View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by City Research Online

City Research Online

City, University of London Institutional Repository

**Citation**: Bischofberger, S., Hiabu, M., Mammen, E. and Nielsen, J. P. ORCID: 0000-0002-2798-0817 (2019). A comparison of in-sample forecasting methods. Computational Statistics and Data Analysis, 137, pp. 133-154. doi: 10.1016/j.csda.2019.02.009

This is the accepted version of the paper.

This version of the publication may differ from the final published version.

**Permanent repository link:** https://openaccess.city.ac.uk/id/eprint/21905/

**Link to published version**: http://dx.doi.org/10.1016/j.csda.2019.02.009

**Copyright and reuse:** City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

City Research Online: http://openaccess.city.ac.uk/ publications@city.ac.uk

# A comparison of in-sample forecasting methods

Stephan M. Bischofberger[a,*], Munir Hiabu[b], Enno Mammen[c], Jens Perch Nielsen[a]

[a] *Cass Business School, City, University of London, 106 Bunhill Row, London, EC1Y 8TZ, United Kingdom*
[b] *School of Mathematics and Statistics, University of Sydney, Camperdown NSW 2006, Australia*
[c] *Institute for Applied Mathematics, Heidelberg University, Im Neuenheimer Feld 205, 69120 Heidelberg, Germany*

**Abstract**

In-sample forecasting is a recent continuous modification of well-known forecasting methods based on aggregated data. These aggregated methods are known as age-cohort methods in demography, economics, epidemiology and sociology and as chain ladder in non-life insurance. Data is organized in a two-way table with age and cohort as indices, but without measures of exposure. It has recently been established that such structured forecasting methods based on aggregated data can be interpreted as structured histogram estimators. Continuous in-sample forecasting transfers these classical forecasting models into a modern statistical world including smoothing methodology that is more efficient than smoothing via histograms. All in-sample forecasting estimators are collected and their performance is compared via a finite sample simulation study. All methods are extended via multiplicative bias correction. Asymptotic theory is being developed for the histogram-type method of sieves and for the multiplicatively corrected estimators. The multiplicative bias corrected estimators improve all other known in-sample forecasters in the simulation study. The density projection approach seems to have the best performance with forecasting based on survival densities being the runner-up.

*Keywords:* age-cohort model, chain ladder method, in-sample forecasting, multiplicative bias correction, nonparametric estimation.

## 1. Introduction

In a period where mathematical statistical fitting of big data via machine learning type of algorithms gets a lot of attention in computational driven advances of prediction, it is worth to remember that some of the most important problems in mathematical statistics are forecasting problems. While a major field of econometrics, mathematical statistics, finance and other fields have researched time series approaches to forecasting, in practice age-cohort methods have often been used as a simpler and more stable alternative to time series. In this paper, we study in-sample forecasting. In-sample forecasting is a recently suggested continuous modification of age-cohort methods that takes advantage of modern smoothing technology. In

*Corresponding author
*Email address:* stephan.bischofberger@cass.city.ac.uk (Stephan M. Bischofberger)

particular, we will present a detailed simulation study comparing several estimators proposed for in-sample forecasting.

In age-cohort models, a cohort is a group of individuals or objects with shared characteristics. Analysis of cohorts is considered in many academic fields, with cohorts representing a common date: date of birth (longevity), admission date to a hospital or prison (longitudinal studies, epidemiology), start date of unemployment (economics), underwriting date of an insurance policy (actuarial science), etc. When modeling different cohorts it is implicitly assumed that individuals in the same cohort have similarities due to a shared environment that differentiates them from other cohorts. In an age-period-cohort model one additionally considers age, i.e., the time from the initial date until onset of an event, and period, i.e., the calendar date of the event. The outcome of interest, $\mu$, for cohort $i$ and age $k$ is modeled log-linearly:

$$\log \mu_{ik} = \alpha_i + \beta_k + \gamma_j, \quad \text{(age-period-cohort model)} \tag{1}$$

where $\alpha$ is the effect of cohort $i$, $\beta$ corresponds to age $k$ and $\gamma$ to period $j$. The parameters $\alpha, \beta, \gamma$ are assumed fixed but unknown and have to be estimated from the data. The dependence on the period, $j$, is implicit via $j = i + k - 1$. Model (1) is omnipresent in a wide array of fields often arising from repeated cross-sectional studies. Recent contributions among many others include aging (Yang, 2011), blood pressure (Tu et al., 2011), health inequalities (Jeon et al., 2016), social capital (Schwadel and Stout, 2012), social acceptability of biotechnology (Rousselière and Rousselière, 2017), household savings (Fukuda, 2006) and obesity epidemic (Reither et al., 2009).

Nested within the age-period-cohort model is the simpler age-cohort model which arises for $\gamma \equiv 0$, meaning that there is no period effect:

$$\log \mu_{ik} = \alpha_i + \beta_k. \quad \text{(age-cohort model)} \tag{2}$$

The age-cohort model in comparison to the age-period-cohort model has two major advantages. Firstly, the parameters are identifiable up to a constant. In contrast, in model (1), a solution $(\alpha, \beta, \gamma)$ can be shifted by an arbitrary linear trend without altering the outcome. This makes interpretation and extrapolation of the parameter estimates difficult. Secondly, forecasting, i.e., estimation for $i+k-1 = j > today$, is possible "in-sample", i.e., without time series extrapolation: Assume that cohorts are observed for $i = 1, \ldots, d_1$, and that age is observed for $k = 1, \ldots, d_2$. Period is given by $i+k$. Once the parameter values $(\alpha_i), (\beta_k)$, $i = 1, \ldots, d_1$, $k = 1, \ldots, d_2$ are fitted, forecasts for the effect $\mu$ for observed cohorts, i.e., $i = 1, \ldots, d_1$, are given up to $d_2$ units ahead via (2). If one further assumes that $d_2$ is an upper bound of age, then complete forecasts are indeed available for all observed cohorts without the need of extrapolation. Clearly, mathematical ease alone can not justify the choice of a model. But in many cases period-effects seem indeed not significant. Hence, often the age-cohort model (2) ensures both a better model fit and mathematical tractability.

The motivating example for the study in this paper is reserving in non-life insurance. Given data of past claims, insurance companies are interested in forecasting the number of future claims for accidents that

have already happened but are not reported yet. This number plays an important role for estimating the reserve: the amount the company sets aside for claims that have to be paid in the following years. The reserve is usually the largest number on the balance sheet of a non-life insurance provider. Estimating the reserve is regulated by law meaning that the mathematical model and the method of estimation have to be approved by regulators. The challenging problem of forecasting the number of so-called IBNR (incurred but not reported) claims is often solved via model (2): For each past claim, one considers the date (cohort $i$) when the accident had happened and the delay (age $k$) there was until the claim was reported to the insurer. Hence, cohort and age satisfy $i + k - 1 \leq today$; given a certain year-wise aggregation. This information is then used to estimate the number of future claims $\mu_{ik}$, $i+k-1 > today$, for accidents in the past, $i \leq today$. Under model (2), the parameters $\alpha_i$ and $\beta_k$ for each cohort $i$ and age $k$ can be estimated from past data. Assuming a maximum delay (usually 7 to 10 years in practice, depending on the business line), the estimates of the parameters can be used to forecast the number of future claims with $i + k - 1 > today$. More details of this age-cohort-reserving example are given in the recent contribution Harnau and Nielsen (2018) and are also included in the highly-cited overview paper of actuarial reserving (England and Verrall, 2002).

Other examples where no significant period effect has been found include among many others cancer studies (Leung et al., 2002; Remontet et al., 2003), returns due to education (Duraisamy, 2002), unemployment numbers (Wilke, 2017), mesothelioma mortality (Peto et al., 1995; Martínez-Miranda et al., 2014).

Given the importance of age-period-cohort models and age-cohort models, it is surprising that continuous versions have not been considered much in the literature. Continuous modeling avoids inefficient pre-smoothing and is in line with recent trends around big data and the drive of modeling and understanding every individual separately. Modeling every individual separately, possibly with additional covariates, results in the estimation of a large number of parameters. An increase of dimension means that data is more sparse so that smoothing methods become necessary. Section 8.3 is devoted to a small simulation study showing how a non-smoothed estimator breaks down when the sample sizes are too small; hence making forecasts unreliable. A series of recent papers introduced several continuous versions of (1), (2) and extensions thereof in what are coined there as in-sample forecasters (Martínez-Miranda et al., 2013; Mammen et al., 2015; Lee et al., 2015; Hiabu et al., 2016; Gámiz et al., 2016; Lee et al., 2017).

This paper is devoted to the continuous analogue of the simple age-cohort model, equation (2):

$$f(x, y) = f_1(x)f_2(y), \qquad x, y \in [0, T], \tag{3}$$

for $T > 0$ and where $f$ is a two-dimensional density function as considered in Martínez-Miranda et al. (2013), Mammen et al. (2015), Hiabu et al. (2016), Gámiz et al. (2016). If $\mu_{ik}$ in (2) denotes occurrence, then (3) arises from (2) by replacing the discrete arguments $(i, k)$ by continuous arguments $(x, y)$. Note that age $x$ and cohort $y$ are values of independent continuous random variables as the effects of cohort $i$ and age $k$ on $\mu_{ik}$ are independent from each other in model (2). Analogue to model (2), $f_1$ represents the effect of age

and $f_2$ that of cohort. Instead of estimating the effects $\alpha_i$ and $\beta_k$ for all $i, k$, we now estimate the marginal distributions $f_1$ and $f_2$ from the data and thus get an estimate for the joint distribution under the assumption of independence. The estimated joint distribution then provides information, without extrapolation, about the future, i.e., density values for $x + y > T$. The estimation problem of model (3) is different to classical statistical literature because observations are not available on the full set $[0, T]^2$, with interest often exactly in the unobserved area, $x + y > T$.

In-sample forecasters generate a unified approach to the class of age-cohort and age-period-cohort models and therefore provide opportunity for a general improvement across disciplines. Generally, consider a distribution on a set $\mathcal{S}$ where data generated from that distribution is only available for observation on a strict subset $\mathcal{S}_1 \subset \mathcal{S}$. Our particular interest is in the density on $\mathcal{S}_2 = \mathcal{S} \setminus \mathcal{S}_1$. An in-sample forecaster is a structured model with the property that the distribution on $\mathcal{S}_2$ is known from the distribution on $\mathcal{S}_1$. In most of the applications we are aware of, $\mathcal{S}_1$ represents the past and $\mathcal{S}_2$ represents the future — hence the term forecasting. One necessary assumption for this methodology to work is that the parameters of the distribution can be estimated from the observations in $\mathcal{S}_1$. For example unspecified nonparametric one-dimensional functions are sufficient to describe the distribution on $\mathcal{S}_2$. More generally, the distribution on $\mathcal{S}_1$ is a function of some components and the distribution on $\mathcal{S}_2$ is another function of the very same components. It is therefore necessary to work inside the world of structured models. Summarizing, the guiding principle of in-sample forecasting is that a forecaster can be constructed from in-sample estimators without further extrapolation. This often seems more intuitive, simpler and more stable than time series forecasting that requires first estimation and then extrapolation. Variations of in-sample forecasters have therefore been developed by practitioners who wish to have a hands-on understanding of all entering components and their relative importance for the forecast. Practitioners often deviate from standard statistical estimation when prior knowledge provide them with extra information. It is of course extremely important that the practitioners understand all entering components to be able to perform such manual corrections in a reliable way. Therefore in-sample forecasting is a powerful methodology in many practical forecasting settings.

Another common two-dimensional application, besides reserving in non-life insurance, appears in medical studies, specifically in the research of the mortality of a disease. Typically, patients enter the study when the disease is diagnosed and they are observed until current calendar time or until some event happens. That event could be death, see for example Martínez-Miranda et al. (2016) forecasting future asbestos related deaths in the UK via a structure as above. Martínez-Miranda et al. (2016) does not coin their methodology in-sample forecasting and they use a discrete non-smooth estimating technique that is common in age-cohort, age-period, period-cohort or age-period-cohort studies. But the structure is the same as the in-sample forecasting methodology considered in this paper and the likelihood based approach of Martínez-Miranda et al. (2016) is referred to as method of sieves in this paper. When this paper explores comparable in-sample forecasting procedures and includes the method of sieves in the optimization considerations, it is

including the vast age-cohort type of studies in the overall comparison. The unsurprising conclusion that the method of sieves, a histogram type estimator, is not efficient leads us to suggest that continuous in-sample forecasting methodology should be introduced more broadly in the vast number of applications in age-period and age-cohort type of studies.

In the above mentioned example about future asbestos related deaths, we have data about past deaths in $\mathcal{S}_1$ and future deaths will happen in $\mathcal{S}_2$. The event under observation is death; future deaths are of course unknown at the day the data collection ends. Only the number of deaths that have already occurred is known. The purpose of the forecasting exercise might be to forecast the number and timing of future deaths in the considered cohort. In this scenario, we have truncated data represented by $(X_i, Y_i)$ where $X_i$ is the date an individual has entered the study and $Y_i$ is time until death. Truncation occurs because $X_i + Y_i$ must be before the day of data collection. The region $\mathcal{S}_2$, where $X + Y$ is after the day of data collection, contains future events only. The typical in-sample forecasting assumes data to be structured in such a way that the distribution of interest depends on one-dimensional components only and that these one-dimensional components can be estimated from the data in $\mathcal{S}_1$.

The aim of this paper is to summarize those methods that solve (3), extend them with multiplicative bias corrected versions and compare them both theoretically and in a simulation study. This should give practitioners and applied researchers guidance when estimation of a continuous age-cohort-model is considered. This study should also be seen as first cornerstone in the understanding of more complex models including continuous analogues of (1) and extensions thereof. We chose to concentrate on the simple model (3) only because optimality is not settled even in this simple continuous age-cohort model. That is the purpose of this study. There are also other interesting generalizations of (3), which are not the continuous analogue to (1). One example would be the model $f(x, y) = f_1(x)f_2(\varphi(x)y)$, modeling an additional operational time term $\varphi$ (Lee et al., 2017). Also such other generalizations need a good fundament of the understanding of the simple age-period model before they can be fully developed.

This paper is organized as follows. We outline the underlying probability model in Section 2. In Sections 3–5 we introduce the different estimators and their multiplicative bias corrected versions are defined in Section 6. Common features of point-wise asymptotic bias and variance are summarized in Section 7. Different problems in finite sample simulation studies and their results are described in Section 8, followed by a conclusion. Asymptotic results for the sieves histogram estimator and their proofs are deferred to the appendix.

## 2. Model

Let $\mathcal{S}$ denote the square $[0, T] \times [0, T]$ for some $T > 0$. We assume a probability space $(\mathcal{S}, \mathfrak{B}(\mathcal{S}), \mathcal{P})$ with the Borel measure $\mathfrak{B}(\mathcal{S})$ on $\mathcal{S}$. Furthermore, let $X$ and $Y$ be two independent random variables with values in $[0, T]$ each such that the distribution of the pair $(X, Y)$ is $\mathcal{P}$. Let the marginal density functions of $X$ and

$Y$ with respect to the Lebesque measure be given by $f_1$ and $f_2$, respectively. Denote the probability density function of the two-dimensional random variable $(X, Y)$ by $f$, which satisfies model (3):

$$f(x, y) = f_1(x)f_2(y), \qquad x, y \in [0, T].$$

In this particular model, the problem of two-dimensional in-sample density forecasting means that we want to estimate $f$ given truncated observations $(X_i, Y_i)$, $i = 1, \ldots, n$, i.e., observations are only available in the subset $\mathcal{S}_1 \subset \mathcal{S}$. Note that these observation have density function $\tilde{f}(x, y) = \mathcal{P}(\mathcal{S}_1)^{-1} f(x, y)\ I_{\mathcal{S}_1}(x, y)$. For simplicity and because of the relevance in application, we set $\mathcal{S}_1 = \{(x, y) \in \mathcal{S}; x+y \leq T\}$, which is the lower left diagonal triangle in $\mathcal{S}$ and which occurs in the examples in the introduction. Hence, it holds $X_i + Y_i \leq T$ for every observation $i$. In the next three sections, we consider three different nonparametric approaches for estimating the marginal densities of the stochastic variables that are truncated to a triangular subset of $\mathcal{S}$. Our estimate for the joint density $f$ will then simply be the product of the estimated marginal densities.

