Econometric Theory, **20**, 2004, 437–463. Printed in the United States of America.
DOI: 10.1017/S0266466604203012

AVERAGE DERIVATIVES FOR HAZARD FUNCTIONS

Tue Gørgens
Australian National University

This paper develops semiparametric kernel-based estimators of risk-specific hazard functions for competing risks data. Both discrete and continuous failure times are considered. The maintained assumption is that the hazard function depends on explanatory variables only through an index. In contrast to existing semiparametric estimators, proportional hazards is not assumed. The new estimators are asymptotically normally distributed. The estimator of index coefficients is root-n consistent. The estimator of hazard functions achieves the one-dimensional rate of convergence. Thus the index assumption eliminates the "curse of dimensionality." The estimators perform well in Monte Carlo experiments.

1. INTRODUCTION

This paper is about modeling failure time data when there is more than one kind of failure. Failure time data are also known in the literature as survival data, duration data, and transition data, and the possibility of several kinds of failure is known as the competing risks problem. Competing risks data are characterized by the presence of two dependent variables: the length of time until failure, Y, and an indicator of the type of failure, S. In addition there may be a q-vector of explanatory variables, X. An example of competing risks is the study by Fallick (1993) of workers' transition from unemployment to employment, where jobs are classified according to industry. Here "failure" refers to the event of finding a job, and the dependent variables are the duration of the unemployment spell and (an indicator of) the industry where employment was taken up.

Competing risks models have long been one of the principal tools in applied econometrics and other fields such as demographics and medical statistics. Other recent studies in economics include, for example, Carling, Edin, Harkman, and Holmlund (1996) on the duration of unemployment spells in Sweden, where spells end with transition into employment, labor market programs, or nonparticipation; Henley (1998) on the duration of residence in own home with transition to other owner occupation, to public-sector rental, private-sector rental

I thank Denise Doiron for stimulating my interest in this research project and Catherine de Fontenay, Hans Christian Kongsted, Lars Muus, seminar participants, and two anonymous referees for comments on an earlier version of the paper. I gratefully acknowledge the hospitality of the University of Aarhus and the University of Copenhagen, where part of this research was undertaken. Address correspondence to: Tue Gørgens, Economics RSSS, Australian National University, Canberra ACT 0200, Australia; e-mail: Tue.Gorgens@anu.edu.au.

 0266-4666/04 $12.00

in the United Kingdom; and Salzberger and Fenn (1999) on the duration of service of judges at the English Court of Appeal ending with retirement or promotion to the House of Lords. Kalbfleisch and Prentice (1980), Cox and Oakes (1984), and Lancaster (1990) give in-depth accounts of the traditional likelihood-based methods of analyzing failure time data, whereas Fleming and Harrington (1991) and Anderson, Borgan, Gill, and Keiding (1993) provide comprehensive treatments of the modern approach based on counting processes.

The concept of a hazard function is central in the analysis of failure time data. In a competing risks setting, a hazard rate is a risk-specific and time-specific failure rate. The hazard rate, $h_s(y|x)$, indicates the rate at which subjects with characteristics x experience failure of type s at time y given that they have not failed before time y. Time here refers to duration, that is, the length of time the subject is at risk of failing. This may be different from calendar time. In the unemployment example, time at risk begins when individuals become unemployed, and the hazard rate at five days refers to the rate at which workers who have been unemployed four days find jobs on the fifth day. (Risk-specific hazard functions are called cause-specific hazard functions by Kalbfleisch and Prentice, 1980, and transition intensities by Lancaster, 1990.)

In many applications interest naturally centers on the relationship between the mean of a dependent variable and a number of explanatory variables. With failure time data, however, it is often more informative to study hazard rates instead of means. Hazard rates will generally vary as time progresses (exhibit duration dependence), and this variation provides important information about the underlying process. For example, although it is useful for economists to know the mean duration of unemployment spells, it is also important for understanding the nature of unemployment to know whether workers become more or less likely to find jobs the longer they have been unemployed. Of course hazard rates may also depend on explanatory variables. It is well documented for instance that the hazard rate out of unemployment depends on the workers' age and education level. As already indicated in the notation, the hazard functions considered in this paper are all conditional on the vector explanatory variables.

This paper develops semiparametric estimation methods for risk-specific hazard functions, assuming that the explanatory variables enter the hazard function through a risk-specific linear combination but otherwise imposing no essential restrictions. Technically, the maintained assumption is that there exist a vector β_s and a function $\hat{h}_s$ such that $h_s(y|x) = \hat{h}_s(y|x'\beta_s)$ for risk s. This is sometimes called an index restriction. The focus of the paper is on estimating the index coefficients, β_s, using an average derivative idea. Once β_s has been estimated, the hazard function function, $\hat{h}_s$, and the integrated hazard function, $\hat{H}_s$, can be estimated using standard nonparametric kernel estimation techniques and $x'\beta_{sn}$ as a proxy for $x'\beta_s$, where β_{sn} denotes an estimator of β_s. This paper considers only continuous explanatory variables; the case of discrete explanatory variables is treated in a companion paper.

Index restrictions are common in many parametric and semiparametric failure time models (some recent examples are Stancanelli, 1999, on unemployment duration in the United Kingdom and Santarelli, 2000, on the duration of firms in the Italian financial sector). Index restrictions are favored by applied researchers, because they are simple and because they facilitate the interpretation of estimation results. Because index coefficients are proportional to the marginal effect of explanatory variables on the hazard function, their relative signs and magnitudes have substantive meaning. The main advantage of the new estimator is that it allows estimation of index coefficients under very weak assumptions. Currently no existing estimator of index coefficients is applicable to competing risks data without assuming additional structure, which is often rejected when tested.

The new estimator can be motivated in a number of ways. The balance of this section describes the relationship between the new estimator and the vast literature on the analysis of failure time data and semiparametric estimation.

Two popular classes of models for failure time data are the accelerated failure time (AFT) models and the proportional hazards (PH) models (also known as Cox models). The AFT models assume that the risk-specific hazard functions have the form $h_s(y|x) = \lambda_s(y\psi_s(x))\psi_s(x)$, where λ_s is known as the baseline hazard and ψ_s is the scale term for the sth kind of failure. The PH models assume that the risk-specific hazard functions have the form $h_s(y|x) = \lambda_s(y)\psi_s(x)$, where again λ_s is known as the baseline hazard and ψ_s is the scale term. Both the AFT and the PH models are often combined with index restrictions. In particular, the most common specifications assume that $\psi_s(x) = e^{x'\beta_s}$ for some vector β_s. AFT and PH index models have been successful in numerous applications, and it is an important property of the new estimators that they are consistent (if not efficient) for these models. The advantage of the new estimator is that the need to choose between alternative classes of index models is avoided. It assumes that the explanatory variables influence the hazard functions through an index but does not otherwise restrict the shape of the hazard functions. It is thus applicable in both AFT and PH settings and also in more general situations.

Currently semiparametric estimators of index coefficients in competing risks models exist only for PH models. Part of the popularity of the PH model is due to the fact that Cox (1972, 1975) has developed a partial likelihood estimator that allows estimation of index coefficients without restricting the shape of the baseline hazard to a particular parametric form. It is, however, still necessary to specify a functional form of the scale term. (Nielsen, Linton, and Bickel, 1998, describe a method for estimating the index coefficients without specifying a functional form of the scale term but instead assuming a particular parametric form for the baseline hazard.) Although the discovery of the partial likelihood estimator greatly simplified modeling, it is not uncommon to test and reject proportionality. Horowitz and Neumann (1992), for example, find evidence of nonproportional hazards in employment duration data. McCall

(1994) analyzes the duration of nonemployment spells and finds strong evidence of nonproportionality. Grambsch and Therneau (1994) find that the treatment effect in survival data for lung cancer patients is not proportional, but diminishes as time passed on. The new estimator extends semiparametric estimation beyond the PH model and further eliminates the need to specify a functional form of the scale term.

Hastie and Tibshirani (1990) extend semiparametric estimation of the PH model in a different direction. Whereas this paper abandons proportionality but maintains the index form, Hastie and Tibshirani generalize the index form but maintain the assumption of proportionality. They replace the standard linearly additive index by a nonlinear additive form. Hastie and Tibshirani provide an example of a data set where the index restriction fails, whereas the generalized form fits well. Although the estimator proposed in the present paper is not suited for such data sets, imposing index restrictions leads to efficiency gains when the restrictions are valid. As mentioned earlier, index restrictions also have the advantage of facilitating the interpretation of estimation results, because index coefficients are proportional to marginal effects. Thus, neither estimator dominates the other.

