(t) + \bar{B}_x(t).$$

where

$$\bar{V}_x(t) = \frac{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)S_{X_i}(s)\,\mathrm{d}M_i(s)}{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)R_i(s)\,\mathrm{d}s}$$

and

$$\bar{B}_x(t) = \frac{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)\{f_{X_i}(s) - f_x(t)\}R_i(s)\,\mathrm{d}s}{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)R_i(s)\,\mathrm{d}s}$$

We first notice that $\bar{B}_x(t)$ 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 $\bar{B}_x(t)$ 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

$$d^{-2}\,\bar{B}_x(t) \xrightarrow{P} \mu_2(K)\{\mathcal{B}_1(f,\phi)(x,t) + \mathcal{B}_2(f,\phi)(x,t)\}$$

where $\mathcal{B}_1(f,\phi)(x,t)$ and $\mathcal{B}_2(f,\phi)(x,t)$ is defined in (11) and (12), respectively.

We can interpret $\bar{V}_x(t)$ 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 $\bar{V}_x(t)$, but not in $V_x(t)$. We therefore see that the asymptotic variance of $\bar{V}_x(t)$ 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):

$$\sqrt{n}d\bar{V}_x(t) \Rightarrow N\{0, \gamma_1(x,t)\}$$

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 $\bar{f}_x^{(d)}(t)$ are equivalent from an asymptotic point of view. First note that

$$\begin{aligned}&\left|\hat{f}_x^{(d,b)}(t) - \bar{f}_x^{(d)}(t)\right|\\ &= \left|\frac{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)\left(\hat{S}^{(b)}_{X_i,(i)}(s) - S_{X_i}(s)\right)\mathrm{d}N_i(s)}{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)R_i(s)\,\mathrm{d}s}\right|\\ &= \left|\frac{\sum_{i=1}^n K_d(x-X_i)\int K_d(t-s)\left(\hat{S}^{(b)}_{X_i,(i)}(s) - S_{X_i}(s)\right)\mathrm{d}N_i(s)}{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)R_i(s)\,\mathrm{d}s}\right|\\ &= \left|\frac{\sum_{i=1}^n K_d(x-X_i)h_i}{\sum_{j=1}^n K_d(x-X_j)}\;\frac{\sum_{j=1}^n K_d(x-X_j)}{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)R_i(s)\,\mathrm{d}s}\right|\\ &\le |\Theta(x)|\left|\sum_{i=1}^n a_i(x)h_i\right|\end{aligned}$$

where $h_i = \int K_d(t-s)\left(\hat{S}^{(b)}_{X_i,(i)}(s) - S_{X_i}(s)\right)\mathrm{d}N_i(s)$, $\Theta(x) = \left|\frac{n^{-1}\sum_{j=1}^n K_d(x-X_j)}{n^{-1}\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)R_i(s)\,\mathrm{d}s}\right|$ 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

$$\begin{aligned}\Xi(x) &= \sum_{i=1}^{n} a_i(x) h_i \\ &= o_P(d^2 + n^{-1/2} d^{-1})\end{aligned}$$

We know that $\ddot{S}^{(b)}_{X_i,(i)}(s) = \hat{S}^{(b)}_{X_i,(i)}(s)$ 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. Therefore it is sufficient to show that

$$\begin{aligned}\ddot{\Xi}(x) &= \sum_{i=1}^{n} a_i(x) \ddot{h}_i \\ &= o_P(d^2 + n^{-1/2} d^{-1})\end{aligned}$$

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 $\left(\sum_{i=1}^n a_i(x)\ddot{h}_i\right)^2 \leq \sum_{i=1}^n a_i(x)\ddot{h}_i^2$ 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 $\int_0^v \ddot{\alpha}^{(b)}_y(u)\,\mathrm{d}u$ and $\int_0^v \alpha^{(b)}_y(u)\,\mathrm{d}u$; 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,

$$\begin{aligned}\mathbb{E}[\ddot{\Xi}^2(x)|X_i] &\le \sum_{i=1}^n a_i(x)\left(g_1(X_i)b^4 + g_2(X_i)n^{-1}b^{-1}\right) \\ &= O_P(b^4 + n^{-1}b^{-1})\end{aligned}$$