The first approach involves survival analysis methods that make use of models specifically designed for truncated and censored observations as described in Section 3. Two of our estimators arise from these methods. The second approach described in Section 4 aims at estimating the filtered joint density $\tilde{f}$ on $\mathcal{S}_1$ first and then projecting it onto a multiplicatively separable subspace of the space of probability density functions on $\mathcal{S}$. A naive re-scaled histogram approach is the basis for the third approach that is outlined in Section 5. We follow a simple algorithm to get a re-scaled histogram despite the truncation on $\mathcal{S}_1$ and smooth the estimator afterwards using a kernel function.

## 3. Survival analysis approach

We first consider the survival analysis approach, where we consider a counting process model in backwards time. Reversing the time scale is a survival analysis trick to change complicated right-truncation to straightforward left-truncation (Ware and DeMets, 1976). Left-truncation is immediately accommodated for in standard counting process theory allowing for standard martingale inference and other standard stochastic process tools to be immediately available.

### *3.1. Survival analysis model*

In this section, we embed the model of Section 2 into a survival analysis setting. We assume a counting process $\{N(t) : t \in [0, T]\}$, i.e., a piecewise constant, nondecreasing càdlàg process with values $0, 1, 2, \ldots$. The intensity $\lambda$ of $N$ with respect to a suitable filtration $\mathcal{F} = \{\mathcal{F}_t : t \in [0, T]\}$ (Andersen et al., 1993, p.60), is defined via

$$\lambda(t) = \lim_{h \downarrow 0} h^{-1} E[N((t+h)-) - N(t-) |\ \mathcal{F}_{t-}], \qquad t \in [0, T],$$

where $N(t-) = \lim_{s \uparrow t} N(s)$ and $\mathcal{F}_{t-}$ is the smallest $\sigma$-algebra containing all $\mathcal{F}_s$ such that $s < t$.

In the sequel, for each observations $(X_i, Y_i)$ we are interested in counting processes $N_1^i$ and $N_2^i$ which are defined as $N_1^i(t) = I(T - X_i \leq t)$ and $N_2^i(t) = I(T - Y_i \leq t)$, $t \in [0, T]$. Let their intensities be given as $\lambda_l^i$, $l = 1, 2$, and we write $\mathcal{F}_{l,t}$ for their natural filtrations. Each set of counting process $\{N_1^i(t) : t \in [0, T], i = 1, \ldots, n\}$ and $\{N_2^i(t) : t \in [0, T], i = 1, \ldots, n\}$ will be used to estimate the marginal distribution of $X$ and $Y$, respectively.

We illustrate the setting for the variable $X$ in the following. Since the problem is symmetric in its covariates, we will obtain the analogous structure for $Y$. Explicitly, the intensity of $N_1^i$ with respect to its natural filtration is given by

$$\lambda_1^i(t) = \alpha_1^R(t|Y_i) I(Y_i \leq t \leq T - X_i), \qquad t \in [0, T], \tag{4}$$

where $\alpha_1^R(t|Y_i) = \lim_{h \downarrow 0} h^{-1} P\left(T - X_i \in [t, t+h) \middle| \, T - X_i \geq t, Y_i(s), s \leq t\right)$ is the conditional hazard of $T - X_i$ given $Y_i$ at $t \in [0, T]$. The structure in equation (4) fulfills Aalen's multiplicative model (Andersen et al., 1993, p. 128).

We refer to $\alpha^R$ as the hazard in reversed time, indicated by the superscript $R$, since $N_1^i$ is defined for $T - X_i$ instead of for $X_i$. The motivation was already mentioned above: Observations in the triangle $\mathcal{S}_1$ correspond to right-truncation, i.e., for every observation $(X_i, Y_i)$ it holds $X_i \leq T - Y_i$, for which we can't derive an intensity as in equation (4). The process in backward time however is left-truncated and allows this representation of the intensity and hence fits into the framework of Aalen's multiplicative model. For more details on this time-reversion see e.g. Hiabu et al. (2016).

In survival analysis, the number of individuals that are at risk is given by the exposure. The exposure at time $t \in [0, 1]$ is defined as $Z_l(t) = \sum_{i=1}^n Z_l^i(t)$, $l = 1, 2$, with $Z_1^i(t) = I(Y_i < t \leq T - X_i)$ and $Z_2^i(t) = I(X_i < t \leq T - Y_i)$.

### *3.2. Survival analysis estimators*

The first two estimators we investigate are one-dimensional kernel estimators arising from the survival analysis approach. Since the natural objects in survival analysis are hazard rates and not densities, we first introduce an estimator of the marginal hazard functions and then transform it into a probability density functions. The second estimator is a straightforward one-dimensional density estimator that has a slightly more advanced structure.

Again, since the counting process estimators are one-dimensional and because of the symmetry in the estimation problem, we denote most of the following only for estimators $\hat{f}_1$ of the marginal density $f_1$. When there is no risk of confusion, we usually leave out the subscript $l$ and just write $\hat{f}$ or $f$, respectively. Clearly, all results also hold for $\hat{f}_2$ being defined analogously.

We focus on local linear estimators and ignore local constant kernels here since our problem is density estimation on a bounded support. Local linear estimators usually perform much better than local constant

kernel density estimators at boundaries (see Fan and Gijbels (1996) and Wand and Jones (1994)). For a bandwidth $h > 0$ and $s, t \in [0, T]$, we define

$$\bar{K}_{t,h}(t-s) = \frac{a_2(t) - a_1(t)(t-s)}{a_0(t)a_2(t) - (a_1(t))^2} K_h(t-s),$$

with

$$a_j(t) = n^{-1} \int K_h(t-s)(t-s)^j Z(s) ds,$$

for $j = 0, 1, 2$, where $K$ is a symmetric kernel function with bounded support and $K_h(t) = h^{-1}K(t/h)$ for $h > 0$. Integration without boundaries denotes integration over the whole support $[0, T]$.

The function $\bar{K}$ can be interpreted as a local linear kernel and will subsequently naturally arise as a solution of a local linear least square criterion.

*3.2.1. One-dimensional hazard estimator*

The first estimator is a transformation of the hazard estimator introduced in Nielsen and Tanggaard (2001). In that setting there was no right-truncation and hence no time-reversion in the estimation process. We first estimate the marginal hazard function of $Y$ in reversed time by the local linear estimator

$$\hat{\alpha}_h^R(t) = n^{-1} \sum_{i=1}^{n} \int \bar{K}_{t,h}(t-s) dN^i(s).$$

For fixed $t \in [0, T]$, this estimator is motivated to be $\hat{\alpha}_h^R(t) = \hat{\theta}_0(t)$, the first component of the minimizer

$$\begin{pmatrix} \hat{\theta}_0(t) \\ \hat{\theta}_1(t) \end{pmatrix} = \arg\min_{\theta_0, \theta_1} \lim_{\varepsilon \to 0} \sum_{i=1}^{n} \int \left[ \left\{ \frac{1}{\varepsilon} \int_s^{s+\varepsilon} dN^i(s) - (\theta_0 + \theta_1(t-s)) \right\}^2 - \xi_1(\varepsilon) \right] K_h(t-s) Z^i(s) ds,$$

with the term $\xi_1(\varepsilon) = \left( \varepsilon^{-1} \int_s^{s+\varepsilon} dN^i(u) \right)^2$ making the expression well-defined.

The transformation of a marginal hazard function $\alpha$ into its corresponding marginal density $f$ is given by

$$f(t) = \alpha(t) \exp\left( - \int_0^t \alpha(s) ds \right), \qquad t \in [0, T].$$

It reflects the equality $\alpha(t) = f(t)/S(t)$ for the survival function $S(t) = 1 - F(t)$ and with $F$ denoting the cumulative density function. For $l = 1, 2$, this motivates the reversed time estimator

$$\hat{f}_{l,h,H}^R(t) = \hat{\alpha}_{l,h}^R(t) \exp\left( - \int_0^t \hat{\alpha}_{l,h}^R(s) ds \right), \qquad t \in [0, T],$$

and finally our density estimator is

$$\hat{f}_{l,h,H}(t) = \hat{f}_{l,h,H}^R(T-t).$$

We do not expect this estimator to perform well because the hazard-density transformation can amplify errors in the estimator of hazard function. Nevertheless, we wanted to compare its performance because the hazard estimator is computationally attractive due to its simple structure.

*3.2.2. One-dimensional counting process estimator*

The second estimator in our comparison is the local linear survival density estimator, see Hiabu et al. (2016) and Nielsen et al. (2009). For every $t \in [0,T]$, it is defined as $\hat{f}^R_{l,h,C}(t) = \hat{\theta}_0(t)$ in

$$\begin{pmatrix} \hat{\theta}_0(t) \\ \hat{\theta}_1(t) \end{pmatrix} = \underset{\theta_0,\theta_1 \in \mathbb{R}}{\arg\min} \sum_{i=1}^n \Bigg[ \int K_h(t-s)\{\theta_0 + \theta_1(t-s)\}^2 Z^i(s)ds \\ -2 \int K_h(t-s)\{\theta_0 + \theta_1(t-s)\}\hat{S}^R(s)Z^i(s)dN^i(s) \Bigg], \tag{5}$$

for a bandwidth $h > 0$. For fixed $t$, $\hat{\theta}_0$ estimates $f_1(T-t)$, i.e., $f_1$ in reversed time. As a pilot estimator for the reversed survival function $S^R(t) = \prod_{s \le t}\{1 - dA^R(s)\}$, we take the Kaplan-Meier product-limit estimator

$$\hat{S}^R(t) = \prod_{s \le t}(1 - \Delta \hat{A}^R(s)) = \prod_{s \le t}\left(1 - \frac{\Delta N(s)}{Z(s)}\right),$$

with the Aalen estimator

$$\hat{A}^R(t) = \sum_{i=1}^n \int_0^t (Z(s))^{-1} dN^i(s)$$

of the integrated hazard function $A^R(t) = \int_0^t \alpha^R(s)ds$. We use the common product-integral notation (Andersen et al., 1993, p.89) for the reversed survival function and in the Kaplan-Meier estimator $\Delta \hat{A}^R(s)$ denotes $A^R(s) - \lim_{u \nearrow s} A^R(u)$, resulting in the product-integral being a product over the finite number of jumps of $\hat{A}^R$ or $N$, respectively.

The minimization criterion (5) can also be motivated as least squares principle via the representation

$$\sum_{i=1}^n \int \left[ \left\{ \frac{1}{\varepsilon} \int_s^{s+\varepsilon} \hat{S}^R(u)dN^i(u) - (\theta_0 + \theta_1(t-s)) \right\}^2 - \xi_2(\varepsilon) \right] K_h(t-s)Z^i(s)ds,$$

in the limit for $\varepsilon \to 0$. The term $\xi_2(\varepsilon) = \left(\varepsilon^{-1} \int_s^{s+\varepsilon} \hat{S}^R(u)dN^i(u)\right)^2$ does not depend on $(\theta_0(t), \theta_1(t))$ and it is needed to make the expression well-defined as in Section 3.2.1.

Solving the minimization (5), the reversed time estimator for $f_1$ at $t \in [0,T]$ with bandwidth $h > 0$ from Hiabu et al. (2016) is given as

$$\hat{f}^R_{l,h,C}(t) = n^{-1} \sum_{i=1}^n \int \bar{K}_{t,h}(t-s)\hat{S}^R_l(s)dN^i_l(s),$$

for $l = 1, 2$ and, finally, we set

$$\hat{f}_{l,h,C}(t) = \hat{f}^R_{l,h,C}(T-t).$$

## 4. Projection approach

The next method we include in this study is a two-dimensional projection approach introduced in Martínez-Miranda et al. (2013) and Mammen et al. (2015). One first estimates the two-dimensional density

on the subspace $\mathcal{S}_1$. This unstructured pilot estimator is then projected onto the space of multiplicatively separable probability density functions on $\mathcal{S}$. Here, as in the survival analysis approach, we propose the local linear estimator.

First, we estimate the joint density on $\mathcal{S}_1$ for every points $z_0 = (x_0, y_0) \in \mathcal{S}_1$ with the estimator $\tilde{f}_h(z_0) = \hat{\Theta}_0(z_0)$ that arises from the local linear minimization:

$$(\hat{\Theta}_0(z_0), \hat{\Theta}_1(z_0))' = \underset{(\Theta_0,\Theta_1)}{\arg\min} \left\{ \lim_{b\to 0} \int_{\mathcal{S}_1} [\tilde{f}_b^{(0)}(z) - \Theta_0 - \Theta_1^t(z_0 - z)]^2 \mathcal{K}_h(z - z_0) dz \right\},$$

where $\tilde{f}_b^{(0)}(z) = (nb_1b_2)^{-1} \sum_{i=1}^n \mathcal{K}_b(z - (X_i, Y_i))$ is a pilot estimator for the two-dimensional density on $\mathcal{S}_1$. Here the bandwidth $b = (b_1, b_2)$ and the kernel $\mathcal{K}_b$ are two-dimensional and we write $z = (x, y)$. For simplicity we take a multiplicative kernel $\mathcal{K}_b(z) = K_{b_1}(x)K_{b_2}(y)$.

Afterwards, we define the projection estimators $\hat{f}_{1,h_1,P}, \hat{f}_{2,h_2,P}$ as the functions minimizing the estimated weighted integrated squared error

$$(\hat{f}_{1,h_1,P}, \hat{f}_{2,h_2,P}) = \underset{(\varphi_1,\varphi_2)}{\arg\min} \left\{ \int_{\mathcal{S}_1} [\tilde{f}_h(x, y) - \varphi_1(x)\varphi_2(y)]^2 w(x, y) d(x, y) \right\},$$

to get estimates for the marginal densities.

We choose the weighting $w(x, y) = \tilde{f}_h^{(0)}(x, y)^{-1}$ and we calculate the solution of the second minimization problem via the following iterative algorithm (see also Martínez-Miranda et al. (2013)):

1. Start with an initial estimator of $f_1$ denoted by $\hat{f}_1^{(0)}$ and let $\hat{f}^{(0)}$ be the unstructured minimizer of the first step.

2. Estimate $f_2$ at $y$ by
$$\hat{f}_2^{(1)}(y) = \int_{\mathcal{S}_{1y}} \hat{f}^{(0)}(x, y)dx \Big/ \int_{\mathcal{S}_{1y}} \hat{f}_1^{(0)}(x)dx,$$
for $\mathcal{S}_{1y} = \{x|(x, y) \in \mathcal{S}_1\}$.

3. Update the estimator for $f_1$ by
$$\hat{f}_1^{(1)}(x) = \int_{\mathcal{S}_{1x}} \hat{f}^{(0)}(x, y)dy \Big/ \int_{\mathcal{S}_{1x}} \hat{f}_2^{(1)}(y)dy,$$
where $\mathcal{S}_{1x} = \{y|(x, y) \in \mathcal{S}_1\}$, using $\hat{f}_2^{(1)}$.

4. Repeat steps 2 and 3 until a certain convergence criterion is achieved.

Under more sophisticated definitions, see Mammen et al. (2001), the estimators $\hat{f}_1^{(1)}$, $\hat{f}_2^{(1)}$ can also be motivated as direct projection of the Dirac delta estimators into the set of multiplicatively separable functions.

