Smooth backfitting of proportional hazards with multiplicative components

Munir Hiabu [*1], Enno Mammen[2], M. Dolores Martínez-Miranda[3] and Jens
P. Nielsen[4]

[1]School of Mathematics and Statistics, University of Sydney, Camperdown
NSW 2006, Australia

[2]Institute for Applied Mathematics, Heidelberg University, Im Neuenheimer
Feld 205, 69120 Heidelberg, Germany

[3]Department of Statistics and Operations Research, University of Granada,
Campus Fuentenueva s/n. 18071 Granada, Spain

[4]Cass Business School, City, University of London, 106 Bunhill Row,
London, EC1Y 8TZ, United Kingdom

January 9, 2020

[*]munir.hiabu@sydney.edu.au (Corresponding author)

Abstract

Smooth backfitting has proven to have a number of theoretical and practical advantages in structured regression. By projecting the data down onto the structured space of interest smooth backfitting provides a direct link between data and estimator. This paper introduces the ideas of smooth backfitting to survival analysis in a proportional hazard model, where we assume an underlying conditional hazard with multiplicative components. We develop asymptotic theory for the estimator. In a comprehensive simulation study we show that our smooth backfitting estimator successfully circumvents the curse of dimensionality and outperforms existing estimators. This is especially the case in difficult situations like high number of covariates and/or high correlation between the covariates, where other estimators tend to break down. We use the smooth backfitter in a practical application where we extend recent advances of in-sample forecasting methodology by allowing more information to be incorporated, while still obeying the structured requirements of in-sample forecasting.

1   Introduction

Nonparametric models suffer from the curse of dimensionality in high dimensional data spaces. Random forests (Breiman, 2001) circumvent the dimensionality problem by assuming that not all variables are relevant and that the function of interest can be approximated well by piecewise constant functions; see Wright and Ziegler (2017) for a recent survival implementation. An alternative is to introduce some structure that stabilizes the system. Introducing structure has the additional advantage that it allows to visualize, interpret, extrapolate and forecast the properties of the underlying data. In this paper, we concentrate on structured models. The smooth backfitting algorithm of Mammen et al. (1999) consideres the simplest nonparametric structure in the regression context - the additive structure. It has many theoretical and practical advantages to earlier approaches of regression backfitting.

The popular regression backfitting approach of Hastie and Tibshirani (1990a) are numerical iterating-procedures estimating one component given the estimates of the rest. In contrast, smooth backfitting is a direct projection of the data down onto the structured space of interest. This direct relationship between data and estimates gives a more solid grip on what is being estimated and the theoretical properties underlying it, see also Nielsen and Sperlich (2005) and Huang and Yu (2019). The purpose of this paper is to introduce smooth backfitting to the field of survival analysis and nonparametric smooth hazard estimation. While the additive structure is the most natural and most widely used in regression, the multiplicative structure seems more natural in hazard estimation. The omnipresent Cox regression model is a proportional hazard model (also known as Cox proportional hazard model) and many extensions and alternatives to the Cox regression model have been formulated in a multiplicative framework. We will therefore consider the multiplicative structure in this paper. It can be used in applications to test Cox regression or other proportional hazard models either visually or quantitatively. But this is beyond the scope of this paper. Multiplicative smooth backfitting is theoretically more challenging than additive smooth backfitting. The smooth backfitting multiplicative regression structure was analysed by Yu et al. (2008) as a special case of generalized additive models. Yu et al. (2008) showed that the multiplicative structure - in contrast to the simpler additive case - provides asymptotic theory with a number of non-trivial interactions between exposure available in different directions. Naturally, the asymptotics provided here for smooth backfitting of multiplicative hazards contains similar interactive components in the asymptotic theory. The survival projection introduced in this paper is different and less intuitive than the nonparametric regression considered in Mammen et al. (1999, 2001). As in density estimation, see Jones (1993), hazard estimation requires a projection of a dirac-delta-sequence related to the jumps of the counting process, see also Nielsen (1998) and Nielsen and Tanggaard (2001). We provide a simple algorithm first projecting the data down onto an unstructured estimator, and then further projecting

3

the unstructured estimator down onto the multiplicative space of interest. Our numerical algorithms are greatly simplified by a new principle of weighting the projection according to the final estimates.

We consider a multiplicatively structured proportional hazard model:

$$\alpha(t, z) = \alpha_0(t)\alpha_1(z_1) \cdots \alpha_d(z_d), \tag{1.1}$$

where α_k, $k = 0, \ldots, d$, are some smooth and positive one-dimensional functions and $z = (z_1, \ldots, z_d) \in \mathbb{R}^d$ are possibly time t-dependent covariates. We do not impose any further structural assumption on α_k. This is in contrast to the semiparametric Cox proportional hazard model where all components, with exception of the baseline hazard α_0, are assumed to take a log-linear shape. We do expect that smooth backfitting of proportional hazard models can be generalized to much the same way as the Cox proportional hazard model is generalized. As one recent example that could be interesting to treat as a smooth backfitting problem, see Hsu et al. (2018).

Estimators for model (1.1) can be categorized in four groups: (i) Therneau et al. (1990) and Grambsch et al. (1995) start with the Cox model and investigate smoothed residual plots; (ii) Hastie and Tibshirani (1990a,b), O'Sullivan (1988, 1993), Sleeper and Harrington (1990) and Huang (1999) consider splines via penalized partial likelihood; (iii) Linton et al. (2003), Honda (2005) build on marginal integration (Linton and Nielsen, 1995) and (iv) Lin et al. (2016) use kernel smoothers starting from a global partial likelihood criterion.

Lin et al. (2016) prove asymptotic efficiency of their estimator and they show by a detailed simulation study that their estimator outperforms the proposals in Huang (1999), Linton et al. (2003), Honda (2005). For this reason in this paper we take their estimator as benchmark. It has been argued that in additive regression models smooth backfitting is less affected by sparseness of high-dimensional data and by strong correlated covariables, see

4

Nielsen and Sperlich (2005). In this paper we will show that this also holds for our smooth backfitter in a multiplicatively structured proportional hazard model. For this purpose we study the smooth backfitter in settings which include challenging high dimensional data and highly correlated covariates. The smooth backfitting approach turns out to show a very good performance compared to all other estimators and to be very robust to both, high dimensional and correlated data. In particular in high-dimensional settings it strongly outperforms the approach of Lin et al. (2016) that already for low dimensions runs into instabilities and numerical problems. In this paper, we do not treat the problem of variable selection. It would in particular be interesting to investigate Lasso approaches or nonparametric tests on the significance of one component. For the development of tests our theory may be used as a first step.

The paper is structured as follows. Section 2 contains the mathematics of the underlying survival model. In Section 3 the smooth backfitting estimator is defined as a projection of unstructured hazard estimators. This is done for unstructured hazard estimators that can be written as a ratio of smooth occurrence and smooth exposure. An example is given by the local constant Nadaraya-Watson estimator. In Section 4 asymptotic properties are outlined for the smooth backfitting estimator. Details are explained in the Supplementary Material of this paper. Section 5 contains our finite sample study illustrating the strong performance of smooth backfitting. In Section 6 we consider a sophisticated version of in-sample forecasting generalising earlier approaches via our proportional hazard model, see Mammen et al. (2015), Hiabu et al. (2016) Lee et al. (2015, 2017, 2018). A smooth extension of the popular actuarial chain ladder model that is used in virtually all non-life insurance companies in the world while estimating outstanding liabilities. In-sample forecasting is possible because of the imposed multiplicative structure.

2 Aalen's multiplicative intensity model

We consider a counting process formulation satisfying Aalen's multiplicative intensity model. It allows for very general observations schemes. It covers filtered observations arising from left truncation and right censoring and also more complicated patterns of occurrence and exposure. In the next section we describe how to embed left truncation and right censoring into this framework. In contrast to Linton et al. (2003) we will hereby allow the filtering to be correlated to the survival time and be represented in the covariate process. We briefly summarize the general model we are assuming.

