DISCUSSION PAPER SERIES

IZA DP No. 8339

A General Semiparametric Approach to Inference with Marker-Dependent Hazard Rate Models

Gerard J. van den Berg
Lena Janys
Enno Mammen
Jens P. Nielsen

July 2014

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

A General Semiparametric Approach to Inference with Marker-Dependent Hazard Rate Models

Gerard J. van den Berg
University of Mannheim, IFAU Uppsala and IZA

Lena Janys
University of Mannheim

Enno Mammen
Heidelberg University and Higher School of Economics, Moscow

Jens P. Nielsen
Cass Business School London

Discussion Paper No. 8339
July 2014

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 8339
July 2014

ABSTRACT

A General Semiparametric Approach to Inference with Marker-Dependent Hazard Rate Models*

We examine a new general class of hazard rate models for survival data, containing a parametric and a nonparametric component. Both can be a mix of a time effect and (possibly time-dependent) marker or covariate effects. A number of well-known models are special cases. In a counting process framework, a general profile likelihood estimator is developed and the parametric component of the model is shown to be asymptotically normal and efficient. The analysis improves on earlier results for special cases. Finite sample properties are investigated in simulations. The estimator is shown to work well under realistic empirical conditions. The estimator is applied to investigate the long-run relationship between birth weight and later-life mortality using data from the Uppsala Birth Cohort Study of individuals born in 1915-1929. The results suggest a relationship that is difficult to capture with simple parametric specifications. Moreover, its shape at higher birth weights differs across gender.

JEL Classification: C41, C14, I12, J13

Keywords: covariate effects, survival analysis, local linear estimation, asymptotic distribution, birth weight, mortality, social class

Corresponding author:

Gerard J. van den Berg
Department of Economics
University of Mannheim
L7, 3–5
68131 Mannheim
Germany
E-mail: gjvdberg@uni-mannheim.de

* We thank participants at conferences in Mannheim and Cambridge and seminars in Bonn, Frankfurt and Bergen for their comments. We are grateful to CHESS (Stockholm University) and its UBCoS board members Ilona Koupil and Denny Vågerö for permission to use the UBCoS Multigen data. Gerard van den Berg thanks the Humboldt Stiftung for financial support through the Alexander von Humboldt Professorship Prize. Gerard van den Berg, Enno Mammen and Lena Janys thank the German Science Foundation for financial support through their FOR916 program. We thank Axel Munk for coordination of this program.

1 Introduction

The analysis of duration or survival data using large samples is widespread in biostatistics, economics, actuarial science, and engineering. In each field, it is of primary concern that the model to be estimated is not unduly restrictive. Semiparametric models provide a balance between flexibility and limited dimensionality. Typically, these models contain a part that is unrestricted and a part that is parametric. Different semiparametric models are estimated with different methods, focusing on different parameters or functions of interest. In survival analysis, it is often natural to take the hazard rate as the focal point of model specification. The most common semiparametric model is the Cox proportional hazard model for the (possibly stochastic) hazard rate $\lambda(t)$,

$$\lambda(t) = \exp(\beta' W(t))\alpha(t) \tag{1}$$

in which the covariate effects β are the parameters of interest and the dependence $\alpha(t)$ of the hazard rate on the elapsed duration or time is unspecified; see Cox(1972). The partial likelihood estimator of the parameter β does not depend on the functional form of α.

However, the estimator requires the assumption that the covariates $W(t)$ affect the hazard rate by way of the parametric functional form $\exp(\beta' W(t))$. Perhaps ironically, there is often more consensus about the functional form of $\alpha(t)$ than about the functional form of the effect of $W(t)$ on the hazard rate. More precisely, there is often more prior knowledge about how the hazard rate varies with the elapsed duration or time t than about how it varies with covariates $W(t)$. For example, in the study of mortality, it is natural to model the effect of age t on the mortality rate by way of the Gompertz specification $\exp(\theta t)$ or by small modifications of this functional form. This functional form is not controversial especially if very high ages are not taken into consideration. At the same time, there is no well-established functional form for the dependence of the mortality rate on socio-economic class, level of education, income and so on. Empirical studies often discretise such explanatory variables into a few categories and estimate effects of corresponding binary indicators using model (1). If the underlying causal pathways are smooth functions of the individual characteristics then the estimated effects may be biased for many values of the covariates.

Another example is provided by the literature on unemployment durations and job durations. Job search theory aims to explain the variation in these durations across individuals by modeling the job search behavior of individuals in labor markets where individual and external circumstances change over time. These theoretical models make precise predictions on how the hazard rates of the unemployment duration and the job duration distributions depend on labor market fluctuations (van den Berg, 2001). The latter can be measured by the ratio of the current unemployment and vacancy rates. This provides a functional form

for the time-varying profile of these hazard rates. It is more difficult to acquire theoretical guidance on how individual characteristics such as work experience, job complexity and number of children affect the level of the hazard rates.

If the functional form of the covariate effects on the hazard rate is unknown then the partial likelihood method used for the estimation of model (1) does not apply. In the current paper we propose a general semiparametric model that does not specify the functional form of covariate effects on the hazard rate, and we develop an estimation method for this model. The model has the form

$$\lambda(t) = \alpha\{X(t);\theta\}g\{Z(t)\} \tag{2}$$

Here, $g(\cdot)$ is unspecified while $\alpha(\cdot;\theta)_{\theta\in\Theta}$ is a parametric class of functions. The vectors $X(t)$ and $Z(t)$ are covariate or marker processes, and their elements may include the elapsed duration (or time) t. We show that this model has many existing semiparametric models as special cases. Note that it also includes nonproportional hazard rate models. In applications, the researcher may be particularly interested in the function g, for example if $Z(t)$ includes a policy instrument or treatment regime or if it includes a marker used to predict future outcomes. However, in other applications θ may be the parameter of interest. In that case, if the functional form of g is unknown, the estimation of a model that assumes a specific functional form for g may result in biased estimates of θ.

The estimator that we develop is a three-stage profile likelihood estimator inspired by the two-stage estimator of Nielsen, Linton and Bickel (1998) for a more restrictive semiparametric model. In our first stage, we estimate g best possible under the assumption that θ is actually known. In the second stage, we use this estimator $\widehat{g}_\theta$ of g in a profile likelihood, recognizing that the stochastic hazard $\alpha\{X(t);\theta\}\widehat{g}_\theta\{Z(t)\}$ actually has a parametric specification family of hazards, enabling the application of standard maximum likelihood methodology; see Borgan (1984). In a third estimation stage, we estimate g by $\widehat{g}_{\widehat{\theta}}$ using local linear kernel hazard regression. A major methodological contribution of our paper is that we improve on the asymptotic analysis in the existing literature for semiparametric hazard rate inference by using the improved asymptotic approximation theory of counting process martingales developed in Mammen and Nielsen (2007). In effect, our estimator of θ is square-root-n consistent, asymptotically normal and efficient.

We apply our newly devised estimation method to the study of the effect of birth weight on longevity. Recently, the interest in long-run effects of conditions in utero has been strongly growing. It has been shown that a range of diseases and death causes at high ages have "developmental origins", i.e. can be caused by conditions in utero. In the epidemiological literature, these conditions are often summarised by birth weight indicators across the full normal range of birth weight values (see e.g. overviews and meta-studies in Poulter et al., 1999, Rasmussen, 2001, Kuh and Ben-Shlomo, 2004, Davey Smith, 2005,

and Huxley et al., 2007). Such indicators have also been used in social sciences as markers of in-utero conditions (see the overview in Almond and Currie, 2011). The studies in these literatures do not estimate the effect of birth weight on morbidity and mortality in a nonparametric fashion. Instead, the effect is postulated to be parametric. Some studies use discrete indicators for whether birth weight lies in one of a small number of exhaustive weight intervals. For example, the landmark study by Leon et al. (1998) distinguishes between four intervals for birth weight in its effect on mortality due to ischaemic heart disease. Osler et al. (2003) use three categories among which 3500+ grams captures the highest birth weight values. Many studies simply use a binary indicator for whether birth weight is "low" (i.e., below 2500 grams) or not. Alternatively, a monotonous parametric continuous functional form is used, e.g. a linear relation between log birth weight and the log of the rate of the occurrence of some adverse health outcome.

Such parametric functional forms may be problematic. First, although most studies are concerned with adverse effects of low birth weight, it is known that a very high birth weight may also give rise to adverse health outcomes later in life. Ahlgren et al. (2007) demonstrate positive associations between birth weight and the rates of almost any type of cancer at higher ages. For certain cancer types the association is especially strong at birth weights above 4000 grams. This suggests that the birth weight effect on mortality is not monotonous.[1] Secondly, the continuity of the underlying biological mechanisms implies that effects of discretised birth weight indicators provide biased estimates of effects at specific birth weights. If medical protocols postulate interventions that condition on birth weight then the benefits of the intervention depend on the accuracy with which the relation between birth weight and outcome are estimated.

This calls for a semiparametric approach in which the long-run effect of birth weight on the mortality rate is not restricted by a parametric functional form. Our method is particularly suitable for this because of the high degree of consensus about the functional form for the age dependence of the mortality rate at ages up to 90. Specifically, we may adopt the Gompertz functional form for this. It is well known that the parameters of this functional form vary by gender and socio-economic class. Contrary to earlier semiparametric estimation methods, our method can deal with this as well as with the possibility that the birth weight effect varies by these personal characteristics.

Clearly, the application requires data of individuals born many decades ago, for whom birth weight and age at death are recorded with high accuracy. We use the Uppsala Birth Cohort Study (UBCoS) which is a lifelong follow-up study of a representative sample of

[1] Suggestive additional evidence for this is provided by the large number of studies reporting a positive association between birth weight and adult obesity (see Parsons et al., 1999, for a literature overview). This association seems to be driven by higher birth weight individuals (see e.g. Rasmussen and Johansson, 1998, and Curhan et al., 1996).

individuals born in 1915–1929. Upon birth, the birth weight was recorded in grams by qualified nurses. The data set contains additional information registered at birth, notably the socio-economic characteristics of the parental household. By now, over half of the sample members have died. We conjecture that this data set provides the best data in the world to relate birth weight and high-age mortality. The estimation results can be used to obtain an improved classification of the birth weight levels that indicate high-age health risks. This may be useful for health policy. In particular, neonatal interventions are often determined by whether the birth weight is above or below 2500 grams (see e.g. Almond, Chay and Lee, 2005). Our results indicate whether this is a sensible cut-off value if interest is in late-life mortality.

The paper is organised as follows. Section 2 presents our semiparametric model and explains how it contains models in the literature as special cases. In Section 3 we introduce the counting process formulation of our model. In Section 4 we define the estimators for the parameter θ and the nonparametric function g. In Section 5 we introduce the asymptotic distribution theory. In Section 6 we derive the local linear version of our estimator g and show the simulation results for the local constant and the local linear estimator to assess their performance under different bandwidth selection techniques. Section 7 contains the empirical application. Section 8 concludes.

2 The semiparametric model

This section presents the semiparametric model and explains how it contains models in the literature as special cases. Our model has the stochastic hazard rate

$$\lambda(t) = \alpha\{X(t); \theta\} g\{Z(t)\} \tag{3}$$

Here, $\alpha(\cdot;\theta)_{\theta\in\Theta}$ is a parametric class of functions whereas $g(\cdot)$ is unspecified apart from smoothness assumptions to be discussed below. Obviously, α and g must be nonnegative. The vectors $X(t)$ and $Z(t)$ are covariate or marker processes with dimensions d_x and d_z, respectively. For sake of exposition, we take $d_x \geq 1$ and $d_z \geq 1$. Note that $d_z = 0$ leads to a fully parametric model while $d_x = 0$ leads to a fully nonparametric model. The elements of $X(t)$ and $Z(t)$ may include the elapsed duration (or time) t. The elements of the vector $X(t)$ can be discretely or continuously distributed. Concerning the elements of $Z(t)$, for obvious reasons, we restrict attention to continuously distributed variables. We discuss exogeneity requirements on the covariate processes below.

The Cox model with a time-varying covariate process is obtained as a special case by taking $Z(t) := t$ and $\alpha\{X(t);\theta\} := \exp\{\theta' X(t)\}$. In this setting, g is the baseline hazard capturing duration dependence of the hazard whereas α is the so-called systematic part

of the hazard. The Stratified Cox model (Kalbfleisch and Prentice, 1980) extends the Cox model by allowing strata to have different baseline hazards. This can be captured in our model by specifying $Z(t) := (W, t)$ with W being discrete and finite, and $\alpha\{X(t); \theta\} := \exp\{\theta' X(t)\}$. Here, different values of W capture different strata.

Nielsen, Linton and Bickel (1998) consider a model with $X(t) := t$ and in which $Z(t)$ has only one element,

$$\lambda(t) = \alpha(t; \theta) g\{Z(t)\}, \tag{4}$$

Clearly, this model is motivated by the same concerns as our own model, as it does not impose a functional form on the covariate effect. However, it is more restrictive in that it does not allow the time effect $\alpha(t; \theta)$ to depend on individual characteristics, and it only deals with one covariate Z. In general, in survival analysis, it is advisable to include all relevant observed covariates in the model, to prevent bias due to omitted unobserved heterogeneity (see the overview in van den Berg, 2001).

Dabrowska (1997) considers a model that can be expressed as

$$\lambda(t) = \exp\{\theta' X(t)\} g\{Z(t)\} \tag{5}$$

in the same notation as above. This is a special case of our model because it assumes a specific functional form for the function α.

Our general model lends itself to other interesting specifications, for example the case where

$$\lambda(t) = \alpha(t; \theta) g_\beta(Z(t)) \tag{6}$$

where g_β is a parametric function that does not necessarily satisfy $g_\beta\{Z(t)\} = \exp\{\beta' Z(t)\}$. One could for example imagine instead that $g_\beta\{Z(t\}) = \beta' Z(t)$.