There is now a substantial literature on estimating index coefficients in situations where the conditional mean of Y given $X = x$ satisfies an index restriction, including methods such as average derivatives for mean functions (Härdle and Stoker, 1989; Powell, Stock, and Stoker, 1989), semiparametric least squares (Ichimura, 1993), maximum rank correlation (Han, 1987; Sherman, 1993), and semiparametric maximum likelihood (Ai, 1997). Generally these estimators are not suited for competing risks data, except under very special circumstances. This is because the conditional mean function need not satisfy any index restriction, even when each risk-specific hazard function does. A simple example is given in Section 5. Existing estimators of index coefficients can be used in applications with uncensored single-risk data, applications with right-censored single-risk data when the censoring mechanism is independent of X, and applications where all risk-specific hazard functions depend on the same index. However, the literature currently contains no semiparametric estimators for multiple-risk data, nor for single-risk data when the censoring mechanism depends on X. The new estimator can be seen as an extension of the average derivative estimator of Powell et al. (1989) to multiple-risk data.

It is possible to estimate hazard functions without imposing any assumptions other than smoothness. References to purely nonparametric estimators are given in Section 4. These estimators can be extremely useful in applications with only one or two explanatory variables, but their rates of convergence decrease dramatically as the number of explanatory variables increases, and they are notoriously unreliable with four or more explanatory variables. This is the now-familiar "curse of dimensionality." Index restrictions are one way of providing sufficient structure to reduce the dimension of the explanatory variables. The new estimator of index coefficients proposed in this paper is root-n consistent, and the rate of convergence of the hazard function estimator is independent of the number of explanatory variables. The new estimator can therefore be seen as an effective means of overcoming the curse of dimensionality in nonparametric estimation.

Finally, right-censored single-risk data, where the censoring mechanism may depend on X, are a special case of multiple-risk data, given that right-censoring formally is equivalent to a separate risk. Therefore the new estimator of β_s has wider applications in the literature on right-censored data. For example, it can be used in the first stage of the Gørgens and Horowitz (1999) semiparametric estimator of the censored transformation (GAFT) model and the Horowitz (1999) semiparametric estimator of the mixed PH model.

The paper is organized as follows. Section 2 considers estimation of index coefficients. Estimation of the variance matrix is described in Section 3. Section 4 discusses estimation of hazard functions. Monte Carlo results are presented in Section 5 and conclusions in Section 6. The Appendix contains the proof of the main theorem.

2. INDEX COEFFICIENT ESTIMATION

To estimate the hazard function for a given risk, the distinction between other risks is not necessary. The subscript s is therefore suppressed until the Monte Carlo section of the paper.

Recall that Y represents the length of time until failure, S is an indicator of the type of failure, and X is a q-vector of explanatory variables. Define the distribution functions

$$F_1(y|x) = \Pr(Y \leq y, S = s | X = x), \tag{1}$$

$$F_2(y|x) = \Pr(Y \geq y | X = x). \tag{2}$$

Using the Stieltjes integral, the integrated hazard function is by definition[1]

$$H(y|x) = \int_0^y \frac{F_1(dv|x)}{F_2(v|x)}. \tag{3}$$

The key assumption of this paper is Assumption 1, which follows. Assumption 1 states that the integrated hazard function satisfies an index restriction. It is easy to show that if the hazard function h is of index form then so is its integral H and vice versa. It is convenient to work with the integrated hazard function, because this avoids the issue of continuous and discrete failure times. The arguments presented here are valid for both continuous and discrete failure times.

Assumption 1. There are a function $\hat{H}$ and a vector β such that $H(y|x) = \hat{H}(y|x'\beta)$ for all y and x.

https://doi.org/10.1017/S0266466604203012 Published online by Cambridge University Press

The new estimator of β is similar to the weighted average derivative estimator by Powell et al. (1989). The idea is simple. If the hazard function is of index form, then[2] $\partial_x H(dy|x) = \partial_2 \hat{H}(dy|x'\beta)\beta$. Let W be a weight function. Then β is proportional to β^* defined by[3]

$$\beta^* = \iint W(y,x)\partial_x H(dy|x)\,dx, \tag{4}$$

provided $\iint W(y,x)\partial_2 \hat{H}(dy|x'\beta)\,dx$ is finite and nonzero. An estimator is defined subsequently by replacing $\partial_x H$ in (4) with a nonparametric estimator.

Let ξ denote the density of X. Define

$$A_1(y,x) = \Pr(Y \le y, S = s|X = x)\xi(x), \tag{5}$$

$$A_2(y,x) = \Pr(Y \ge y|X = x)\xi(x). \tag{6}$$

Note that by equation (3)

$$H(y|x) = \int_0^y \frac{A_1(dv,x)}{A_2(v,x)}. \tag{7}$$

This paper considers the case where $W(y,x) = w(y,x)A_2(y,x)^2$ and w is another weight function. This choice is convenient because it avoids random denominators in the estimation formula for $\partial_x H(dy|x)$. Because

$$\partial_x H(dy|x) = \frac{\partial_x A_1(dy,x)}{A_2(y,x)} - \frac{\partial_x A_2(y,x)A_1(dy,x)}{A_2(y,x)^2}, \tag{8}$$

it follows that

$$\beta^* = \iint w(y,x)A_2(y,x)\partial_x A_1(dy,x)\,dx - \iint w(y,x)\partial_x A_2(y,x)A_1(dy,x)\,dx, \tag{9}$$

provided C is finite and nonzero, where

$$C = \iint w(y,x)A_2(y,x)^2\partial_2 \hat{H}(dy|x'\beta)\,dx. \tag{10}$$

Choosing the weight function w is not complicated. In fact, in many applications w can be omitted. The main purpose of the weight function is to provide a way of ensuring that C is finite and nonzero. In all experiments presented in Section 5, the Monte Carlo section, this is satisfied with $w(y,x) = 1$ for all y and x.

The estimator proposed here consists of replacing unknown functions in (9) by sample analogs based on kernel estimation. The sample available for analysis is assumed to consist of n independent observations $(Y_i, S_i, X_i')'$, $i = 1,2,\ldots,n$. Let b be a bandwidth parameter and let $K: R^q \to R$ be a kernel function. Define $K_b(x) = b^{-q}K(b^{-1}x)$. Then define the estimators

$$A_{1n}(y,x) = \frac{1}{n}\sum_{i=1}^{n} K_b(x - X_i)1(Y_i \le y)1(S_i = s), \tag{11}$$

$$A_{2n}(y,x) = \frac{1}{n}\sum_{i=1}^{n} K_b(x - X_i)1(Y_i \ge y). \tag{12}$$

The estimator of β^* is

$$\begin{aligned}\beta_n^* &= \iint w(y,x)A_{2n}(y,x)\partial_x A_{1n}(dy,x)\,dx - \iint w(y,x)\partial_x A_{2n}(y,x)A_{1n}(dy,x)\,dx \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\int \partial_x K_b(x-X_i)K_b(x-X_j)w(Y_i,x)1(Y_j \ge Y_i)1(S_i = s)\,dx \\ &\quad - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\int K_b(x-X_i)\partial_x K_b(x-X_j)w(Y_i,x)1(Y_j \ge Y_i)1(S_i = s)\,dx.\end{aligned} \tag{13}$$

Computing β_n^* involves evaluating a q-dimensional integral. It is possible to simplify this to q one-dimensional integrals with closed-form solutions by using a polynomial product kernel and weight function, as in the Monte Carlo experiments in Section 5.

Uniform consistency and asymptotic normality of β_n^* are established in Theorem 1, which follows. Assumption 2 defines the sample.

Assumption 2. The sequence $\{(Y_i, S_i, X_i')'\}_{i=1}^n$ is a random sample.

The derivation of the limiting distribution depends on applications of the mean value theorem and Taylor series expansions. Hence, the underlying functions must be smooth. Sufficient conditions are listed as Assumption 3.[4]

Assumption 3. For $k \in N_+$ given subsequently, the following conditions are satisfied.

(1) The q-vector X is absolutely continuous and has density ξ with respect to Lebesgue measure.
(2) ξ is bounded.
(3) $\int |\partial_x^j A_1(dy,\cdot)|$ exist and are bounded and continuous for $j = 1,\ldots,1+k$.
(4) $\partial_x^j A_2$ exist and are bounded and continuous for $j = 1,\ldots,1+k$.

A researcher who wishes to use the estimators must choose a weight function, a bandwidth, and a kernel function. To establish consistency and asymp-

https://doi.org/10.1017/S0266466604203012 Published online by Cambridge University Press

totic normality, it is necessary to restrict the choices. Sufficient conditions are given in Assumptions 4 and 5 and in the theorem itself.