We observe n *iid* copies of the stochastic processes $(N(t), Y(t), Z(t))$, $t \in [0, R_0]$, $R_0 > 0$. Here, N denotes a right-continuous counting process which is zero at time zero and has jumps of size one. The process Y is left-continuous and takes values in $\{0, 1\}$ where the value 1 indicates that the individual is under risk. Finally, Z is a d-dimensional left-continuous covariate process with values in a hyperrectangle $\prod_{j=1}^{d}[0, R_j] \subset \mathbb{R}^d$. The multivariate process $((N_1, Y_1, Z_1), \ldots, (N_n, Y_n, Z_n))$, $i = 1, \ldots, n$, is adapted to the filtration $\mathcal{F}_t$ which satisfies les conditions habituelles (the usual conditions), see Andersen et al. (1993) (pp. 60). Now we assume that N_i satisfies Aalen's multiplicative intensity model, that is

$$\lambda_i(t) = \lim_{h \downarrow 0} h^{-1} E[N_i((t+h)-) - N_i(t-)| \ \mathcal{F}_{t-}] = \alpha(t, Z_i(t))Y_i(t). \tag{2.1}$$

The deterministic function $\alpha(t, z)$ is called hazard function and it is the failure rate of an individual at time t given the covariate $Z(t) = z$.

2.1 Left truncation and right censoring time as covariates

The most prominent example for Aalen's multiplicative intensity model is filtered observation due to left truncation and right censoring. We now show how to embed model (1.1)

with covariate, Z, possibly carrying truncation and censoring information into Aalen's mul-tiplicative intensity model. Every covariate coordinate can carry individual truncation infor-mation as long as it corresponds to left truncation. That is, we observe (T, Z) if and only if $(T, Z(T)) \in \mathcal{I}$, where the set $\mathcal{I}$ is compact and it holds that if $(t_1, Z(t_1)) \in \mathcal{I}$ and $t_2 \geq t_1$, then $(t_2, Z(t_2)) \in \mathcal{I}$, $a.s..$ The set $\mathcal{I}$ is allowed to be random but is independent of T given the covariate process Z. Furthermore, T can be subject to right censoring with censoring time C. We assume that also T and C are conditional independent given the covariate process Z. This includes the case where the censoring time equals one covariate coordinate. In con-clusion, we observe n iid copies of $(\widetilde{T}, Z^*, \mathcal{I}, \delta)$, where $\delta = \mathbb{1}(T^* < C)$, $\widetilde{T} = \min(T^*, C)$, and (T^*, Z^*) is the truncated version of (T, Z), i.e, (T^*, Z^*) arises from (T, Z) by conditioning on the event $(T, Z(T)) \in \mathcal{I}$.

Then, for each subject, $i = 1, \ldots, n$, we can define a counting process N_i as $N_i(t) = \mathbb{1}\left\{\widetilde{T}_i \leq t, \ \delta_i = 1\right\}$, with respect to the filtration $\mathcal{F}_{i,t} = \sigma\left(\left\{\widetilde{T}_i \leq s, \ Z_i^*(s), \ \mathcal{I}_i, \ \delta_i : \ s \leq t\right\} \cup \mathcal{N}\right)$, where $\mathcal{N}$ is a class of null-sets that completes the filtration. After straightforward computa-tions one can conclude that under the setting above, Aalen's multiplicative intensity model (2.1) is satisfied with

$$\alpha_z(t) = \alpha(t, z) = \lim_{h \downarrow 0} h^{-1}\mathrm{Pr}\{T_i \in [t, t+h) | \ T_i \geq t, \ Z_i(t) = z\},$$

$$Y_i(t) = \mathbb{1}\left\{(t, Z_i^*(t)) \in \mathcal{I}_i, \ t \leq \widetilde{T}_i\right\}.$$

3 Estimation

We tackle the problem in two steps. First the data are projected down onto an unstructured space resulting in an unstructured estimator of the $d+1-$dimensional hazard function $\alpha(t, z)$; see (1.1). In the second step, the unstructured estimator is projected further down onto the multiplicative space of interest resulting in $d + 1$ one-dimensional smooth backfitting

estimators of the multiplicative components, $(\alpha_0(t), \alpha_1(z_1), \ldots, \alpha_d(z_d))$. For the first step we assume to have an unstructured estimator with simple ratio of smoothed occurrence and smoothed exposure. Our theory will encompass all estimators with this simple ratio structure. We discuss in Section 4 and in the accompanied Supplementary Material that our estimation procedure works under quite general assumptions. In particular we do not need the unstructured estimator of the first step to be consistent. The final structured estimator circumvents the curse of dimensionality even if consistency is not assured in the first step. This is reassuring noting that the unstructured estimator will most probably have exponentially deteriorating performance with growing dimension d.

We introduce the notation $X_i(t) = (t, Z_i(t))$. We also set $x = (t, z)$, with coordinates $x_0 = t$, $x_1 = z_1, \ldots, x_d = z_d$, and write the hazard as $\alpha(t, z) = \alpha(x)$.

3.1 First step: The unstructured estimator

To estimate the components of the structured hazard in (3.2) below, we will need an unstructured pilot estimator of the hazard α first. We propose the local constant kernel estimator, $\widetilde{\alpha}^{LC}$. Its value in x is defined as

$$\widetilde{\alpha}^{LC}(x) = \lim_{\varepsilon \to 0} \arg\min_{\theta_0 \in \mathbb{R}} \sum_{i=1}^{n} \int \left\{ \frac{1}{\varepsilon} \int_{s}^{s+\varepsilon} \mathrm{d}N_i(u) - \theta_0 \right\}^2 \tag{3.1}$$
$$\times \kappa_n(X_i(s)) K_b(x - X_i(s)) Y_i(s) \, \mathrm{d}s.$$

In the following, we restrict ourselves to a multiplicative kernel, i.e., for , $(u_0, \ldots, u_d) \in \mathbb{R}^{d+1}$, $K(u_0, \ldots, u_d) = \prod_{j=0}^{d} k(u_j)$, and a one-dimensional bandwidth b with boundary correction $\kappa_n(x) = \prod_{j=0}^{d} \left(\int K_b(x_j - u_j) \, du_j \right)^{-1}$ and $K_b(u) = \prod_{j=0}^{d} b^{-1} k(b^{-1} u_j)$, where for simplicity of notation the bandwidth $b > 0$ does not depend on j. More general choices would have been possible with the cost of extra notation.

8

We get $\widetilde{\alpha}^{LC}(x) = \widehat{O}^{LC}(x)/\widehat{E}^{LC}(x)$, with smoothed occurrence and smoothed exposure given by

$$\widehat{O}^{LC}(x) = \sum_{i=1}^{n} \int \kappa_n(X_i(s)) K_b(x - X_i(s)) \mathrm{d}N_i(s),$$

$$\widehat{E}^{LC}(x) = \sum_{i=1}^{n} \int \kappa_n(X_i(s)) K_b(x - X_i(s)) Y_i(s) \mathrm{d}s.$$

Under standard smoothing conditions, if b is chosen of order $n^{-1/(4+d+1)}$, then the bias of $\widetilde{\alpha}^{LC}(x)$ is of order $n^{-2/(4+d+1)}$ and the variance is of order $n^{-4/(4+d+1)}$, which is the optimal rate of convergence in the corresponding regression problem, see Stone (1982). For an asymptotic theory of these estimators see Linton et al. (2003).

3.2   Second step: The structured smooth backfitting estimator

In this section we will project the unstructured estimator of the previous section down onto the multiplicative space of interest. Other choices that have a simple ratio structure of occurrence and exposure are possible. Due to filtering, observations are assumed to be only available on a subset of the full support, $\mathcal{X} \subseteq \mathcal{R} = \prod_{j=0}^{d}[0, R_j]$. Our estimators are restricted to this set and detailed assumptions on $\mathcal{X}$ and the data generating functions are given in the Supplementary Material. Our calculations simplify via a new principle we call solution-weighted minimization. We assume that we have the solution and use it strategically in the least squares weighting. While the definition is not explicit, it is made feasible by defining it as an iterative procedure. In the sequel we will assume a multiplicative structure of the hazard α, i.e.,

$$\alpha(x) = \alpha^* \prod_{j=0}^{d} \alpha_j(x_j), \tag{3.2}$$

where α_j, $j = 0, \ldots, d$, are some functions and α^* is a constant. For identifiability of the components, we make the following further assumption:

$$\int \alpha_j(x_j) w_j(x_j) \, \mathrm{d}x_j = 1, \quad j = 0, \ldots, d,$$

where the w_j's are some weight functions.