## 5. Smoothed structured histogram approach (sieves estimator)

Martínez-Miranda et al. (2013) also proposed smoothing a structured histogram with a kernel function as another approach. The discrete histogram estimator is constructed from column-wise proportions in a table of aggregated data. The histogram estimators are known to actuaries as forward factors or development factors and are calculated in every non-life insurance company as part of the omnipresent chain ladder method. The chain ladder method is a simple algorithm that is widely used to solve the reserving problem mentioned in Section 1 and can be used in our model from Section 2 with aggregated data. Hence, our motivation for the fourth estimator is to enhance a method that is well-known to practitioners in the insurance industry by kernel smoothing. The estimator can certainly be applied to every problem satisfying the model in Section 2 as e.g. in the medical study example from the introduction.

Before specifying how our estimator is defined, we provide a few words about the way data is aggregated in the setting where the chain ladder method is usually applied. Instead of observing and aggregating $(X_i, Y_i)$, $i = 1, \ldots, n$, we only observe $(X_i, X_i + Y_i)$ after aggregation. Naive aggregation of the observations $(X_i, Y_i)$ would result in the square $\mathcal{S}$ being split into equidistant rectangular bins. The entries in the diagonal which includes the date where data collection ended would then overlap with the unobserved area (the future) making forecasting more tricky. Hence, as is done in practice and outlined in Appendix A.1, we divide $\mathcal{S}$ into parallelograms (and triangles for the first age column).

For a definition of forward factors and a concrete algorithm on how to get the histogram, we follow England and Verrall (2002). Let $\delta > 0$ be a bin width such that $m_\delta = T\delta^{-1}$ is an integer. We assume our parallelogram grid consists of $m_\delta$ bins with edge length $\delta$ and we count the numbers of observations in this grid in an $(m \times m)$-matrix $C$. Let $C_{ij}$ denote the number of events for which $Y$ is in bin $i$ and $X$ is in bin $j$. Then the cumulative numbers of events with respect to $X$ are given by $D_{ij} = \sum_{k=1}^{j} C_{ik}$. Now the forward factors $\{\lambda_j : j = 1, \ldots, m-1\}$ are defined as

$$\hat{\lambda}_j = \frac{\sum_{i=1}^{m-j+1} D_{ij}}{\sum_{i=1}^{m-j+1} D_{i,j-1}} = \frac{\sum_{i=1}^{m-j+1} \sum_{k=1}^{j} C_{ik}}{\sum_{i=1}^{m-j+1} \sum_{k=1}^{j-1} C_{ik}},$$

and they give an averaged proportion by how much the number of observations increased from one bin to the next. We can use the forward factors to construct a histogram $\{\hat{p}_1^\delta(j) : j = 1, \ldots, m\}$ with bin width $\delta$ for the distribution of $X$ via

$$\hat{p}_1^\delta(1) = \frac{1}{\prod_{k=1}^{m-1} \hat{\lambda}_k}, \qquad \hat{p}_1^\delta(j) = \frac{\hat{\lambda}_{j-1} - 1}{\prod_{k=j-1}^{m-1} \hat{\lambda}_k}, \qquad j = 2, \ldots, m.$$

Hence, $\hat{p}_1^\delta(j)$ is an estimator for the probability of an event being in bin $j$ (with respect to $X$) giving the proportion of observations in the $j$th parallelogram column, with a naively estimated correction for truncation. See Appendix A.1 for a more technical description of $\hat{\lambda}_k$ and a more detailed derivation of $\hat{p}_1^\delta$.

The first formal model definition of forward factors is given in Kuang et al. (2009). The histogram estimator $\hat{p}_2^\delta$ for $Y$ is obtained analogously. Now, on the same discrete grid, the one-dimensional sieves estimator for $f_l$ at $t \in [0, T]$ is defined as

$$\hat{f}_{l,h,S}^\delta(t) = \sum_{s=1}^{m_\delta} K_h(t - s\delta)\hat{p}_l^\delta(s\delta),$$

Thus, $\hat{p}_l^\delta(s)$ is a normalized histogram and we "smooth over it".

Note that we use the kernel $K$ to get a local constant estimator and not the linear kernel $\bar{K}_{t,h}$ as in the cases before to maintain the histogram nature of this estimator. Moreover, this facilitates the notation in the proof of the asymptotic that is given in Proposition 1 in Appendix A.

This approach describes a sieves estimator since we get less aggregated histograms as pre-estimators with decreasing bandwidth $\delta$ but the choice $\delta = 0$ is not possible. See Martínez-Miranda et al. (2013) for a review of other sieves methods for two-dimensional multiplicative in-sample forecasting and Gámiz et al. (2016) for a pre-binned local linear hazard estimator.

### *5.1. Smoothed histogram as counting process estimator*

At the grid points $t \in \{kb : k = 1, \ldots, m_b\}$, the sieves estimator $\hat{f}_{l,h,S}^\delta(t)$ equals the continuous one-dimensional survival analysis estimator defined via

$$\hat{f}_{l,h,S^*}(T - t) = \sum_{i=1}^{n} \int K_h(t - s) \left(Z_l(s)\right)^{-1} \hat{S}_l^R(s) dN_l^i(s).$$

Hence, $\hat{f}_{l,h,S^*}$ can be interpreted as a continuous generalization of $\hat{f}_{l,h,S}^\delta$. In the simulation study in Section 8, we use the same bandwidth $\delta$ for the histogram $\hat{p}_l^\delta$ and the discretized grid on which we evaluate continuous functions. Therefore, $\hat{f}_{l,h,S}^\delta(t)$ and $\hat{f}_{l,h,S^*}(T - t)$ coincide in the results at every evaluated point $t$.

The motivation for this identification is the connection with the forward factor-based histogram and the Kaplan-Meier estimator. As shown in Appendix B.1, $\hat{f}_{l,h,S}^\delta$ can be identified with the reversed time counting process estimator

$$\hat{f}_{l,h,S}^{\delta,R}(T - t) = \sum_{i=1}^{n} \int K_h(t - s) \left(Z_l^\delta(s)\right)^{-1} \hat{S}_l^\delta(s) dN_l^{i,\delta}(s),$$

for the pre-binned counting processes $N^{i,\delta}$, the corresponding exposure $Z^\delta$ and survival estimator $\hat{S}^\delta$ that are defined in Appendix A.1. The counting process estimator $\hat{f}_{l,h,S^*}$ has the same asymptotic behavior as $\hat{f}_{l,h,S}$ for $\delta \to 0$ fast enough. For completeness it is given in Proposition 3 in the appendix.

See also Hiabu (2017) for a detailed clarification of the relationship between the backward time survival analysis in Section 3 and actuarial forward factors.

## 6. Multiplicatively bias corrected estimators

The bias correction we are using was introduced for nonparametric regression in Linton and Nielsen (1994) and it was applied to density estimation in Jones et al. (1995). The aim is to increase the asymptotic order of the bias term by "dividing the bias out". One advantage of this particular bias correction is the simple implementation. As expected, the bias corrected estimators only perform better for large sample sizes and in practice their bias is often worse than that of the unmodified estimator for very small finite sample sizes because of constants in the asymptotic bias being larger.

Multiplicative bias correction is motivated by the identity $f(t) = \hat{f}(t)g(t)$ for $g(t) = f(t)/\hat{f}(t)$, i.e., the inverse of the multiplicative bias of $\hat{f}(t)$. Thus, by multiplying $\hat{f}$ by an estimator $\hat{g}$ of $g$, we end up with an estimator $\hat{f}^{BC}(t) = \hat{f}(t)\hat{g}(t)$ of $f(t)$ for each $t \in [0, T]$.

Note that all estimators $\hat{f}$ and $\hat{g}$ are kernel estimators with bandwidth $h_f$ and $h_g$, respectively, and we take the same bandwidths $h_f = h_g$ for both of them. As explained in Jones et al. (1995), the bandwidths have to be of the same asymptotic order for the bias cancellation to work and thus $h_f = ch_g$, $c > 0$. Hence, the choice of just one bandwidth throughout, i.e., $c = 1$, is the obvious one.

The bias corrected estimators have the following representations. We also give minimization criteria that motivate some of them. Again, we illustrate the estimators for $g$ just for the covariate $X$ and suppress the indices $l$.

### *6.1. One-dimensional hazard estimator*

The multiplicative bias corrected hazard estimator was presented in Nielsen and Tanggaard (2001). Analogously to the density case, we motivate the corrected estimator via $\alpha(t) = \hat{\alpha}(t)g_H(t)$ and, for a bandwidth $h > 0$, we estimate $g_H(t) = \alpha(t)/\hat{\alpha}(t)$ in reversed time via

$$\hat{g}^R_{h,H}(t) = n^{-1} \sum_{i=1}^{n} \int \bar{K}_{t,h}(t-s)\{\hat{\alpha}^R_h(s)\}^{-1} dN^i(s),$$

which minimizes

$$\begin{pmatrix} \hat{\theta}_0(t) \\ \hat{\theta}_1(t) \end{pmatrix} = \underset{\theta_0,\theta_1 \in \mathbb{R}}{\arg\min} \sum_{i=1}^{n} \int \left[ \left\{ \frac{1}{\varepsilon} \int_s^{s+\varepsilon} dN^i(u) - (\theta_0 + \theta_1(t-s))\alpha^R_h(s) \right\}^2 - \xi_1^{BC}(\varepsilon) \right] K_h(t-s) Z^i(s) ds,$$

in the first component $\hat{g}^R_{h,H}(t) = \hat{\theta}_0(t)$ for fixed $t$. The term $\xi_1^{BC}(\varepsilon) = (\alpha^R_h(s))^2 \xi_1(\varepsilon)$, where $\xi_1(\varepsilon)$ was defined in Section 3.2.1, makes the expression well-defined. The hazard in backward time is then estimated via

$$\hat{\alpha}^{R,BC}_h(t) = \hat{\alpha}^R_h(t)\hat{g}^R_{h,H}(t),$$

and afterwards transformed into a density in forward time,

$$\hat{f}_{h,H}^{BC}(t) = \hat{\alpha}_h^{R,BC}(T-t)\exp\left(-\int_0^t \hat{\alpha}_h^{R,BC}(T-s)ds\right), \qquad t \in [0,T].$$

### 6.2. One-dimensional counting process estimator

The multiplicative bias corrected version of the counting process estimator was first introduced in Nielsen et al. (2009). Here $\hat{g}_h$ is defined as the first component $\hat{g}_h(t) = \hat{\theta}_0(t)$ of the minimizer

$$\begin{pmatrix}\hat{\theta}_0(t)\\ \hat{\theta}_1(t)\end{pmatrix} = \underset{\theta_0,\theta_1\in\mathbb{R}}{\arg\min}\sum_{i=1}^n \Bigg[\int K_h(t-s)\{(\theta_0+\theta_1(t-s))\hat{f}_{h,C}^R(s)\}^2 Z^i(s)ds$$
$$-2\int K_h(t-s)\{(\theta_0+\theta_1(t-s))\hat{f}_{h,C}^R(s)\}\hat{S}^R(s)Z^i(s)dN^i(s)\Bigg],$$

or equivalently the first component of the minimizer of

$$\sum_{i=1}^n \int \left[\left\{\frac{1}{\varepsilon}\int_s^{s+\varepsilon}\hat{S}^R(u)dN^i(u) - (\theta_0+\theta_1(t-s))\hat{f}_{h,C}^R(s)\right\}^2 - \xi_2^{BC}(\varepsilon)\right] K_h(t-s)Z^i(s)ds,$$

with respect to $(\theta_0,\theta_1)$ in the limit for $\varepsilon \to 0$, respectively, which results in

$$\hat{g}_{h,C}^R(t) = n^{-1}\sum_{i=1}^n \int \bar{K}_{t,h}(t-s)\hat{S}^R(s)\{\hat{f}_{h,C}^R(s)\}^{-1}dN^i(s), \qquad t\in[0,T].$$

The term $\xi_2^{BC}(\varepsilon) = (\hat{f}_{h,C}^R(s))^2\xi_2(\varepsilon)$, with $\xi_2(\varepsilon)$ from Section 3.2.2, is independent of $\theta_0,\theta_1$ and we need it to make the integral well-defined. The Kaplan-Meier estimator $\hat{S}^R$ was also introduced in Section 3.2.2. For the bias corrected estimator we then set

$$\hat{f}_{h,C}^{BC}(t) = \hat{f}_{h,C}^R(T-t)\hat{g}_{h,C}^R(T-t).$$

### 6.3. Projection estimator

The bias correction of the projection estimators $\hat{f}_{l,h,P}$, $l=1,2$, is also described in Martínez-Miranda et al. (2013). For a bandwidth $h>0$, we get a projection estimator $\hat{g}_{h,P}$ of $g$ from the first component $\hat{\Theta}_0(z_0)$ of the minimizer

$$(\hat{\Theta}_0(z_0),\hat{\Theta}_1(z_0)) = \underset{(\Theta_0,\Theta_1)}{\arg\min}\left\{\lim_{b\to 0}\int_{\mathcal{S}_1}[\tilde{f}_b(z) - (\Theta_0 - \Theta_1^t(z_0-z))\hat{f}_{h,P}(z)]^2 K_h(z-z_0)dz\right\},$$

and we set

$$\hat{f}_{h,P}^{BC}(t) = \hat{f}_{h,P}^{BC}(t)\hat{g}_{h,P}(t), \qquad t\in[0,T]. \tag{6}$$

The multiplicatively bias corrected estimators can be obtained from (6) using a similar approach as in Section 4, see Martínez-Miranda et al. (2013).

*6.4. Sieves estimator*

The adaption for the smoothed histogram estimator is done analogously to the other one-dimensional approaches. We estimate $g$ by

$$\hat{g}^{\delta}_{h,S}(t) = n^{-1} \sum_{s=1}^{m_b} K_h(t - sb)\hat{p}(s)\{\hat{f}^{\delta}_{h,S}(s)\}^{-1},$$

for bandwidth $h > 0$ and bin size $\delta$, and then set

$$\hat{f}^{\delta,BC}_{h,S}(t) = \hat{f}_{h,S}(t)\hat{g}^{\delta}_{h,S}(t).$$

The point-wise asymptotic behavior of $\hat{f}^{\delta,BC}_{h,S}(t)$ is given in Proposition 2 in Appendix A. The analogous counting process adaption of $\hat{g}^{\delta}_{h,S}$ in backward time would now be

$$\hat{g}^{R}_{h,S^*}(t) = n^{-1} \sum_{i=1}^{n} \int K_h(t - s)\hat{S}^R(s)(Z(s))^{-1}\{\hat{f}^{R}_{l,h,S^*}(s)\}^{-1} dN^i(s),$$

and again the resulting estimator $\hat{f}^{BC}_{h,S^*}(t) = \hat{f}^{R}_{l,h,S^*}(T - t)\hat{g}^{R}_{h,S^*}(T - t)$ coincides with $\hat{f}^{\delta,BC}_{h,S}$ on the grid we have used for our computations.

## 7. Theoretical comparison

Point-wise asymptotic normality with a bias term of order $O(h^2)$ is known for all non-bias corrected estimators. The multiplicative bias corrected versions of all estimators have a bias of order $O(h^4)$ in their point-wise asymptotics. The asymptotic orders and the main terms of the point-wise asymptotic bias and variance of every estimator in this paper are given in Table 1 for $T = 1$. The results hold under usual regularity conditions (see references).

We use the following notations in Table 1: $\kappa_2 = \int_{-1}^{1} s^2 K(s)ds$; $\kappa_2^{\delta} = \int_{-1}^{1} s^2 K(s)d\mu^{\delta}(s)$; $U(t) = \{nh\gamma(t)\}^{-1} f(t)F(t)$; $U^{\delta}(t) = \{nh\gamma^{\delta}(t)\}^{-1} f^{\delta}(t)F^{\delta}(t) = U(t)+o(1)$; $U_{\alpha}(t) = \{nh\gamma(t)\}^{-1}\alpha(t)$; $v_1 = \int K^2(s)ds$; $v_1^{\delta} = \int K^2(s)d\mu^{\delta}(s) = v_1 + o(1)$; $v_2 = \int \Gamma_K^2(s)ds$, with $\Gamma_K(s) = 2K - (K * K)(s)$; $v_2^{\delta} = \int \Gamma_K^2(s)d\mu(s) = v_2 + o(1)$; $v_3(t) = (\int_0^{1-t} w(t,s)f(t,s)ds)^{-2} \int_0^{1-t} w^2(t,s)f(t,s)ds$ for the weighting $w = (nh^2)^{-1} \sum_{i=1}^{n} \mathcal{K}_b(z - (X_i, Y_i))$, $b = (h, h)$. The operator $*$ denotes convolution.