Assumption 4. The weight function $w: R^{1+q} \to R$ satisfies the following conditions.

1. C defined in (10) is finite and nonzero.
2. w is bounded.
3. $\partial_x w$ and $\partial_x^2 w$ exist and are bounded.

Assumption 5. For $k \in N_+$ given subsequently, the kernel function $K: R^q \to R$ satisfies the following conditions.

1. K is a bounded kernel with support $[-1,1]^q$ and the order of K is at least k. That is, $\int K(x)\,dx = 1$ and $\int x^j K(x)\,dx = 0$ for $j = 1,2,\ldots,k-1$.
2. $\partial_x K$ exists and is bounded and continuous on R^q.

These are standard assumptions in the literature on semiparametric estimation. To state the theorem, define[5]

$$\Phi(\mathbf{y},\mathbf{s},\mathbf{x}) = 2\int w(v,\mathbf{x})1(\mathbf{y} \geq v)\partial_x A_1(dv,\mathbf{x}) - 2w(\mathbf{y},\mathbf{x})\partial_x A_2(\mathbf{y},\mathbf{x})1(\mathbf{s} = s)$$
$$+ \int \partial_x w(v,\mathbf{x})1(\mathbf{y} \geq v)A_1(dv,\mathbf{x}) - \partial_x w(\mathbf{y},\mathbf{x})A_2(\mathbf{y},\mathbf{x})1(\mathbf{s} = s)$$
$$- 2\beta^*. \tag{14}$$

It is straightforward to verify that $E\Phi(Y,S,X) = 0$. Define the variance matrix $\Sigma = E\Phi(Y,S,X)\Phi(Y,S,X)'$.

THEOREM 1. *Suppose Assumptions 1–5 hold. Then*

i. *If* $nb^{2q+2} \to \infty$, *then* $|\beta_n^* - E\beta_n^* - n^{-1}\sum_{i=1}^n \Phi(Y_i,S_i,X_i)| = o_p(n^{-1/2})$ *and* $n^{1/2}(\beta_n^* - E\beta_n^*) \to^d N(0,\Sigma)$ *as* $n \to \infty$.
ii. *If* $nb^{2k} \to 0$, *then* $n^{1/2}(E\beta_n^* - \beta^*) \to 0$ *as* $n \to \infty$.
iii. *If* $nb^{2q+2} \to \infty$ *and* $nb^{2k} \to 0$, *then* $|\beta_n^* - \beta^* - n^{-1}\sum_{i=1}^n \Phi(Y_i,S_i,X_i)| = o_p(n^{-1/2})$ *and* $n^{1/2}(\beta_n^* - \beta^*) \to^d N(0,\Sigma)$ *as* $n \to \infty$.

Given the first approximation result in part i of the theorem, asymptotic normality in part i follows from the Lindeberg–Lévy central limit theorem and the Cramér–Wold theorem. Part iii follows immediately from parts i and ii.

The most important conclusions of Theorem 1 are that β_n^* converges at the root-n rate, which is the familiar rate from parametric estimation, and that β_n^* is asymptotically normally distributed. These nice properties are not unexpected, because they are shared with the index coefficient estimators listed in the introduction.

It is worth pointing out that the condition $nb^{2q+2} \to \infty$ in the theorem is determined by the "diagonal" terms where $i = j$ in (13). Examination of the proof of the theorem shows that if these diagonal terms were omitted, the con-

https://doi.org/10.1017/S0266466604203012 Published online by Cambridge University Press

dition $nb^{2q+2} \to \infty$ could be weakened to $nb^{q+2} \to \infty$, which is the same as the requirement in Powell et al. (1989).

3. VARIANCE ESTIMATION

In empirical research it is important to have a measure of the uncertainty associated with an estimate. This section describes how to estimate the variance matrix Σ of β_n^*. For convenience, define $T = (Y, S, X')'$ and $\mathbf{t} = (\mathbf{y}, \mathbf{s}, \mathbf{x}')'$. Then define

$$\rho_0(\mathbf{t}_i, \mathbf{t}_j) = \int \partial_x K_b(x - \mathbf{x}_i) K_b(x - \mathbf{x}_j) w(\mathbf{y}_i, x) 1(\mathbf{y}_j \geq \mathbf{y}_i) 1(\mathbf{s}_i = s)\, dx - \int K_b(x - \mathbf{x}_i) \partial_x K_b(x - \mathbf{x}_j) w(\mathbf{y}_i, x) 1(\mathbf{y}_j \geq \mathbf{y}_i) 1(\mathbf{s}_i = s)\, dx \quad \textbf{(15)}$$

and

$$r_n(\mathbf{t}) = \frac{1}{n} \sum_{j=1}^{n} (\rho_0(\mathbf{t}, T_j) + \rho_0(T_j, \mathbf{t})). \quad \textbf{(16)}$$

It is easy to verify, using equation (13), that

$$\beta_n^* = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \rho_0(T_i, T_j) = \frac{1}{n} \sum_{i=1}^{n} \frac{r_n(T_i)}{2}. \quad \textbf{(17)}$$

Examination of the proof of Theorem 1 suggests estimating Σ by

$$\Sigma_n = \frac{1}{n} \sum_{i=1}^{n} (r_n(T_i) - 2\beta_n^*)(r_n(T_i) - 2\beta_n^*)'. \quad \textbf{(18)}$$

The estimator Σ_n is easily computed alongside β_n^* and does not require additional choice of bandwidths. The properties of Σ_n are investigated through simulation in Section 5, where it performs well.

4. HAZARD FUNCTION ESTIMATION

Once the index coefficients have been estimated, the next question is how to estimate the integrated hazard function, H, and the hazard function itself, h. This section describes how to estimate H by combining the index restriction with an existing nonparametric estimator of H. Estimation of h is not explicitly discussed, but it is straightforward to adapt existing nonparametric estimators in a similar way.

A purely nonparametric estimator of H was first proposed by Beran (1981), who built on the well-known Nelson–Aalen estimator of the unconditional integrated hazard function. Weak convergence and uniform consistency of Beran's estimator are established by Dabrowska (1987, 1989). McKeague and Utikal (1990) extend Beran's estimator to the case of time-dependent explanatory variables and propose an estimator of h based on smoothing of the estimator of H. Nielsen and Linton (1995) propose alternative estimators of H and h that are analogues to the Nadaraya–Watson estimator for nonparametric regression. Li and Doss (1995) and Nielsen (1998) consider local linear estimators; in the framework of local polynomial estimation, the earlier estimators are all local constant estimators.[6]

Any of the aforementioned estimators may be combined with an index restriction. To keep things as simple as possible this paper focuses on Beran's estimator. To estimate H, Beran simply replaces A_1 and A_2 in (7) by the estimators A_{1n} and A_{2n} given in (11) and (12). Specifically,

$$H_n^{NP}(y|x) = \int_0^y \frac{A_{1n}(dv,x)}{A_{2n}(v,x)}. \tag{19}$$

As mentioned in the introduction, generally H_n^{NP} has good properties in applications with just a few explanatory variables, but like other nonparametric estimators H_n^{NP} suffers from the curse of dimensionality and is useless in applications with four or more explanatory variables. The index restriction can be used to eliminate the curse of dimensionality.

Define $Z = X'\beta$. The key to utilizing the index restriction is the fact that, under Assumption 1, $\hat{H}(y|x'\beta)$ is not only the conditional integrated hazard function of Y given $X = x$ but also the conditional integrated hazard function given $Z = x'\beta$. This may not be immediately obvious. To prove it, define the distribution functions

$$\hat{F}_1(y|z) = \Pr(Y \le y, S = s|Z = z), \tag{20}$$

$$\hat{F}_2(y|z) = \Pr(Y \ge y|Z = z). \tag{21}$$

Note that in general $F_j(y|x) \neq \hat{F}_j(y|x'\beta)$ for $j = 1, 2$, with uncensored single-risk data as a notable exception. Let P_z denote the conditional distribution of X given $Z = z$; then $\hat{F}_j(y|z) = \int F_j(y|x)\,P_z(dx)$. By definition $H(dy|x) = F_1(dy|x)/F_2(y|x)$, and by assumption $H(dy|x) = \hat{H}(dy|x'\beta)$. It follows that $F_1(dy|x) = \hat{H}(dy|x'\beta)F_2(y|x)$, whence

$$\begin{aligned}\hat{F}_1(dy|z) &= \int F_1(dy|x)\,P_z(dx)\\ &= \int \hat{H}(dy|x'\beta)F_2(y|x)\,P_z(dx)\\ &= \hat{H}(dy|z)\int F_2(y|x)\,P_z(dx)\\ &= \hat{H}(dy|z)\hat{F}_2(y|z).\end{aligned} \tag{22}$$

https://doi.org/10.1017/S0266466604203012 Published online by Cambridge University Press

It follows that

$$\hat{H}(y|z) = \int_0^y \frac{\hat{F}_1(dv|z)}{\hat{F}_2(v|z)}, \tag{23}$$

whence $\hat{H}$ must be the conditional integrated hazard function given $Z = x'\beta$. The result (23) is important because it implies that estimators of $\hat{H}$ and $\hat{h}$ can be based on probabilities conditional on Z instead of X. This simplifies estimation considerably and allows the use of existing nonparametric estimators.