We also need the following notation:

$$F_t(z) = Pr\left(Z_1(t) \le z | \, Y_1(t) = 1\right), \quad y(t) = E[Y_1(t)].$$

By denoting $f_t(z)$ the density corresponding to $F_t(z)$ with respect to the Lebesgue measure, we also define $E(x) = f_t(z)y(t)$ and $O(x) = E(x)\alpha(x)$.

We define the estimators $\widehat{\alpha}^*$ and $\widehat{\alpha} = (\widehat{\alpha}_0, \ldots, \widehat{\alpha}_d)$ of the hazard components in (3.2) as solution of the following system of equations:

$$\widehat{\alpha}_k(x_k) = \frac{\int_{\mathcal{X}_{x_k}} \widehat{O}(x)\mathrm{d}x_{-k}}{\int_{\mathcal{X}_{x_k}} \widehat{\alpha}^* \prod_{j \ne k} \widehat{\alpha}_j(x_j)\widehat{E}(x)\mathrm{d}x_{-k}}, \quad k = 0, \ldots, d, \tag{3.3}$$

$$\int \widehat{\alpha}_k(x_k) w_k(x_k) \, \mathrm{d}x_k = 1, \quad k = 0, \ldots, d. \tag{3.4}$$

Here $\mathcal{X}_{x_k}$ denotes the set $\{(x_0, \ldots, x_{k-1}, x_{k+1}, \ldots, x_d)| \, (x_0, \ldots, x_d) \in \mathcal{X}\}$, and $x_{-k} = (x_0, \ldots, x_{k-1}, x_{k+1}, \ldots, x_d)$. Furthermore, $\widehat{E}$ and $\widehat{O}$ are some full-dimensional estimators of E and O – not necessarily the one provided in the previous section. We will discuss in the Supplementary Material that the system above has a solution with probability tending to one. In the next section and in the Supplementary Material we will show asymptotic properties of the estimator. We will see that we do not require that the full-dimensional estimators $\widehat{E}$ and $\widehat{O}$ are consistent. We will only need asymptotic consistency of marginal averages of the estimators. This already highlights that our estimator efficiently circumvents

10

the curse of dimensionality.

In practice, system (3.3) can be solved by the following iterative procedure:

$$\widehat{\alpha}_k^{(r+1)}(x_k) = \frac{\int_{\mathcal{X}_{x_k}} \widehat{O}(x)\mathrm{d}x_{-k}}{\int_{\mathcal{X}_{x_k}} \prod_{j=0}^{k-1} \widehat{\alpha}_j^{(r+1)}(x_j) \prod_{j=k+1}^{d+1} \widehat{\alpha}_j^{(r)}(x_j) \widehat{E}(x)\mathrm{d}x_{-k}}, \quad k = 0, \ldots, d \qquad (3.5)$$

After a finite number of cycles or after a termination criterion applies, the last values of $\widehat{\alpha}_k^{(r+1)}(x_k)$, $k = 0, ..., d$, are multiplied by a factor such that the constraint (3.4) is fulfilled. This can always be achieved by multiplication with constants. This gives the backfitting approximations of $\widehat{\alpha}_k(x_k)$ for $k = 0, ..., d$.

3.3   Interpretation as direct projection

The strength of our smooth backfitting estimator is that it can be motivated directly from a least squares criterium on the data without ad-hoc adjustment. The estimator $\widehat{\alpha}$ can be motivated as solution of

$$\lim_{\varepsilon \to 0} \arg\min_\theta \sum_{i=1}^n \int \int \left\{ \frac{1}{\varepsilon} \int_s^{s+\varepsilon} \mathrm{d}N_i(u) - \theta(x) \right\}^2 \times K_b(x - X_i(s)) Y_i(s) \, \mathrm{d}s \, \mathrm{d}\nu(x), \quad (3.6)$$

where θ runs over some space of smooth multiplicative functions of the form $\theta = \prod_{j=0}^d \theta_j(x_j)$. To see this, consider the estimator $\overline{\alpha} = (\overline{\alpha}^*, \overline{\alpha}_0, \ldots, \overline{\alpha}_d)$ that minimizes

$$\arg\min_{\overline{\alpha}} \int_{\mathcal{X}} \left\{ \widetilde{\alpha}^{LC}(x) - \overline{\alpha}^* \prod_{j=0}^d \overline{\alpha}_j(x_j) \right\}^2 w(x)\mathrm{d}x. \qquad (3.7)$$

For $\nu(x) = w(x)/\widehat{E}^{LC}(x)dx$ the solution of (3.7) is exactly (3.6). With that choice, we get

$$\overline{\alpha}^* = \frac{\int_{\mathcal{X}} \widetilde{\alpha}(x) \prod_{j=0}^d \overline{\alpha}_j(x_j) w(x)\mathrm{d}x}{\int_{\mathcal{X}} \left\{ \prod_{j=0}^d \overline{\alpha}_j(x_j) \right\}^2 w(x)\mathrm{d}x},$$

11

and $(\overline{\alpha}_0, \ldots, \overline{\alpha}_d)$ can be described via the backfitting equation

$$\overline{\alpha}_k(x_k) = \frac{\int_{\mathcal{X}_{x_k}} \widetilde{\alpha}(x) \prod_{j \neq k} \overline{\alpha}_j(x_j) w(x) \mathrm{d}x_{-k}}{\int_{\mathcal{X}_{x_k}} \overline{\alpha}^* \left\{ \prod_{j \neq k} \overline{\alpha}_j(x_j) \right\}^2 w(x) \mathrm{d}x_{-k}}, \quad k = 0, \ldots, d. \tag{3.8}$$

The asymptotic variance of kernel estimators of α is proportional to $\alpha(x)/E(x)$, see e.g. Linton and Nielsen (1995). This motivates the choice $w(x) = E(x)/\alpha(x)$. However, this choice is not possible because $E(x)$ and $\alpha(x)$ are unknown. One could use $w(x) = \check{E}(x)/\check{\alpha}(x)$ where $\check{E}(x)$ and $\check{\alpha}(x)$ are some pilot estimators of E and α. We follow another idea and we propose to weight the minimization (3.7) with its solution. We choose

$$w(x) = \frac{\widehat{E}(x)}{\prod_i \widehat{\alpha}_i(x)}, \tag{3.9}$$

and heuristically, by putting $\overline{\alpha}_j = \widehat{\alpha}_j$ and by plugging (3.9) into (3.8), we get (3.3).

4 Asymptotic properties of the smooth backfitter of multiplicative hazards

In the Supplementary Material we show that $\widehat{\alpha}_j$ converges to the true α_j with optimal one dimensional nonparametric rate $n^{-2/5}$, given that the bandwidth b is chosen of order $n^{-1/5}$. This means in particular that the asymptotic rate does not depend on the dimension d. Under regularity assumptions, in Theorem 3 in the Supplementary Material we show that

$$n^{2/5}\{(\widehat{\alpha}_j - \alpha_j)(x_j) - \alpha_j(x_j)B_j(x_j))\} \to \mathrm{N}(0, \alpha_j^2(x_j)\sigma_j^2(x_j)),$$

where $B_j(x_j)$ can be formally defined as the j-th component of a projection of the bias of the unconstrained estimator onto the multiplicative space and $\sigma_j^2(x_j)$ is the variance of a

weighted average of the unconstrained estimator where the other components are integrated out.

In the simulation study of the next section we show that estimation seems to work well even when d is of similar order as n. This is, we believe, a major strength of our smooth backfitting estimator.

A few notes on the proof. The estimator $\widehat{\alpha}_j$ is defined as solution of a nonlinear operator equation. Asymptotic properties are derived by approximating the estimators of this equation by a linear equation that can be interpreted as one equation that arises in nonparametric additive regression models (Mammen et al., 1999), and then one shows that the solution of the linear equation approximates the estimation error in the multiplicative model. The linear equation and its solution is well understood from the theory of additive models. This is our essential step to arrive at an asymptotic understanding of our estimator $\widehat{\alpha}_j$. Assumptions are of standard nature in marker dependent hazard papers and they can be verified for the local constant estimators we are interested in, see in particular Nielsen and Linton (1995), Nielsen (1998) and Linton et al. (2003) for related arguments. However, the conditions are formulated more general and they are not restricted to the local constant smoothers. They are not even tight to kernel smoothers. Any smoother could be used as long as it obeys the structure of being a ratio of a smoothed occurrence and a smoothed exposure.