In general, the inclusion of t as an element of $X(t)$ and/or $Z(t)$ allows for nonproportional hazard specifications, that is, specifications where the hazard effects of t on the one hand and the covariates on the other are not multiplicative. Allowing for nonproportionality is useful, as proportionality is often hard to justify. For example, in the study of mortality, where it is natural to model the parametric effect of age t on the hazard by way of $\exp(\theta t)$, the coefficient θ varies with the gender of the individual. In the study of unemployment durations, the hazard rate of interest is the transition rate out of unemployment into employment. Economic-theoretical models predict that the decrease of this rate with the elapsed unemployment duration is stronger if aggregate labor market conditions are unfavorable (Blanchard and Diamond, 1994) or if the difference between the unemployment insurance level and the welfare level is large (van den Berg, 1990, 2001). Such studies warrant survival models that allow for nonproportional hazard specifications.

At this stage we should emphasise that our model does not rule out that $X(t)$ and $Z(t)$ have common elements. As we shall see, the estimation method deals with this automatically. If for example the hazard rate equals $\exp(\theta t) \cdot t^2$ and if we specify $\alpha(t;\theta) = \exp(\theta t)$ then the estimator $\widehat{g}(t)$ will converge to t^2 asymptotically. Scale parameters of α and g are not identified from each other, meaning that one should impose an arbitrary and innocuous normalization on these.

Semiparametric models are typically developed in conjunction with estimation methods tailored to the model. The Cox model and the corresponding partial likelihood estimation method are a case in point. It is useful to discuss some key properties of the estimators developed for the semiparametric models of Nielsen et al. (1998) and Dabrowska (1997) and other models and contrast them with properties of the estimator developed in our paper. Nielsen et al. (1998) show that their estimator of θ in (4) is efficient. This estimator has two stages. In the first stage, they estimate g best possible under the assumption that θ is actually known. In the second stage, they use this estimator $\widehat{g}_\theta$ of g in a profile likelihood, recognizing that the stochastic hazard

$$\widehat{\lambda}(t) = \alpha_\theta(t)\widehat{g}_\theta\{Z(t)\} \tag{7}$$

actually has a parametric specification family of hazards, enabling the application of standard maximum likelihood methodology; see Borgan (1984). Clearly, our estimator is valid under weaker conditions than in Nielsen et al. (1998), since our model is more general. As we shall see, one reason that we are able to achieve this is that we use the improved asymptotic approximation theory of counting process martingales developed in Mammen and Nielsen (2007). In effect, our estimator of θ is square-root-n consistent, asymptotically normal and efficient.

Dabrowska (1997) proves asymptotic square-root-n consistency and asymptotic normality of her estimator of θ. However, she does not achieve efficiency as we do with our approach.

We end this section by mentioning fully nonparametric approaches to statistical inference as an alternative approach to nonparametric inference. As we shall see, our estimator for the function g will be inspired by the nonparametric estimators developed in Nielsen and Linton (1995) and Nielsen (1998). These studies develop local constant and local linear kernel hazard estimators, respectively, for a model framework where the stochastic hazard is fully unspecified as a function of a vector $Z(t)$ which may include t. As methods for statistical inference on hazard rates, such estimators have the advantage that they do not rely on arbitrary functional-form assumptions, but the disadvantage that they suffer from the curse of dimensionality. Of course, this also applies to other estimators for nonparametric survival models, such as the estimators of Dabrowska (1987) and Spierdijk (2008). The advantage and disadvantage also apply (but to a lesser extent) to the estimator of

Linton, Nielsen and van de Geer (2003) for a model that is a hybrid between a semiparametric and nonparametric model; it assumes that the stochastic hazard is a multiplicative or additive function of unspecified functions of single elements of $Z(t)$. In this paper we do not consider this model. Neither do we consider semiparametric transformation models for survival data, since these are difficult to interpret in terms of hazard rate properties. See Dabrowska (2006) for an example of an estimator for such a model. Towards the end of the paper we briefly discuss semiparametric models with single-index structures for the dependence of the hazard rate on $Z(t)$.

3 Counting process formulation of the model

We follow the model formulations of Mammen and Nielsen (2007) and restrict ourselves to an independent identically distributed sampling and the one-jump counting process case. Let $N(t) = (N_1(t), ..., N_n(t))$ be an n-dimensional collection of n one-jump counting processes with respect to an increasing, right-continuous, complete filtration $\mathcal{F}_t \in [0, T]$; that is, N is adapted to the filtration and has components N_i taking values in $\{0, 1\}$, indicating, by the value 1, whether or not an observed jump has been registered for the i th individual. The N_i's are right-continuous step functions, zero at time zero, with jumps of size one. The variable $N_i(t)$ records the number of observed failures for the i'th individual during the time interval $[0, t]$ and is defined over the whole period $[0, T]$, where T is finite. Suppose that N_i has predictable intensity (see Andersen et al., 1993),

$$\lambda_i(t)dt = E\{dN_i(t)|\mathcal{F}_{t-}\} = \alpha\{X_i(t); \theta_0\}g\{Z_i(t)\}Y_i(t)dt \tag{8}$$

where Y_i is a predictable process taking values in $\{0, 1\}$ indicating, by the value 1, when the ith individual is at risk, whereas X_i is a d_x dimensional and Z_i a d_z dimensional predictable covariate process with support in some compact set $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ and $\mathcal{Z} \subseteq \mathbb{R}^{d_z}$, respectively.

We assume that the stochastic processes $(N_1, X_1, Z_1, Y_1), ..., (N_n, X_n, Z_n, Y_n)$ are independent and identically distributed for the n individuals. Let
$\mathcal{F}_{t,i} = \sigma\{N_i(u), X_i(u), Z_i(u); u \leqslant t\}$ and $\mathcal{F}_t = \vee_{i=1}^n \mathcal{F}_{t,i}$. It follows that λ_i is predictable with respect to $\mathcal{F}_{t,i}$ and hence $\mathcal{F}_t$, and the processes $M_i(t) = N_i(t) - \Lambda_i(t)$, $i = 1, ..., n$, with compensators $\Lambda_i(t) = \int_0^t \lambda_i(u)du$, are square integrable martingales with respect to $\mathcal{F}_{t,i}$ on the time interval $[0, T]$. Hence, $\Lambda_i(t)$ is the compensator of $N_i(t)$ with respect to both the filtration $\mathcal{F}_{t,i}$ and the filtration $\mathcal{F}_t$.

4 Definition of the estimators of θ and g

4.1 Three-step approach

To estimate θ we use a semiparametric profile likelihood method. In general (i.e. in parametric models) profile likelihood estimation allows the researcher to "profile out" the nuisance parameter, i.e. not the parameter that is of primary interest. In essence, one evaluates the likelihood function for the parameter of interest at all values for the nuisance parameter (i.e. the likelihood function depends on the nuisance parameter as if it were known). This has the advantage of being able to construct confidence intervals for the parameter of interest without needing to construct confidence intervals for the nuisance parameter. Semiparametric profile likelihood estimation allows efficient estimation (i.e. the usual square-root-n rate of convergence) under some restrictions on the bandwidth b (see for example Davison, 2002). For expositional reasons we follow the notation of Nielsen, Linton and Bickel (1998) as closely as possible. This implies formulating the procedure via a Nadaraya-Watson type estimator, which we call the local constant estimator. The approach immediately generalises to the notationally slightly more burdensome local linear approach. The finite sample analyses in our paper illustrates that the local linear methodology performs on average better in practice than the local constant approach.

We generalise the approach of Nielsen, Linton and Bickel (1998) for our semiparametric setting.

Step (i). First, the nonparametric function g is estimated under the assumption that the true parameter θ is known. This estimator of g depends on θ and on a smoothing parameter b. We make use of a leave-one-out version denoted by $\widehat{g}_{b,\theta,-i}(z)$ if the i-th observation is left out.

Step (ii). Second, we derive the likelihood function for the observable data assuming that the true g is known. The parameter θ is now estimated from the pseudolikelihood that arises when g is replaced by $\widehat{g}_\theta(z)$. This estimator depends on a value b_1 of the bandwith b and we therefore denote the estimator by $\widehat{\theta}_{b_1}$.

Step (iii). The final estimator of g is now calculated by assuming that $\widehat{\theta}$ is the true parameter and by using kernel smoothing using a bandwidth b_2. Therefore, the final estimator of g is of the form $\widehat{g}_{b_2,\widehat{\theta}_{b_1}}(z)$.

The two bandwidth vectors b_1 and b_2 should be not chosen identically. In order to obtain an asymptotically unbiased estimator of θ we need an undersmoothing bandwidth b_1. Thus b_1 should be of smaller order than b_2. In our empirical application, we will choose the tuple (b_1, b_2) jointly data-adaptively such that an overall cross-validation criterion is minimised; see Subsection B.2 in the appendix.

4.2 Definition of $\widehat{g}_{\theta}$

In this subsection we present the local constant estimator of the nonparametric function g. For any value of θ, we use the following leave-one-out procedure:

$$\widehat{g}_{b,\theta,-i}(z) = \frac{\sum_{j\neq i}\int K_b\{z - Z_j(u)\}dN_j(u)}{\sum_{j\neq i}\int K_b\{z - Z_j(u)\}\alpha\{X_j(u);\theta\}Y_j(u)du},$$

where K is a multivariate kernel function with $K_b(\cdot) = b_{prod}^{-1}K(B^{-1}\cdot)$ for any multivariate $b = (b_1^0, ..., b_{d_z}^0)^T$. Here, B is the diagonal matrix with diagonal entries $b_1^0, ..., b_{d_z}^0$ and $b_{prod} = b_1^0 \cdot ... \cdot b_{d_z}^0$. We will not always indicate dependence on the bandwidth b and write $\widehat{g}_{\theta,-i}(z)$ instead of $\widehat{g}_{b,\theta,-i}(z)$. Under our regularity conditions, we have that $\widehat{g}_{\theta,-i}(z) - \widehat{g}_{\theta}(z) = o_P(1)$, uniformly in θ, i and z, where

$$\widehat{g}_{\theta}(z) = \widehat{g}_{b,\theta}(z) = \frac{\sum_{j=1}^{n}\int K_b\{z - Z_j(u)\}dN_j(u)}{\sum_{j=1}^{n}\int K_b\{z - Z_j(u)\}\alpha\{X_j(u);\theta\}Y_j(u)du}.$$

Furthermore, $\widehat{g}_{\theta_0}$ consistently estimates $g(z)$ (see Nielsen and Linton, 1995), and, away from the true parameter value,

$$\widehat{g}_{\theta}(z) \to_p g_{\theta}(z) \equiv \frac{g(z)e_{\theta_0}(z)}{e_{\theta}(z)}, \tag{9}$$

where $e_{\theta}(z) = \int \alpha\{x;\theta\}f_u(x,z)y(u)du\ dx$ with $y(u) = P(Y_i(u) = 1)$. Let

$$g^*_{\theta,-i}(z) = \frac{\sum_{j\neq i}\int K_b\{z - Z_j(u)\}\lambda_j(u)du}{\sum_{j\neq i}\int K_b\{z - Z_j(u)\}\alpha\{X_j(u);\theta\}Y_j(u)du} \tag{10}$$

and note that

$$\widehat{g}_{\theta,-i}(z) - g^*_{\theta,-i}(z) = \frac{\sum_{j\neq i}\int K_b\{z - Z_j(u)\}dM_j(u)}{\sum_{j\neq i}\int K_b\{z - Z_j(u)\}\alpha\{X_j(u);\theta\}Y_j(u)du}. \tag{11}$$

As we show below, this quantity can be analyzed by martingale methods. We call $g^*_{\theta,-i}(z) - g_{\theta,-i}(z)$ the stable and $\widehat{g}_{\theta,-i}(z) - g^*_{\theta,-i}(z)$ the variable part of $\widehat{g}_{\theta,-i}(z)$.

4.3 Definition of $\widehat{\theta}$

In this subsection we present the expression for the estimator $\widehat{\theta}$ of the parameter θ. We use maximum likelihood. The standard (conditional on Y, X and Z) log-likelihood for a counting process is $\sum_{i=1}^{n}\int \ln\lambda_i(u)dN_i(u) - \sum_{i=1}^{n}\lambda_i(u)du$ (see Aalen, 1978). If $g(z)$ were known, we would maximise the following likelihood function over θ

$$\ell(\theta) = \sum_{n=1}^{n}\int \mu_{\theta}\{X_i(u), Z_i(u)\}dN_i(u) - \sum_{n=1}^{n}\int \exp[\mu_{\theta}\{X_i(u), Z_i(u)\}]Y_i(u)du \tag{12}$$

where $\mu_\theta(x,z) = \ln[\alpha(x;\theta)g(z)]$ is the logarithmic hazard. Consequently, the maximum likelihood estimator for θ given known g (denoted as $\widehat{\theta}_g$) is given by

$$\widehat{\theta}_g = \arg\max_\theta \sum_{n=1}^{n} \int \mu_\theta\{X_i(u), Z_i(u)\} dN_i(u) - \sum_{n=1}^{n} \int \exp[\mu_\theta\{X_i(u), Z_i(u))\}] Y_i(u) du \quad (13)$$

Since $g(z)$ is not known, we substitute $\hat{\mu}_{\theta,-i}(x,z)$ for $\mu_\theta(x,z)$ where $\hat{\mu}_{\theta,-i}(x,z) = \ln[\alpha(x;\theta)\widehat{g}_{\theta,-i}(z)]$:

$$\hat{\ell}(\theta) = \sum_{n=1}^{n} \int \hat{\mu}_{\theta,-i}\{X_i(u), Z_i(u)\} dN_i(u) - \sum_{n=1}^{n} \int \exp[\hat{\mu}_{\theta,-i}\{X_i(u), Z_i(u)\}] Y_i(u) du. \quad (14)$$

The pseudo-maximum likelihood estimator $\widehat{\theta}$ is defined as

$$\widehat{\theta} = \arg\max_{\theta \in \mathcal{N}_0} \hat{\ell}(\theta). \quad (15)$$

Here, $\mathcal{N}_0$ is a fixed compact subset of Θ having θ_0 as an interior point.