## 8. Simulation study

The focus of this paper is on the finite sample performance of the estimators we introduced in the previous sections to show how useful they are in practice, especially to weed out unstable methods that are at risk of breaking down completely in challenging problems. The idea is to discover the best estimator among our selection of density estimators whose bias terms are of the same asymptotic order and to find a rule of thumb for the number of observations that are necessary for the multiplicative bias correction to improve

| Estimator | Bias order | leading terms | Variance order | asympt. value | Annotations |
|---|---|---|---|---|---|
| $\hat{f}_{h,C}$ | $h^2$ | $(1/2)h^2\kappa_2 f''(t)$ | $(nh)^{-1}$ | $v_1 U(t)$ | 1 |
| $\hat{f}_{h,P}$ | $h^2$ | – | $(nh)^{-1}$ | $v_3(t)$ | 2 |
| $\hat{f}_{h,H}$ | $h^2$ | $(1/2)h^2\kappa_2\alpha''(t)$ | $(nh)^{-1}$ | $v_1 U_\alpha(t)$ | 3 |
| $\hat{f}^{\delta}_{h,S}$ | $h^2$ | $(1/2)h^2\kappa_2^{\delta} f_l''(t) + o(\delta^2)$ | $(nh)^{-1}$ | $v_1^{\delta} U^{\delta}(t)$ | 4 |
| $\hat{f}^{BC}_{h,C}$ | $h^4$ | $(1/4)h^4\kappa_2^2 f(t)(f''/f)''(t)$ | $(nh)^{-1}$ | $v_2 U(t)$ | 1 |
| $\hat{f}^{BC}_{h,P}$ | $h^4$ | – | $(nh)^{-1}$ | – | 5 |
| $\hat{f}^{BC}_{h,H}$ | $h^4$ | $(1/4)h^4\kappa_2^2\alpha(t)(\alpha''/\alpha)''(t)$ | $(nh)^{-1}$ | $v_2 U_\alpha(t)$ | 3 |
| $\hat{f}^{\delta,BC}_{h,S}$ | $h^4$ | $(1/4)h^4(\kappa_2^{\delta})^2 f^{\delta}(t)((f^{\delta})''/f^{\delta})''(t) + o(\delta^2)$ | $(nh)^{-1}$ | $v_2^{\delta} U^{\delta}(t)$ | 6 |

Table 1: Main terms of point-wise asymptotic bias and variance terms under regularity assumptions. All bias and variance terms contain an additional error of lower order $o(h^2)$ or $o(h^4)$, respectively. The symbol "–" denotes that there is no closed form solution.

Annotations: [1]see Nielsen et al. (2009); [2]no closed form solution for bias, see Mammen et al. (2015); [3]asymptotic theory for the hazard estimator $\hat{\alpha}$ and not for the resulting $\hat{f}_{h,H}$, see Nielsen and Tanggaard (2001); [4]Proposition 1, see notation in Appendix A.1 , $\delta$ is width of histogram bins; [5]no closed form solution, see Mammen et al. (2015); [6]Proposition 2, see notation in Appendix A.1 , $\delta$ is width of histogram bins.

bias. As pointed out in Section 7, the bias corrected estimators have a leading asymptotic bias term of order $O(h^4)$ instead of $O(h^2)$, however, a higher order of convergence usually leads to larger bias for small finite samples and the estimators behave differently despite their common order of convergence because of different constants in the bias. Clearly, different performance can be due to pure noise as well and hence we run 1000 simulations for each estimation.

In four different settings on the unit square, i.e., for $T = 1$ we compare the best-case performance of all eight density estimators with respect to the integrated squared error

$$\mathrm{ISE}(\hat{f}_l, f_l) = \int_0^1 [\hat{f}_l(t) - f_l(t)]^2 dt,$$

for $l = 1, 2$. "Best-case" means that we choose the best possible bandwidth with respect to the ISE which can be calculated exactly since the true distribution is given. Hence, we avoid the problem of choosing a bandwidth given data.

For some of the estimators bandwidth selection has already been investigated. Martínez-Miranda et al. (2013) apply the projection approach on real data and determine an optimal bandwidth for $\hat{f}_{h,P}$ and $\hat{f}^{BC}_{h,P}$ by cross-validation. The asymptotic behavior of bandwidths from cross-validation (see Rudemo (1982), Bowman (1984) and Hall (1983)) and double one-sided cross-validation (DO-validation, see Hart and Yi (1998), Martínez-Miranda et al. (2009)) for the counting process estimator $\hat{f}_{h,C}$ is given in Hiabu et al. (2016) where the estimator is applied on data with a data-driven bandwidth. Cross-validation and DO-validation for the hazard estimator $\hat{\alpha}_{h,R}$ that is used for the computation of $\hat{f}_{h,H}$ and full asymptotics of the resulting bandwidths are given in Gámiz et al. (2016). However, data-driven bandwidth selection for the structured histogram approach and for the bias corrected versions of the counting process density and hazard estimators

are not covered yet in literature. In particular, a comparison of bandwidth selection results for the whole range of estimators in this study has not been done yet. Being beyond the scope of this paper, it will be part of future work.

The settings we chose are motivated by practical relevance in the application of actuarial reserving and by challenging distributions that point out weaknesses of the estimators.

For $f_1$ we take mixtures of truncated normal distributions. A mixture of $\mathcal{N}(0.2, 0.1)$, $\mathcal{N}(0.5, 3)$ and $\mathcal{N}(0.7, 0.2)$ with equal weights truncated to $[0, 1]$ is motivated by the empirical distributions of real data sets and referred to as the "truncated mixed normal" distribution in the following. To make the estimation at the boundary more challenging, we have chosen a mixture of $\mathcal{N}(0.2, 0.1)$, $\mathcal{N}(0.5, 3)$ and $\mathcal{N}(1, 0.05)$ with equal weights truncated to $[0, 1]$ as a variation and call it the "boundary challenge" distribution. Note that we try distributions for $X$ with mass at the boundaries to investigate weaknesses because some estimators tend to values close to 0 at the edges. The issue of problems at the boundaries is well-known and local linear kernel density estimators are known to perform better than local constant estimators, see e.g. Jones (1993).

We investigate the following distributions for $Y$ as $f_2$. A beta distribution with parameters $\alpha = 1$ and $\beta = 4$ is taken as an empirically motivated "decreasing beta" distribution and we take a mixture of beta distributions with parameters $(2, 5)$, $(3, 10)$ and $(9, 4)$ and equal weights for a more complex example in which there are less observations with values of $Y$ close to 0 and 1, respectively which results in less observations in both corners of the triangle $\mathcal{S}_1$.

With $f_2$ decreasing towards 0 at the boundaries of the interval $[0, 1]$ in every investigated scenario, the results in Table 3 reflect aforementioned problems of the estimation at boundaries: The ISE for $f_1$ (which always satisfies $f_1(0), f_1(1) > 0$) is much larger than that in $f_2$ (with $f_2(0) = f_2(1) = 0$) in every single case.

The shapes of the probability density functions are given in Figure 1. We take all combinations of these distributions and label the four scenarios in Table 2. For each scenario 1000 random samples of sizes 100, 1000 and 10000 were generated.

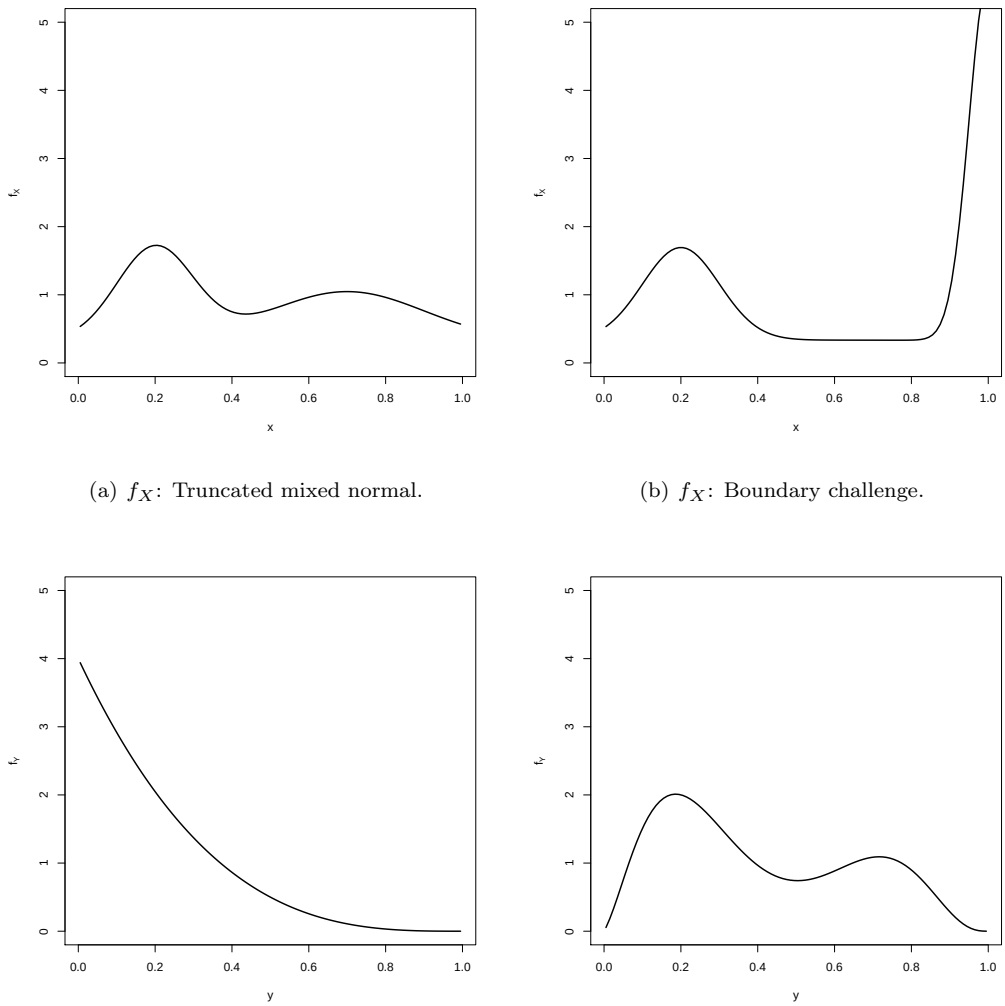
All simulated observations $(X_i, Y_i)$ satisfy $X_i + Y_i \leq 1$. Recalling the model in Section 2, this right-truncation defines our observed subset $\mathcal{S}_1 = \{(x, y) \in \mathcal{S}; x + y \leq 1\}$ of the full support $\mathcal{S} = [0, 1]^2$ of $(X, Y)$. There are no observations in the complement $\mathcal{S}_2 = [0, 1]^2 \setminus \mathcal{S}_1$.

The interval $[0, 1]$ is discretized as a grid with 100 points and we take the corresponding approximation of the ISE. As mentioned in Section 5, we take the same bandwidth $\delta = 1/100$ for the sieves estimators $\hat{f}^{\delta}_{h,S}$ and $\hat{f}^{\delta,BC}_{h,S}$.

We analyze two problems: First, we want to estimate the distributions of $X$ and $Y$ and we measure the results by the ISE in each component. The other problem is to estimate the mass $r = \int_{\mathcal{S}_2} f(x, y) d(x, y)$ of the distribution of $(X, Y)$ in $\mathcal{S}_2$. We use the bandwidth minimizing the ISE for both problems since a bandwidth minimizing $r$ results in massively over-smoothed density estimators because it picks the best bandwidth for

| Scenario | Variable | Distribution |
|---|---|---|
| 1 | $X$ | truncated mixed normal |
| | $Y$ | decreasing beta |
| 2 | $X$ | boundary challenge |
| | $Y$ | decreasing beta |
| 3 | $X$ | truncated mixed normal |
| | $Y$ | mixture of betas |
| 4 | $X$ | boundary challenge |
| | $Y$ | mixture of betas |

Table 2: Scenarios in the simulation study.




(a) $f_X$: Truncated mixed normal.

(b) $f_X$: Boundary challenge.

(c) $f_Y$: Decreasing beta.

(d) $f_Y$: Mixture of betas.

Figure 1: The probability distribution functions of the distributions used for the simulation study.

the bottom left corner in the triangle $\mathcal{S}_1$ — the area with the biggest impact on the estimate of $r$. To find the best bandwidth we compute the ISE for all bandwidths $h \in \{k/100 : k = 1, 2, \ldots, 50\}$ and take the minimizer. Each kernel estimator is computed using the Epanechnikov kernel $K(t) = 0.75(1-t^2)I(-1 \leq t \leq 1)$, $t \in [0, 1]$.

For the projection estimators $f_{1,h,P}$, $f_{2,h,P}$, the algorithm stops after $k$ iterations if the criterion

$$\frac{1}{m}\sum_{j=1}^{m}\frac{|f_1^{(k)}(s_j) - f_1^{(k-1)}(s_j)|}{f_1^{(k-1)}(s_j)} < 0.001,$$

for the discretization $s_1, \ldots, s_m$ of $[0, 1]$ is fulfilled or if we have reached the defined maximum number of $k = 20$ iterations. This method needs up to five times as long to be computed per bandwidth compared to the other estimators. The crucial part, however, is not the recursive algorithm, but the fact that being two-dimensional, it uses two-dimensional bandwidths. For $n_{h_1}$ and $n_{h_2}$ being the number of bandwidths that are compared to estimate $f_1$ and $f_2$, respectively, the computation time of the two-dimensional estimator is hence of computational order $O(n_{h_1} n_{h_2})$ whereas the one-dimensional estimators have computation times of order $O(n_{h_1} + n_{h_2})$. Thus, the total computation time over all bandwidths is of much higher order than that of the one-dimensional estimators.

### *8.1. Density estimation*

The results of the comparison of the ISE are given in Table 3. In the sequel, the median and mean values of the ISE are always taken over 1000 simulation runs.