5   Simulation study

In this section we present detailed simulations of the smooth backfitter. The Supplementary Material contains additional simulation results for the setting discussed in Honda (2005) and Lin et al. (2016). In these latter settings, the performance of the smooth backfitter is similar to the estimator of Lin et al. (2016). And both estimators outperform the estimators considered in Honda (2005).

In the simulation study below, and with additional results in The Supplementary Material, we compare our estimator with the estimator of Lin et al. (2016). The models of these simulations contain high-dimensional and correlated covariates and in particular, we will show that our estimator – in contrast to Lin et al. (2016) – works in high dimensions where the dimension of the covariates is of similar order as the sample size and also in cases with higher correlation between the covariates. To highlight the impact of increasing dimension we will consider the dimensions $d = 2$ and $d = 9$. Further results for dimensions $d = 30$ and $d = 99$ can be found in the Supplementary Material. The low dimensional setting, $d = 2$, with uncorrelated covariates is similar to the setting of Honda (2005) and Lin et al. (2016), and the difference between our estimator and the estimator of Lin et al. (2016) are indeed marginal in this case. However, below and in the Supplementary Material we show that this changes drastically with increasing dimension and/or correlation.

5.1 The setting

Since the estimator of Lin et al. (2016) is based on a partial likelihood approach, it does not estimate the baseline hazard, α_0. We will consider the sub-model

$$\alpha(x) = \prod_{k=0}^{d} \alpha_k(x_k) = \exp\left\{\sum_{k=1}^{d} \eta_k(x_k)\right\},$$

i.e., we assume a constant baseline hazard, $\alpha_0 \equiv 1$. More specifically, we assume that the survival times T_i follow an exponential distribution with parameter value $\alpha(X_i)$, $(i = 1, \ldots, n)$. We add right censoring with censoring variables C_i that follow an exponential distribution with parameter $\frac{4}{7}\alpha(X_i)$, $(i = 1, \ldots, n)$. We will compare the estimators for $\eta_k, k = 1, \ldots, d$. Our proposed estimator is derived as $\widehat{\eta}_k^{SBF} = \log(\widehat{\alpha}_k)$ and we compare it to, $\widehat{\eta}_k^{\text{Lin et al.}}$, proposed in Lin et al. (2016).

14

We used the following two models:

$$\text{Model 1: } \eta_k(z_k) = \begin{cases} -z_k & \text{if k is odd,} \\ \\ 2z_k & \text{if k is even.} \end{cases}$$

$$\text{Model 2: } \eta_k(z_k) = \begin{cases} 2\sin(\pi z_k) & \text{if k is odd,} \\ \\ 2z_k & \text{if k is even.} \end{cases}$$

The covariates $(Z_{i1}, \ldots, Z_{id})$ are generated as follows. We first simulate $(\widetilde{Z}_{i1}, \ldots, \widetilde{Z}_{id})$ from a d-dimensional multi-normal distribution with mean equal 0 and $\text{Corr}(Z_{ij}, Z_{il}) = \rho$ if $j \neq l$, else 1. Afterwards we set

$$Z_{ik} = 2.5\pi^{-1}\arctan(\widetilde{Z}_{ik}).$$

This has been independently repeated for every individual $i = 1, \ldots, n$. After trying several bandwidths, if not said otherwise, we present the results for a bandwidth $b = 0.3$ for both estimators in every model. As kernel function k, we used the Epanechnikov kernel. Performance is measured via the integrated squared error evaluated at the observed points:

$$ISE_k = n_{obs}^{-1} \sum_{i_{obs}} \left(\eta_k(Z_{i_{obs}k}) - \widehat{\eta}_k(Z_{i_{obs}k}) \right)^2,$$

$$ISE_{odd} = \{ISE_k | k = odd\}, \quad ISE_{even} = \{ISE_k | k = even\},$$

where i_{obs} runs over all observed individuals and $n_{obs} < n$ is the number of observed individuals. The next two subsections discuss the outcomes of the simulations for $d = 2$ and $d = 9$.

5.2 Dimension $d = 2$

In low dimensions both estimators perform satisfactorily. Already in dimension $d = 2$, for Model 1, it occurred that the algorithm of Lin et al. (2016) stopped before the final calculation of the estimator. This happened for three simulation samples out of 1200 (6 settings $\times 200$ samples each) runs. The algorithm stopped at a step where a matrix has to be inverted that is nearly singular. In contrast, our estimator does not need to invert matrices and we did not have any singularity or convergence issues for our estimator. Performance-wise, we refer to Figure 2 displaying boxplots of the integrated squared errors of both estimators. The two estimators perform similar with a small advantage towards our smooth backfitting estimator. The smooth backfitting estimator performs especially better for the smaller sample size $n = 200$. While Lin et al. (2016) performed better in Model 2 with no correlation, the performance of the smooth backfitting estimator seems more stable when correlation is added. Figure 1 shows the 200 sample estimates of the first component α_1 of Model 2 with $\rho = 0.8$. The estimator of Lin et al. (2016) struggles especially at the boundaries and this is more pronounced with a small bandwidth (left panel). If bandwidth is increased (right panel), the estimate of Lin et al. (2016) seems over-smoothed and it is not able to replicate the full magnitude of the local extrema at -0.5 and -0.5.

5.3 Dimension $d = 9$

When the dimension d is increased to 9, the estimator of Lin et al. (2016) breaks down considerably more often than in the case $d = 2$, i.e., in 59 (4+5+48+1+1 out of 1200) cases in Model 1 and 9 times in Model 2; see Table 1. Note that in the more extreme cases of $d = 99$, provided in the Supplementary Material, nearly all simulations of Lin et al. (2016) (780 and 607 out of 800 for Model 1 and Model 2, respectively) broke down. In contrast, our estimator converged in all cases considered. Performance-wise we refer to Figure 4
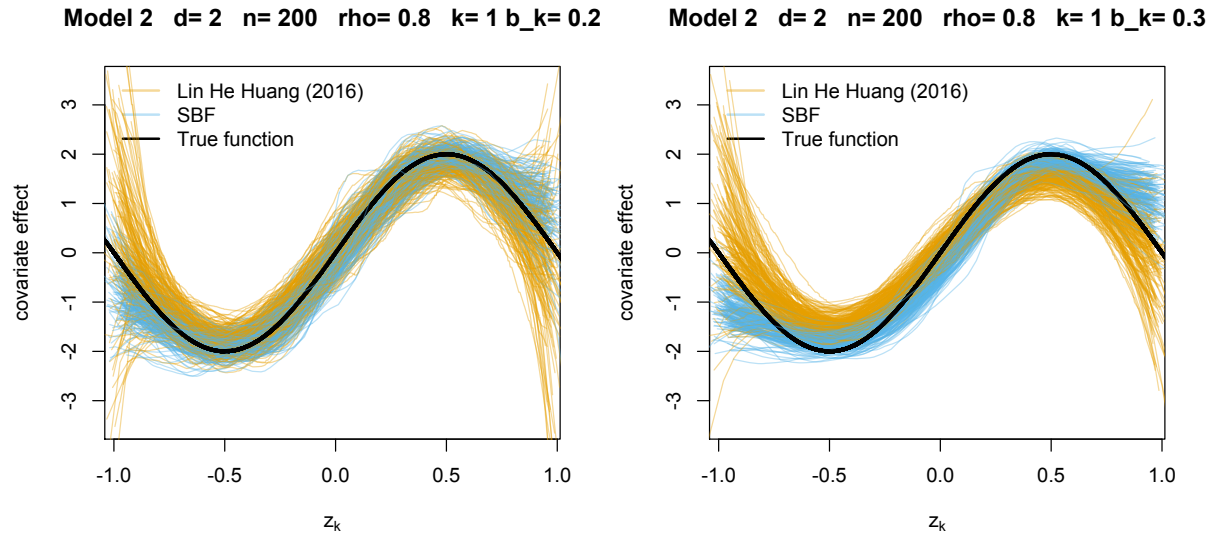
16

Figure 1: Estimates of α_1 from 200 simulations. Left panel uses a bandwidth of 0.2 and right panel a bandwidth of 0.3.

displaying boxplots of the integrated squared errors of both estimators. We see that for $d = 9$ the smooth backfitter performs better in every setting. The better performance is more pronounced when more correlation is present. Figure 3 shows 200 sample estimates of the first component α_1 of Model 2 with $\rho = 0.8$. The results are similar as in the case $d = 2$, but now more pronounced: The estimator of Lin et al. (2016) struggles at the boundaries and if bandwidth is increased (right panel), the estimate of Lin et al. (2016) seems too smooth and is not able to replicate to full magnitude of the local extrema at -0.5 and -0.5.