5 Asymptotic distribution theory

We will show that $Q_n(\theta) = n^{-1}\{\hat{\ell}(\theta) - \hat{\ell}(\theta_0)\}$ converges in probability, uniformly in a neighborhood $\mathcal{N}_0$ of θ_0, to a nonrandom function $Q(\theta)$ that is uniquely maximised at θ_0. In fact, we will first show that $Q_n(\theta)$ can be approximated by $\overline{Q}_n(\theta) = n^{-1}\{\overline{\ell}(\theta) - \overline{\ell}(\theta_0)\}$, where

$$\overline{\ell}(\theta) = \sum_{i=1}^{n} \int \overline{\mu}_\theta\{X_i(u), Z_i(u)\} dN_i(u) - \sum_{i=1}^{n} \int \exp[\overline{\mu}_\theta\{X_i(u), Z_i(u)\}] Y_i(u) du \quad (16)$$

with $\overline{\mu}_\theta(x,z) = \ln[\alpha(x,\theta)g_\theta(z)]$. We show in the appendix that $\overline{Q}_n(\theta)$ approaches

$$Q(\theta) = \int\int \left[\ln\left\{ \frac{\alpha(x;\theta)e_{\theta_0}(z)}{\alpha(x;\theta_0)e_\theta(z)} \right\} - \frac{\alpha(x,\theta)e_{\theta_0}(z)}{\alpha(x;\theta_0)e_\theta(z)} + 1 \right] \alpha(x;\theta) f_u(x,z) y(u) du \; dz, \quad (17)$$

in probability, uniformly over any compact neighborhood of θ_0. This will imply consistency of $\widehat{\theta}$.

In a next step we will show asymptotic normality of $\widehat{\theta}$. Let $\hat{s}_\theta$ (the score vector) and $\hat{H}_{\theta\theta}$ (the Hessian matrix) be the first and second derivatives of the pseudolikelihood $\hat{\ell}$ standardised by sample size:

$$
\begin{aligned}
\hat{s}_\theta(\theta) &= \frac{1}{n}\sum_{i=1}^{n}\int \frac{\partial \hat{\mu}_{\theta,-i}}{\partial\theta}\{X_i(u),Z_i(u)\}dN_i(u) \qquad (18)\\
&\quad -\frac{1}{n}\sum_{i=1}^{n}\int \frac{\partial \hat{\mu}_{\theta,-i}}{\partial\theta}\{X_i(u),Z_i(u)\}\alpha\{X_i(u);\theta\}\widehat{g}_{\theta,-i}\{Z_i(u)\}Y_i(u)du,\\
\hat{H}_{\theta\theta}(\theta) &= n^{-1}\sum_{i=1}^{n}\int \frac{\partial^2 \hat{\mu}_{\theta,-i}}{\partial\theta\partial\theta^T}\{X_i(u),Z_i(u)\}dN_i(u)\\
&\quad -n^{-1}\sum_{i=1}^{n}\int \left\{\frac{\partial^2 \hat{\mu}_{\theta,-i}}{\partial\theta\partial\theta^T}+\frac{\partial \hat{\mu}_{\theta,-i}}{\partial\theta}\frac{\partial \hat{\mu}_{\theta,-i}}{\partial\theta^T}\right\}\{X_i(u),Z_i(u)\}\alpha\{X_i(t);\theta\}\widehat{g}_{\theta,-i}\{Z_i(t)\}Y_i(u)du.
\end{aligned}
$$

By the mean value theorem

$$
0 = n^{1/2}\hat{s}_\theta(\theta_0) + \hat{H}_{\theta\theta}(\breve{\theta})n^{1/2}(\widehat{\theta}-\theta_0), \qquad (19)
$$

where $\breve{\theta}$ lies between θ_0 and $\widehat{\theta}$. We first analyze the pseudoscore vector evaluated at the true θ_0, using (18) with $\theta=\theta_0$

$$
\begin{aligned}
\hat{s}_\theta(\theta_0) &= \frac{1}{n}\sum_{i=1}^{n}\int \frac{\partial \hat{\mu}_{\theta_0,-i}}{\partial\theta}\{X_i(u),Z_i(u)\}dM_i(u) + \frac{1}{n}\sum_{i=1}^{n}\int \frac{\partial \hat{\mu}_{\theta_0,-i}}{\partial\theta}\{X_i(u),Z_i(u)\}d\Lambda_i(u)\\
&\quad -\frac{1}{n}\sum_{i=1}^{n}\int \frac{\partial \hat{\mu}_{\theta_0,-i}}{\partial\theta}\{X_i(u),Z_i(u)\}\alpha\{X_i(u);\theta_0\}g\{Z_i(u)\}Y_i(u)du \qquad (20)\\
&\quad -\frac{1}{n}\sum_{i=1}^{n}\int \frac{\partial \hat{\mu}_{\theta_0,-i}}{\partial\theta}\{X_i(u),Z_i(u)\}\alpha\{X_i(u);\theta_0\}\left[\widehat{g}_{\theta_0,-i}\{Z_i(u)\}-g\{Z_i(u)\}\right]Y_i(u)du.
\end{aligned}
$$

Here we have substituted N by $M+\Lambda$ and $\widehat{g}_{\theta_0,-i}$ by $g+\widehat{g}_{\theta_0,-i}-g$. By the definition of Λ_i, we find that the second and third term on the right hand side of (20) cancel. We then break $\widehat{g}_{\theta_0,-i}-g$ into stable and variable terms. Using the decomposition (11), we find, after interchanging the order of summation and integration, that

$$
\begin{aligned}
&\sum_{i=1}^{n}\int \frac{\partial \hat{\mu}_{\theta_0,-i}}{\partial\theta}\{X_i(u),Z_i(u)\}\alpha\{X_i(u);\theta_0\}\{\widehat{g}_{\theta_0,-i}-g^*_{\theta_0,-i}\}\{Z_i(u)\}Y_i(u)du\\
&= \sum_{i=1}^{n}\int \frac{\partial \hat{\mu}^*_{\theta_0,-i}}{\partial\theta}\{Z_i(u)\}dM_i(u),
\end{aligned}
$$

where

$$
\frac{\partial \hat{\mu}^*_{\theta_0,-i}}{\partial\theta}\{Z_i(u)\} = \sum_{j\neq i}^{n}\int \frac{(\partial \hat{\mu}_{\theta_0,-j}/\partial\theta)\{X_j(t),Z_j(t)\}\alpha\{X_j(t);\theta_0\}Y_j(t)K_b\{Z_j(t)-Z_i(u)\}}{\sum_{k\neq j}\int K_b\{Z_j(t)-Z_k(r)\}\alpha\{X_k(r);\theta_0\}Y_k(r)dr}dt.
$$

Now substitute $\partial\overline{\mu}_{\theta_0}/\partial\theta + \partial \ln \widehat{g}_{\theta_0,-i}/\partial\theta - \partial \ln g_{\theta_0,-i}/\partial\theta$ for $\partial\hat{\mu}_{\theta_0,-i}/\partial\theta$ in the first term on the right hand side of (20). Collecting everything together we obtain that

$$\begin{aligned}
\hat{s}_\theta(\theta_0) &= n^{-1}\sum_{i=1}^{n}\int \frac{\partial\overline{\mu}_{\theta_0,-i}}{\partial\theta}\{X_i(u),Z_i(u)\}dM_i(u) \\
&\quad -n^{-1}\sum_{i=1}^{n}\int \frac{\partial\hat{\mu}^*_{\theta_0,-i}}{\partial\theta}\{Z_i(u)\}dM_i(u) \\
&\quad +n^{-1}\sum_{i=1}^{n}\int \left\{\frac{\partial \ln \widehat{g}_{\theta_0,-i}}{\partial\theta} - \frac{\partial \ln g_{\theta_0}}{\partial\theta}\right\}\{X_i(u)\}dM_i(u) \\
&\quad -n^{-1}\sum_{i=1}^{n}\int \frac{\partial\hat{\mu}_{\theta_0,-i}}{\partial\theta}\{X_i(u),Z_i(u)\}\alpha\{X_i(u);\theta_0\}\{g^*_{\theta_0,-i}-g\}\{Z_i(u)\}Y_i(u)du.
\end{aligned} \tag{21}$$

We have written $\hat{s}_\theta$ as a sum of four terms: the last term is a stochastic average of $g^*_{\theta_0,-i} - g$ that arises from the bias obtained in the estimation of g: it is asymptotically negligible if a sufficiently small bandwidth is chosen. Undersmoothing is necessary in many semiparametric estimation problems; see Bickel et al. (1993) for a discussion. In the appendix we will show that the second and third term on the right hand side of (21) are also $o_p(n^{-1/2})$. Because the integrands converge to zero, in probability, this would immediately follow if the integrands are predictable. But the latter is not the case, and therefore the formal proof is more complicated (see the appendix). The proof makes use of the approach to the predictability issue developed in Mammen and Nielsen (2007). We have that

$$\begin{aligned}
n^{1/2}\hat{s}_\theta(\theta_0) &= n^{1/2}s^e_\theta(\theta_0) + o_p(1), \text{ where} \\
s^e_\theta(\theta_0) &= n^{-1}\sum_{i=1}^{n}\int \frac{\partial\overline{\mu}_{\theta_0}}{\partial\theta}\{X_i(u),Z_i(u)\}dM_i(u).
\end{aligned} \tag{22}$$

since $\partial \ln \overline{\mu}_\theta\{X_i(u),Z_i(u)\}/\partial\theta$ is a predictable process, we can apply Rebolledo's martingale central limit theorem to $s^e_\theta(\theta_0)$ and we get that

$$n^{1/2}s^e_\theta(\theta_0) \to N(0,\mathcal{I}_0), \text{ in distribution,} \tag{23}$$

where

$$\mathcal{I}_0 = \int\int \frac{\partial\overline{\mu}_{\theta_0}}{\partial\theta}\frac{\partial\overline{\mu}_{\theta_0}}{\partial\theta^T}(x,z)\alpha(x,\theta_0)g(z)f_u(x,z)y(u)du\ dz$$

with

$$\frac{\partial\overline{\mu}_{\theta_0}}{\partial\theta}(x,z) = \frac{\partial \ln \alpha}{\partial\theta}(x,\theta_0) - \frac{\partial \ln e_{\theta_0}}{\partial\theta}(z).$$

In the appendix, we also show that the Hessian matrix $\hat{H}_{\theta\theta}(\theta)$ satisfies

$$\sup_{\theta\in\mathcal{N}_n}|\hat{H}_{\theta\theta}(\theta) - \mathcal{I}_0| \to_p 0, \tag{24}$$

for $\mathcal{N}_n = \{\theta : |\theta - \theta_0| \leq \delta_n\}\delta_n \to 0$ is a shrinking neighborhood of θ_0. In conclusion, we get from (19), (22), (23) and (24) that $n^{1/2}(\widehat{\theta} - \theta_0) \to N(0, \mathcal{I}_0^{-1})$, in distribution.

The following theorem summarises our discussion. The assumptions and the proof of the theorem are in the appendix.

Theorem 1. *(i) Make the assumptions (A1)–(A4). With probability tending to one, there exists a maximiser $\widehat{\theta}$ in (15). All (measurable) choices of the maximiser result in a consistent estimator: $\widehat{\theta} \xrightarrow{p} \theta_0$.*

(ii) Make the additional assumptions (A5)–(A8). Then

$$n^{1/2}(\widehat{\theta} - \theta_0) \xrightarrow{d} N(0, \mathcal{I}_0^{-1}). \tag{25}$$

(iii) The asymptotic covariance matrix $\mathcal{I}_0^{-1}$ is consistently estimated by $\hat{H}_{\theta\theta}^{-1}(\widehat{\theta})$.

We now argue that our estimator achieves the semiparametric efficiency bound. For this purpose consider the following parametric specification of the hazard function:

$$\lambda_i(t;\theta) = \alpha\{X_i(t);\theta\} g_\theta\{Z_i(t)\} Y_i(t). \tag{26}$$

The pseudo-maximum likelihood estimator in the model is given as maximiser $\overline{\theta}$ of the likelihood function $\overline{\ell}(\theta)$. By classical theory one gets that

$$n^{1/2}(\overline{\theta} - \theta_0) = \mathcal{I}_0^{-1} n^{-1/2} \sum_{i=1}^{n} \int \frac{\partial \overline{\mu}_{\theta_0}}{\partial \theta}\{X_i(u), Z_i(u)\} dM_i(u) + o_P(1).$$

Thus, $\overline{\theta}$ has the same asymptotic limit distribution as $\widehat{\theta}$ and the specification (26) is the hardest parametric submodel of our semiparametric model. In particular, we get that $\mathcal{I}_0$ is the semiparametric information matrix.

In our simulations and in our empirical application we also use a local linear estimator of the functions g_θ. It can be shown that this also leads to efficient estimation of θ.

In the final estimation step an estimator of g is calculated. This can be done by $\widehat{g}_{b^*,\widehat{\theta}}(z)$ where $\widehat{\theta}$ is plugged in for the parameter θ. As discussed above, the bandwidth vector b^* should differ from b. We also consider a local linear estimator $\widehat{g}^{LL}_{b^*,\widehat{\theta}}(z)$. For a definition of $\widehat{g}^{LL}_{b^*,\widehat{\theta}}(z)$ see Appendix B.1.