As an example, consider estimating H by combining the index restriction with Beran's nonparametric estimator. Let ζ denote the density of Z and define the functions

$$\hat{A}_1(y,z) = \Pr(Y \leq y, S = s | Z = z)\zeta(z), \tag{24}$$

$$\hat{A}_2(y,z) = \Pr(Y \geq y | Z = z)\zeta(z). \tag{25}$$

Then by Assumption 1 and equation (23)

$$H(y|x) = \int_0^y \frac{\hat{A}_1(dv|x'\beta)}{\hat{A}_2(v|x'\beta)}. \tag{26}$$

To estimate H, define $Z_{in} = X_i'\beta_n$, where β_n is an estimator of β (e.g., the estimator discussed in the previous section). Let b_z be a bandwidth parameter and let $K_z: R \to R$ be a kernel function. Define $K_{zn}(z) = b_z^{-1} K_z(b_z^{-1} z)$ and the estimators

$$\hat{A}_{1n}(y,z) = \frac{1}{n}\sum_{i=1}^{n} K_{zn}(z - Z_{in}) 1(Y_i \leq y) 1(S_i = s), \tag{27}$$

$$\hat{A}_{2n}(y,z) = \frac{1}{n}\sum_{i=1}^{n} K_{zn}(z - Z_{in}) 1(Y_i \geq y). \tag{28}$$

Then $H(y|x)$ can be estimated by

$$H_n(y|x) = \int_0^y \frac{\hat{A}_{1n}(dv, x'\beta_n)}{\hat{A}_{2n}(v, x'\beta_n)} = \frac{1}{n}\sum_{i=1}^{n} \frac{K_{zn}(x'\beta_n - Z_{in}) 1(Y_i \leq y) 1(S_i = s)}{\hat{A}_{2n}(Y_i, x'\beta_n)}. \tag{29}$$

It is not hard to prove that H_n is uniformly consistent over a compact set, is asymptotically normally distributed, and achieves the rate of convergence for a one-dimensional explanatory variable. Moreover, the asymptotic variance of H_n can be estimated using standard estimators, because root-n consistency of β_n implies that the randomness in H_n that stems from using β_n instead of β disappears asymptotically.

5. MONTE CARLO EXPERIMENTS

This section reports the results of Monte Carlo experiments undertaken to investigate the properties of the new estimators in samples of moderate size. The new estimator (New) is compared to two alternative estimators: the standard implementation of the Cox (1972) partial likelihood-based estimator (Cox), where the scale term is the exponential function (see the introduction), and the average mean derivative estimator of Powell et al. (1989) (PSS). The PSS estimator is implemented using $\ln Y$ instead of Y as the dependent variable to reduce the effect of heteroskedasticity. (The new estimator and the Cox's estimator are invariant to monotone transformations of Y.)

The first set of experiments compares the three estimators in a competing risks setting with two risks and proportional hazards. The second risks can be thought of as censoring, and the focus is on the effect of the degree of censoring. Because the structure imposed by the Cox estimator is valid in these experiments, the Cox estimator is expected to perform best. In contrast, the PSS estimator is inconsistent with censored data. The purpose of these experiments is to get an idea of the efficiency loss associated with the new estimator when the data are known to satisfy PH restrictions.

The Monte Carlo designs are as follows. There are two explanatory variables. The first is standard normally distributed, and the second is distributed chi-square with three degrees of freedom. The explanatory variables are independent. There are two risks with hazard functions $h_s(y|x) = e^{\gamma_s + x'\beta_s}$ and coefficients $\beta_1 = (1/\sqrt{2}, 1/\sqrt{2})'$ and $\beta_2 = (1,0)'$. Varying γ_s changes the expected proportion of failures of the first and second kind. Results are presented for experiments with $\gamma_1 = 0$ and $\gamma_2 = -\infty, -1, 0, 1$. The corresponding proportion of failures of the second kind (the expected degree of censoring) is approximately 0%, 25%, 50%, and 75%.

To compute the new estimator it is necessary to choose a weight function, a kernel function, and a bandwidth. In all experiments, the weight function is simply $w(y,x) = 1$. The kernel function is the product kernel $K((x_1,x_2)') = K_u(x_1)K_u(x_2)$, where K_u is the univariate fourth-order ($k = 4$) kernel

$$K_u(u) = \frac{105}{64}(1 - 5u^2 + 7u^4 - 3u^6)1(|u| \le 1). \tag{30}$$

The bandwidth is chosen separately for each experiment to approximately minimize the statistic D, a measure of error described subsequently, on a 12-point grid ranging from 0.5 to 4.0. Note that the optimal bandwidth is specific to the error measure and that the bandwidth that minimizes D is typically different from the bandwidth that minimizes, say, the root mean square error of the ratio of the coefficient estimates.[7] Developing an optimal data-driven bandwidth selection procedure for the new estimator is left for future research.[8]

Deciding on an appropriate loss function for evaluating and comparing different estimators is complicated by the fact that β_s is identified only up to scale.[9]

Tables 1 and 2 report bias, root mean square error, and mean absolute error for estimates of $\beta^*_{12}/\beta^*_{11}$, $\beta^*_{11}/\|\beta^*_1\|$, and $\beta^*_{12}/\|\beta^*_1\|$ and also the mean D of $\|(\beta^*_{1n}/\|\beta^*_{1n}\|) - (\beta^*_1/\|\beta^*_1\|)\|$, where β^*_{1j} denotes the jth component of β^*_1 and $\|\cdot\|$ denotes the euclidean norm.

The sample size is 200. There are 5,000 Monte Carlo replications in each experiment. The computations are carried out using GAUSS and GAUSS′ pseudo-random number generators. As indicated in the tables, the Cox estimator did not converge in all samples. The results for the Cox estimator exclude these samples.

The first set of simulation results is presented in Table 1. There are three main conclusions. First, the Cox estimator has the lowest bias, root mean square error, and mean absolute error in all experiments. Thus, if the data are known to come from a PH model, the Cox estimator should be used. Second, the performance of all estimators deteriorates with higher levels of censoring. This effect is due to the fact that with, say, 50% censoring only 50% the observations contain information about β_1. It is the number of observations with failure of type 1 that determines the precision of the estimator, not the total number of observations. Third, whereas the PSS estimator performs slightly better than

TABLE 1. Monte Carlo comparison of estimators, proportional hazards models

	β_{12}/β_{11}			$\beta_{11}/\|\beta_1\|$			$\beta_{12}/\|\beta_1\|$			
Est.	B	R	M	B	R	M	B	R	M	D
I: 0% censoring										
New	−0.13	0.24	0.20	0.05	0.09	0.07	−0.06	0.11	0.08	0.11
Cox	0.01	0.12	0.09	0.00	0.04	0.03	0.00	0.04	0.03	0.05
PSS	−0.02	0.23	0.18	0.01	0.08	0.06	−0.02	0.08	0.07	0.09
II: 24% censoring										
New	−0.09	0.22	0.18	0.04	0.08	0.06	−0.05	0.09	0.07	0.09
Cox	0.01	0.14	0.11	0.00	0.05	0.04	0.00	0.05	0.04	0.05
PSS	*−0.27*	*0.32*	*0.29*	*0.10*	*0.12*	*0.11*	*−0.13*	*0.16*	*0.13*	*0.17*
III: 50% censoring										
New	0.08	0.31	0.22	−0.02	0.09	0.07	0.01	0.09	0.07	0.10
Cox	0.02	0.19	0.14	0.00	0.06	0.05	0.00	0.06	0.05	0.07
PSS	*−0.50*	*0.52*	*0.50*	*0.19*	*0.19*	*0.19*	*−0.27*	*0.29*	*0.27*	*0.33*
IV: 75% censoring										
New	0.30	0.79	0.44	−0.07	0.15	0.12	0.04	0.13	0.10	0.16
Cox	0.05	0.31	0.22	−0.01	0.09	0.07	0.00	0.09	0.07	0.10
PSS	*−0.73*	*0.74*	*0.73*	*0.25*	*0.26*	*0.25*	*−0.45*	*0.46*	*0.45*	*0.52*

Note: B, R, and M denote bias, root mean square error, and mean absolute error, respectively, whereas D is the mean of $\|(\beta^*_{1n}/\|\beta^*_{1n}\|) - (\beta^*_1/\|\beta^*_1\|)\|$. Italics indicate experiments where the PSS estimator is inconsistent. The results for the Cox estimator in experiment IV exclude 21 samples where the estimator did not converge.