6 In-sample forecasting of outstanding loss liabilities

The so-called chain ladder method is a popular approach to estimate outstanding liabilities. It started off as a deterministic algorithm, and it is used today for almost every single insurance policy over the world in the business of non-life insurance. In many developed
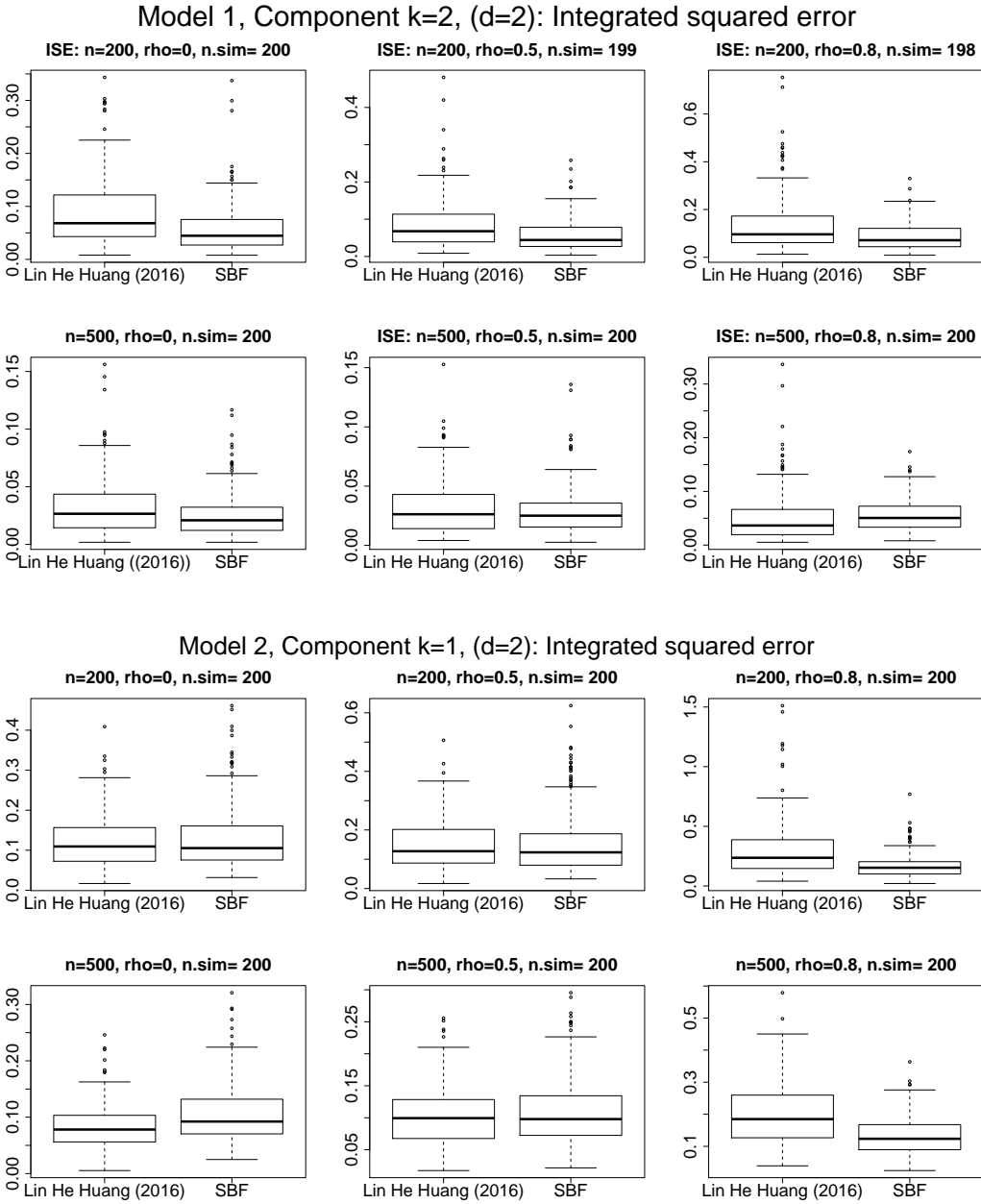
Figure 2: Boxplots of the integrated squared errors. Simulations are taken out where the algorithm for the calculation of the estimator in Lin et al. (2016) stopped without calculation of all values of the estimator. The value $n.sim$ is the number of the remaining simulations, i.e., 200 minus number of break downs.

Number of breakdowns in Lin et al. (2016) for $d = 9$					
(out of 200 simulations)					
	Model 1			Model 2		
	$\rho = 0$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0$	$\rho = 0.5$	$\rho = 0.8$
n=200	4	5	48	0	0	9
n=500	0	1	1	0	0	0

Table 1: Number of breakdowns in the algorithm of Lin et al. (2016) out of 200 simulations for dimension $d = 9$.
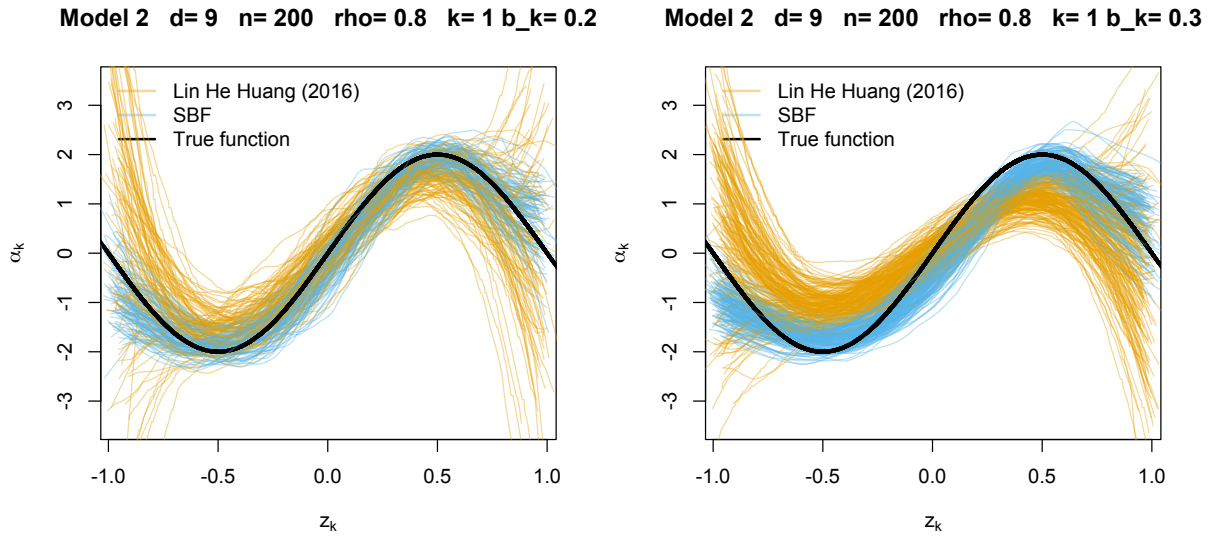

Figure 3: Estimates of α_1 from 200 simulations. Left panel uses a bandwidth of 0.2 and right panel a bandwidth of 0.3.