Our first observation is that the multiplicative bias correction works in practice. As indicated by the theoretical results, for a big enough number of observations the bias corrections result in a smaller bias than that of non-bias corrected estimators. Considering the ISE of both $f_1$ and $f_2$, there are only 4 out of 32 cases where the median bias of the non-corrected version of an estimator is better than its bias correction for a sample size of 10000 and those cases only occur in the challenging Scenarios 3 and 4. Three out of these four cases occur with the sieves estimator $\hat{f}^{\delta}_{h,S}$ (with the ISE increased by less than 14% if we use bias correction) and one with the hazard estimator $\hat{f}_{h,H}$ (where the bias correction is less than 5% worse with respect to the ISE). Also measured by the empirical mean integrated squared error, there are only 4 out of 32 cases where the bias correction does not work for sample size 10000. Here the crucial methods are the projection approach $\hat{f}_{h,P}$ in Scenario 2 and the $\hat{f}^{\delta}_{h,S}$ in Scenarios 3 and 4. Besides, this indicates that challenging estimation problems increase the number of observations that are needed for asymptotic bias improvements to show. For 25 out of 32 cases the bias correction is already better for 1000 observations in the median and in the empirical mean. Here, the observation can be made in all four approaches and in all settings except for Scenario 1.

| Scenario | Estimator | $n = 100$ | | | | $n = 1000$ | | | | $n = 10000$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ISE$(\cdot, f_1)$ | | ISE$(\cdot, f_2)$ | | ISE$(\cdot, f_1)$ | | ISE$(\cdot, f_2)$ | | ISE$(\cdot, f_1)$ | | ISE$(\cdot, f_2)$ | |
| | | Median | Mean (s.d.) | Median | Mean (s.d.) | Median | Mean (s.d.) | Median | Mean (s.d.) | Median | Mean (s.d.) | Median | Mean (s.d.) |
| 1 | $\hat{f}_{h,C}$ | 5.63 | 6.50 (3.73) | 0.98 | 1.92 (2.49) | 1.04 | 1.18 (0.65) | 0.28 | 0.40 (0.39) | 0.18 | 0.20 (0.10) | 0.07 | 0.09 (0.06) |
| | $\hat{f}^{BC}_{h,C}$ | 5.93 | 7.10 (4.49) | 1.91 | 3.09 (3.47) | 1.03 | 1.17 (0.63) | 0.17 | 0.28 (0.31) | 0.15 | 0.17 (0.09) | 0.02 | 0.04 (0.04) |
| | $\hat{f}_{h,P}$ | 4.80 | 5.47 (3.26) | 0.89 | 2.09 (2.88) | 0.94 | 1.12 (0.71) | 0.22 | 0.38 (0.44) | 0.17 | 0.20 (0.11) | 0.06 | 0.08 (0.07) |
| | $\hat{f}^{BC}_{h,P}$ | 4.69 | 5.47 (3.47) | 1.23 | 2.11 (2.55) | 0.77 | 0.93 (0.62) | 0.10 | 0.17 (0.21) | 0.12 | 0.15 (0.09) | 0.02 | 0.04 (0.04) |
| | $\hat{f}_{h,H}$ | 5.88 | 6.72 (4.08) | 18.99 | 20.24 (7.51) | 1.35 | 1.49 (0.75) | 7.97 | 8.15 (1.85) | 0.38 | 0.40 (0.13) | 5.15 | 5.16 (0.53) |
| | $\hat{f}^{BC}_{h,H}$ | 4.32 | 5.37 (3.68) | 16.02 | 17.32 (7.05) | 1.03 | 1.14 (0.52) | 6.72 | 6.88 (1.74) | 0.32 | 0.33 (0.10) | 4.72 | 4.74 (0.51) |
| | $\hat{f}^{\delta}_{h,S}$ | 4.86 | 5.67 (4.01) | 2.55 | 3.91 (3.72) | 0.95 | 1.17 (0.81) | 0.67 | 0.82 (0.57) | 0.18 | 0.20 (0.11) | 0.17 | 0.18 (0.09) |
| | $\hat{f}^{\delta,BC}_{h,S}$ | 5.04 | 6.28 (4.94) | 1.27 | 2.33 (2.67) | 0.88 | 1.10 (0.80) | 0.50 | 0.63 (0.45) | 0.14 | 0.16 (0.10) | 0.13 | 0.15 (0.08) |
| 2 | $\hat{f}_{h,C}$ | 12.54 | 18.96 (18.19) | 0.88 | 1.76 (2.37) | 3.03 | 3.92 (3.17) | 0.27 | 0.40 (0.41) | 0.60 | 0.81 (0.64) | 0.07 | 0.09 (0.06) |
| | $\hat{f}^{BC}_{h,C}$ | 12.09 | 18.19 (18.31) | 1.62 | 2.80 (3.58) | 2.95 | 3.53 (2.62) | 0.17 | 0.28 (0.31) | 0.44 | 0.73 (0.80) | 0.02 | 0.03 (0.03) |
| | $\hat{f}_{h,P}$ | 9.39 | 15.96 (17.81) | 0.50 | 1.18 (1.79) | 2.39 | 3.27 (3.01) | 0.08 | 0.12 (0.14) | 0.51 | 0.68 (0.53) | 0.03 | 0.03 (0.02) |
| | $\hat{f}^{BC}_{h,P}$ | 6.29 | 12.20 (15.83) | 0.98 | 1.69 (2.17) | 1.55 | 2.07 (1.96) | 0.11 | 0.17 (0.17) | 0.31 | 0.45 (0.38) | 0.02 | 0.04 (0.04) |
| | $\hat{f}_{h,H}$ | 11.01 | 16.46 (16.59) | 19.20 | 20.34 (7.56) | 2.90 | 3.77 (2.74) | 8.00 | 8.17 (1.87) | 0.75 | 0.94 (0.66) | 5.17 | 5.19 (0.56) |
| | $\hat{f}^{BC}_{h,H}$ | 9.58 | 15.58 (16.46) | 15.62 | 16.94 (6.93) | 2.77 | 3.42 (2.41) | 6.64 | 6.84 (1.76) | 0.56 | 0.78 (0.69) | 4.76 | 4.77 (0.53) |
| | $\hat{f}^{\delta}_{h,S}$ | 20.99 | 29.27 (24.99) | 2.49 | 3.61 (3.49) | 2.85 | 5.01 (5.62) | 0.65 | 0.82 (0.59) | 0.45 | 0.78 (0.88) | 0.16 | 0.18 (0.10) |
| | $\hat{f}^{\delta,BC}_{h,S}$ | 17.03 | 25.91 (24.32) | 1.16 | 2.08 (2.39) | 2.18 | 4.08 (4.86) | 0.48 | 0.62 (0.48) | 0.32 | 0.56 (0.67) | 0.13 | 0.15 (0.08) |
| 3 | $\hat{f}_{h,C}$ | 8.67 | 10.05 ( 6.17) | 6.59 | 8.13 (6.05) | 2.22 | 2.58 (1.52) | 1.06 | 1.20 (0.68) | 0.54 | 0.62 (0.40) | 0.28 | 0.29 (0.10) |
| | $\hat{f}^{BC}_{h,C}$ | 8.74 | 11.45 (11.09) | 6.97 | 8.93 (7.65) | 2.00 | 2.42 (1.56) | 1.06 | 1.24 (0.76) | 0.41 | 0.49 (0.33) | 0.19 | 0.20 (0.10) |
| | $\hat{f}_{h,P}$ | 6.07 | 6.91 ( 4.27) | 4.61 | 5.58 (4.45) | 1.72 | 1.84 (0.92) | 0.87 | 0.99 (0.58) | 0.52 | 0.59 (0.32) | 0.26 | 0.28 (0.10) |
| | $\hat{f}^{BC}_{h,P}$ | 5.98 | 7.42 ( 6.15) | 5.00 | 6.14 (4.95) | 1.51 | 1.71 (0.98) | 0.90 | 1.01 (0.57) | 0.48 | 0.56 (0.35) | 0.19 | 0.20 (0.08) |
| | $\hat{f}_{h,H}$ | 9.87 | 11.64 ( 7.84) | 7.11 | 8.59 (6.56) | 2.95 | 3.30 (1.81) | 1.61 | 1.75 (0.81) | 0.85 | 0.98 (0.51) | 0.57 | 0.58 (0.15) |
| | $\hat{f}^{BC}_{h,H}$ | 6.99 | 9.42 (11.26) | 5.97 | 7.96 (7.50) | 1.73 | 2.00 (1.13) | 1.20 | 1.39 (0.80) | 0.57 | 0.62 (0.25) | 0.42 | 0.44 (0.14) |
| | $\hat{f}^{\delta}_{h,S}$ | 6.53 | 7.60 ( 5.62) | 6.47 | 7.82 (5.53) | 1.91 | 2.93 (4.25) | 1.33 | 1.48 (0.74) | 0.49 | 0.94 (2.33) | 0.32 | 0.33 (0.11) |
| | $\hat{f}^{\delta,BC}_{h,S}$ | 7.37 | 9.21 ( 7.50) | 6.69 | 8.08 (5.73) | 1.99 | 3.21 (5.13) | 1.45 | 1.56 (0.69) | 0.50 | 0.97 (2.72) | 0.36 | 0.37 (0.11) |
| 4 | $\hat{f}_{h,C}$ | 37.08 | 66.06 (61.73) | 6.16 | 7.08 (4.73) | 6.75 | 12.76 (14.90) | 0.94 | 1.04 (0.53) | 2.77 | 3.32 ( 2.87) | 0.27 | 0.28 (0.10) |
| | $\hat{f}^{BC}_{h,C}$ | 33.33 | 62.84 (59.47) | 6.40 | 7.65 (5.49) | 6.71 | 10.65 (12.53) | 0.90 | 1.04 (0.57) | 2.73 | 3.09 ( 2.11) | 0.17 | 0.19 (0.09) |
| | $\hat{f}_{h,P}$ | 17.31 | 53.54 (61.54) | 3.96 | 4.72 (3.37) | 6.10 | 9.66 (10.95) | 0.72 | 0.81 (0.44) | 2.84 | 3.81 ( 3.14) | 0.24 | 0.25 (0.09) |
| | $\hat{f}^{BC}_{h,P}$ | 13.73 | 42.52 (49.28) | 4.44 | 5.25 (3.70) | 4.62 | 8.28 (10.87) | 0.78 | 0.88 (0.49) | 1.80 | 2.36 ( 2.04) | 0.17 | 0.19 (0.08) |
| | $\hat{f}_{h,H}$ | 27.47 | 62.02 (61.13) | 6.63 | 7.57 (4.95) | 7.68 | 12.39 (13.47) | 1.46 | 1.57 (0.69) | 2.78 | 3.42 ( 2.42) | 0.55 | 0.56 (0.14) |
| | $\hat{f}^{BC}_{h,H}$ | 26.81 | 50.81 (49.75) | 5.34 | 6.64 (5.09) | 6.89 | 10.14 (11.39) | 1.08 | 1.19 (0.59) | 2.90 | 3.40 ( 2.32) | 0.40 | 0.42 (0.13) |
| | $\hat{f}^{\delta}_{h,S}$ | 99.38 | 95.48 (47.72) | 5.96 | 6.83 (4.25) | 29.70 | 35.93 (28.16) | 1.18 | 1.30 (0.60) | 9.54 | 13.13 (12.09) | 0.30 | 0.31 (0.10) |
| | $\hat{f}^{\delta,BC}_{h,S}$ | 94.17 | 94.17 (56.06) | 6.04 | 7.01 (4.44) | 22.98 | 32.08 (28.29) | 1.26 | 1.37 (0.57) | 7.20 | 10.71 (10.33) | 0.34 | 0.35 (0.10) |

Table 3: Median, mean and standard deviation of ISE$(\hat{f}_1, f_1)$ and ISE$(\hat{f}_2, f_2)$ for 100, 1000 and 10000 observations. The statistics are taken over 1000 simulation runs for each setting and the ISE was evaluated on the same grid with 100 points which was used before.

The second observation is that the two-dimensional estimators $\hat{f}_{h,P}$ and $\hat{f}_{h,P}^{BC}$ globally perform best for small samples sizes of not more than 10000. Especially for very small sample sizes ($< 1000$) it outperforms the other estimators with and without bias correction. There is only one out of 32 cases where the best estimator for 100 observations is neither $\hat{f}_{h,P}$ nor $\hat{f}_{h,P}^{BC}$ measured by median and only two cases measured by the mean. Besides, it also competes well with the other three approaches for sample sizes of 1000 and 10000. However, with increasing sample sizes, the counting process estimator $\hat{f}_{h,C}$ and especially its bias corrected version $\hat{f}_{h,C}^{BC}$ leads to the best results with respect to the integrated squared error in one of the four scenarios. For even larger samples sizes ($n = 10^6, 10^7$) which we investigated but which are not illustrated here, the bias corrected one-dimensional counting process estimator performed best.

The sieves estimators $\hat{f}_{h,S}^{\delta}$, $\hat{f}_{h,S}^{\delta,BC}$ perform surprisingly well and lead to satisfying results except for a complete breakdown in the estimation of $f_1$ with only 100 observations in Scenarios 2 and 4 which contain the boundary challenge problem. However, in most cases, the counting process approach or the projection approach leads to a smaller ISE than both sieves estimators.

As expected, globally the hazard estimators $\hat{f}_{h,H}$ and $\hat{f}_{h,H}^{BC}$ can't compete either with the two best approaches since the transformation from a hazard to a density function is not stable enough. Especially for the decreasing beta distribution in Scenarios 1 and 2 the hazard estimators failed at estimating $f_2$. However, they perform well and they can compete with the other estimators in Scenarios 3 and 4 and especially in the estimation of $f_1$ for small sample sizes.

In contrast to the hazard estimators, all other estimators perform extraordinary in Scenarios 1 and 2 with respect to the estimation of $f_2$, the mixed beta distribution. For the non-hazard estimators, Scenario 4 with the combination of the boundary challenge distribution and the mixture of beta distributions the estimation of $f_1$ is the most challenging estimation problem with by far the biggest ISE throughout and especially for $n = 100$ observations. The "boundary challenge" distribution also leads to slightly bigger ISE in $f_1$ in Scenario 2. In Scenario 1 we achieve the best results in the estimation of the distribution of $X$.

### *8.2. Application: Aggregated forecast*

The main application we compare here is based on the estimation of the probability mass of the distribution of $(X, Y)$ in the unobserved region $\mathcal{S}_2$, i.e., we investigate an estimator

$$\hat{r}(\hat{f}_1, \hat{f}_2) = \int_{\mathcal{S}_2} \hat{f}_1(x)\hat{f}_2(y)d(x, y).$$

of $r = \int_{\mathcal{S}_2} f(x, y)d(x, y)$. Recall that we assume $f(x, y) = f_1(x)f_2(y)$. Therefore, we estimate $f(x, y)$ in $r$ by $\hat{f}_1(x)\hat{f}_2(y)$. However, we are particularly interested in the estimation of the ratio between the probability mass in $\mathcal{S}_2$ and that in $\mathcal{S}_1$, weighted by the number of observations, i.e., in the object $R_n = nr/(1 - r)$. We estimate $R_n$ by

$$\hat{R}_n(\hat{f}_1, \hat{f}_2) = \frac{n\hat{r}(\hat{f}_1, \hat{f}_2)}{1 - \hat{r}(\hat{f}_1, \hat{f}_2)}.$$

Following the interpretation of the survival analysis approach, we want to know how many events will occur in the future, given the number of events in the past. This is illustrated through our main application of actuarial reserving mentioned in Section 1:

**Example 1** (Reserving in non-life insurance; IBNR numbers)**.** *One has observed $n$ past claims $(X_i, Y_i)$, $i = 1, \ldots, n$, where $X_i$ denotes the year the accident of claim $i$ happened and $Y_i$ is the delay until it was reported to the insurance company. We want to estimate the number of claims for accidents that have already happened but will be reported in the future. The density of $(X, Y)$ is assumed to be $f$ and $X$ and $Y$ are independent.*

*The data fits into the triangular form of Section 2, since by assumption all observations of $X$ occurred less than $T$ years ago. The time of data collection is represented by the diagonal on the scaled square $\mathcal{S} = [0, T]^2$ since only claims that have incurred and that also have been reported before that day are observed.*

*In our model, and in particular assuming a maximum delay of $T$ years, there is an unknown total number of claims but we know that all claims are contained in the squared support $[0, T]^2$. Hence, an estimate for the number of outstanding claims is $R_n$, the ratio of the probability of a claim to be in the future divided by the probability of already being observed times the number of observations $n$ that are already observed.*

*Another direct application would be the estimation of the RBNS claims, the claims that have been reported but not yet settled, i.e., the final payments have not been made yet for these claims.*

Another good illustration is given through aforementioned medical study.

**Example 2** (Medical study)**.** *Study about patients who go infected with a deadly disease during the last $T$ years. Assuming patients are only included into the data set after their deaths, one wants to forecast the total number of deaths in the next years not knowing the number of infected people. We only assume the time until death $Y$ is at most $T$ years and we have data about the time of death $(X_i + Y_i)$ and time of infection $(X_i)$ of $n$ patients ($i = 1, \ldots, n$) that have died during the last $T$ years. The number of future deaths in this group of people is then given by $R_n$ if time until death and time of infection are independent with densities $f_1$ and $f_2$, respectively. 3*

The estimated probability mass $\hat{r}(\hat{f}_1, \hat{f}_2)$ just depends on the marginal estimators $\hat{f}_1, \hat{f}_2$. However, in many cases estimates of $R_n$ are more accurate than density estimators since errors can "cancel each other out". In practice, over-smoothed density estimator often lead to a more stable estimation of this number. Because of the practical relevance of this estimation problem, we hence also use the fit of $\hat{R}_n$ as a measure of goodness for our estimators.