Corollary 1. *Suppose that assumptions (A1)-(A8) hold, and that $n^{1/5}b^*$ converges to a limit $0 \leq \gamma < \infty$. Then*

$$\begin{aligned} n^{2/5}\{\widehat{g}_{b^*,\widehat{\theta}}(z) - g(z)\} &\to N(\beta(z), \nu(z)), \\ n^{2/5}\{\widehat{g}^{LL}_{b^*,\widehat{\theta}}(z) - g(z)\} &\to N(\beta^{LL}(z), \nu(z)), \end{aligned}$$

where

$$\begin{aligned}\beta(z) &= \frac{\gamma^2}{2}\mu_2(K)\left\{2\frac{\partial g}{\partial z}(z)\frac{\partial \ln e_{\theta_0}}{\partial z}(z)+\frac{\partial^2 g}{\partial z^2}(z)\right\},\\ \beta^{LL}(z) &= \frac{\gamma^2}{2}\mu_2(K)\frac{\partial^2 g}{\partial z^2}(z),\\ \nu(z) &= \gamma^{-1}||K||^2\frac{g(z)}{e_{\theta_0}(z)}\end{aligned}$$

with $\mu_2(K)=\int K(t)^2dt$. *Furthermore,*

$$\hat{\nu}(z)=\frac{nb^*\sum_{i=1}\int K_{b^*}\{z-Z_i(u)\}^2dN_i(u)}{[K_{b^*}\{z-Z_i(u)\}\alpha\{X_i(u);\widehat{\theta}\}Y_i(u)du]^2}$$

is a consistent estimator of $\nu(z)$, *i.e.*

$$\hat{\nu}(z)\to_p \nu(z).$$

6 Simulation study

To study the performance of our estimator, we simulate data from the following models:

$$\begin{aligned}&\textbf{Model 1:} && \lambda(t)=\exp\{\theta t\}\gamma\times z(1-z),\\ &\textbf{Model 2:} && \lambda(t)=\gamma\theta t^{\theta-1}\exp\left\{-\frac{1}{2}\cos(2\pi z)-\frac{3}{2}\right\},\\ &\textbf{Model 3:} && \lambda(t)=\exp\{\theta t\}\exp\left\{-\frac{1}{2}\cos(2\pi z)-\frac{3}{2}\right\},\\ &\textbf{Model 4:} && \lambda(t)=t^{\theta-1}(1-t)z(1-z).\end{aligned} \tag{27}$$

with $\theta=1.5$ and $\gamma=1$. The two-dimensional hazards as functions of t and z are shown in Figure 1.

We report the estimation results from 100 simulated samples using a discretised version of the local constant estimator and the local linear estimator (see Subsection B.3 in the appendix). We simulate on a grid $R\times R'$ with size 100×100 (i.e., $R=100, R'=100$) which seems sufficient for our purposes. The sample size is either $n=10000$ or $n=5000$ observations. Our estimator is evaluated along three dimensions: (1) **bandwidth selection:** We evaluate whether feasible bandwidth selection methods work to choose the two bandwidths b_1 and b_2, (2) **parameter estimate:** we compare the true parameter θ with its estimate, and (3) **Integrated Squared Error (ISE):** we evaluate the integrated squared error of our estimator of the function g. Table 1 reports the results for the cross-validated

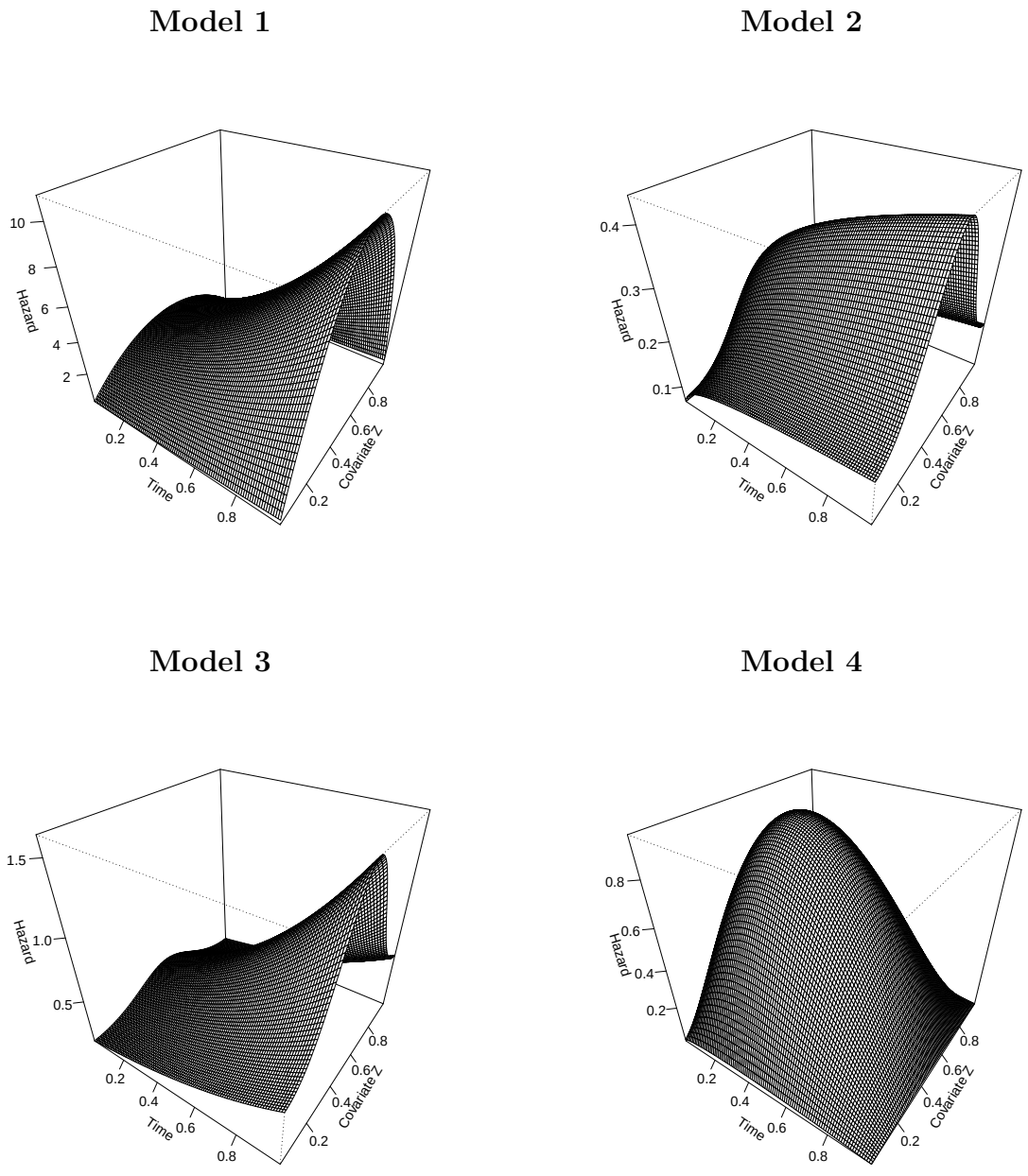


Figure 1: *The two-dimensional hazard functions of Models 1–4, see (27).*

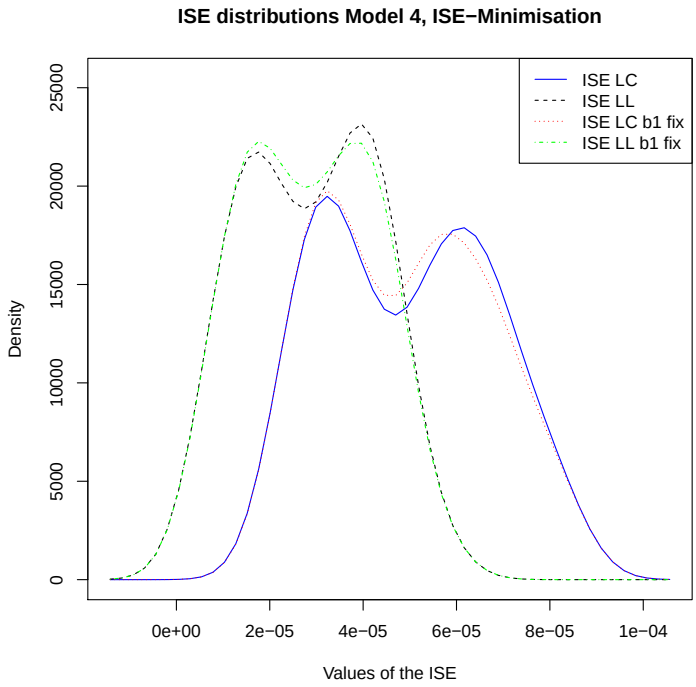


Figure 2: *The empirical distributions for the ISE with an optimally chosen b_1 (solid line and dashed line) and with a fixed b_1 (the dotted line and the dot-dash line, for the local constant and the local linear estimator, respectively.)*

bandwidths, the ISE bandwidths and the resulting parameter estimates in terms of the average absolute deviation from the true parameter. In general, the estimator performs well regardless of the true form of the hazard and independent of whether we use the ISE bandwidths or the bandwidths selected by minimising the cross-validation criterion. The parameter is estimated with precision, regardless of the method, and the parameter estimates are in general not sensitive to bandwidth choice. It seems that the local constant is as good or even better as the local linear estimator for estimating the parameter, although the differences are small. Overall, the local linear estimator performs slightly better than the local constant estimator (in terms of the distribution of the ISE), which suggests that the local linear is better suited to capture the nonparametric function, which is not a surprising result, considering the well-known shortcomings of the local constant estimator in boundary regions.

In almost all cases, the standard errors on b_1 are rather large, at least compared with the standard errors on b_2. This result reflects how little the parameter estimate depends on the bandwidth choice. This suggests that applied researchers might find it practical to fix b_1 to be very small and only consider different bandwidths for b_2. To illustrate this, we

calculated the ISE for 100 samples for the case where we fix b_1 at the smallest bandwidth. In Figure 2 we plot the empirical distribution of the ISE for both when b_1 is fixed and when b_1 is optimally chosen. When comparing the two curves it is clear that (at least in this case) fixing b_1 does not lead to a strong decrease in performance.

Figure 3 visualizes the empirical distribution of the integrated squared error for all 100 samples, for both sample sizes and the two different estimators and for all four models. In general, the local linear estimator performs better than the local constant estimator. Increasing the sample size leads on average to a reduction of the ISE and a reduction in the variance of the distribution of the ISE. However, while we can retrieve the parameter with relative precision, cross-validation tends towards undersmoothing in many of the cases that were considered. While this is not surprising, better feasible bandwidth selection methods, such as "do-validation" (Gámiz Pérez et al., 2013) might improve performance.

We also compare the ISE for the nonparametric local constant and local linear estimators to the ISE for our semiparametric estimator. In our simulation setting, the former estimators target a function that has a dimensionality that exceeds the dimensionality of the function g with one. We find that if the semiparametric model is true, then - unsuprisingly - it improves estimation accuracy enormously to impose this semiparametric structure from the outset rather than using a fully nonparametric approach. In all cases, local linear estimation performs significantly better than local constant estimation, irrespectively of whether a semiparametric or a nonparametric is considered. These results are unsurprising and the results are not listed here; however, they do provide us with a helpful sanity check of the estimation and modelling approach of this paper.

7 Empirical application: the effect of birth weight on later-life mortality

7.1 The Uppsala Birth Cohort Study data

The Uppsala Birth Cohort Study is a lifelong follow-up study of birth cohorts of individuals born in Uppsala in 1915–1929.[2] Information on early-life characteristics of these newborns and social characteristics of their parents was retrieved from the neonatal register of the hospital in Uppsala. Mortality is observed from parish records and national death registers. Loss of follow-up due to emigration is observed from censuses (starting with the 1960 census), routine administrative registers (starting in 1961 or later), and archives. In the data at our disposal, individuals are followed over time up to the end of 2002, so that the

[2]Rajaleid, Manor and Koupil (2008) demonstrate that it is representative of birth cohorts in Sweden in the years 1915–1929.