countries, the non-life insurance industry has revenues amounting to around 5%. It is therefore comparable to - but smaller than - the banking industry. In every single product sold, the chain ladder method (because actuaries hardly use other methods) comes in, estimating the outstanding liabilities that eventually aggregate to the reserve - the single biggest number of most non-life insurers balance sheets. The insurers liabilities often amount to many times the underlying value of the company. In Europe alone those outstanding liabilities are
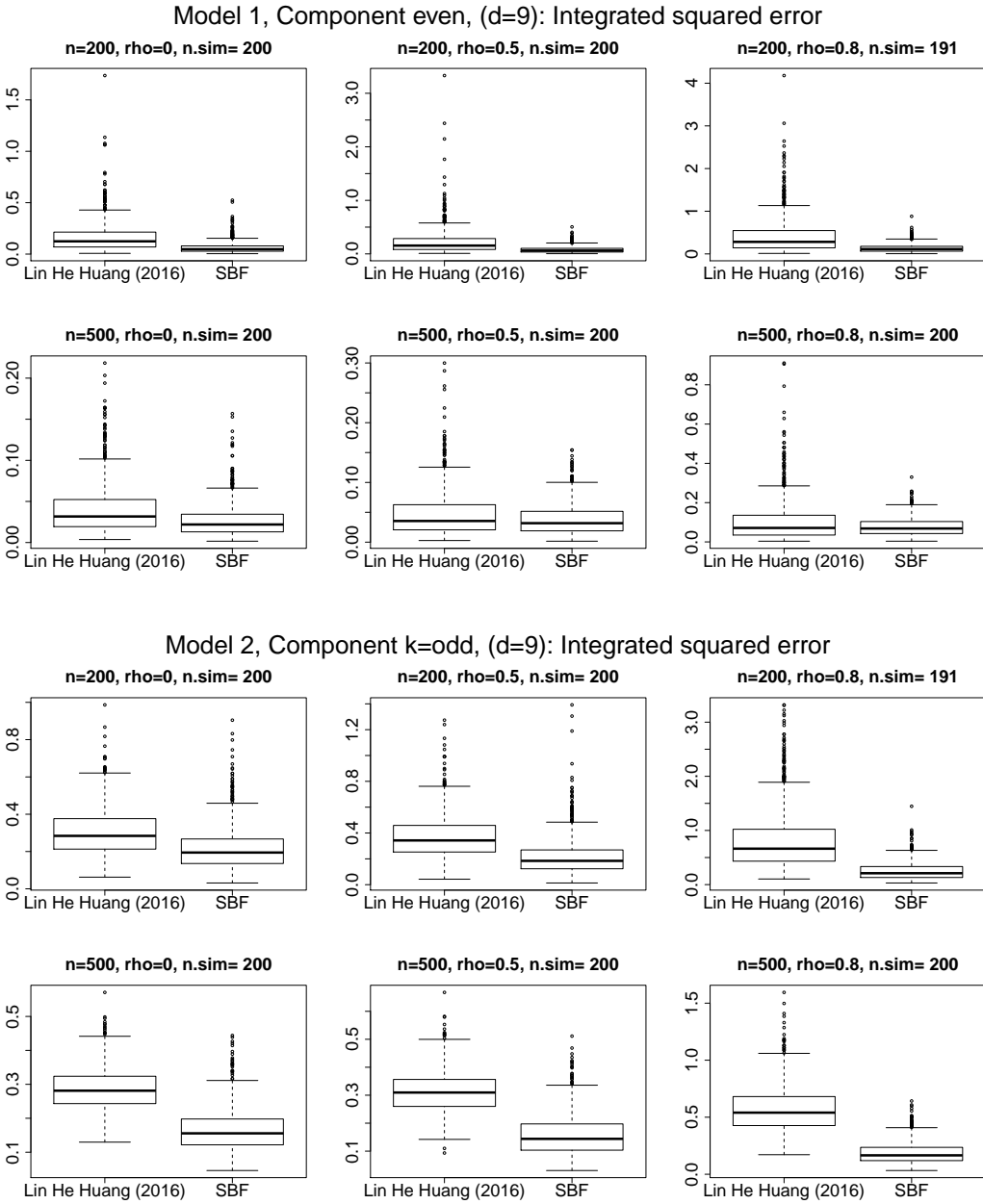
Figure 4: Boxplots of the integrated squared errors. Simulations are taken out where the algorithm for the calculation of the estimator in Lin et al. (2016) stopped without calculation of all values of the estimator. The value *n.sim* is the number of the remaining simulations, i.e., 200 minus number of break downs.

estimated to accumulate to around € 1trn. It is therefore of obvious importance that this estimate is not too far from the best possible estimate. We describe in this section how the methodology introduced in this paper can be applied to provide a solution to this challenging problem.

We analyze reported claims from a motor business line in Cyprus. The same data set has been used by Hiabu et al. (2016) and it consists of the number of claims reported between 2004 and 2013. During these 10 years (3654 days), $n = 58180$ claims were reported. The data are given as $\{(T_1, Z_1), \ldots, (T_n, Z_n)\}$, where Z_i denotes the underwriting date of claim i, and T_i the time between underwriting date and the date of report of a claim in days, also called reporting delay. Hence, in the notation of the previous sections, the covariate underwriting date, $Z(t) = Z$, does not depend on time and has dimension $d = 1$. The data exist on a triangle, with $T_i + Z_i \leq 31$ December 2013 $= R_0$, which is a subset of the full support $\mathcal{R} = [0, R_0]^2$ ($0 = 1$ January 2004). The aim is to forecast the number of future claims from contracts written in the past which have not been reported yet. Figure 5 shows the observed data that lie on a triangle, while the forecasts are required on the triangle that added to the first completes a square. Here it is implicitly assumed that the maximum reporting delay of a claim is 10 years. Actuaries call this assumption that the triangle is fully run off. In our data set, this is a reasonable assumption looking at Figure 5.

The classical chain ladder method is able to provide a simple solution to the above problem. Recently, Martínez-Miranda et al. (2013) have pointed out that this method can be viewed as a multiplicative density method: the original, un-truncated random variable (T, Z) having density $f(t, z) = f_1(t)f_2(z)$; and the authors suggested to embed the method in a more standard mathematical statistical vocabulary to engage mathematical statisticians in future developments. In particular, Martínez-Miranda et al. (2013) showed that one could consider the traditional chain ladder estimator as a multiplicative histogram in a continuous framework, and presented an alternative by projecting an unconstrained local linear density
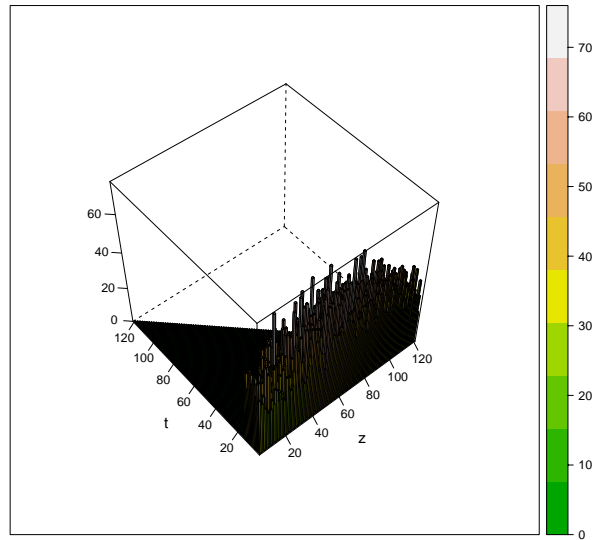
21

Figure 5: Histogram of claim numbers of a motor business line between 2004 and 2013. Axis z represents the underwriting time (in months) and axis t the reporting delay (in months).

down onto a multiplicative subspace. This approach was called continuous chain ladder and it has been further analyzed by Mammen et al. (2015); Lee et al. (2015, 2017), providing full asymptotic theory of the underlying density components. A related approach by Hiabu et al. (2016); Hiabu (2017) proposes to transform the two-dimensional multiplicative continuous chain ladder problem to two one-dimensional continuous hazard estimation problems via an elegant time-reverting trick. The application considered in this paper generalizes the most important of these reversed hazards to a two-dimensional multiplicatively structured hazard. In this way the continuous chain ladder is improved and generalized allowing more flexibility for the estimation of outstanding liabilities in the insurance business.