which gives

$$\begin{aligned}\mathbb{E}[\ddot{\Xi}^2(x)] &= \mathbb{E}\left[\mathbb{E}\left[\ddot{\Xi}^2(x)|X_i\right]\right] \\ &= O_P(b^4 + n^{-1}b^{-1})\end{aligned}$$

and hence

$$\begin{aligned}\left|\hat{f}_x^{(b_1,b_2)}(t) - \bar{f}_x^{(b_1)}(t)\right| &= O_P(b^2 + n^{-1/2}b^{-1/2}) \\ &= o_P(d^2 + n^{-1/2}d^{-1})\end{aligned}$$

where the last equality holds when $d > b$ and $d^2 < b$. □

Appendix B: Proof of Theorem 2

The Proof of Theorem 2 is analogous to the Proof of Theorem 1 in Appendix A. We define

$$\bar{g}_x^{(d)}(t) = h_x(t) \frac{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)S_{X_i}(s)\{h_{X_i}(s)\}^{-1}\,\mathrm{d}N_i(s)}{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)R_i(s)\,\mathrm{d}s},$$

and divide the error of $\bar{g}_x^{(d)}(t)$ into its variable part $\hat{V}(t)$ and its stable part $\hat{B}(t)$:

$$\bar{g}_x^{(d,b)}(t) - f_x(t) = \hat{V}(t) + \hat{B}(t)$$

where

$$\hat{V}_x(t) = \frac{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)S_{X_i}(s)h_x(t)\{h_{X_i}(s)\}^{-1}\,\mathrm{d}M_i(s)}{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)R_i(s)\,\mathrm{d}s}$$

and

$$\begin{aligned}\hat{B}_x(t) &= \frac{h_x(t)\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)\left[S_{X_i}(s)\{h_{X_i}(s)\}^{-1}\alpha_{X_i}(s) - \frac{f_x(t)}{h_x(t)}\right]R_i(s)\,\mathrm{d}s}{\sum_{i=1}^n \int K_d(t-s)K_d(x-X_i)R_i(s)\,\mathrm{d}s}\\ &= h_x(t)B^*(t)\end{aligned}$$

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 $\hat{V}_x(t)$ corresponds to $\bar{V}_x(t)$ in the Proof of Theorem 1 in Appendix A, but with an extra term $h_x(t)\left\{h_{X_i}(s)\right\}^{-1}$ that enters in front of $\mathrm{d}M_i(s)$. But $h_x(t)\left\{h_{X_i}(s)\right\}^{-1}$ is asymptotically equivalent to 1, and the asymptotics of the variable part of $\hat{g}_x^{(d,b)}(t)$ is therefore identical to the asymptotics of the variable part of $\hat{f}_x^{(d,b)}(t)$.

When it comes to the stable part $\hat{B}_x(t)$, we note that $B_x^*(t)$ corresponds to $\bar{B}_x(t)$ in the Proof of Theorem 1 in Appendix A, but with c instead of f. The final asymptotics of $\hat{B}_x(t)$ is therefore identical to the asymptotics of $\bar{B}_x(t)$, but with c replacing f. We therefore have

$$d^{-2}\,\hat{B}_x(t) \xrightarrow{P} \mu_2(K)h_x(t)\{\mathcal{B}_1(c,\phi)(x,t) + \mathcal{B}_2(c,\phi)(x,t)\}$$

The second part of the proof, where we have to show that $\hat{g}_x^{(d,b)}(t)$ and $\bar{g}_x^{(d)}(t)$ are equivalent from an asymptotic point of view, corresponds to the Proof of Theorem 1 in Appendix A. □

Appendix C: Proof of Theorem 3

The Proof of Theorem 3 is based on a combination of the Proof of Theorem 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 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 with the variance in Theorem 3. □

Appendix D: Proof of Theorem 4

The Proof of Theorem 4 is based on a straightforward combination of the Proof of Theorem 2 and the Proof of Theorem 3 and we leave it out. □

Appendix E: Maximum likelihood parameters for the Champernowne distribution

The following describes the procedure for estimating the parameters of the Champernowne distribution (14) by a maximum likelihood procedure taking filtering into account.