Integrated Squared Error							
Model	n	Bandwidths				Parameter, $\overline{e}$	
		LC		LL		LC	LL
		b1	b2	b1	b2		
1	10000	0.3084 (0.1709)	0.078 (0.0126)	0.2244 (0.2269)	0.1476 (0.0179)	0.011	0.031
	5000	0.2872 (0.1858)	0.0872 (0.0126)	0.1688 (0.1985)	0.1674 (0.0237)	0.02	0.035
2	10000	0.1044 (0.1701)	0.1346 (0.0249)	0.1024 (0.1761)	0.1398 (0.0244)	0.01	0.01
	5000	0.0418 (0.1082)	0.1466 (0.0246)	0.0456 (0.1214)	0.1448 (0.0254)	0.027	0.027
3	10000	0.4049 (0.0998)	0.1235 (0.016)	0.4571 (0.0974)	0.1305 (0.0147)	0.014	0.022
	5000	0.325 (0.1888)	0.1314 (0.0266)	0.3474 (0.1965)	0.1432 (0.0264)	0.015	0.023
4	10000	0.1642 (0.2643)	0.1628 (0.0258)	0.15 (0.2373)	0.2468 (0.0357)	0.03	0.03
	5000	0.136 (0.1839)	0.1684 (0.038)	0.1302 (0.1703)	0.2354 (0.0495)	0.054	0.054

Cross-Validation							
Model	n	Bandwidths				Parameter, $\overline{e}$	
		LC		LL		LC	LL
		b1	b2	b1	b2		
1	10000	0.177 (0.1554)	0.026 (0.0117)	0.276 (0.225)	0.029 (0.0164)	0.028	0.03
	5000	0.104 (0.1159)	0.021(0.0034)	0.19 (0.2028)	0.021 (0.0052)	0.038	0.039
2	10000	0.035 (0.0643)	0.022 (0.0095)	0.036 (0.0727)	0.022 (0.0095)	0.011	0.011
	5000	0.193 (0.2427)	0.021 (0.0001)	0.691 (0.0192)	0.626 (0.009)	0.031	0.034
3	10000	0.182 (0.1505)	0.038 (0.0154)	0.213 (0.1792)	0.039 (0.0154)	0.036	0.036
	5000	0.171 (0.1577)	0.032 (0.0061)	0.202 (0.1946)	0.032 (0.0064)	0.033	0.033
4	10000	0.092 (0.1809)	0.032 (0.006)	0.082 (0.1692)	0.032 (0.0058)	0.031	0.031
	5000	0.101 (0.1508)	0.031 (0.0044)	0.098 (0.1469)	0.031 (0.0039)	0.055	0.055

Table 1: *Simulation results for the models in equation (27), with* $\theta_0 = 1.5$ *as the true parameter, for two different sample sizes (5000, 10000). The numbers are averages over 100 simulated samples. The upper panel shows the results for bandwidths chosen by the infeasible strategy of minimizing the ISE.* $b1$ *and* $b2$ *refer to the two associated bandwidths. Standard errors are in parentheses. The lower panel shows the results for bandwidths chosen by the feasible bandwidth selection criterion of minimizing the cross-validation score (CV). LC and LL refer to the use of the local constant and the local linear estimator, respectively. In the last column, the parameter estimate is reported in terms of the average of the estimation error* $\overline{e} = abs(\widehat{\theta}_{b_1} - \theta_0)$ *over 100 samples.*

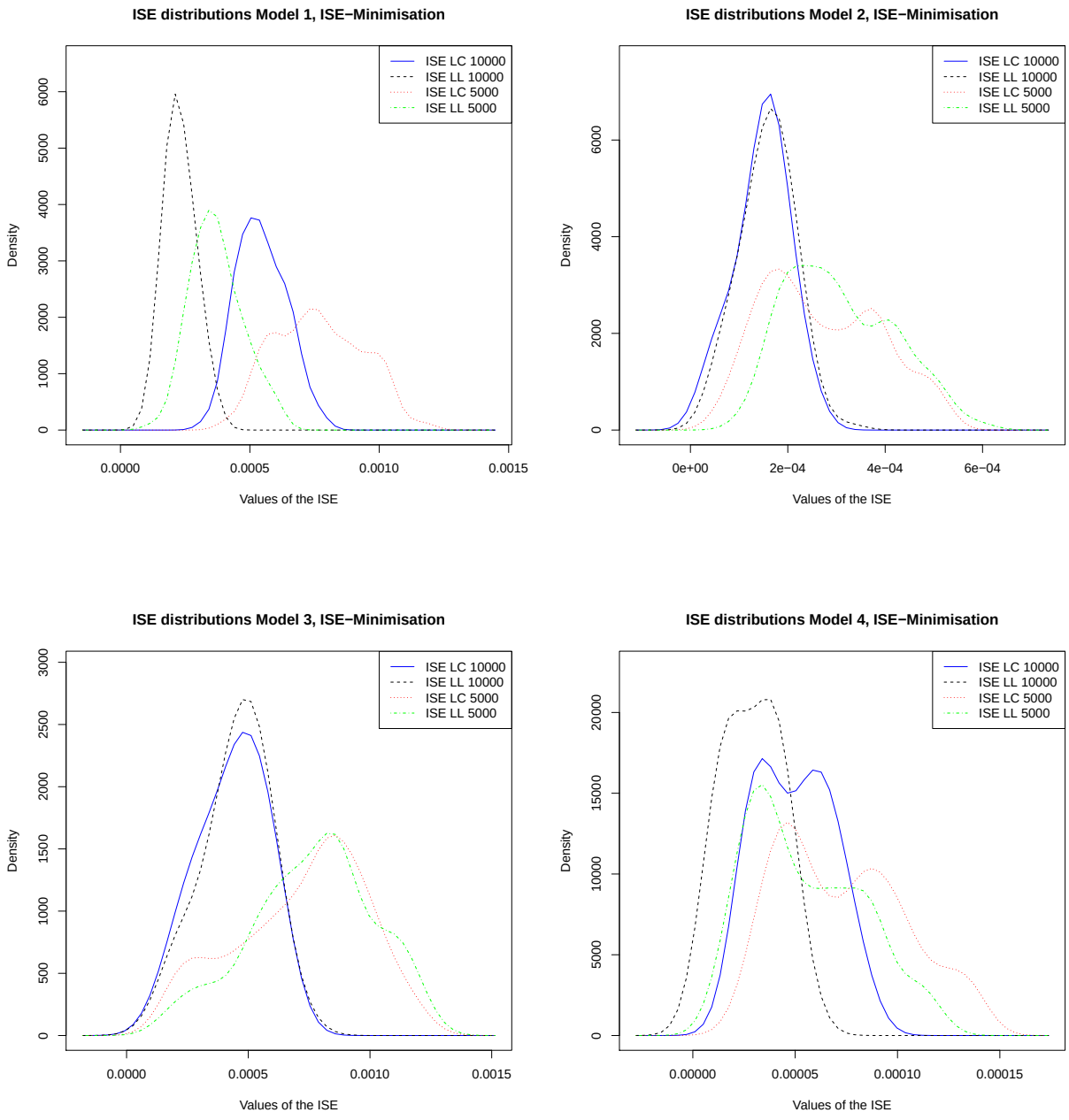


Figure 3: *The smoothed distribution of the integrated squared error over 100 samples. The solid line represents the local constant estimator with* $n = 10000$*, the dashed line represents the local linear estimator with* $n = 10000$*. The dotted line represents the local constant estimator with* $n = 5000$ *and the dot-dash line represents the local linear estimator with* $n = 5000$*.*

highest observed death age is 87. Leon et al. (1998) and Rajaleid, Manor and Koupil (2008) provide detailed descriptions of the data.

The birth and death dates and the resulting individual lifetime durations are observed in days. Not all variables are observed for all of individuals, but birth date, lifetime duration (or time until loss of follow-up) and birth weight are observed for virtually every individual. We omit all individuals who were stillborn or died within one day. This leads to a sample size of 13668 individuals.

Birth weight was recorded in grams. We trim the data by discarding 2 observations with birth weight below 1000 g and 27 observations with birth weight above 5000 g. For 13 of the remaining individuals, birth weight is not observed. This leads to the final sample size of $n = 13639$ individuals. The socio-economic status or social class at birth is a grouped hierarchically ordered version of the Swedish SEI code which in turn is based on the occupation of the main breadwinner in the household. The values run from 1 (highest class) to 7.

In the sample, 50% are observed to die before 2002 and 50% have right-censored lifetime durations (almost all of the latter are still alive at the end of 2002). Table 2 gives some sample statistics of the main variables that were made accessible for our study.

To interpret the results it is useful to emphasise that living conditions in Sweden in the birth years 1915–1929 were relatively good in comparison to most other countries at the time and in comparison to many developing countries today. Life expectancy was among the highest in the world, and infant mortality among the lowest (around 5%). The public health care system was modern, with institutionalised maternal and child health care in urban areas. At the time of birth, most individuals in our data resided in or around the city of Uppsala. In the years 1915–1929, the population of the city of Uppsala was stable at the level of around 30,000 inhabitants. The two largest sectors in the city's labor market were manufacturing and trade, occupying 45% and 25% of the workforce, respectively. Electricity was available everywhere. Lobell, Schön and Krantz (2007) provide details of the Swedish economy in these years and the surrounding decades. National Central Bureau of Statistics (1969) provide detailed descriptions of demographic developments. Sundin and Willner (2007) contains a detailed history of public health in Sweden. Modin (2002) describes local conditions in Uppsala around the 1920s. Notice that contemporary birth weight values are in the same ball park as those in the data.

The data have been used by a number of studies on long-run effects of birth weight. All of these estimate Cox Proportional Hazard models with partial likelihood. Leon et al. (1998) and Rajaleid, Manor and Koupil (2008) use discrete birth weight indicators based on a small number of weight intervals. van den Berg and Modin (2013) assume that the log cardiovascular mortality rate is a linear function of the log birth weight.

variable	10^{th} percentile	mean	90^{th} perc.
right-censored durations		0.50	
duration (years) if uncensored	0.9	54.6	80.0
duration (years) if censored	73.7	77.6	84.6
birth weight (grams)	2750	3416	4080
birth year	1916	1922.6	1928
social class at birth (1 to 7: high to low)	2	4.2	6
male		0.52	
male birth weight	2810	3478	4140
female birth weight	2700	3349	4000

Table 2: *Summary statistics of the sample*

7.2 Model specification and results

As a parametric baseline function we choose the Gompertz function that has been shown to accurately model the age dependence of mortality for the ages covered by our observation window (up to age 87). Specifically,

$$\lambda(t) = \exp\{\theta t\} g(z) \tag{28}$$

with z being the birth weight. In model extensions we allow θ to vary with other covariates x (see below).

We discretise the time dimension in 150 intervals and the covariate in 100 intervals ($R = 150$, $R' = 100$). We use the Epanechnikov Kernel given by $K(u) = \frac{3}{4}(1-u^2)\mathbb{I}_{|u\leq 1|}$. As a robustness check we also use the kernel used in Nielsen and Tanggaard (2001), but the choice of kernel does not alter our results in any substantial way. The confidence intervals are calculated using the bootstrap procedure for kernel hazard estimators introduced by Fledelius et al. (2004).

The estimate of the shape parameter θ is practically unaffected by the bandwidth choice. The first line of Table 3 gives estimates for the local linear estimator.

Figure 4 shows the estimates of the nonparametric function g, using the local linear estimator. The confidence bands are calculated using the bootstrap method of Fledelius et al. (2004). The x-axis depicts birth weight and the y-axis shows the estimated values of the nonparametric part of the hazard function. The estimated function varies over z in an

inverted J- or a U-shape, indicating that mortality risk decreases as birth weight increases and then increases again at very high birth weights. The results can be interpreted in the following way: In relation to an infant born in the optimal birth weight range of about 3000-3500g, the relative risk is about $2\frac{1}{2}$ as high as for an infant born with $1000g$ and $1\frac{1}{2}$ times as high as for an infant born weighting 5000g.

The local constant estimator does not perform satisfactorily. Specifically, it loses structure very quickly and becomes flat when the bandwidth is increased. This did not occur in the simulations and may be due to the scarcity of observations in the boundary regions in the application (we use a rescaled covariate z to lie on the unit interval $[0, 1]$, according to the formula $z_u = (z - z^{min})/(z^{max} - z^{min})$). Using a local linear framework is therefore strictly preferred.

The association between birth weight and mortality at low ages may be strongly affected by medical interventions in the first years of life. In contrast, at higher ages, biological mechanisms may drive the association. At the same time, survival up to late adulthood means survival into the 1960s and beyond, allowing the individual to benefit from medical innovations in the mid 20th century. A single function g is not necessarily able to fit such widely differing explanations. It is therefore interesting to see whether the estimated g changes if we truncate longevity from below at, say, age 40. Figure 7 plots the shape of g for that case. The results do not fundamentally differ from those in Figure 4. The mortality rate at very low birth weights is now point-estimated to be lower. This may be due to improvements in medical technology in the mid 20th century. Alternatively, dynamic selection may cause the frailest individuals among those with low birth weight to have died before age 40, causing an attenuation of the association beyond age 40. The "dynamic selection" explanation is at odds with the model that does not allow for systematic ex ante unobserved heterogeneity. However, note that the truncation of low longevities does not in fact entail the kind of simple attenuation of birth weight effects that one may expect to observe in case of dynamic selection. Specifically, the point estimate for the "optimal birth weight" shifts slightly to the right and lies now at about 4000g. In any case, whatever the differences between Figures 4 and 7, one should keep in mind that the confidence bands in Figure 7 are wider than in Figure 4, especially at extreme birth weight values.