In Hiabu et al. (2016) it is assumed that T and Z are independent, which means that the underwriting date of a claim has no effect on the reporting delay. We are not going to impose such a strong restriction. In order to discuss the independence assumption, consider Figure 6. The points in the plots are derived by first transforming the data into a triangle
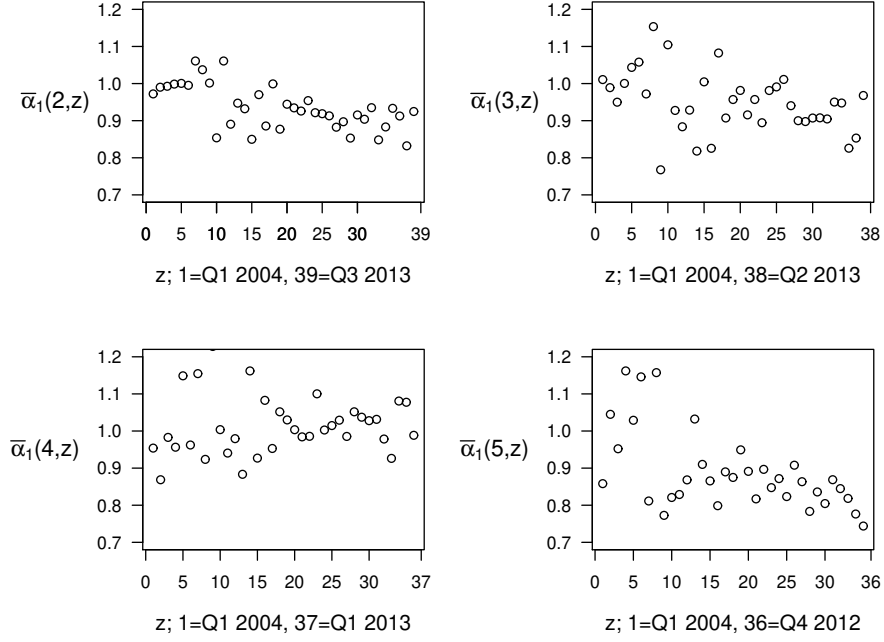
22

Figure 6: Scaled quarterly hazard rates of the first four development quarters.

with dimension 3654×3654,

$$\mathcal{N}_{t,z} = \sum_{i=1}^{n} I\big(T_i = t,\ Z_i = z\big), \quad (t,z) \in \{1, 2, \ldots, 3654\}^2, t + z \leq 3654$$

and then aggregating the data into a quarterly triangle, $(\mathcal{N}_{t,z}^Q)$, with dimension 40×40, see also Hiabu (2017). Then, for $t = 2, \ldots, 5$, one derives the quarterly hazard rate as ratio of occurrence and exposure, $\overline{\alpha}(t,z) = \mathcal{N}_{t,z}^Q / \sum_{l=1}^{z} \mathcal{N}_{t,l}^Q$. These values are then scaled by an eye-picked norming factor, $\overline{\alpha}_0(t), t = 2, \ldots, 5$, letting $\overline{\alpha}(t,z)$ start at around 1 as a function of z with fixed t. The final values, $\overline{\alpha}_1(t,z) = \overline{\alpha}_0(t)\overline{\alpha}(t,z)$, are displayed in Figure 6. We only show plots for $t \leq 5$ since almost all claims are reported after five quarters.

If the independence assumption of Hiabu et al. (2016) is satisfied, the points should lie around a horizontal line in each plot. If the multiplicative hazard assumption of this paper is

satisfied, then any smooth shape is allowed, but all four graphs must be equal after correction for noise. This is because under the model which will be defined below, the graphs, $\overline{\alpha}_1(\cdot, \cdot)$, with the first component fixed, mimic a quarterly version of α_1.

Inspecting the four plots, one can argue to see a negative drift of similar magnitude in each graph, the values decaying from around 1 to 0.8. This indicates that the approach of this paper should give a better fit to the data compared to the model of Hiabu et al. (2016).

From this discussion we now continue with embedding our observations in the proportional hazard framework. Afterwards we will show how the hazard estimate can be used to forecast the number of outstanding claims. First note that we cannot apply the approach of this paper directly, since in this application we only observe T if $T \leq R_0 - Z$, which is a right truncation. Analogue to Hiabu et al. (2016), we transform the random variable T to $T^R = R_0 - T$. This has the result that the right truncation truncation becomes left truncation, $T^R \geq Z$. Thus we consider the random variable T^R as our variable of interest. With the notation considered in Section 2.1, we now have $T = T^R, d = 1, Z(t) = Z, \delta = 1, \mathcal{I} = \{(t^R, z) \in \mathcal{R} | 0 \leq z \leq t^R\}$. We conclude that the counting process $N_i(t^R) = \mathbb{1}\left\{T_i^R \leq t^R\right\}$, satisfies Aalen's multiplicative intensity model with respect to the filtration given in Section 2.1 and

$$\alpha_z(t^R) = \alpha(t^R, z) = \lim_{h \downarrow 0} h^{-1} \Pr\{T^R \in [t^R, t^R + h] | \ T^R \geq t^R, \ Z = z\},$$

$$Y_i(t^R) = \mathbb{1}\left\{(t^R, Z_i) \in \mathcal{I}, \ t^R \leq T_i^{R,*}\right\}.$$

In Section 3.1 we suggested a local constant estimator as pilot. In this application we prefer a local linear estimator because we anticipate high mass at the boundaries. The local linear estimator (Nielsen, 1998) has an automated boundary correction. We estimate the unstructured hazard as ratios of occurrence and exposure from a local linear estimation

(Gámiz et al., 2013):

$$\widehat{O}^{LL}(x) = n^{-1} \sum_{i=1}^{n} \int \left\{ 1 - (x - X_i(s))D(x)^{-1}c_1(x) \right\} K_b(x - X_i(s))\mathrm{d}N_i(s),$$

$$\widehat{E}^{LL}(x) = n^{-1} \sum_{i=1}^{n} \int \left\{ 1 - (x - X_i(s))D(x)^{-1}c_1(x) \right\} K_b(x - X_i(s))Y_i(s)\mathrm{d}s,$$

where the components of the $(d+1)$ - dimensional vector c_1 are

$$c_{1j}(x) = n^{-1} \sum_{i=1}^{n} \int K_b(x - X_i(s))(x_j - X_{ij}(s))Y_i(s)ds, \quad j = 0, \ldots, d,$$

and the entries (d_{jk}) of the $(d+1) \times (d+1)$ - dimensional matrix $D(x)$ are given by

$$d_{jk}(x) = n^{-1} \sum_{i=1}^{n} \int K_b(x - X_i(s))(x_j - X_{ij}(s))(x_k - X_{ik}(s))Y_i(s)\mathrm{d}s.$$

Note that we have $\mathcal{X} = \mathcal{I}$. The components of the multiplicative conditional hazard are then computed as in (3.5). These estimators require the choice of the bandwidth parameter, which was assumed to be scalar in order to simplify the notation in this paper. In this application we generalize this restriction allowing for different smoothing levels in each dimension, namely reporting delay and underwriting time. The bandwidth parameter is then a vector $b = (b_0, b_1)$ and we estimate it using cross-validation, see further details in the Supplementary Material. To alleviate the computational burden of cross-validation we aggregated the data triangle $\mathcal{N}_{t,z}$ considering bins of two days when applying a discrete version of the estimators described in the Supplementary Material. After several trials we run the cross-validation minimization over $b_0 \in \{1300, 1400, 1500, 1600, 1700, 1800\}$ and $b_1 \in \{2, 3, 4, 5\}$. The cross-validated bandwidth components were $b_0 = 1600$ and $b_1 = 3$ (unit=2 days).