Let $(\tilde{Y}_i, X_i, T_i, D_i)_{i=1,\ldots,n}$ be the data set to which we want to estimate a Champernowne distribution, where $\tilde{Y}_i = Y_i \wedge C_i$ 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 $N_i(s) = I(\tilde{Y}_i, D = 1)$ be the corresponding counting process with intensities $\lambda_i(s)$, and let $R_i(s) = I(T_i < s < \tilde{Y}_i)$ 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

$$L(\theta) = \left(\prod_{0<t\leq\infty} \alpha(t\theta)^{dN.(t)}\right) \exp\left(\int_0^{\tau} \alpha(u,\theta)R.(u)\,\mathrm{d}u\right)$$

where $N.(s) = \sum_{i=1}^{n} N_i(s)$ and $R.(s) = \sum_{i=1}^{n} R_i(s)$.

We therefore determine the parameters of the Champernowne distribution by maximizing the log likelihood function with respect to θ

$$\log L(\theta) = \sum_{i=1}^{n} \log\{\alpha(\tilde{Y}_i, \theta)\} D_i - \int_0^{\infty} \alpha(u, \theta) R_i(u)\,\mathrm{d}u$$

Acknowledgments We would like to thank Dorota M. Dabrowska, Jianqing Fan and Thomas Mikosch, for helpful comments.

References

Andersen, P. K., Borgan, Ø., Gill, R. D., Keiding, N. (1993). *Statistical models based on counting processes*. Springer Series in Statistics. New York: Springer-Verlag.

Bagkavos, D. (2008). Transformations in hazard rate estimation. *Journal of Nonparametric Statistics, 20*(8), 721–738.

Bagkavos, D. (2009). Local linear hazard rate estimation and bandwidth selection. *The Annals of the Institute of Statistical Mathematics* (in press).

Bagkavos, D., Patil, P. N. (2008). Local polynomial fitting in failure rate estimation. *IEEE Transactions on Reliability, 56*, 126–163.

Beran, R. (1981). Nonparametric regression with randomly censored survival data. Unpublished technical report, Berkeley: University of California.

Bolancé, C., Guillén, M., Nielsen, J. P. (2003). Kernel density estimation of actuarial loss functions. *Insurance: Mathematics and Economics, 32*, 19–36.

Bolancé, C., Guillén, M., Nielsen, J. P. (2008). Inverse beta transformation in kernel density estimation. *Statistics and Probability Letters, 78*, 1757–1764.

Bouaziz, O., Lopez, O. (2010). Conditional density estimation in a censored single-index regression model. *Bernoulli, 16*(2), 514–542.

Buch-Kromann, T. (2009). Comparison of tail performance of the champernowne transformed kernel density estimator, the generalized pareto distribution and the g-and-h distribution. *The Journal of Operational Risk, 4*(2), 43–67.

Buch-Kromann, T., Englund, M., Gustafsson, J., Nielsen, J. P., Thuring, F. (2007). Non-parametric estimation of operational risk losses adjusted for under-reporting. *Scandinavian Actuarial Journal, 4*, 293–304.

Buch-Kromann, T., Guillén, M., Linton, O., Nielsen, J. P. (2009). Multivariate density estimation using dimension reducing information and tail flattening transformations. Under revision to Insurance: Mathematics and Economics.

Buch-Larsen, T., Nielsen, J. P., Guillén, M., Bolancé, C. (2005). Kernel density estimation for heavy-tailed distributions using the Champernowne transformation. *Statistics, 39*(6), 503–518.

Clements, A., Hurn, S., Lindsay, K. (2003). Mobius-like mappings and their use in kernel density estimation. *Journal of the American Statistical Association, 98*(464), 993–1000.

Dabrowska, D. M. (1987). Nonparametric regression with censored survival time data. *Scandinavian Journal of Statistics Theory and Applications, 14*(3), 181–197.

Embrechts, P., Klüppelberg, C., Mikosch, T. (1997). *Modelling extremal events*. Applications of Mathematics (New York) (Vol. 33). Berlin: Springer.

Guillen, M., Gustafsson, J., Nielsen, J. P., Pritchard, P. (2007). Using external data in operational risk capital. *The Geneva papers, 32*, 178–189.

Gustafsson, J. (2006). Modelling operational risk with kernel density estimation using the champernowne transformation. *The ICFAI Journal of Risk & Insurance, 3*(4), 39–75.