TABLE 2. Monte Carlo comparison of estimators, nonproportional hazards models

	β_{12}/β_{11}			$\beta_{11}/\|\beta_1\|$			$\beta_{12}/\|\beta_1\|$			
Est.	B	R	M	B	R	M	B	R	M	D
V: $\mu_1(z) = [-\ln(1+\exp(-z)) - 0.806]/0.477$										
New	−0.13	0.24	0.19	0.05	0.09	0.07	−0.06	0.11	0.08	0.11
Cox	*−0.33*	*0.37*	*0.34*	*0.12*	*0.14*	*0.13*	*−0.16*	*0.19*	*0.16*	*0.21*
PSS	−0.13	0.25	0.21	0.05	0.09	0.08	−0.07	0.11	0.09	0.12
VI: $\mu_1(z) = [\ln(1+\exp(z/0.15)) - 2.69]/4.26$										
New	−0.01	0.27	0.20	0.01	0.09	0.07	−0.02	0.09	0.07	0.10
Cox	*0.29*	*0.37*	*0.30*	*−0.09*	*0.11*	*0.09*	*0.08*	*0.09*	*0.08*	*0.12*
PSS	0.01	0.28	0.21	0.00	0.09	0.07	−0.02	0.10	0.07	0.10
VII: $\mu_1(z) = [1/(1+\exp(-z/0.50)) - 0.484]/0.308$										
New	−0.09	0.21	0.17	0.03	0.08	0.06	−0.04	0.09	0.07	0.09
Cox	*−0.17*	*0.25*	*0.21*	*0.06*	*0.09*	*0.08*	*−0.08*	*0.12*	*0.09*	*0.12*
PSS	−0.08	0.22	0.17	0.03	0.08	0.06	−0.04	0.09	0.07	0.09
VIII: $f_\epsilon(u) = (1/\sqrt{2\pi})\exp(-(\frac{1}{2})u^2)$										
New	−0.12	0.27	0.22	0.05	0.10	0.08	−0.06	0.12	0.09	0.12
Cox	*0.08*	*0.22*	*0.16*	*−0.02*	*0.07*	*0.06*	*0.02*	*0.06*	*0.05*	*0.08*
PSS	−0.02	0.23	0.18	0.01	0.08	0.06	−0.02	0.09	0.07	0.09
IX: $f_\epsilon(u) = (1+\exp(u)/0.2)^{-0.2-1}\exp(u)$										
New	−0.10	0.24	0.19	0.04	0.08	0.07	−0.05	0.10	0.08	0.10
Cox	*0.14*	*0.36*	*0.27*	*−0.04*	*0.11*	*0.08*	*0.02*	*0.10*	*0.08*	*0.12*
PSS	−0.02	0.23	0.18	0.01	0.08	0.06	−0.02	0.08	0.07	0.09
X: $f_\epsilon(u) = 0.5/\sqrt{2\pi}\exp(-\frac{1}{2}\ln(u+\sqrt{u^2+1})^2 0.5^2)/\sqrt{u^2+1}$										
New	−0.06	0.10	0.08	0.02	0.04	0.03	−0.02	0.04	0.03	0.04
Cox	*0.06*	*0.27*	*0.18*	*−0.01*	*0.08*	*0.06*	*0.01*	*0.09*	*0.06*	*0.09*
PSS	−0.02	0.54	0.18	0.01	0.08	0.06	−0.02	0.08	0.06	0.09
XI: $\sigma(z) = 0.71\exp(z/2)$										
New	−0.09	0.21	0.17	0.04	0.08	0.06	−0.04	0.09	0.07	0.09
Cox	*−0.32*	*0.45*	*0.36*	*0.11*	*0.15*	*0.12*	*−0.18*	*0.27*	*0.19*	*0.23*
PSS	−0.10	0.20	0.16	0.04	0.07	0.06	−0.05	0.09	0.07	0.09
XII: $\sigma(z) = 0.38(1+z^2)$										
New	−0.07	0.16	0.13	0.03	0.06	0.05	−0.03	0.07	0.05	0.07
Cox	*−0.38*	*0.64*	*0.42*	*0.13*	*0.17*	*0.14*	*−0.22*	*0.31*	*0.23*	*0.27*
PSS	−0.11	0.18	0.15	0.04	0.07	0.05	−0.05	0.08	0.06	0.08
XIII: $\sigma(z) = 2.22/(1+\exp(1-z/0.5))$										
New	−0.08	0.17	0.14	0.03	0.06	0.05	−0.04	0.07	0.05	0.07
Cox	*−0.19*	*0.34*	*0.29*	*0.07*	*0.12*	*0.10*	*−0.10*	*0.17*	*0.13*	*0.17*
PSS	−0.07	0.19	0.15	0.03	0.07	0.05	−0.04	0.08	0.06	0.08

Note: See Table 1. Italics indicate that the Cox estimator is inconsistent in all experiments. The results for the Cox estimator exclude 1, 0, 11, 0, 1, 6, 2, 12, and 1 samples, respectively, where the estimator did not converge.

the new estimator on the uncensored (single-risk) data in experiment I, the inconsistency of the PSS estimator with censored (multirisk) data is obvious in experiments II–IV. In contrast, the new estimator performs well in all experiments.

To see why the PSS estimator is inconsistent in a competing risks framework, it is instructive to present the standard PH model in regression form. This also serves as an introduction to the design of the next set of experiments. For data of the nature considered in this paper, it is well known that there exist latent risk-specific failure times Y_1^* and Y_2^* such that $Y = \min(Y_1^*, Y_2^*)$, the hazard function for Y_s^* is h_s, and Y_1^* and Y_2^* are conditionally independent given X (see, e.g., Cox and Oakes, 1984, Sect. 9.2). This representation is sometimes known as the independent competing risks model.[10] It is also well known that $U_s = H_s(Y_s^*|X)$ has a unit exponential distribution (see, e.g., Kalbfleisch and Prentice, 1980). The standard PH model assumes that $H_s(Y_s^*|X) = \Lambda_s(Y_s^*)e^{X'\beta_s}$, where Λ_s is the integrated baseline hazard. It follows that

$$\ln \Lambda_s(Y_s^*) = -X'\beta_s + V_s, \tag{31}$$

where $V_s = \ln U_s$ has a type 1 extreme value distribution, whence

$$Y = \min[\Lambda_1^{-1}(\exp(-X'\beta_1)V_1), \Lambda_2^{-1}(\exp(-X'\beta_2)V_2)]. \tag{32}$$

Note that both indices, $X'\beta_1$ and $X'\beta_2$, appear on the right-hand side and that the conditional mean function, $E(Y|X)$, therefore does not satisfy a (single-) index restriction. In this context, the PSS estimator estimates a weighted average of β_1 and β_2.

The second set of experiments investigates the properties of the estimators in a more general model[11]

$$\ln \Lambda_s(Y_s^*) = -\mu_s(X'\beta_s) + \sigma_s(X'\beta_s)\epsilon_s. \tag{33}$$

This model is empirically relevant because several commonly used non-PH models are special cases of (33), including the AFT model, the GAFT model of Ridder (1990), and the mixed PH model. The Cox estimator assumes the data satisfy (31) and will, in general, be inconsistent if (31) is not satisfied. In contrast, the new semiparametric estimator is consistent when the data satisfy (33).

For simplicity, there is only one risk in these experiments, and the effective sample size is thus equal to the total number of observations. Because the Cox estimator and the new semiparametric estimator are invariant to monotone transformations of Y, it is assumed that $\Lambda_s(y) = y$. To facilitate comparisons, μ_s is normalized such that $\mu_s(X'\beta_s)$ has mean zero and variance one, ϵ_s is standardized to have mean zero and variance one, and σ_s is normalized such that $\sigma_s(X'\beta_s)\epsilon_s$ has mean zero and variance one.