The results of the estimation procedure are given in Figure 7 and Figure **??**. Figure 7 shows the estimated components of the multiplicatively structured hazard estimator. The
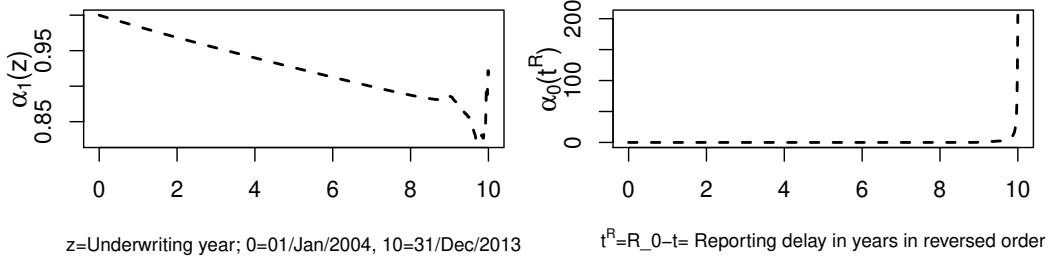
25

Figure 7: The estimated multiplicative hazard components .

function α_2 which captures the underwriting date effect, seems, up to a possible boundary effect, linear. This suggests that a semiparametric approach with nonparametric baseline hazard and linear covariate effect might be suitable. While we do not investigate this point further in our data illustration, this particular case illustrates nicely how our nonparametric approach could be employed for model selection.

Finally the total number of outstanding claims, i.e. the reserve, can be estimated as

$$\text{Reserve} = \sum_{i=1}^{n} \frac{\int_{R_0-Z_i}^{R_0} \widehat{f}_{Z_i}(t)\mathrm{d}t}{\int_{0}^{R_0-Z_i} \widehat{f}_{Z_i}(t)\mathrm{d}t}, \quad \widehat{f}_z(t) = \widehat{\alpha}_0(R_0-t)\widehat{\alpha}_1(z)\exp\left\{-\int_{0}^{R_0-t} \widehat{\alpha}_0(s)\widehat{\alpha}_1(z)\mathrm{d}s\right\}.$$

Note that $\widehat{f}_z(t)$ is an estimator of the conditional density of the survival time T. The reserve can be also decomposed further to provide the 'cash flow' of the next periods. If the future is divided into M periods, each of them with length $\delta = R_0/M$, then the amount of claims forthcoming in the ath $(a = 1, 2, \ldots, M)$ period can be estimated by

$$\text{Reserve}_P(a) = \sum_{i=1}^{n} \frac{\int_{(R_0-Z_i+a\delta-1)\wedge R_0}^{(R_0-Z_i+a\delta)\wedge R_0} \widehat{f}_{Z_i}(t)\mathrm{d}t}{\int_{0}^{R_0-Z_i} \widehat{f}_{Z_i}(t)\mathrm{d}t}.$$

Table 2 shows the estimated number of of outstanding claims for future quarters. We compare the approach of this paper with the results derived by Hiabu et al. (2016) and the traditional chain ladder method. The two latter approaches have in common that they

26

Table 2: Number of outstanding claims for future quarters; $1 = 2014\,Q1, \ldots, 39 = 2022\,Q3$. The backfitting approach in this paper (SBF) is compared with the chain ladder method (CLM) and the approach in Hiabu et al. (2016).

Future quarter	1	2	3	4	5	6	7	8	9	10	11	$12-39$	Tot.
Hiabu et al. (2016)	970	684	422	166	14	5	3	2	1	1	1	0	2270
CLM	948	651	387	148	12	5	3	2	1	1	1	0	2160
SBF	872	621	400	130	53	7	4	3	2	1	1	1	2193

assume independence between underwriting date, Z, and reporting delay, T. We see that while all approaches estimate a similar total number of outstanding claims (reserve), those two approaches have distributions over the quarters that are very different from the results obtained by the method proposed in this paper. It seems that the violation of the independence assumption has not a big influence on the reserve, since it balances the different development patterns arising from different periods out. However, the problem becomes quite serious if one is interested in more detailed estimates like the cash flow.

References

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

Breiman, L. (2001). Random forests. *Machine learning 45*, 5–32.

Gámiz, M. L., L. Janys, M. D. Martínez-Miranda, and J. P. Nielsen (2013). Bandwidth selection in marker dependent kernel hazard estimation. *Comput. Stat. Data An. 68*, 155–169.

Grambsch, P. M., T. M. Therneau, and T. R. Fleming (1995). Diagnostic plots to reveal functional form for covariates in multiplicative intensity models. *Biometrics 51*, 1469–1482.

Hastie, T. J. and R. J. Tibshirani (1990a). *Generalized Additive Models*. London: Chapman and Hall.

Hastie, T. J. and R. J. Tibshirani (1990b). Exploring the nature of covariate effects in the proportional hazards model. *Biometrics 46*, 1005–1016.

Hiabu, M. (2017). On the relationship between classical chain ladder and granular reserving. *Scand. Actuar. J. 2017*, 708–729.

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

Honda, T. (2005). Estimation in additive cox models by marginal integration. *Annals of the Institute of Statistical Mathematics 57*, 403–423.

Hsu, L., M. Gorfine, and D. Zucker (2018). On estimation of the hazard function from population-based case–control studies. *Journal of the American Statistical Association 113*, 560–570.

Huang, J. (1999). Efficient estimation of the partly linear additive cox model. *The Annals of Statistics 27*, 1536–1563.

Huang, L.-S. and C.-H. Yu (2019). Classical backfitting for smooth-backfitting additive models. *Journal of Computational and Graphical Statistics 28*, 386–400.

Jones, M. C. (1993). Simple boundary correction for kernel density estimation. *Statistics and computing 3*, 135–146.

Lee, Y. K., E. Mammen, J. P. Nielsen, and B. U. Park (2015). Asymptotics for in-sample density forecasting. *Ann. Stat. 43*, 620–651.

Lee, Y. K., E. Mammen, J. P. Nielsen, and B. U. Park (2017). Operational time and in-sample density forecasting. *Ann. Stat. 45*, 1312–1341.

Lee, Y. K., E. Mammen, J. P. Nielsen, and B. U. Park (2018). In-sample forecasting: A brief review and new algorithms. *ALEA-Latin American Journal of Probability and Mathematical Statistics 15*, 875–895.

Lin, H., Y. He, and J. Huang (2016). A global partial likelihood estimation in the additive cox proportional hazards model. *J. Statist. Plann. Inference 169*, 71–87.

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

Linton, O. B., J. P. Nielsen, and S. Van de Geer (2003). Estimating multiplicative and additive hazard functions by kernel methods. *Ann. Stat. 31*, 464–492.

Mammen, E., O. B. Linton, and J. P. Nielsen (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Ann. Stat 27*, 1443–1490.

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

Mammen, E., M. D. Martínez-Miranda, and J. P. Nielsen (2015). In-sample forecasting applied to reserving and mesothelioma. *Insurance Math. Econom. 61*, 76–86.

Martínez-Miranda, M. D., J. P. Nielsen, S. Sperlich, and R. Verrall (2013). Continuous chain ladder: Reformulating and generalising a classical insurance problem. *Expert. Syst. Appl. 40*, 5588–5603.

Nielsen, J. P. (1998). Marker dependent kernel hazard estimation from local linear estimation. *Scand. Actuar. J. 1998*, 113–124.

Nielsen, J. P. and O. B. Linton (1995). Kernel estimation in a non-parametric marker dependent hazard model. *Ann. Stat. 23*, 1735–1748.

Nielsen, J. P. and S. Sperlich (2005). Smooth backfitting in practice. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*, 43–61.

Nielsen, J. P. and C. Tanggaard (2001). Boundary and bias correction in kernel hazard estimation. *Scand. J. Stat. 28*, 675–698.

O'Sullivan, F. (1988). Nonparametric estimation of relative risk using splines and cross-validation. *SIAM Journal on Scientific and Statistical Computing 9*, 531–542.

O'Sullivan, F. (1993). Nonparametric estimation in the cox model. *The Annals of Statistics 21*, 124–145.

Sleeper, L. A. and D. P. Harrington (1990). Regression splines in the cox model with application to covariate effects in liver disease. *Journal of the American Statistical Association 85*(412), 941–949.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Stat. 10*, 1040–1053.

Therneau, T. M., P. M. Grambsch, and T. R. Fleming (1990). Martingale-based residuals for survival models. *Biometrika 77*, 147–160.

Wright, M. N. and A. Ziegler (2017). ranger: A fast implementation of random forests for high dimensional data in c++ and r. *Journal of Statistical Software, Articles 77*, 1–17.

Yu, K., B. U. Park, and E. Mammen (2008). Smooth backfitting in generalized additive models. *Ann. Stat. 36*, 228–260.