Gustafsson, J., Nielsen, J. P. (2008). A mixing model for operational risk. *The Journal of Operational Risk, 3*(3), 25–37.

Gustafsson, J., Nielsen, J. P., Pritchard, P., Roberts, D. (2006a). Quantifying operational risk guided by kernel smoothing and continuous credibility. *The ICFAI Journal of Financial Risk Management, 3*(2), 23–47.

Gustafsson, J., Nielsen, J. P., Pritchard, P., Roberts, D. (2006b). Quantifying operational risk guided by kernel smoothing and continuous credibility: a practioner's view. *The Journal of operational risk, 1*(1), 43–55.

Hastie, T. J., Tibshirani, R. J. (1990). *Generalized additive models*. Monographs on statistics and applied probability (Vol. 43). London: Chapman & Hall Ltd.

Heuchenne, C., Van Keilegom, I. (2007a). Location estimation in nonparametric regression with censored data. *Journal of Multivariate Analysis, 98*, 1558–1582.

Heuchenne, C., Van Keilegom, I. (2007b). Nonlinear regression with censored data. *Technometrics, 49*, 34–44.

Jacobsen, M. (1982). *Statistical analysis of counting processes*. Lecture notes in statistics (Vol. 12). New York: Springer.

Koekemoer, G., Swanepoel, J. W. H. (2008a). A semi-parametric method for transforming data to normality. *Statistics and Computing, 18*(3), 241–257.

Koekemoer, G., Swanepoel, J. W. H. (2008b). Transformation kernel density estimation with applications. *Journal of Computational and Graphical Statistics, 17*(3), 750–769.

Li, G., Doss, H. (1995). An approach to nonparametric regression for life history data using local linear fitting. *The Annals of Statistics, 23*(3), 787–823.

Linton, O., Nielsen, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika, 82*(1), 93–100.

Linton, O., Nielsen, J. P., van de Geer, S. (2003). Estimating multiplicative and additive hazard functions by kernel methods. *The Annals of Statistics, 31*(2), 464–492, dedicated to the memory of Herbert E. Robbins.

Linton, O., Mammen, E., Nielsen, J. P., Van Keilegom, I. (2010). Nonparametric regression with filtered data. *Bernoulli* (to appear).

Mammen, E., Nielsen, J. P. (2007). A general approach to predictability issue in survival analysis with applications. *Biometrika, 94*(4), 873–892.

Martinussen, T., Scheike, T. H. (2006). *Dynamic regression models for survival data*. Statistics for biology and health. New York: Springer.

McKeague, I. W., Utikal, K. J. (1990). Inference for a nonlinear counting process regression model. *The Annals of Statistics, 18*(3), 1172–1187.

Nielsen, J. P. (1998). Marker dependent kernel hazard estimation from local linear estimation. *Scandinavian Actuarial Journal, 2*, 113–124.

Nielsen, J. P., Linton, O. (1995). Kernel estimation in a nonparametric marker dependent hazard model. *The Annals of Statistics, 23*(5), 1735–1748.

Nielsen, J. P., Tanggaard, C. (2001). Boundary and bias correction in kernel hazard estimation. *Scandinavian Journal of Statistics Theory and Applications, 28*(4), 675–698.

Nielsen, J. P., Jones, C., Tanggaard, C. (2009). Local linear density estimation for filtered survival data. *Statistics, 2*(43), 167–186.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Monographs on statistics and applied probability. London: Chapman & Hall.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics, 8*(6), 1348–1360.

Van Keilegom, I., Akritas, M. G. (1999). Transfer of tail information in censored regression models. *The Annals of Statistics, 27*(5), 1745–1784.

Van Keilegom, I., Veraverbeke, N. (2001). Hazard rate estimation in nonparametric regression with censored data. *Annals of the Institute of Statistical Mathematics, 53*(4), 730–745.

Van Keilegom, I., Veraverbeke, N. (2002). Density and hazard estimation in censored regression models. *Bernoulli, 8*(5), 607–625.

Wand, M. P., Marron, J. S., Ruppert, D. (1991). Transformations in density estimation. *Journal of the American Statistical Association, 86*(414), 343–361, with discussion and a rejoinder by the authors.

Yang, L., Marron, J. S. (1999). Iterated transformation-kernel density estimation. *Journal of the American Statistical Association, 94*(446), 580–589.