Results are presented in Table 2. The benchmark is experiment I in Table 1, that is, the standard Cox PH model where $\mu_1(z) = z$, $\sigma_1(z) = 1$, and ϵ_1 has a type 1 extreme value distribution. Deviations from the benchmark model are indicated in Table 2. The function μ_1, which is linear in the benchmark exper-

https://doi.org/10.1017/S0266466604203012 Published online by Cambridge University Press

iment, is concave and increasing, convex and increasing, and shaped like a typical distribution function in experiments V–VII, as illustrated in Figure 1. (When reading the figure, note that the 1st and the 99th percentiles of $X'\beta_1$ are -2.0 and 2.8, respectively.) The distribution of ϵ_1, which is type 1 extreme value in the benchmark experiment, is normal in experiment VIII, log Burr with parameter 0.20 in experiment IX, and $\epsilon_1 = \sinh(2N)$ in experiment X, where N is standard normal. The resulting distribution of ϵ_1 in experiment X is highly leptokurtic. As mentioned before, the densities of ϵ_1 are scaled such that ϵ_1 has mean zero and variance one, as shown in Figure 2. Finally, the function σ_1, which is unity in the benchmark experiment, is convex and increasing, shaped like a convex parabola, and shaped like a distribution function in experiments XI–XIII, as illustrated in Figure 3.

Examining Table 2 reveals several interesting results. First of all, the Cox estimator has the poorest properties of all three estimators. This is not surprising, because the Cox estimator is inconsistent in these experiments. The only experiment in which the Cox estimator does significantly better than the other estimators is IX, where ϵ_1 is normally distributed. Reflecting the inconsistency, the performance of the Cox estimator is much worse in experiment IX than in the benchmark experiment I. However, the difference between the normal density and the type 1 extreme value density is sufficiently small that the Cox esti-

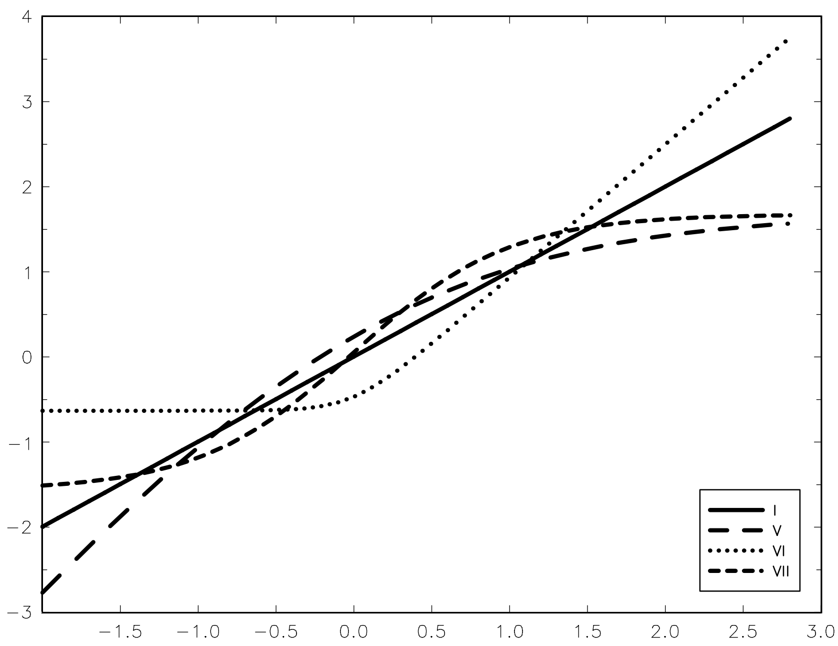

Figure 1. Monte Carlo μ functions.

https://doi.org/10.1017/S0266466604203012 Published online by Cambridge University Press

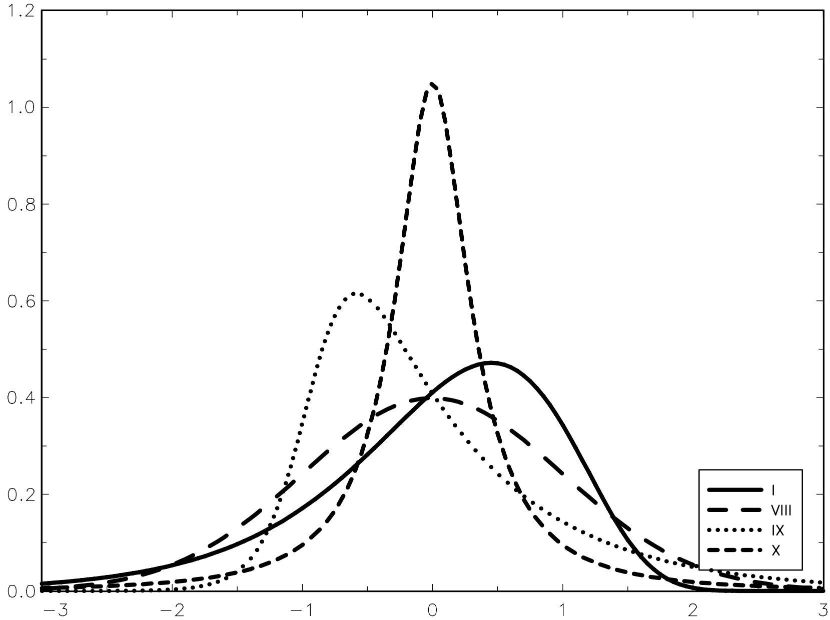

FIGURE 2. Monte Carlo densities of ϵ.

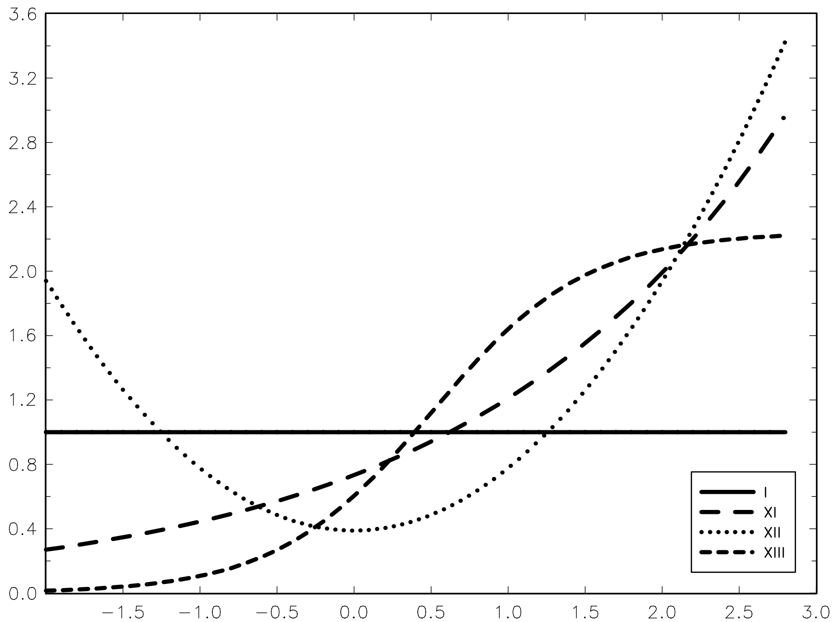

FIGURE 3. Monte Carlo σ functions.

https://doi.org/10.1017/S0266466604203012 Published online by Cambridge University Press

mator outperforms the new estimator. Clearly this "anomaly" is a result of the relatively small sample size considered here. In sufficiently large samples the new estimator will have better properties than the Cox estimator.

Another interesting result is the striking similarity between the properties of the new estimator and the PSS estimator. In most of experiments, the bias, root mean square error, and mean absolute error differ by no more than 0.02. In experiments IX and XII the bias of the estimator of β_{12}/β_{11} is slightly larger, but the remaining losses are similar. Only in experiments VIII and X are there significant differences, and the PSS is preferred in VIII, whereas the new estimator is preferred in X. Interestingly, experiments VIII and X correspond to the simple linear regression model with symmetrically distributed errors; in VIII the errors are normally distributed, whereas in X they are leptokurtic. As the latter can be interpreted as data with a relatively high proportion of "outliers," the conclusion is that the new hazard-based estimator is less sensitive to outliers than the mean-based PSS estimator. This is similar to the well-known result from robust estimation theory that the sample mean is a less efficient (and the sample median a more efficient) estimator of location the more leptokurtic the data.