7.3 Comparison to a parametric specification for g

To compare the performance of the estimator with a parametric specification, we replace the nonparametric function g with a quadratic polynomial,

$$g(z; \beta) = \exp\{\beta_0 + \beta_1 z + \beta_2 z^2\} \tag{29}$$

The parameters β_0, β_1 and β_2 are estimated with maximum likelihood. The estimates

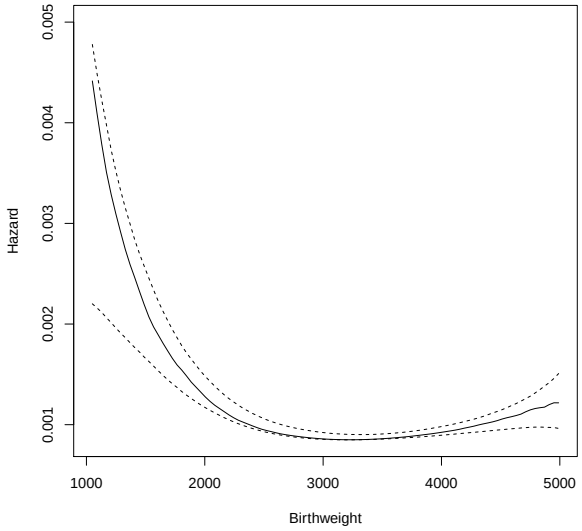


Figure 4: *Estimation results for g using the local linear estimator. The y-axis reports the estimated hazard, the x-axis depicts birth weights.*

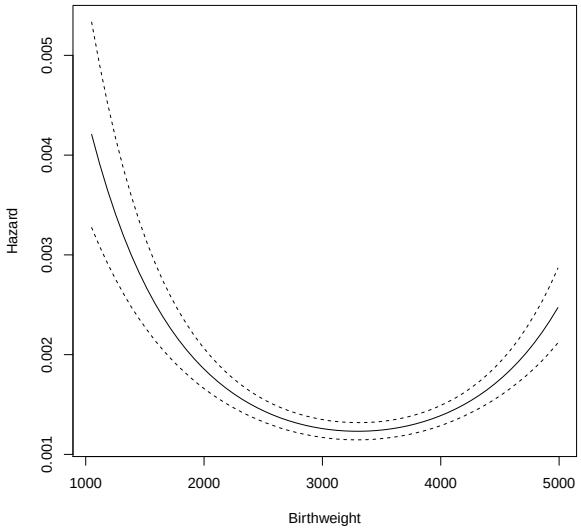


Figure 5: *Estimated parametric covariate hazard function.*

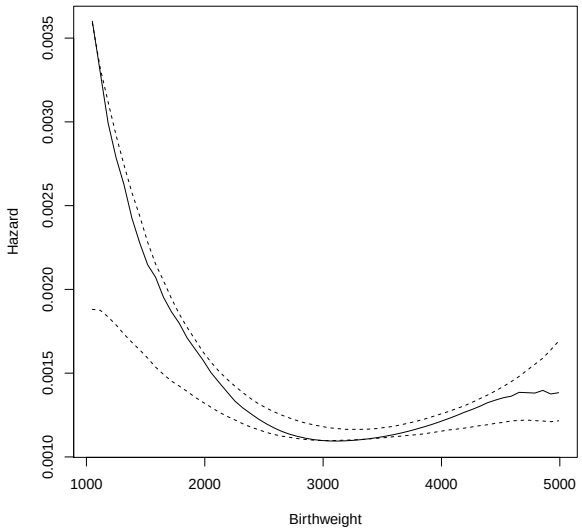


Figure 6: *The estimated nonparametric function with social class contained in the covariate vector, included in the baseline hazard, full range of birth weights.*

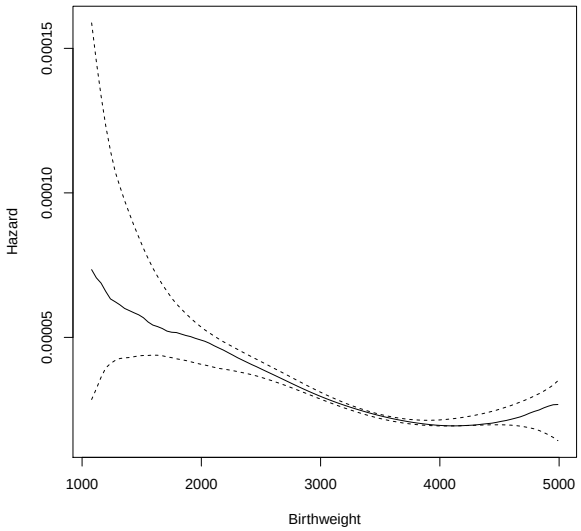


Figure 7: *The estimated g when using a truncated sample of individuals with longevity exceeding age 40.*

	$\widehat{\theta}_1$		$\widehat{\theta}_2$	
Semiparametric Model (only θ_2)	–		0.000096	(0.000000000003)
Semiparametric Model	0.041	(0.000042)	0.00095	(0.000000003)
Parametric Model	0.042	(0.0000048)	0.00095	(0.0000000003)
Men	0.04	(0.000007)	0.0001	(0.000000000005)
Women	0.036	(0.0001)	0.001	(0.00000000008)

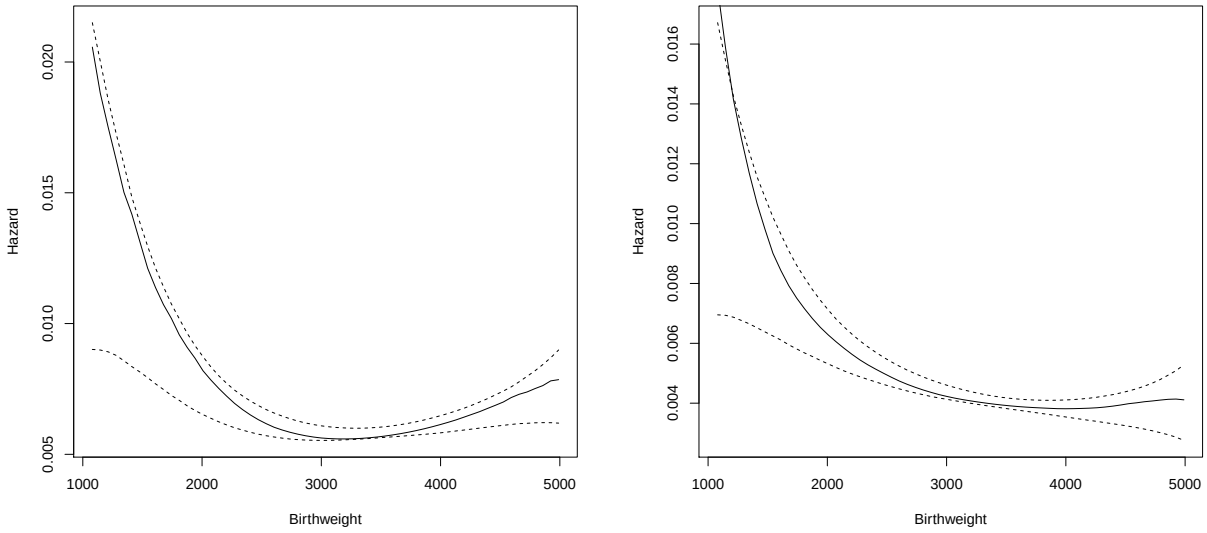
Table 3: *Parameter estimates for θ_1 and θ_2 in a model with parental social class at birth, standard errors are shown in parentheses.*

(standard errors) for β_1 and β_2 are: $-4.64(0.65)$ and $3.98(0.44)$, respectively. Further, $\widehat{\theta} = 9.56e-05(2.67e-12)$, which is very close to what we find in the semiparametric analysis. The results are shown in Figure 5. While the differences are not large, the parametric analysis overestimates the mortality risk at high birth weights. The larger point is, of course, that it is not possible, ex-ante, to know the exact parametric form of the hazard.

One may argue that the inclusion of z^2 as a covariate in the parametric $g(z;\beta)$ in (29) is likely to lead to a bad fit at very high values of z. As an alternative, we replace z and z^2 in $g(z;\beta)$ by $\log z$ and $(\log z)^2$. However, it turns out that the estimation results do not add new insights to those above.

7.4 Including additional covariates

Social class at birth. Our approach allows us to extend the vector $X(t)$ to include more covariates than just time t. As mentioned above, it is important to avoid omitted covariates in order to prevent unobserved heterogeneity bias. To proceed, we parameterise our parametric function as $\alpha_\theta(X(t)) = \exp\{\theta_1 X^d + \theta_2 t\}$, where X^d denotes parental social class at the birth of the individual. The relevant estimation results are depicted in Figure 6. The shape of the estimated risk is not materially different from the estimate ignoring social class. The parameter estimates are reported in the first line in Table 3. Belonging to a lower social class increases mortality hazard. For a fully parametric model the results are in row 2 in Table 3. The parameter estimates are very similar to those for our semiparametric model.



Males. *Females.*

Figure 8: *Semiparametric estimation results for g using the local linear estimator, controlling for social class at birth and stratified by gender.*

Stratifying by gender. Gender is known to have a large effect on mortality. We stratify our empirical analysis by gender and estimate the impact of birth weight, age and social class separately for men and women. The parameter estimates are shown in Table 3. The estimates for the impact of social class (θ_1) do not differ substantially between men and women, whereas the age dependence estimate (θ_2) is larger for men than for women. For birth weight, the effects differ by gender; see Figure 8. The left panel depicts the effect for men and the right panel for women. The increased risk at high birth weight is much more pronounced for men than for women. Apparently, birth weights above 4000g present a risk factor for men but not for women.

8 Conclusion

In the paper we specify a very general class of semiparametric survival models and we develop an estimation technique for these models. The class of models includes models in which the hazard rate is a nonparametric function of covariates. We argue that our paper serves a need for estimation methods for such models, since they cannot be recast in the Cox model. Indeed, our class of models is more general than other semiparametric model

classes studied in the literature. We prove that our estimator is consistent *and* efficient. In simulations we show that our estimator performs well with sample sizes that are common in epidemiology and econometrics. In the estimation procedure, we recommend to use local linear kernel estimation for the nonparametric function of the covariates.

We apply the estimator to study the association between birth weight and late-life mortality, which is seen as an issue of great interest due to its relevance for the "developmental origins" theory of late-life health. This application allows us to assess the performance of the estimator under realistic empirical conditions, with a sample size of about 13,000 individuals of which about half have right-censored lifetimes. We find a nonmonotonic relationship. This is preserved if we control for social class at birth. The relationship cannot be captured with a simple parametric polynomial, confirming the usefulness of our approach. Separate analyses by gender show that the nonmonotonicity is mostly due to an increased later-life mortality risk for men with high birth weight.

The application very much focuses on the flexible estimation of covariate effects. We should point out that our approach is also useful if one aims to estimate a parametric part of the hazard rate in the presence of some covariates whose effects cannot be captured parametrically because there is insufficient prior knowledge on their functional form. The effects of such covariates are then nuisance functions, but they nevertheless need to be taken into account when estimating the parameters of interest. Our approach deals with that.

As an obvious topic for further research one may consider the inclusion of unobserved heterogeneity or frailty terms in the individual hazard rates. This is potentially important because a model specification with many covariates leads to a curse of dimensionality, while at the same time the omission of covariates without controlling for unobserved heterogeneity may lead to biased inference. A different but related topic for further research may be to reduce the dimensionality of the model by assuming a single-index structure for the parametric part of the hazard rate as a function of covariates and markers.

References

Aalen, O. (1978), Nonparametric inference for a family of counting processes, *The Annals of Statistics* 6(4), 701–726.

Ahlgren, M., J. Wohlfahrt, L.W. Olsen, T.I.A. Sørensen and M. Melbye (2007), Birth weight and risk of cancer, *Cancer* 110, 412–419.

Almond, D., K. Chay and D. Lee (2005), The costs of low birth weight, *Quarterly Journal of Economics* 120, 1031–1083.

Almond, D. and Currie, J. (2011), Killing me softly: The fetal origins hypothesis, *Journal of Economic Perspectives* 25, 153–172.

Andersen, P.K., Borgan,Ø., Gill, R.D. and Keiding, N. (1993), Statistical models based on counting processes, *Springer*.

van den Berg, G.J. (1990), Nonstationarity in job search theory, *Review of Economic Studies* 57, 255–277.

van den Berg, G.J. (2001), Duration models: specification, identification, and multiple durations, in: J.J. Heckman and E. Leamer (eds.), *Handbook of Econometrics, Volume V*, North-Holland, Amsterdam.

van den Berg, G.J. and B. Modin (2013), Economic conditions at birth, birth weight, ability, and the causal path to cardiovascular mortality, Working paper, IZA Bonn.

Blanchard, O.J. and Diamond, P. (1994), Ranking, unemployment duration, and wages, *Review of Economic Studies* 61, 417–434.

Bickel, P.J., Klaassen, C.A.J., Ritov, Y. and Wellner, J.A. (1993), Efficient and adaptive estimation for semiparametric models, *Baltimore: Johns Hopkins University Press*.

Borgan, Ø. (1984), Maximum likelihood estimation in parametric counting process models, with applications to censored failure time data, *Scandinavian Journal of Statistics* 11 (1) 1–16.

Cox, D.R. (1972), Regression models and life tables, *Journal of the Royal Statistical Society. Series B* 34 (2), 187–220.

Curhan, G.C., Chertow, G.M., Willett, W.C., Spiegelman, D., Colditz, G.A., Manson, J.E., Speizer, F.E. and Stampfer, M.J. (1996), Birth weight and adult hypertension and obesity in women, *Circulation* 94, 1310–1315.

Dabrowska, D.M. (1987), Non-parametric regression with censored survival time data, *Scandinavian Journal of Statistics* 14, 181–197.

Dabrowska, D.M. (1997), Smoothed Cox regression, *The Annals of Statistics* 25(4), 1510–1540.

Dabrowska, D.M. (2006), Estimation in a class of semiparametric transformation models, *IMS Lecture Notes – Monograph Series 2nd Lehmann Symposium – Optimality* 49, 131–169.

Davey Smith, G. (2005), Epidemiological Freudianism, *International Journal of Epidemiology* 34, 1–2.

Davison, A.C. (2002), Statistical Models, *Cambridge Series in Statistical and Probabilistic Mathematics*.

Fledelius, P., Guillen, M., Nielsen, J.P. and Vogelius, M. (2004), Two-dimensional hazard estimation for longevity analysis, *Scandinavian Actuarial Journal* 2, 133–156.

Gámiz Pérez, M.L., Janys, L., Martínez Miranda, M.D. and Nielsen, J.P. (2013), Bandwidth selection in marker dependent kernel hazard estimation, *Computational Statistics and Data Analysis* 68, 155–169.

Huxley, R., C.G. Owen, P.H. Wincup, D.G. Cook, J. Rich-Edwards, G. Davey Smith and R. Collins (2007), Is birth weight a risk factor for ischemic heart disease in later life?, *American Journal of Clinical Nutrition* 85, 1244–1250.