To compare the approaches, we use the relative error

$$\mathrm{err}(\hat{f}_1, \hat{f}_2) = \frac{\hat{R}_n(\hat{f}_1, \hat{f}_2) - R_n}{R_n},$$

as well as its squared value $\mathrm{err}^2$ as measures of goodness in our simulations.

The results of the estimation of $R_n$ are given in Table 4. We state the median, mean and standard deviation of $\text{err}^2$ in the 1000 simulation runs.

Here, clearly the bias corrected version $\hat{f}_{h,P}^{BC}$ of the projection estimator can be recommended as an overall winner. In ten of the twelve cases, it is the best estimator of $R_n$ in the mean integrated squared error. The regular projection estimator $\hat{f}_{h,P}$ performed best in two cases (in one of which $\hat{f}_{h,P}$ and $\hat{f}_{h,P}^{BC}$ have the same error) and the bias corrected counting process estimator $\hat{f}_{h,C}^{BC}$ wins in one of the twelve cases. The results for the median are the same except for one case where $\hat{f}_{h,C}$ slightly outperforms $\hat{f}_{h,C}^{BC}$.

In two thirds of the scenarios the best three estimators for the reserve measured by the median arise from the counting process and the projection approach. A sieves estimate only ranks second or twice in four cases whereas the hazard estimators only rank third in two cases. Therefore, it seems clear that the method of sieves is not smoothed enough, and that the hazard estimation approach is the wrong transformation of the unknown one-dimensional functions for in-sample forecasting. We are therefore left with the projected density approach and the backward survival density approach as our two test winners for in-sample forecasting. In the next section, we will further illustrate why the method of sieves might be too simple and why modern smoothing approaches are indeed necessary when working with in-sample forecasting problems. This is not an established fact or a well-known insight. For example, in the excellent book Martinussen and Scheike (2006) the point of view has been taken to concentrate estimation on integrated quantities that can be estimated without smoothing at $\sqrt{n}$-rates of convergence. While this obviously simplifies a lot of things in many important cases and without harming the applied statistician when interpreting results, this point of view does not seem to work when considering in-sample forecasting. There is more information on this in the next section.

### *8.3. Necessity of smoothing in the aggregated forecast*

To illustrate the importance of smoothing in the estimation problem in the last section, we compare the results of $\hat{f}_{h,P}$ and $\hat{f}_{h,P}^{BC}$ (which we have identified as the estimators from the most promising approach) with the discrete non-smoothed histogram estimator $\hat{p}^{\delta}$ which is widely used in practice in the insurance industry, the so-called chain ladder estimators for the actuarial reserve. It occurs in the application explained in the last section and it is derived from the forward factors mentioned in Section 5 as explained in e.g. England and Verrall (2002).

The method only uses the accumulation in the direction of $X$ as given by the formulas in Section 5. The development factors $\hat{\lambda}_j$ are used to extrapolate forecasts for the values of the cumulative matrix $D$ in the unobserved region via

$$\hat{D}_{i,n-i+2} = D_{i,n-i+1}\hat{\lambda}_{n-i+2},$$
$$\hat{D}_{i,k} = \hat{D}_{i,k-1}\hat{\lambda}_k, \qquad k = n-i+3, n-i+4, \ldots, n.$$

| Scenario | Estimator | $n = 100$ | | $n = 1000$ | | $n = 10000$ | |
|---|---|---|---|---|---|---|---|
| | | Median | Mean | Median | Mean | Median | Mean |
| 1 | $\hat{f}_{h,C}$ | 0.0444 | 0.1004 (0.1558) | 0.0071 | 0.0139 (0.0202) | 0.0008 | 0.0020 (0.0030) |
| | $\hat{f}^{BC}_{h,C}$ | 0.0498 | 0.1113 (0.1740) | 0.0054 | 0.0117 (0.0163) | 0.0005 | 0.0013 (0.0019) |
| | $\hat{f}_{h,P}$ | 0.0392 | 0.1004 (0.1711) | 0.0062 | 0.0147 (0.0231) | 0.0008 | 0.0020 (0.0030) |
| | $\hat{f}^{BC}_{h,P}$ | 0.0378 | 0.0921 (0.1494) | 0.0051 | 0.0109 (0.0163) | 0.0006 | 0.0014 (0.0023) |
| | $\hat{f}_{h,H}$ | 0.0561 | 0.1522 (0.2652) | 0.0128 | 0.0263 (0.0365) | 0.0101 | 0.0125 (0.0097) |
| | $\hat{f}^{BC}_{h,H}$ | 0.0486 | 0.1380 (0.2538) | 0.0104 | 0.0218 (0.0297) | 0.0097 | 0.0117 (0.0087) |
| | $\hat{f}^{\delta}_{h,S}$ | 0.0422 | 0.1306 (0.2873) | 0.0075 | 0.0169 (0.0274) | 0.0008 | 0.0020 (0.0031) |
| | $\hat{f}^{\delta,BC}_{h,S}$ | 0.0468 | 0.1399 (0.3316) | 0.0072 | 0.0169 (0.0267) | 0.0008 | 0.0020 (0.0030) |
| 2 | $\hat{f}_{h,C}$ | 0.0438 | 0.0792 (0.0984) | 0.0072 | 0.0162 (0.0221) | 0.0011 | 0.0027 (0.0044) |
| | $\hat{f}^{BC}_{h,C}$ | 0.0399 | 0.0851 (0.1181) | 0.0058 | 0.0120 (0.0160) | 0.0008 | 0.0018 (0.0026) |
| | $\hat{f}_{h,P}$ | 0.0291 | 0.0592 (0.0829) | 0.0047 | 0.0099 (0.0132) | 0.0008 | 0.0020 (0.0031) |
| | $\hat{f}^{BC}_{h,P}$ | 0.0156 | 0.0496 (0.0843) | 0.0023 | 0.0060 (0.0104) | 0.0005 | 0.0010 (0.0013) |
| | $\hat{f}_{h,H}$ | 0.0516 | 0.0998 (0.1400) | 0.0118 | 0.0233 (0.0305) | 0.0056 | 0.0075 (0.0075) |
| | $\hat{f}^{BC}_{h,H}$ | 0.0408 | 0.1047 (0.2160) | 0.0072 | 0.0165 (0.0233) | 0.0032 | 0.0052 (0.0059) |
| | $\hat{f}^{\delta}_{h,S}$ | 0.0708 | 0.1469 (0.2242) | 0.0081 | 0.0170 (0.0246) | 0.0008 | 0.0018 (0.0027) |
| | $\hat{f}^{\delta,BC}_{h,S}$ | 0.0604 | 0.1183 (0.1670) | 0.0061 | 0.0146 (0.0221) | 0.0006 | 0.0016 (0.0026) |
| 3 | $\hat{f}_{h,C}$ | 0.0518 | 0.1086 (0.1826) | 0.0085 | 0.0168 (0.0227) | 0.0017 | 0.0037 (0.0052) |
| | $\hat{f}^{BC}_{h,C}$ | 0.0509 | 0.1430 (0.3777) | 0.0072 | 0.0166 (0.0246) | 0.0011 | 0.0025 (0.0037) |
| | $\hat{f}_{h,P}$ | 0.0319 | 0.0965 (0.2160) | 0.0041 | 0.0097 (0.0133) | 0.0015 | 0.0030 (0.0038) |
| | $\hat{f}^{BC}_{h,P}$ | 0.0363 | 0.1081 (0.2558) | 0.0044 | 0.0097 (0.0138) | 0.0007 | 0.0017 (0.0024) |
| | $\hat{f}_{h,H}$ | 0.0551 | 0.1861 (0.5970) | 0.0126 | 0.0299 (0.0455) | 0.0062 | 0.0098 (0.0107) |
| | $\hat{f}^{BC}_{h,H}$ | 0.0553 | 0.2310 (0.7446) | 0.0097 | 0.0239 (0.0366) | 0.0048 | 0.0068 (0.0070) |
| | $\hat{f}^{\delta}_{h,S}$ | 0.0430 | 0.1156 (0.2466) | 0.0104 | 0.0342 (0.1006) | 0.0016 | 0.0089 (0.0649) |
| | $\hat{f}^{\delta,BC}_{h,S}$ | 0.0563 | 0.1441 (0.3040) | 0.0123 | 0.0401 (0.1225) | 0.0017 | 0.0100 (0.0772) |
| 4 | $\hat{f}_{h,C}$ | 0.1579 | 0.2325 (0.2242) | 0.0174 | 0.0369 (0.0512) | 0.0047 | 0.0083 (0.0097) |
| | $\hat{f}^{BC}_{h,C}$ | 0.1717 | 0.3348 (1.2380) | 0.0145 | 0.0330 (0.0488) | 0.0024 | 0.0050 (0.0068) |
| | $\hat{f}_{h,P}$ | 0.0792 | 0.1693 (0.1910) | 0.0095 | 0.0228 (0.0362) | 0.0032 | 0.0061 (0.0075) |
| | $\hat{f}^{BC}_{h,P}$ | 0.0686 | 0.1525 (0.1811) | 0.0080 | 0.0222 (0.0390) | 0.0017 | 0.0044 (0.0070) |
| | $\hat{f}_{h,H}$ | 0.1198 | 0.1915 (0.2052) | 0.0159 | 0.0330 (0.0431) | 0.0072 | 0.0112 (0.0125) |
| | $\hat{f}^{BC}_{h,H}$ | 0.1219 | 0.2303 (0.5166) | 0.0108 | 0.0262 (0.0407) | 0.0029 | 0.0058 (0.0078) |
| | $\hat{f}^{\delta}_{h,S}$ | 0.2955 | 0.3260 (0.3824) | 0.0839 | 0.1493 (0.2394) | 0.0197 | 0.0687 (0.1865) |
| | $\hat{f}^{\delta,BC}_{h,S}$ | 0.3123 | 0.3423 (0.3641) | 0.0776 | 0.1197 (0.1515) | 0.0175 | 0.0453 (0.1056) |

Table 4: Median, mean and standard deviation of the squared relative errors $\text{err}^2$ for 100, 1000 and 10000 observations. The statistics are taken over 1000 simulation runs for each setting.

The forecast for $R_n$, i.e., the so-called actuarial reserve in aforementioned application, is then given by the aggregated estimated data:

$$\hat{R}_n^{CLM} = \sum_{i+j>n} \hat{D}_{i,j},$$

where the summation is over all valid indices $i, j = 1, \ldots, n$ such that $i + j > n$.

The relative errors err of the results are given in Table 5. In nine out of twelve cases, the bias corrected projection estimators $\hat{f}^{BC}_{h,P}$ lead to the smallest absolute error in the estimation of $R_n$ whereas the non bias corrected $\hat{f}_{h,P}$ scored best once. Although the discrete chain ladder method has the smallest absolute error in two cases, it is striking that the standard deviation in the estimation is more than twice as high as the

standard deviation of the other methods in more than half of the cases. Hence, it turns out to be very unstable and therefore more risky than smoothed methods in practice. This reflects the typical bias-variance trade-off in nonparametric estimation. Without smoothing the discrete chain ladder method can result in small bias, however not enough smoothing always leads to high variance.

The sign of the bias is especially essential for the application in industry since it tells whether the costs will be under- or overestimated. The discrete chain ladder method tends to over-estimate the cost more often than the smoothed method. This behavior has already been described in Hiabu (2017).

In the results in Table 5, we omitted the cases in which the algorithm of the chain ladder method failed due to not enough observations. For only 100 observations, in each scenario the chain ladder method failed in 2 out 1000 simulation runs. It gave an invalid output because of the algorithm of the chain ladder method itself which is not stable if there are too many cells with value 0 in the occurring matrices $C$ and $D$. This issue also underlines the riskiness of the chain ladder method in extreme cases where one would need a very stable estimator.

We admit that the comparison is a bit unfair because the chain ladder method was originally designed for aggregated data and we aggregate by the same steps as our grid points, i.e., the discrete estimators are evaluated on as many grid points as the a smooth estimators in our simulation. Clearly, pre-aggregation in broader bins prevents invalid outputs. However, it underlines that smoothing is necessary in certain problems.

Table 6 shows that density estimation without smoothing (i.e. with the discrete chain ladder method) cannot compete with smoothed estimators — even though the aggregated forecast of the non-smoothed estimator is often reasonable.

| Scenario | n | Mean (s.d.) of err | | |
|---|---|---|---|---|
| | | $\hat{f}_{h,P}$ | $\hat{f}^{BC}_{h,P}$ | $\hat{p}^{\delta}$ |
| 1 | 100 | 0.091 (0.304) | 0.016 (0.303) | 0.078 (0.599) |
| | 1000 | 0.036 (0.116) | 0.010 (0.104) | 0.033 (0.154) |
| | 10000 | 0.017 (0.041) | 0.008 (0.037) | 0.031 (0.049) |
| 2 | 100 | -0.021 (0.242) | -0.082 (0.207) | 0.078 (0.759) |
| | 1000 | 0.031 (0.094) | -0.011 (0.077) | 0.037 (0.188) |
| | 10000 | 0.028 (0.035) | 0.004 (0.031) | 0.035 (0.056) |
| 3 | 100 | 0.061 (0.305) | 0.051 (0.325) | 0.003 (0.512) |
| | 1000 | 0.029 (0.094) | -0.004 (0.099) | 0.025 (0.302) |
| | 10000 | 0.033 (0.043) | -0.002 (0.042) | 0.020 (0.156) |
| 4 | 100 | -0.212 (0.353) | -0.202 (0.335) | -0.272 (0.800) |
| | 1000 | -0.071 (0.133) | -0.069 (0.132) | -0.019 (0.563) |
| | 10000 | -0.039 (0.067) | 0.000 (0.066) | 0.057 (0.434) |

Table 5: Mean and standard deviation of the relative error err of the projection estimators compared to the structured histogram $\hat{p}^{\delta}$ with $\delta = 0.01$.

| Scenario | n | Mean (s.d.) of ISE$(\cdot, f_1)$ | | |
|---|---|---|---|---|
| | | $\hat{f}_{h,P}$ | $\hat{f}^{BC}_{h,P}$ | $\hat{p}^{\delta}$ |
| 1 | 100 | 5.47 (3.26) | 5.47 (3.47) | 133.39 (166.19) |
| | 1000 | 1.12 (0.71) | 0.93 (0.62) | 14.86 (6.61) |
| | 10000 | 0.20 (0.11) | 0.15 (0.09) | 1.48 (0.48) |
| 2 | 100 | 15.96 (17.81) | 12.20 (15.83) | 337.42 (401.25) |
| | 1000 | 3.27 (3.01) | 2.07 (1.96) | 39.59 (29.78) |
| | 10000 | 0.68 (0.53) | 0.45 (0.38) | 3.78 (2.58) |
| 3 | 100 | 6.91 (4.27) | 7.42 (6.15) | 4.52 (32.29) |
| | 1000 | 1.84 (0.92) | 1.71 (0.98) | 18.84 (137.46) |
| | 10000 | 0.59 (0.32) | 0.56 (0.35) | 14.58 (96.51) |
| 4 | 100 | 53.54 (61.54) | 42.52 (49.28) | 19.37 (181.36) |
| | 1000 | 9.66 (10.95) | 8.28 (10.87) | 101.55 (331.53) |
| | 10000 | 3.81 (3.14) | 2.36 (2.04) | 179.29 (381.48) |

Table 6: Mean and standard deviation of the integrated squared error in the estimation of $f_1$ of the projection estimators compared to the structured histogram $\hat{p}^{\delta}$ with $\delta = 0.01$.