The third and last set of experiments shows the properties of various test statistics based on the estimator of the covariance matrix described in Section 3. The main concern is the magnitude of the size distortion that is expected to occur in small samples. Because the scale of β_1 is not identified, the only interesting statistical hypotheses concern the statistical significance of each of the explanatory variables and the ratio of the explanatory variables. Accordingly, simulation results are presented for t-tests of the hypotheses $H_0: \beta_{11} = 0$ and $H_0: \beta_{12} = 0$ and for an F-test of the hypothesis $H_0: \beta_{11} = \beta_{12}$. To focus on size distortions, the true parameters are $\beta_1 = (0,1)'$, $\beta_1 = (1,0)'$, and $\beta_1 = (1/\sqrt{2},1/\sqrt{2})'$, respectively. The experimental designs are otherwise as described earlier. The test statistics are based on the unnormalized estimates, because normalization (division by a random variable) often leads to (additional) bias, variance, skewness, and generally ill-behaved moments.

Tables 3–5 show the proportion of samples where the test statistic is smaller than the theoretical asymptotic quantiles from the $N(0,1)$ distribution and the $\chi^2(1)$ distribution, respectively. Thus, the first entry in the body of Table 3, .04, indicates that the (t-) test statistic was smaller than -1.645 in 4% of the 5,000 samples. Asymptotically this number should approach .05, as indicated in the column heading. Similarly, the last entry in the first row, .99, indicates that in 99% of the samples the (F-) test statistic was smaller than 6.635, the 99th percentile of the $\chi^2(1)$ distribution.

It can be seen from Table 3 that the simulated probabilities are very close to the asymptotic probabilities for the new estimator. Only in a few cases is the difference larger than 3–4 percentage points. This suggests that asymptotic critical values can be used when testing hypotheses about the index coefficients, at least in samples of 200 observations or more.

https://doi.org/10.1017/S0266466604203012 Published online by Cambridge University Press

TABLE 3. Monte Carlo probabilities using asymptotic quantiles, new estimator

	$H_0: \beta_{11} = 0$				$H_0: \beta_{12} = 0$				$H_0: \beta_{11} = \beta_{12}$		
	0.05	0.10	0.90	0.95	0.05	0.10	0.90	0.95	0.90	0.95	0.99
I	0.04	0.09	0.91	0.96	0.05	0.10	0.91	0.96	0.86	0.93	0.99
II	0.06	0.11	0.91	0.96	0.06	0.11	0.91	0.95	0.84	0.91	0.98
III	0.06	0.11	0.91	0.95	0.06	0.11	0.90	0.95	0.89	0.94	0.99
IV	0.06	0.11	0.90	0.95	0.06	0.12	0.90	0.95	0.87	0.93	0.99
V	0.03	0.08	0.93	0.97	0.05	0.10	0.91	0.96	0.87	0.94	0.99
VI	0.05	0.10	0.90	0.95	0.04	0.09	0.91	0.95	0.89	0.95	0.99
VII	0.04	0.09	0.91	0.96	0.05	0.10	0.91	0.95	0.89	0.95	0.99
VIII	0.06	0.11	0.91	0.96	0.06	0.10	0.90	0.95	0.87	0.94	0.99
IX	0.04	0.09	0.92	0.96	0.04	0.09	0.91	0.95	0.90	0.95	0.99
X	0.02	0.06	0.95	0.98	0.03	0.08	0.92	0.96	0.92	0.97	1.00
XI	0.03	0.07	0.93	0.97	0.04	0.10	0.91	0.96	0.92	0.96	1.00
XII	0.03	0.07	0.93	0.97	0.05	0.10	0.91	0.95	0.92	0.97	1.00
XIII	0.03	0.06	0.94	0.98	0.05	0.09	0.91	0.95	0.93	0.97	1.00

Note: In the first four columns of results, labeled $H_0: \beta_{11} = 0$, the true parameter is $\beta_1 = (0,1)'$. In the next four columns of results, labeled $H_0: \beta_{12} = 0$, the true parameter is $\beta_1 = (1,0)'$. In the last three columns, labeled $H_0: \beta_{11} = \beta_{12}$, the true parameter is $\beta_1 = (1/\sqrt{2}, 1/\sqrt{2})'$. The Monte Carlo designs are identical to the previous in all other respects.

Table 4 shows similar numbers for the Cox estimator using the usual variance estimator based on the inverse information matrix. As expected the simulated probabilities are virtually identical to the asymptotic probabilities in experiments I–IV where the Cox estimator is correctly specified and consistent. Perhaps more surprisingly, the probabilities are not far off for the *t*-tests in experiments V–VIII. However, the simulated probabilities are very different from the asymptotic in all other cases where the Cox estimator is inconsistent. Note that the number of samples where the Cox estimator failed to converge is extremely large in some experiments.

Table 5 contains results for the PSS estimator using the variance estimator suggested by Powell et al. (1989). The simulated probabilities here are remarkably close to asymptotic probabilities in all experiments where the PSS estimator is consistent and indeed are slightly better than simulated probabilities for the new estimator. Note that the PSS estimator is also consistent in the censored experiments II–IV when $\beta_1 = (1,0)'$, because the index coefficients are the same for both risks. For the experiments where the PSS estimator is inconsistent, the simulated probabilities are very different from the asymptotic, as expected.

The conclusion of these Monte Carlo simulations is straightforward. If the risk-specific hazard rates are known to be proportional, use Cox's partial likelihood estimator. If such information is not available, but the data are either

TABLE 4. Monte Carlo probabilities using asymptotic quantiles, Cox estimator

	$H_0 : \beta_{11} = 0$				$H_0 : \beta_{12} = 0$				$H_0 : \beta_{11} = \beta_{12}$		
	0.05	0.10	0.90	0.95	0.05	0.10	0.90	0.95	0.90	0.95	0.99
I	0.05	0.11	0.90	0.95	0.05	0.09	0.89	0.94	0.90	0.95	0.99
II	0.05	0.10	0.90	0.95	0.05	0.10	0.90	0.94	0.90	0.95	0.99
III	0.06	0.11	0.90	0.95	0.05	0.10	0.89	0.94	0.90	0.95	0.99
IV	0.06	0.11	0.90	0.95	0.04	0.11	0.89	0.94	0.89	0.94	0.99
V	*0.06*	*0.12*	*0.89*	*0.94*	*0.04*	*0.09*	*0.89*	*0.95*	*0.26*	*0.35*	*0.54*
VI	*0.06*	*0.12*	*0.88*	*0.94*	*0.05*	*0.10*	*0.87*	*0.93*	*0.51*	*0.62*	*0.82*
VII	*0.06*	*0.12*	*0.89*	*0.94*	*0.05*	*0.10*	*0.89*	*0.94*	*0.63*	*0.73*	*0.87*
VIII	*0.08*	*0.13*	*0.87*	*0.93*	*0.07*	*0.12*	*0.86*	*0.92*	*0.80*	*0.87*	*0.95*
IX	*0.12*	*0.18*	*0.83*	*0.89*	*0.10*	*0.14*	*0.83*	*0.89*	*0.67*	*0.75*	*0.87*
X	*0.22*	*0.28*	*0.72*	*0.78*	*0.19*	*0.23*	*0.69*	*0.76*	*0.46*	*0.53*	*0.67*
XI	*0.11*	*0.18*	*0.84*	*0.89*	*0.09*	*0.14*	*0.83*	*0.90*	*0.37*	*0.44*	*0.57*
XII	*0.13*	*0.20*	*0.81*	*0.87*	*0.11*	*0.15*	*0.83*	*0.89*	*0.31*	*0.37*	*0.49*
XIII	*0.13*	*0.19*	*0.81*	*0.87*	*0.11*	*0.16*	*0.80*	*0.88*	*0.50*	*0.58*	*0.72*

Note: See Table 3. Italics indicate experiments where the Cox estimator is inconsistent. The results for $H_0 : \beta_{11} = 0$ exclude 0, 0, 0, 20, 563, 0, 518, 0, 2, 88, 633, 1,211, and 3 samples, respectively, and the results for $H_0 : \beta_{11} = \beta_{12}$ exclude 0, 0, 0, 21, 1, 0, 11, 0, 1, 6, 2, 12, and 1 sample because the estimator did not converge.