Kalbfleisch, J.D. and Prentice, R.L. (1980), The Statistical Analysis of Failure Time Data, *Wiley*, New York.

Kuh, D. and Ben-Shlomo, Y. (2004), A Life Course Approach to Chronic Disease Epidemiology, *Oxford University Press*, Oxford.

Leon, D.A., Lithell, H.O., Vågerö, D., Koupilová, I., Mohsen, R., Berglund, L., Lithell, U.B. and McKeigue P.M. (1998), Reduced fetal growth rate and increased risk of death from ischaemic heart disease: cohort study of 15000 Swedish men and women born 1915–29, *British Medical Journal* 317, 241–245.

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

Lobell, H., Schön, L. and Krantz, O. (2007), Observations from the new Swedish Historical National Accounts, Working paper, Lund University.

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

Modin, B. (2002), Setting the Scene for Life: Longitudinal Studies of Early Social Disadvantage and Later Life Chances, *Centre for Health Equity Studies*, Stockholm.

National Central Bureau of Statistics (1969), *Historical Statistics of Sweden*, Part 1. Population, Second edition, 1720–1967, National Central Bureau of Statistics, Stockholm.

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

Nielsen, J.P., Linton, O. and Bickel, P.J. (1998), On a semiparametric survival model with flexible covariate effect, *The Annals of Statistics*, 215–241.

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

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

Osler, M., Andersen, A.M.N., Due, P., Lund, R., Damsgaard, M.T. and Holstein, B.E. (2003), Socioeconomic position in early life, birth weight, childhood cognitive function, and adult mortality. A longitudinal study of Danish men born in 1953, *Journal of Epidemiology and Community Health* 57, 681–686.

Parsons, T.J., Power, C., Logan, S. and Summerbell, C.D. (1999), Childhood predictors of adult obesity: a systematic review, *International Journal of Obesity* 23, 1–107.

Poulter, N.R., Chang, C.L., MacGregor, A.J., Snieder, H. and Spector, T.D. (1999), Association between birth weight and adult blood pressure in twins: historical cohort study, *British Medical Journal* 319, 1330–1333.

Rajaleid, K., Manor, O. and Koupil, I. (2008), Does the strength of the association between foetal growth rate and ischaemic heart disease mortality differ by social circumstances in early or later life?, *Journal of Epidemiology and Community Health* 62(5), e6.

Rasmussen, F. and Johansson, M. (1998), The relation of weight, length and ponderal index at birth to body mass index and overweight among 18-year-old males in Sweden, *European Journal of Epidemiology* 14, 373–380.

Rasmussen, K.M. (2001), The "fetal origins" hypothesis: challenges and opportunities for maternal and child nutrition, *Annual Review of Nutrition* 21, 73–95.

Spierdijk, L. (2008), Nonparametric conditional hazard rate estimation: a local linear approach, *Computational Statistics and Data Analysis* 52, 2419–2434.

Sundin, J. and S. Willner (2007), Social Change and Health in Sweden: 250 Years of Politics and Practice, Swedish National Institute of Public Health, Östersund.

van der Vaart, A.W. (2000), *Asymptotic Statistics*, Cambridge University Press.

Appendix A Technical Appendix

A.1 Assumptions

We make use of the following assumptions:

(A1) For $0 \leq t \leq 1$ it holds that $Pr\{Z_i(t) \in \mathcal{Z}\} = 1$ and $Pr\{X_i(t) \in \mathcal{X}_1 \times \mathcal{X}_2\} = 1$ for compact subsets $\mathcal{Z}$, $\mathcal{X}_1$ of $\mathbb{R}^{d_z}$ or $\mathbb{R}^{d_x^1}$, respectively, and for a finite set $\mathcal{X}_2 \subset \mathbb{R}^{d_x^2}$ with $d_z \geq 1$, $d_x^1, d_x^2 \geq 0$ and $d_x := d_x^1 + d_x^2 \geq 1$. The sets $\mathcal{Z}$, $\mathcal{X}_1$ and $\mathcal{X}_2$ do not depend on t. The covariate vector $(X_i(t), Z_i(t))$ has a density $f_t(x, z)$ with respect to $\nu = \nu_x \times \nu_z$ where ν_z is the Lebesgue measure on $\mathbb{R}^{d_z}$ and ν_x is a product of a d_x^2-dimensional Lebesgue measure and the counting measure on $\mathcal{X}_1$. For a neighborhood $\mathcal{N}_0$ of θ_0 we assume that for fixed x_2 the functions $g(z)$, $\alpha(x_1, x_2; \theta)$ and $f_t(x_1, x_2, z)$ are strictly positive and continuous on $\mathcal{Z}$, $\mathcal{X}_1 \times \mathcal{N}_0$, and $[0, T] \times \mathcal{X}_1 \times \mathcal{Z}$, respectively. Furthermore, for $\theta \in \mathcal{N}_0$ and $z \in \mathcal{Z}$ the second derivatives of $g(z)$ and of $e_\theta(z)$ with respect to z exist and are continuous in θ and z. For the definition of e_θ see equation (9).

(A2) The function $\alpha(x; \theta)$ is Lipschitz continuous w.r.t. θ with a Lipschitz constant that does not depend on x, i.e. there exists a constant $C > 0$ such that $|\alpha(x; \theta_1) - \alpha(x; \theta_2)| \leq C\|\theta_1 - \theta_2\|$ for all $\theta_1, \theta_2 \in \mathcal{N}_0$ and $x \in \mathcal{X}_1 \times \mathcal{X}_2$.

(A3) The kernel K is a multivariate kernel function $K(y) = k(y_1) \cdot ... \cdot k(y_{d_z})$ where k is a symmetric, Lipschitz continuous probability density function with compact support, say $[-1, 1]$. It holds that $b_{max} := \max\{b_1^0, ..., b_{d_z}^0\} \to 0$ and that $(nb_{prod})^{-1}(\log n) \to 0$.

(A4) For all $\theta \in \mathcal{N}_0$ it holds that $\alpha(x_1, x_2; \theta)/e_\theta(z) \neq \alpha(x_1, x_2; \theta_0)/e_{\theta_0}(z)$ with positive ν-measure.

(A5) It holds that $b_{max} = o(n^{-1/4})$.

(A6) The function $\alpha(x; \theta)$ is twice differentiable w.r.t. θ and the derivative is Lipschitz continuous w.r.t. θ with a Lipschitz constant that does not depend on x, i.e. there exists a constant $C > 0$ such that $\|\frac{\partial^2}{\partial\theta\partial\theta^T}\alpha(x; \theta_1) - \frac{\partial^2}{\partial\theta\partial\theta^T}\alpha(x; \theta_2)\| \leq C\|\theta_1 - \theta_2\|$ for all $\theta_1, \theta_2 \in \mathcal{N}_0$ and $x \in \mathcal{X}_1 \times \mathcal{X}_2$.

(A7) The semiparametric information matrix $\mathcal{I}_0$ is finite and nonsingular.

(A8) θ_0 is an interior point of Θ.

A.2 Proof of Theorem 1

A.2.1 Proof of (i).

We will show that

$$\sup_{\theta\in\mathcal{N}_0}|Q_n(\theta)-Q(\theta)|=o_P(1). \tag{30}$$

We now argue that this implies the claim of (i). Put

$$d_\theta(x,z)=(\alpha(x_1,x_2;\theta)e_{\theta_0}(z))/(\alpha(x_1,x_2;\theta_0)e_\theta(z)).$$

From (A1) and (A4) we get that $\ln d_\theta(x,z)-d_\theta(x,z)+1\neq 0$ with positive ν-measure for $\theta\in\mathcal{N}_0$ with $\theta\neq\theta_0$. Note that $\ln(x)-x+1<0$ for $x\neq 1$. Thus we have that $Q(\theta)<Q(\theta_0)$ for $\theta\in\mathcal{N}_0$ with $\theta\neq\theta_0$. Since $Q(\theta)$ is continuous in θ we get the statement of (i), see e.g. Theorem 5.7 in van der Vaart (2000). It remains to show (30). We will show that

$$\sup_{\theta\in\mathcal{N}_0}\left|Q_n(\theta)-\overline{Q}_n(\theta)\right|=o_P(1), \tag{31}$$

$$\sup_{\theta\in\mathcal{N}_0}\left|\overline{Q}_n(\theta)-Q(\theta)\right|=o_P(1). \tag{32}$$

Claim (32) follows by a uniform law of large numbers. Note that $\overline{Q}_n(\theta)$ is an average of i.i.d. summands that are continuous in θ and uniformly bounded. For the proof of (31) it suffices to show that

$$sup_{\theta\in\mathcal{N}_0}n^{-1}\sum_{i=1}^n\int\left[\ln\widehat{g}_{\theta,-i}\{Z_i(u)\}-\ln g_\theta\{Z_i(u)\}\right]dN_i(u)\to_p 0,$$

$$sup_{\theta\in\mathcal{N}_0}n^{-1}\sum_{i=1}^n\int\alpha\{X_i(u);\theta\}\left[\widehat{g}_{\theta,-i}\{Z_i(u)\}-g_\theta\{Z_i(u)\}\right]Y_i(u)du\to_p 0;$$

These two claims follow from

$$sup_{1\leq i\leq n,\theta\in\mathcal{N}_0,z\in\mathcal{Z}}|\widehat{g}_{\theta,-i}(z)-g_\theta(z)|=O_P((nb_{prod})^{-1/2}(\log n)^{1/2}+b_{max})=o_P(1),$$

see Condition (A3). The result on the uniform convergence of $\widehat{g}_{\theta,-i}$ follows by standard kernel smoothing theory. One uses that $|\widehat{g}_{\theta,-i}(z)-\widehat{g}_\theta(z)|=O_P((nb_{prod})^{-1})$, uniformly for $1\leq i\leq n,\theta\in\mathcal{N}_0,z\in\mathcal{Z}$. Then one argues that it suffices to prove uniform convergence over a grid of points θ and z values where the number of grid points increases polynomially. At this point one uses Lipschitz continuity of the kernel K and α, see (A1) and (A2). Then one shows uniform convergence over this grid by application of an exponential inequality for $\widehat{g}_{\theta,-i}(z)-g_\theta(z)$. □

A.2.2 Proof of (ii).

As outlined in Section (5) we have to show (22) and (24). For the proof of (22) it suffices to show the following claims, see also (21).

$$n^{-1/2}\sum_{i=1}^{n}\int\left\{\frac{\partial\hat{\mu}_{\theta_0,-i}}{\partial\theta}-\frac{\partial\overline{\mu}_{\theta_0}}{\partial\theta}\right\}\{X_i(u),Z_i(u)\}dM_i(u) \quad\to_p\quad 0, \tag{33}$$

$$n^{-1/2}\sum_{i=1}^{n}\int\frac{\partial\hat{\mu}^*_{\theta_0,-i}}{\partial\theta}\{Z_i(u)\}dM_i(u) \quad\to_p\quad 0, \tag{34}$$

$$n^{-1/2}\sum_{i=1}^{n}\int\frac{\partial\hat{\mu}_{\theta_0,-i}}{\partial\theta}\{X_i(u),Z_i(u)\}[\alpha\{X_i(u);\theta_0\}(g^*_{\theta_0,-i}-g)\{Z_i(u)\}Y_i(u)du \quad\to_p\quad 0. \tag{35}$$

Claim (35) follows from $\sup_{z\in\mathcal{Z},1\le i\le n}|(g^*_{\theta_0,-i}-g)(z)|=O_P(b^2)=o_P(n^{-1/2})$, see (A5). For the proof of (33) we apply the results in Mammen and Nielsen (2007). For the determination of $\frac{\partial\hat{\mu}_{\theta_0,-i}}{\partial\theta}(x,z)$ one has to calculate

$$\begin{aligned}
\hat{v}_{\theta,-i}(z) &= n^{-1}\sum_{j\neq i}\int K_b\{z-Z_j(u)\}dN_j(u),\\
\hat{w}^0_{\theta,-i}(z) &= n^{-1}\sum_{j\neq i}\int K_b\{z-Z_j(u)\}\alpha\{X_j(u);\theta\}Y_j(u)du,\\
\hat{w}^1_{\theta,-i}(z) &= n^{-1}\sum_{j\neq i}\int K_b\{z-Z_j(u)\}\frac{\partial\alpha}{\partial\theta}\{X_j(u);\theta\}Y_j(u)du.
\end{aligned}$$

Define $\frac{\partial\hat{\mu}^c_{\theta_0,-i}}{\partial\theta}(x,z)$ as $\frac{\partial\hat{\mu}_{\theta_0,-i}}{\partial\theta}(x,z)$ but with $\hat{v}_{\theta,-i}(z),\hat{w}^0_{\theta,-i}(z),\hat{w}^1_{\theta,-i}(z)$ replaced by

$$\begin{aligned}
\hat{v}^c_{\theta,-i}(z) &= \min\left\{c^{-1},\max\left\{c,n^{-1}\sum_{j\neq i}\int K_b\{z-Z_j(u)\}dN_j(u)\right\}\right\},\\
\hat{w}^{0,c}_{\theta,-i}(z) &= \min\left\{c^{-1},\max\left\{c,n^{-1}\sum_{j\neq i}\int K_b\{z-Z_j(u)\}\alpha\{X_j(u);\theta\}Y_j(u)du\right\}\right\},\\
\hat{w}^{1,c}_{\theta,-i}(z) &= \min\left\{c^{-1},\max\left\{c,n^{-1}\sum_{j\neq i}\int K_b\{z-Z_j(u)\}\frac{\partial\alpha}{\partial\theta}\{X_j(u);\theta\}Y_j(u)du\right\}\right\}.
\end{aligned}$$