## 9. Conclusion

In this paper we have introduced multiplicative bias correction for all known nonparametric in-sample forecasting estimators. Furthermore, for the first time asymptotic theory has been established for the kernel smoothed structured histogram which has been identified as a method of sieves. The first conclusion from the finite sample simulation study presented in this paper is that multiplicative bias correction almost always leads to superior performance. The two-dimensional density projection approach of Mammen et al. (2015) and Lee et al. (2015) resulted in the best estimates with the reversed time survival density approach of Hiabu et al. (2016) coming in as a competitive runner-up. The method of sieves does not seem to be competitive and the transformation of a hazard estimator into a density estimator performs badly in this setting. We also establish that smoothing seems to be crucial for in-sample forecasting. This is thought provoking in an academic and, in particular, in a practical environment where discrete histogram type models with fixed sub-optimal bin choices are omnipresent. We conclude that the epidemiological, actuarial, econometric, engineering, forecasting and other common approaches to discrete age-period or age-cohort type of forecasting could be significantly improved by introducing continuous models and smoothing. We therefore believe continuous in-sample forecasting to have an increasing impact in the future.

## Appendix A. Asymptotic results

To derive the asymptotic behavior of the smoothed histogram estimator we illustrate the assumptions and proofs for the covariate $X$. For simplicity of illustration we leave out subscripts $l$ and dependence on the bandwidth $h$ where the interpretation is clear.

First we define aggregated observations $X_i^{\delta}$ that approximate the real but unknown continuous observations $X_i$ and describe the assumptions we need. We identify a counting process through the aggregated

observations and derive its intensity and hazard rate. With this new notation we can state the assumptions for our results.

*Appendix A.1. Aggregated observations and corresponding counting processes*

The aggregation is analogous to the one in Hiabu (2017) and we adapt the notation from there. For simplicity of notation let $\delta > 0$ be such that $\delta^{-1}$ is an integer. Observations $(X_i, Y_i) \in \mathcal{S}_1 = \{(x, y) \in [0, T]^2 : x + y \leq T\}$ are aggregated to $(X_i^\delta, Y_i^\delta)$ on

$$\mathcal{S}_1{}^\delta = \{(x_j, y_k) = ((j + 0.5)\delta, (k + 0.5)\delta) : j, k = 0, 1, \ldots, T\delta^{-1} - 1, j + k \leq T\},$$

via

$$(X_i^\delta, Y_i^\delta) = (x_j, y_k) \Leftrightarrow Y_i \in [k\delta, (k + 1)\delta) \text{ and } X_i + Y_i \in [(j + k)\delta, (j + k + 1)\delta).$$

We define the time reversed counting process $N_i^\delta(t) = I(T - X_i^\delta \leq t)$ with respect to the filtration $\mathcal{F}_t^{i,\delta} = \sigma\left(\left\{(t - X_i^\delta) \leq s : s \leq t\right\} \cup \mathcal{N}\right)$ for a null set $\mathcal{N}$. Let $\mu^\delta$ be the counting measure that is defined via $\mu^\delta(A) = \delta\#\{j : (j + 0.5\delta) \in A, j = -T\delta^{-1} + 1, \ldots, -1, 0, 1, \ldots, T\delta^{-1} - 1\}$, $A \in \mathcal{B}$, instead of the Lebesgue measure. Note that $\mu^\delta$ needs to be defined for values below zero for technical reasons only. The density of $X^\delta$ with respect to $\mu^\delta$ is given via

$$f^\delta(t) = \begin{cases} 0 & \text{if } t \neq (j + 0.5)\delta \\ \delta^{-1} \int_{j\delta}^{(j+1)\delta} f(s)ds & \text{if } t = (j + 0.5)\delta, \end{cases}$$

with $f$ being the Lebesgue density of $X$. The intensity of $N_i^\delta$ is given by $\lambda_i^\delta(t) = \alpha_\delta(t)Z_i^\delta(t)$ where $\alpha_\delta(t) = f^\delta(T - t)/(1 - F^\delta(T - t))$ is the hazard rate of $X^\delta$ in the reversed time scale counting measure. The exposure is given by $Z_i^\delta(t) = I(Y_i^\delta \leq t \leq T - X_i)$. Furthermore let $N^\delta(t) = \sum_{i=1}^n N_i^\delta(t)$ and $Z^\delta(t) = \sum_{i=1}^n Z_i^\delta(t)$ and we write $S^\delta(t) = 1 - F^\delta(T - t)$ for the survival function in reversed time.

*Appendix A.2. Asymptotic properties*

To derive the asymptotic behavior of the sieves histogram estimator, we make the following assumptions.

**A1** It holds $f \in C^2([0, T])$ and $f(t) > 0$ for all $t \in [0, T]$.

**A2** The kernel $K$ is symmetric, has bounded support $[-1, 1]$, has bounded seconded moment and satisfies $K \in C^2([-1, 1])$.

**A3** The bandwidth $h = h(n)$ satisfies $h \to 0$ and $n^{1/4}h \to \infty$ as $n \to \infty$.

**A4** There exists a continuous function $\gamma$ such that $\sup_{s\in[0,1]}|Z^\delta(s)/n - \gamma(s)| = o_P(1)$ and $\gamma(t) > 0$ for every $t$.

Furthermore, we introduce the notation $\kappa_0^\delta = \int K(s)d\mu^\delta(s)$. Through a Taylor expansion it can be easily shown that Assumption **A2** implies $\kappa_0^\delta = 1 + o(\delta^2)$. Moreover, it holds $\int sK(s)d\mu^\delta(s) = 0$ under **A2**. We also define $\kappa_2^\delta(s) = \int K_h(s)s^2 d\mu^\delta(s)$ and $\Theta_K = \frac{1}{2}\int_{-1}^{1} K(s)K'(s)ds$.

**Proposition 1.** *Under Assumptions **A1**–**A4**, for $t \in (0,T)$, it holds for every $\delta > 0$ that*

$$(nh)^{1/2}\left\{\hat{f}^\delta_{l,h,S}(t) - f_l(t) - B_l(t)\right\} \to \mathcal{N}(0, \sigma_l^2(t))$$

*in distribution as $n \to \infty$, where $B_l(t) = 0.5h^2\kappa_2^\delta f_l''(t) + o_P(h^2) + o(\delta^2)$, $\sigma_l^2(t) = R^\delta(K)f_l^\delta(t)F_l^\delta(t)(\gamma_l^\delta(t))^{-1}$. Moreover, it holds $\kappa_2^\delta \to \int u^2K(u)du$, $R^\delta(K) = \int K(u)^2du + o_p(\delta^2)$ for $\delta \to 0$.*

To prove asymptotic normality for $\hat{f}^{\delta,BC}_{h,S}$ we need a stronger assumption on the density.

**A1**' Let $f \in C^4([0,T])$ such that $f(t), f''(t) > 0$ for all $t \in [0,T]$.

**Proposition 2.** *Under Assumptions **A1'**, **A2**–**A4**, for $t \in (0,T)$, it holds for every $\delta > 0$ that*

$$(nh)^{1/2}\left\{\hat{f}^{\delta,BC}_{l,h,S}(t) - f_l(t) - B_l^{BC}(t)\right\} \to \mathcal{N}(0, \sigma^2_{l,BC}(t))$$

*in distribution as $n \to \infty$, where $B_l^{BC}(t) = (1/4)h^4(\kappa_2^\delta)^2 f_l^\delta(t)((f_l^\delta)''/f_l^\delta)''(t) + o_P(h^2) + o(\delta^2)$, $\sigma^2_{l,BC}(t) = f_l^\delta(t)F_l^\delta(t)(\gamma_l^\delta(t))^{-1}\int \Gamma_K^2(u)d\mu^\delta(u)$ with $\Gamma_K(u) = 2K - (K * K)(u)$. Moreover, it holds $\kappa_2^\delta \to \int u^2K(u)du$, $\int \Gamma_K^2(u)d\mu^\delta(u) \to \int \Gamma_K^2(u)du$ for $\delta \to 0$.*

The asymptotic behavior of $\hat{f}_{S^*}$ and $\hat{f}^{BC}_{S^*}$ was already described in Nielsen et al. (2009) for arbitrary weightings $W$ with asymptotic bias and variance independent of $W$. The estimators in said paper coincide with ours for the choice $W(s) = (Z_l(s)/n)^{-1}$. We make the following additional assumptions for $l = 1, 2$:

**B1** $\sup_{s\in[0,1]}|Z_l(s)/n - \gamma_l(s)| = o_P(1)$.

**B2** $\sup_{s\in[0,1]}|\hat{S}_l^R(s) - S_l^R(s)| = O_P(n^{-1/2})$.

**Proposition 3.** *Under Assumptions **A1**–**A3**, **B1**–**B2**, for $t \in (0,T)$, it holds*

$$(nh)^{1/2}\left\{\hat{f}_{l,h,S^*}(t) - f(t) - B(t)\right\} \to \mathcal{N}(0, \sigma^2(t))$$

*in distribution as $n \to \infty$, where $B_l(t) = \kappa_2 f_l''(t)h^2/2 + o(h^2)$, $\sigma^2(t) = R(K)f_l(t)F_l(t)\gamma_l(t)^{-1}$ and $\gamma_l(t) = \mathbb{P}(Z_l^1(t) = 1)$ with the notations $\kappa_2 = \int s^2K(s)ds$ and $R(K) = \int K^2(s)ds$.*

## Appendix B. Proofs

The proof of Proposition 3 is mainly based on the following central limit theorem for martingales which was proved in Ramlau-Hansen (1983).

**Theorem 1** (Ramlau-Hansen (1983))**.** *Let $W_n(t)$ be a predictable processes and let there be some $\sigma^2 \geq 0$ such that for every $\varepsilon > 0$*

$$\int_0^1 W_n^2(s)\Lambda_n(s) = \sigma^2 + o_p(1), \qquad \int_0^1 W_n^2(s) I\left(W_n^2(s) > \varepsilon\right)\Lambda_n(s) = o_p(1).$$

*Let $(N_n)_n$ be a sequence of counting processes on $[0,1]$ with corresponding sequence of martingales given by $M_n(t) = N_n(t) - \int_0^t \Lambda_n(s)ds$ where $(\Lambda_n(s))_n$ is the sequence of intensity processes. Then it holds that $\int_0^1 W_n(s)dM_n(s) \to \mathcal{N}(0,\sigma^2)$ in distribution as $n \to \infty$.*

*Appendix B.1. Proof of Proposition 1*

First we observe that our histogram estimator can be represented through a Kaplan-Meier type estimator $\hat{S}^\delta$ of the aggregated cumulative hazard function that is defined through a Nelson-Aalen type estimator — just with respect to the counting measure $\mu^\delta$ for aggregated observations instead of the Lebesgue measure. Then we derive the asymptotic theory for $\hat{S}^\delta$ analogously to Andersen et al. (1993). Next we estimate the error that arises through estimation of the aggregated density in grid points instead of the real Lebesgue density in exact points. Finally, Proposition 1 is proved with a standard counting process martingale proof involving Theorem 1.

*Appendix B.1.1. Forward factors and structured histogram*

As outlined in Hiabu (2017), the forward factors from the chain ladder method can be represented through estimators of exposure and occurrence and identified with transformed hazard estimators. Explicitly, we can write the forward factors as

$$\hat{\lambda}_j = \frac{1}{1 - \delta\hat{\alpha}^{H,\delta}(x_j)},$$

for

$$\hat{\alpha}^{H,\delta}(t) = \frac{\int_{x_{j-1}}^{x_j} dN^\delta(s)}{\delta\int_{x_{j-1}}^{x_j} Z^\delta(s)d\mu^\delta(s)} = \frac{O^{H,\delta}(x_k)}{E^{H,\delta}(x_k)},$$

where $j$ is such that $t \in [x_{j-1}, x_j)$.

We now motivate how to transform the forward factors into a histogram. Let $n_1$ denote the number of observations in the first bin. With the forward factors $\hat{\lambda}_j$ being proportions of cumulative data, we get that the cumulative number of observations in bin $j$ is $(\hat{\lambda}_j \cdots \hat{\lambda}_1)n_1$ whereas the actual number of observations in bin $j$ is $(\hat{\lambda}_j \cdots \hat{\lambda}_1)n_1 - (\hat{\lambda}_{j-1} \cdots \hat{\lambda}_1)n_1 = (\hat{\lambda}_j - 1)(\hat{\lambda}_{j-1} \cdots \hat{\lambda}_1)n_1$. To get a histogram, i.e. proportions, we divide the last equation by the number of total observations, i.e. the last bin for cumulative observations $(\hat{\lambda}_{m-1} \cdots \hat{\lambda}_1)n_1$ to get

$$\hat{p}(1) = \frac{1}{\prod_{k=1}^{m-1}\hat{\lambda}_k}, \qquad \hat{p}(j) = \frac{\hat{\lambda}_{j-1} - 1}{\prod_{k=j-1}^{m-1}\hat{\lambda}_k}, \qquad j = 2, \ldots, m.$$

This leads to the representation of the structured histogram as

$$\begin{aligned}\hat{p}(1) &= \frac{1}{\prod_{k=1}^{m-1}\hat{\lambda}_k} \\ &= \prod_{k=1}^{m-1}\left(1-\delta\hat{\alpha}^{H,\delta}(T-x_k)\right) \\ &= \delta\hat{\alpha}^{H,\delta}(T-x_1)\hat{S}^{\delta}(x_1),\end{aligned}$$

with $\hat{\alpha}^{H,\delta}(T-x_1)=\delta^{-1}$, and

$$\begin{aligned}\hat{p}(j) &= \frac{\hat{\lambda}_{j-1}-1}{\prod_{k=j-1}^{m-1}\hat{\lambda}_k} \\ &= \frac{\delta\hat{\alpha}^{H,\delta}(T-x_j)}{1-\delta\hat{\alpha}^{H,\delta}(T-x_j)}\prod_{k=j-1}^{m-1}\left(1-\delta\hat{\alpha}^{H,\delta}(T-x_k)\right) \\ &= \delta\hat{\alpha}^{H,\delta}(T-x_j)\hat{S}^{\delta}(x_j),\end{aligned}$$

$j=1,\ldots,m$, where we used the survival estimator $\hat{S}^{\delta}(x_j)=\prod_{k=j}^{m-1}\left(1-\delta\hat{\alpha}^{H,\delta}(T-x_k)\right)$.

These expressions help us to represent the estimator as a counting process estimator. First note that intervals $[x_{j-1},x_j)$ satisfy $|x_j-x_{j-1}|<\delta$ and hence $\mu^{\delta}([x_{j-1},x_j))=\delta$. Moreover, the aggregated counting process only jumps exactly once per bin (given there is at least one observation per bin). Hence, we get the identity

$$\delta\hat{\alpha}^{H,\delta}(t)=\begin{cases} Z^{\delta}(t)^{-1} & N^{\delta} \text{ jumps at } t \\ 0 & \text{else.}\end{cases} \tag{B.1}$$

This enables us to write

$$\begin{aligned}\hat{f}_S^{\delta}(t) &= \sum_{s=1}^{m_b} K_h(t-sb)\hat{p}(sb) \\ &= n^{-1}\int K_h(t-s)(Z^{\delta}(s)/n)^{-1}\hat{S}^{\delta}(s)dN^{\delta}(s).\end{aligned} \tag{B.2}$$

*Appendix B.1.2. Identification with Kaplan-Meier estimator and asymptotic behavior of $\hat{S}^{\delta}$*

To show convergence of $\hat{S}^{\delta}$ towards $S^{\delta}$, we first illustrate that $\hat{S}^{\delta}$ corresponds to the Kaplan-Meier estimator in our setting. We introduce the notations

$$\begin{aligned}\hat{A}_{\delta}(t) &= \int_0^t Z^{\delta}(s)^{-1}dN^{\delta}(s), \\ A_{\delta}(t) &= \int_0^t \alpha_{\delta}(s)d\mu^{\delta}(s), \\ A_{\delta}^{*}(t) &= \int_0^t J_{\delta}(s)\alpha_{\delta}(s)d\mu^{\delta}(s), \\ J_{\delta}(t) &= I(Z^{\delta}(t)>0),\end{aligned}$$

and we use the convention $0/0 = 0$. The integrated hazard estimator $\hat{A}_\delta(t)$ is the Nelson-Aalen estimator for the aggregated counting process $N^\delta$. Equation (B.1) implies furthermore $\Delta\hat{A}_\delta(t) = Z^\delta(t)^{-1} = \delta\hat{\alpha}^{H,\delta}(t)$ if $N^\delta$ jumps at $t$ and $\Delta\hat{A}_\delta(t) = 0 = \delta\hat{\alpha}^{H,\delta}(t)$ otherwise.