If $c>0$ is chosen small enough one can check that $\hat{v}^c_{\theta,-i}(z)=\hat{v}_{\theta,-i}(z)$, $\hat{w}^{0,c}_{\theta,-i}(z)=\hat{w}^0_{\theta,-i}(z)$, $\hat{w}^{1,c}_{\theta,-i}(z)=\hat{w}^1_{\theta,-i}(z)$ for all $1\le i\le n$, $\theta\in\mathcal{N}_0$ and $z\in\mathcal{Z}$, with probability tending to one. Thus, $\frac{\partial\hat{\mu}^c_{\theta_0,-i}}{\partial\theta}(x,z)=\frac{\partial\hat{\mu}_{\theta_0,-i}}{\partial\theta}(x,z)$ for all $z\in\mathcal{Z}$ and $x\in\mathcal{X}_1\times\mathcal{X}_2$, with probability tending to one. We now apply Corollary 2 in Mammen and Nielsen (2007) with $h_i^{(n)}\{X_i(u),Z_i(u)\}$ equal to the leave-one-out version $n^{-1/2}\left\{\frac{\partial\hat{\mu}_{\theta_0,-i}}{\partial\theta}-\frac{\partial\overline{\mu}_{\theta_0}}{\partial\theta}\right\}\{X_i(u),Z_i(u)\}$ and with $h_{i,j}^{(n)}$ in

this corollary equal to two-leave-out analogues. Then Corollary 2 implies (33) if one verifies that

$$\sum_{i=1}^{n} \rho_i^2 + n \sum_{i=1}^{n} \delta_i^2 \to 0, \tag{36}$$

where $\rho_i^2 = E[\int h_i^{(n)}\{X_i(u), Z_i(u)\}^2 \alpha\{X_i(u);\theta\} g\{Z_i(u)\}\, du]$ and $\delta_i^2 = \max_{1\le j\le n} E[\int \{h_i^{(n)} - h_{i,j}^{(n)}\}\{X_i(u), Z_i(u)\}^2 \alpha\{X_i(u);\theta\} g\{Z_i(u)\}\, du]$. Now, (36) can be easily verified because of $\max_{1\le i\le n} \rho_i^2 = O(n^{-2} b_{prod}^{-1})$ and $\max_{1\le i\le n} \delta_i^2 = O(n^{-3} b_{prod}^{-1})$. Thus, we get (33).

For the proof of (22) it remains to check (34). Note first that

$$n^{-1/2} \sum_{i=1}^{n} \int \frac{\partial \mu_{\theta_0}^*}{\partial \theta}\{Z_i(u)\} dM_i(u) \quad \to_p \quad 0, \tag{37}$$

$$n^{-1/2} \sum_{i=1}^{n} \int \left\{ \frac{\partial \hat{\mu}_{\theta_0}^*}{\partial \theta} - \frac{\partial \mu_{\theta}^*}{\partial \theta} \right\}\{Z_i(u)\} dM_i(u) \quad \to_p \quad 0, \tag{38}$$

where

$$\frac{\partial \mu_{\theta_0}^*}{\partial \theta}\{Z_i(u)\} = e_{\theta_0}^{-1}\{Z_i(u)\} \int \frac{\partial \overline{\mu}_{\theta_0}}{\partial \theta}\{x, Z_i(u)\} \alpha\{x;\theta_0)\} f_t\{Z_i(u), x\} y(t) dt\; dx. \tag{39}$$

Since

$$\frac{\partial \overline{\mu}_{\theta_0}}{\partial \theta}(x,z) = \frac{\partial \ln \alpha}{\partial \theta}(x;\theta) - \frac{\partial \ln e_\theta}{\partial \theta}(z),\; \frac{\partial e_\theta}{\partial \theta}(z) = \int \frac{\partial \alpha}{\partial \theta}(x;\theta) f(z,x) y(u) du$$

we have, on substituting this into (39) and using

$$\int \frac{\partial \alpha}{\partial \theta}\{X_i(t);\theta_0\} f\{Z_i(u), X_i(t)\} y(t) dt = e_{\theta_0}^{-1} \frac{\partial e_{\theta_0}}{\partial \theta}\{Z_i(u)\} \int \alpha\{X_i(t);\theta_0\} f\{Z_i(u), X_i(t)\} y(t) dt,$$

that $\partial \mu_{\theta_0}^*\{Z_i(u)\}/\partial\theta = 0$ and (37) holds immediately. The final proof of (38) is very similar to that of (33) above.

For the proof of statement (ii) of the theorem it remains to check (24). We will show the following expansions for sequences δ_n with $\delta_n \to 0$. These expansions immediately imply (24).

$$\sup_{|\theta-\theta_0|\le\delta_n} \left| \hat{H}_1(\theta) + \int\int \frac{\partial \overline{\mu}_\theta}{\partial \theta} \frac{\partial \overline{\mu}_\theta}{\partial \theta^T}(x,z) \alpha(x,\theta) g_\theta(z) f_u(x,z) y(u) dz\; dx\; du \right| = o_p(1), \tag{40}$$

$$\sup_{|\theta-\theta_0|\le\delta_n} \left| \hat{H}_j(\theta) \right| = o_p(1), \text{ for } j = 2, ..., 5 \tag{41}$$

where

$$
\begin{aligned}
\hat{H}_1(\theta) &= -n^{-1}\sum_{i=1}^{n}\int \frac{\partial\hat{\mu}_\theta}{\partial\theta}\frac{\partial\hat{\mu}_\theta}{\partial\theta^T}\{X_i(u),Z_i(u)\}\alpha\{X_i(u);\theta\}g_\theta\{Z_i(u)\}Y_i(u)du,\\
\hat{H}_2(\theta) &= n^{-1}\sum_{i=1}^{n}\int \frac{\partial^2\hat{\mu}_\theta}{\partial\theta\partial\theta^T}\{X_i(u),Z_i(u)\}\\
&\quad\times[\alpha\{X_i(u);\theta_0\}g\{Z_i(u)\ -\alpha\{X_i(u);\theta\}g_\theta\{Z_i(u)\}]Y_i(u)du,\\
\hat{H}_3(\theta) &= n^{-1}\sum_{i=1}^{n}\int \left[\frac{\partial^2\hat{\mu}_\theta}{\partial\theta\partial\theta^T}-\frac{\partial^2\hat{\mu}_{\theta_0}}{\partial\theta\partial\theta^T}\right]\{X_i(u),Z_i(u)\}dM_i(u),\\
\hat{H}_4(\theta) &= n^{-1}\sum_{i=1}^{n}\int \frac{\partial^2\hat{\mu}_{\theta_0}}{\partial\theta\partial\theta^T}\{X_i(u),Z_i(u)\}dM_i(u),\\
\hat{H}_5(\theta) &= -n^{-1}\sum_{i=1}^{n}\int \left\{\frac{\partial^2\hat{\mu}_\theta}{\partial\theta\partial\theta^T}+\frac{\partial\hat{\mu}_\theta}{\partial\theta}\frac{\partial\hat{\mu}_\theta}{\partial\theta^T}\right\}\{X_i(u),Z_i(u)\}\\
&\quad\times\alpha\{X_i(u);\theta\}\{\widehat{g}_\theta-g_\theta\}\{Z_i(u)\}Y_i(u)du.
\end{aligned}
$$

Note that $\hat{H}_{\theta\theta}(\theta)=\sum_{j=1}^{5}\hat{H}_j(\theta)$. For the proof of (40)–(41) one uses results on the uniform convergence of $\widehat{g}_\theta$ and its first two partial derivatives w.r.t. θ and uniform laws of large numbers. Compare also the proof of part (i) of the theorem for the proof of (41) for $j=4$. □

A.2.3 Proof of (iii).

This follows immediately from (24) and the consistency of $\widehat{\theta}$. □

A.3 Proof of Corollary 1

The asymptotic distribution of $\widehat{g}$ follows directly from Nielsen, Linton and Bickel (1998).

Appendix B The local linear estimator and the discretised estimator

B.1 The local linear estimator

In this subsection we give a definition of the local linear estimator $\widehat{g}^{LL}_{b,\widehat{\theta}}(z)$. This estimator is defined as γ_0 where (γ_0, γ_1) solves

$$\begin{aligned} 0 \quad &\stackrel{!}{=} \quad \sum_{i=1}^{n} \int_0^T \begin{pmatrix} 1 \\ z - Z_i(s) \end{pmatrix} \alpha_\theta(X(s)) K_b(z - Z_i(s)) \alpha_\theta(X(s))^{-1} dN_i(s) \\ &- \quad \sum_{i=1}^{n} \int_0^T \begin{pmatrix} 1 & z - Z_i(s) \\ z - Z_i(s) & (z - Z_i(s))^2 \end{pmatrix} \begin{pmatrix} \gamma_0 \\ \gamma_1 \end{pmatrix} \alpha_\theta(X(s))^2 K_b(z - Z_i(s)) \alpha_\theta(X(s))^{-1} Y_i(s) ds. \end{aligned}$$

Thus we have that

$$\widehat{g}^{LL}_{b,\widehat{\theta}}(z) = \frac{a_{22}(\theta) y_1 - a_{12}(\theta) y_2}{a_{11}(\theta) a_{22}(\theta) - a_{12}(\theta)^2} \tag{42}$$

with $y_1 = \sum_{i=1}^n \int_0^T K_b(z - Z_i(s)) dN_i(s) ds$, $y_2 = \sum_{i=1}^n \int_0^T (z - Z_i(s)) K_b(z - Z_i(s)) dN_i(s) ds$, $a_{11}(\theta) = \sum_{i=1}^n \int_0^T K_b(z - Z_i(s)) \alpha_\theta(X(s)) Y_i(s) ds$, $a_{12}(\theta) = \sum_{i=1}^n \int_0^T (z - Z_i(s)) K_b(z - Z_i(s))$ $\alpha_\theta(X(s)) Y_i(s) ds$, and $a_{22}(\theta) = \sum_{i=1}^n \int_0^T (z - Z_i(s))^2 K_b(z - Z_i(s))$ $\alpha_\theta(X(s)) Y_i(s) ds$.

B.2 Bandwidth selection: integrated squared error and cross-validation

Our estimated stochastic hazard depends on two bandwidths b_1, b_2 that we wish to select from the data. In the following notation $\widehat{g}_\theta$ is equal to $\widehat{g}^{LL}_\theta$ and

$$\widehat{\lambda}(s) = \alpha_{\widehat{\theta}}(X_i(s)) \widehat{g}_{\widehat{\theta}}(Z_i(s)) \tag{43}$$

where $\widehat{\theta}$ depends on b_1, while $\widehat{g}$ depends on $\widehat{\theta}$ and therefore on both bandwidths $b1$ and b_2.

We want to choose b_1, b_2 to minimise the cross-validation score, as introduced by Nielsen and Linton (1995):

$$Q_{CV}(b_1, b_2) = n\{\sum_{i=1}^{n} \int \widehat{\lambda}_i^2(X_i(s), Z_i(s)) Y_i(s) ds - 2 \sum_{i=1}^{n} \int \widehat{\lambda}_i^{i-1}(X_i(s), Z_i(s)) dN_i(s)\} \tag{44}$$

where $\widehat{\lambda}_i^{i-1}(X_i(s), Z_i(s))$ is the leave-one-out version of the estimator.

B.3 Discretised estimators

We use a discrete version of the pseudolikelihood equation (14). Let $E_{rr'}$ be the number of exposures at the point rr' in the two-dimensional grid (with $R \times R'$ gridpoints) and $O_{rr'}$ the number of occurrences (or failures).In case of local constant estimation of g, the discrete estimator for g is:

$$\widehat{g}_\theta(z) = \frac{\sum_{r'=1}^{R'} \sum_{r=1}^{R} K_b(z - Z_{r'}(r)) O_{rr'}}{\sum_{r'=1}^{R'} \sum_{r=1}^{R} K_b(z - Z_{r'}(r)) \alpha(X(r); \theta) E_{rr'}} \tag{45}$$

The discrete estimator for θ follows from the discrete version of the likelihood function:

$$\hat{\ell}(\theta) = \sum_{r'=1}^{R'} \sum_{r=1}^{R} \{\ln[\alpha(X(r); \theta) \widehat{g}_\theta(z)] O_{rr'}\} - \sum_{r'=1}^{R'} \sum_{r=1}^{R} \{[\alpha(X(r); \theta) \widehat{g}_\theta(z)] E_{rr'}\} \tag{46}$$

in which (45) can be inserted. This can in turn be straightforwardly modified into a discretised leave-one-out estimator. Bandwidth selection is accordingly modified, along the lines of Subsection B.2.

With local linear estimation of g, the discrete estimator for g is specified completely analogously. In notation analogous to above:

$$y_1 = \sum_{r'=1}^{R'} \sum_{r=1}^{R} K_b(z - Z_{r'}(r)) O_{rr'}, \qquad y_2 = \sum_{r'=1}^{R'} \sum_{r=1}^{R} (z - Z_{r'}(r)) K_b(z - Z_{r'}(r)) O_{rr'}$$

$$a_{11} = \sum_{r'=1}^{R'} \sum_{r=1}^{R} K_b(z - Z_{r'}(r)) \alpha_\theta(X(r)) E_{rr'}$$

$$a_{12} = \sum_{r'=1}^{R'} \sum_{r=1}^{R} (z - Z_{r'}(r)) K_b(z - Z_{r'}(r)) \alpha_\theta(X(r)) E_{rr'}$$

$$a_{22} = \sum_{r'=1}^{R'} \sum_{r=1}^{R} (z - Z_{r'}(r))^2 K_b(z - Z_{r'}(r)) \alpha_\theta(X(r)) E_{rr'}$$

Again, this can be straightforwardly modified into a discretised leave-one-out estimator.