Thus, we can identify $\prod_{k=j}^{m-1}\left(1-\delta\hat{\alpha}^{H,\delta}(T-x_k)\right)$ with the Kaplan-Meier estimator $\prod_{s\le t}\left(1-\Delta\hat{A}_\delta(s)\right)$. The following proposition describes the convergence of the Kaplan-Meier estimator in the aggregated setting.

**Proposition 4.** *Let $t \in [0,T]$ and assume that, for every $\varepsilon > 0$ and $\delta > 0$ , it holds*

$$n\int_0^t \frac{J_\delta(s)}{Z^\delta(s)}\alpha_\delta(s)d\mu^\delta(s) \xrightarrow{P} \sigma^2(t), \tag{B.3}$$

$$n\int_0^t \frac{J_\delta(s)}{Z^\delta(s)} I\left(\frac{J_\delta(s)}{Z^\delta(s)} > \varepsilon\right)\alpha_\delta(s)d\mu^\delta(s) \xrightarrow{P} 0, \tag{B.4}$$

$$n^{1/2}\int_0^t (1-J_\delta(s))\alpha_\delta(s)d\mu^\delta(s) \xrightarrow{P} 0. \tag{B.5}$$

*as $n\to\infty$ and for a continuous function $\sigma \ge 0$ with $\sigma(0) = 0$. Then, as $n\to\infty$, for $\delta > 0$ fixed, it holds*

$$\sup_{s\in[0,t]} |\hat{S}_\delta(s) - S_\delta(s)| = O_P(n^{-1/2}).$$

*Proof.* The proposition follows with Lenglart's inequality and the functional delta method (Andersen et al., 1993, pp. 86, 111), since $\hat{S}_\delta$ and $S_\delta$ are functionals of $\hat{A}_\delta$ and $A_\delta$, respectively. Indeed, it holds $S_\delta(t) = \prod_{s\le t}(1-\Delta A_\delta(s))$.

We first show the convergence of $n^{1/2}\sup_{s\in[0,T]}|\hat{A}_\delta(s) - A^*_\delta(s)|$. Lenglart's inequality yields

$$P\left(n^{1/2}\sup_{s\in[0,T]}|\hat{A}_\delta(s) - A^*_\delta(s)| > \eta\right) \le \frac{\delta}{\eta} + P\left(n^{1/2}\langle\hat{A}_\delta - A^*_\delta\rangle(T) > \delta\right),$$

for every $\delta > 0$ and every $\eta > 0$. Now we use a limit theorem to show $n^{1/2}\langle\hat{A}_\delta - A^*_\delta\rangle(T) = O_P(1)$. Note that we have $n^{1/2}\left(\hat{A}_\delta(t) - A^*_\delta\right)(t) = n^{1/2}\int_0^t \frac{J_\delta(s)}{Z^\delta(s)}dM_\delta(s)$ with $dM_\delta(s) = dN_\delta(s) - \alpha_\delta(s)Z^\delta(s)d\mu^\delta(s)$ and $M_\delta$ being a local square integrable martingale. Thus, $\hat{A}_\delta(t) - A^*_\delta$, is a local square integrable martingale as well. Hence, the convergence follows from the martingale central limit theorem in Rebolledo (1980) (see also (Andersen et al., 1993, p. 83)), under the condition

$$n^{1/2}\langle\hat{A}_\delta - A^*_\delta\rangle(t) \xrightarrow{P} \sigma^2(t), \tag{B.6}$$

for all $t \in [0,T]$ and a continuous function $\sigma \ge 0$ with $\sigma(0) = 0$ and the Lindeberg condition we assume in (B.5). Condition (B.6) follows from (B.3) since

$$n^{1/2}\langle\hat{A}_\delta - A^*_\delta\rangle(t) = n\int_0^t \left(\frac{J_\delta(s)}{Z^\delta(s)}\right)^2 Z^\delta(s)\alpha_\delta(s)d\mu^\delta(s) = n\int_0^t \frac{J_\delta(s)}{Z^\delta(s)}\alpha_\delta(s)d\mu^\delta(s).$$

The conclusion of the proposition then follows by $n^{1/2}\sup_{s\in[0,T]}|A^*_\delta(s) - A_\delta(s)| \xrightarrow{P} 0$ which is ensured through assumption (B.5). □

*Appendix B.1.3. Estimation error through aggregation*

We derive the error that occurs from estimating $f$ at the grid points $t_j = (j+0.5)\delta$ for $j\delta < t \leq (j+1)\delta$. Under Assumption **A1**, through a Taylor expansion, we get

$$f(t_j) - f(t) = (t_j - t)f'(t) + o(\delta^2).$$

Moreover, the deviation between $f_\delta$ and $f$ in grid points can be bounded by

$$f^\delta(t_j) - f(t_j) = \delta^{-1} \int_{j\delta}^{(j+1)\delta} [f(u) - f(t_j)]\, du = o(\delta^2)$$

through another Taylor expansion. Concluding, it holds

$$f^\delta(t_j) - f(t) = (t_j - t)f'(t) + o(\delta^2). \tag{B.7}$$

*Appendix B.1.4. Proof of Proposition 1*

*Proof of Proposition 1.* With the representation of $\hat{f}_S^\delta$ in equation (B.2), we can prove Proposition 1 with a standard procedure for counting process estimators. First we evaluate our estimator at the closest grid point $t_j$ to $t$. Then we split $\hat{f}_S^\delta(t_j) - f(t) = V(t) + B(t)$ into a deterministic bias part $B(t) = f_\delta^*(t_j) - f(t)$ and a martingale part $V(t) = \hat{f}_S^\delta(t_j) - f_\delta^*(t_j)$ with the definition

$$f_\delta^*(t) = \int K(t-s)\hat{S}^\delta(s)\alpha^\delta(s)d\mu^\delta(s).$$

The conditions (B.3)–(B.5) of Proposition 4 are satisfied and hence we get that the bias term satisfies

$$\begin{aligned} B(t) &= \int K_h(t_j - s)\left[f^\delta(s) - f(t)\right] d\mu^\delta(s) + O_P(n^{-1/2}) + (1-\kappa_0^\delta)f(t) \\ &= \int K_h(t_j - s)\left[f(s_k) + (s_k - s)f'(s) - f(t)\right] d\mu^\delta(s) + (1-\kappa_0^\delta)f(t) + O_P(n^{-1/2}) + o(\delta^2) \\ &= \frac{1}{2}h^2\kappa_2^\delta f''(t) + o_P(h^2) + o(\delta^2), \end{aligned}$$

where we have used the approximation in equation (B.7) as well as a second order Taylor expansions of $f$ and Assumptions **A2**. Note that it furthermore holds $O_P(n^{-1/2}) = o_P(h^2)$ under Assumption **A3**. We write $s_k$ for the grid point closest to $s$.

Under Assumption **A4** and with Proposition 4, it holds for the martingale term that

$$\begin{aligned} V(t) &= \hat{f}_S(t_j) - f^*(t_j) \\ &= \frac{1}{n}\int K_h(t-s)(\gamma^\delta(s))^{-1}S^\delta(s)dM^\delta(s) + O_P(n^{-1/2}), \end{aligned}$$

where we have used that $dM_\delta(s) = dN_\delta(s) - \alpha_\delta(s)Z^\delta(s)d\mu^\delta(s)$ and $M_\delta$ is a local square integrable martingale as in the proof of Proposition 4. With Theorem 1, we can conclude $(nh)^{1/2}V(t) \to \mathcal{N}\left(0, \sigma^2(t)\right)$ for $\sigma^2(t) = R^\delta(K)f^\delta(t)F^\delta(t)(\gamma^\delta(t))^{-1}$, where $R^\delta(K) = \int K(s)^2 d\mu^\delta(s)$.

Convergence of the factor $R^\delta(K)$ in the variance for $\delta \to 0$ can be shown with another Taylor expansion of $K$. □

*Appendix B.2. Proofs of Proposition 2 and 3*

*Proof of Proposition 2.* Following the proofs in Linton and Nielsen (1994) and Nielsen and Tanggaard (2001) and using the representations from the proof of Proposition 1, we get the result. □

*Proof of Proposition 3.* The proof is analogous to the last part of the proof of Proposition 1 with $N^\delta$ replaced by the non-aggregated counting process $N$. □

**Acknowledgment**

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) through the Research Training Group RTG 1953.

**References**

Andersen, P., Borgan, O., Gill, R., Keiding, N., 1993. Statistical Models Based on Counting Processes. Springer, New York.

Bowman, A.W., 1984. An alternative method of cross-validation for the smoothing of density estimates. Biometrika 71, 353–360.

Duraisamy, P., 2002. Changes in returns to education in India, 1983–94: by gender, age-cohort and location. Economics of Education Review 21, 609–622.

England, P.D., Verrall, R.J., 2002. Stochastic claims reserving in general insurance. British Actuarial Journal 8, 443–544.

Fan, J., Gijbels, I., 1996. Local polynomial modelling and its applications. Chapman and Hall, London.

Fukuda, K., 2006. Age-period-cohort decomposition of aggregate data: an application to US and Japanese household saving rates. Journal of Applied Econometrics 21, 981–998.

Gámiz, M.L., Mammen, E., Martínez-Miranda, M.D., Nielsen, J.P., 2016. Double one-sided cross-validation of local linear hazards. Journal of the Royal Statistical Society: Series B 78, 755–779.

Hall, P., 1983. Large sample optimality of least squares cross-validation in density estimation. The Annals of Statistics 11, 1156–1174.

Harnau, J., Nielsen, B., 2018. Over-dispersed age-period-cohort models. Journal of the American Statistical Association 0, 1–11.

Hart, J.D., Yi, S., 1998. One-sided cross-validation. Journal of the American Statistical Association 93, 620–631.

Hiabu, M., 2017. On the relationship between classical chain ladder and granular reserving. Scandinavian Actuarial Journal 2017, 708–729.

Hiabu, M., Mammen, E., Martínez-Miranda, M.D., Nielsen, J.P., 2016. In-sample forecasting with local linear survival densities. Biometrika 103, 843–859.

Jeon, Y.J., Kim, C.R., Park, J.S., Choi, K.H., Kang, M.J., Park, S.G., Park, Y.J., 2016. Health inequalities in hypertension and diabetes management among the poor in urban areas: a population survey analysis in South Korea. BMC public health 16, 492.

Jones, M.C., 1993. Simple boundary correction for kernel density estimation. Statistics and Computing 3, 135–146.

Jones, M.C., Linton, O.B., Nielsen, J.P., 1995. A simple bias reduction method for density estimation. Biometrika 82, 327–338.

Kuang, D., Nielsen, B., Nielsen, J.P., 2009. Chain-ladder as maximum likelihood revisited. Annals of Actuarial Science 4, 105–121.

Lee, Y.K., Mammen, E., Nielsen, J.P., Park, B.U., 2015. Asymptotics for in-sample density forecasting. The Annals of Statistics 43, 620–651.

Lee, Y.K., Mammen, E., Nielsen, J.P., Park, B.U., 2017. Operational time and in-sample density forecasting. The Annals of Statistics 45, 1312–1341.

Leung, G.M., Thach, T.Q., Lam, T.H., Hedley, A.J., Foo, W., Fielding, R., Yip, P.S.F., Lau, E.M.C., Wong, C.M., 2002. Trends in breast cancer incidence in Hong Kong between 1973 and 1999: an age-period-cohort analysis. British Journal of Cancer 87, 982.

Linton, O.B., Nielsen, J.P., 1994. A multiplicative bias reduction method for nonparametric regression. Statistics & Probability Letters 19, 181–187.

Mammen, E., Marron, J.S., Turlach, B.A., Wand, M.P., 2001. A general framework for constrained smoothing. Statistical Science 16, 232–248.

Mammen, E., Martínez-Miranda, M.D., Nielsen, J.P., 2015. In-sample forecasting applied to reserving and mesothelioma. Insurance: Mathematics and Economics 61, 76–86.

Martínez-Miranda, M.D., Nielsen, B., Nielsen, J.P., 2014. Inference and forecasting in the age-period-cohort model with unknown exposure with an application to mesothelioma mortality. Journal of the Royal Statistical Society: Series A 178, 29–55.

Martínez-Miranda, M.D., Nielsen, B., Nielsen, J.P., 2016. Simple benchmark for mesothelioma projection for Great Britain. Occupational and Environmental Medicine , 561–563.

Martínez-Miranda, M.D., Nielsen, J.P., Sperlich, S., 2009. One sided cross-validation for density estimation with an application to operational risk, in: von Gregoriou, G.N. (Ed.), Operational Risk Towards Basel III: Best Practices and Issues in Modelling. Management and Regulation. John Wiley and Sons, New Jersey, pp. 177–195.

Martínez-Miranda, M.D., Nielsen, J.P., Sperlich, S., Verrall, R.J., 2013. Continuous chain ladder: Reformulating and generalising a classical insurance problem. Expert Systems with Applications 40, 5588–5603.

Martinussen, T., Scheike, T.H., 2006. Dynamic regression models for survival data. Springer, New York.

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

Nielsen, J.P., Tanggaard, C., Jones, M.C., 2009. Local linear density estimation for filtered survival data, with bias correction. Statistics 43, 167–186.

Peto, J., Matthews, F.E., Hodgson, J.T., Jones, J.R., 1995. Continuing increase in mesothelioma mortality in Britain. The Lancet 345, 535–539.

Ramlau-Hansen, H., 1983. Smoothing counting process intensities by means of kernel functions. The Annals of Statistics 11, 453–466.

Rebolledo, R., 1980. Central limit theorems for local martingales. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 51, 269–286.

Reither, E.N., Hauser, R.M., Yang, Y., 2009. Do birth cohorts matter? Age-period-cohort analyses of the obesity epidemic in the United States. Social science & medicine 69.

Remontet, L., Estève, J., Bouvier, A.M., Grosclaude, P., Launoy, G., Menegoz, F., Exbrayat, C., Tretare, B., Carli, P.M., Guizard, A.V., et al., 2003. Cancer incidence and mortality in France over the period 1978–2000. Revue d'épidémiologie et de santé publique 51, 3–30.

Rousselière, D., Rousselière, S., 2017. Decomposing the effects of time on the social acceptability of biotechnology using age-period-cohort-country models. Public Understanding of Science 26, 650–670.

Rudemo, M., 1982. Empirical choice of histograms and kernel density estimators. Scandinavian Journal of Statistics 9, 65–78.

Schwadel, P., Stout, M., 2012. Age, period and cohort effects on social capital. Social Forces 91, 233–252.

Tu, Y.K., Smith, G.D., Gilthorpe, M.S., 2011. A new approach to age-period-cohort analysis using partial least squares regression: the trend in blood pressure in the Glasgow alumni cohort. PLOS One 6, e19401.

Wand, M.P., Jones, M.C., 1994. Kernel Smoothing. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Taylor & Francis.

Ware, J.H., DeMets, D.L., 1976. Reanalysis of some baboon descent data. Biometrics 32, 459–463.

Wilke, R.A., 2017. Forecasting macroeconomic labour market flows: What can we learn from micro-level analysis? Oxford Bulletin of Economics and Statistics 80, 822–842.

Yang, Y., 2011. Aging, cohorts, and methods, in: Handbook of Aging and the Social Sciences (Seventh Edition). Elsevier, pp. 17–30.