TABLE 5. Monte Carlo probabilities using asymptotic quantiles, PSS estimator

	$H_0 : \beta_{11} = 0$				$H_0 : \beta_{12} = 0$				$H_0 : \beta_{11} = \beta_{12}$		
	0.05	0.10	0.90	0.95	0.05	0.10	0.90	0.95	0.90	0.95	0.99
I	0.06	0.10	0.90	0.95	0.05	0.10	0.91	0.96	0.89	0.94	0.98
II	*0.00*	*0.00*	*0.02*	*0.04*	0.05	0.11	0.90	0.95	*0.57*	*0.68*	*0.85*
III	*0.00*	*0.00*	*0.00*	*0.00*	0.05	0.10	0.90	0.95	*0.12*	*0.19*	*0.38*
IV	*0.00*	*0.00*	*0.00*	*0.00*	0.05	0.10	0.90	0.95	*0.00*	*0.01*	*0.03*
V	0.04	0.09	0.92	0.97	0.05	0.10	0.91	0.96	0.86	0.93	0.99
VI	0.05	0.10	0.90	0.95	0.04	0.10	0.91	0.96	0.90	0.95	0.99
VII	0.04	0.09	0.91	0.96	0.05	0.10	0.91	0.96	0.89	0.95	0.99
VIII	0.05	0.11	0.90	0.95	0.05	0.10	0.90	0.95	0.88	0.93	0.98
IX	0.05	0.11	0.90	0.95	0.05	0.10	0.90	0.95	0.88	0.94	0.99
X	0.05	0.11	0.90	0.95	0.05	0.11	0.90	0.95	0.88	0.93	0.98
XI	0.04	0.09	0.92	0.96	0.05	0.10	0.91	0.96	0.88	0.94	0.99
XII	0.04	0.09	0.91	0.97	0.05	0.10	0.91	0.96	0.85	0.93	0.99
XIII	0.03	0.08	0.93	0.97	0.05	0.11	0.91	0.96	0.93	0.97	1.00

Note: See Table 3. Italics indicate experiments where the PSS estimator is inconsistent.

https://doi.org/10.1017/S0266466604203012 Published online by Cambridge University Press

single-risk or the censoring mechanism is known to be independent of the explanatory variables and the data are relatively free of outliers, there are two reasons to favor the PSS estimator over the new estimator: the PSS estimator is much faster to compute, and its variance estimator appears to have slightly better properties. In other cases, such as nontrivial competing risks data or data with outliers, the new estimator is preferable.

6. CONCLUSION

This paper proposed a new semiparametric estimator of index coefficients for a risk-specific hazard function. It is assumed that the hazard function satisfies an index restriction, but no other essential assumptions are imposed. In particular, the assumption of proportional hazards is avoided. The new estimator is applicable to competing risks data. Currently no other semiparametric estimator is applicable to competing risks data without the assumption of proportional hazards. The estimator described in this paper requires that the explanatory variables be continuous; however, an estimator for discrete explanatory variables is developed in a companion paper.

The index coefficient estimator was shown to be root-n consistent and asymptotically normally distributed, similarly to index coefficient estimators developed for other settings. A consistent estimator of the variance matrix was proposed. The paper also described how the index restriction can be used to eliminate the curse of dimensionality in nonparametric estimation.

The Monte Carlo simulations suggested that the new estimator performs well in samples of 200 observations. The new estimator was compared with Cox's partial likelihood estimator and the weighted average derivative estimator of Powell et al. (1989). Based on the Monte Carlo simulations it was possible to formulate guidelines for when to use which estimator.

One large problem, the question of optimal bandwidth selection, was left for future research. The approach taken by Härdle and Tsybakov (1993) and Powell and Stoker (1996) for the weighted average derivative estimator of Powell et al. (1989) seems promising, and it may be to worthwhile to investigate whether it can be successfully adapted to the present index coefficient estimator.

NOTES

1. This expression captures the case where Y is discrete in addition to the continuous case. If Y is discrete, (3) becomes $H(y|x) = \sum_j 1(\gamma_j \le y)(F_1(\gamma_j|x) - F_1(\gamma_{j-1}|x))/F_2(\gamma_j|x)$, where $\gamma_1, \gamma_2, \dots$ are the support points of Y and $F_1(\gamma_0|x) = 0$. The corresponding hazard function is $h(y|x) = \sum_j 1(\gamma_j = y)(F_1(\gamma_j|x) - F_1(\gamma_{j-1}|x))/F_2(\gamma_j|x)$. If Y is continuous, (3) becomes $H(y|x) = \int_0^y \partial_1 F_1(v|x)/F_2(v|x)\,dv$, where $\partial_1 F_1$ is the density corresponding to F_1. The corresponding hazard function is $h(y|x) = \partial_1 F_1(y|x)/F_2(y|x)$.

2. If f is a function, let $\partial_i^j f$ denote the jth-order partial derivative of f with respect to its ith argument. A vertical bar (as in $F_1(y|x)$) is equivalent to a comma (as in $A_1(y,x)$) when counting arguments, and $\partial_i^j fg$ means $(\partial_i^j f)g$, not $\partial_i^j(fg)$. With a bit of notational abuse, let $\partial_x^j f$ denote the

https://doi.org/10.1017/S0266466604203012 Published online by Cambridge University Press

q-vector of jth-order partial derivatives with respect to whichever argument represents the q-vector of explanatory variables. Let $\partial_{x'}^{j} f$ denote the transpose of $\partial_{x}^{j} f$.

3. Throughout the paper, the range is not indicated whenever integration is over an entire euclidean space.

4. For simplicity of exposition, derivatives are assumed to exist everywhere on the domain of the original functions. The result of Theorem 1 continues to hold even if a function is not differentiable everywhere, provided w is chosen to avoid "edge effects" in the kernel smoothing. That is, if the kernel estimates involve smoothing over X_i near x then $A_1(y,\cdot)$ and $A_2(y,\cdot)$ must be smooth on $[x-b, x+b]$ for all b small.

5. Throughout the paper boldface lowercase letters are used as placeholders and integration dummies for the corresponding uppercase random variables.

6. The theory of estimation of the unconditional hazard function has also progressed in recent years. See, e.g., Nielsen and Tanggaard (2001).

7. The same kernel and bandwidth selection procedure is used to compute the PSS estimator.

8. Optimal bandwidth selection procedures were developed for the original weighted average derivative estimator of Powell et al. (1989) by Härdle and Tsybakov (1993) and Powell and Stoker (1996). Their ideas can be extended to the present setting.

9. The scale of β_s is not nonparametrically identified. This is true even in a proportional hazard framework, where $h(y|x) = \lambda_s(y)\psi_s(x'\beta_s)$, because the scale of β_s can be subsumed into ψ_s. The scale is identified only when a specific parametric form of ψ_s is assumed, as in the standard implementation of the Cox estimator, where $\psi_s(z) = e^z$.

10. Although substantial progress has been made on developing models with dependent risks, the independent competing risks models continue to play an important role in economic and econometric analysis. For example, the studies mentioned at the beginning of the introduction all assume independent risks.

11. In a study of nonemployment data, Koop and Ruhm (1993) find some support for using a cumulative distribution function instead of the exponential function in the standard PH model.

REFERENCES

Ai, C. (1997) A semiparametric maximum likelihood estimator. *Econometrica* 65, 933–963.

Andersen, P.K., Ø. Borgan, R.D. Gill, & N. Keiding (1993) *Statistical Models Based on Counting Processes*. Springer-Verlag.

Beran, R. (1981) Nonparametric Regression with Randomly Censored Survival Data. Technical report, Department of Statistics, University of California, Berkeley.

Carling, K., P.-A. Edin, A. Harkman, & B. Holmlund (1996) Unemployment duration, unemployment benefits, and labor market programs in Sweden. *Journal of Public Economics* 59, 313–334.

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

Cox, D.R. (1975) Partial likelihood. *Biometrika* 62, 269–276.

Cox, D.R. & D. Oakes (1984) *Analysis of Survival Data*. Chapman and Hall.

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

Dabrowska, D.M. (1989) Uniform consistency of the kernel conditional Kaplan-Meier estimate. *Annals of Statistics* 17, 1157–1167.

Fallick, B.C. (1993) The industrial mobility of displaced workers. *Journal of Labor Economics* 11, 302–323.

Fleming, T.R. & D.P. Harrington (1991) *Counting Processes and Survival Analysis*. Wiley.

Gørgens, T. & J. Horowitz (1999) Semiparametric estimation of a censored regression model with an unknown transformation of the dependent variable. *Journal of Econometrics* 90, 155–191.

Grambsch, P.M. & T.M. Therneau (1994) Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 81, 515–526.

Han, A.K. (1987) Non-parametric analysis of a generalized regression model. *Journal of Econometrics* 35, 303–316.

Härdle, W. & T. M. Stoker (1989) Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association* 84, 986–995.

Härdle, W. & A.B. Tsybakov (1993) How sensitive are average derivatives? *Journal of Econometrics* 58, 31–48.

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

Henley, A. (1998) Residential mobility, housing equity, and the labour market. *Economic Journal* 108, 414–427.

Horowitz, J.L. (1999) Semiparametric estimation of a proportional hazard model with unobserved heterogeneity. *Econometrica* 67, 1001–1028.

Horowitz, J.L. & G.R. Neumann (1992) A generalized moments specification test of the proportional hazards model. *Journal of the American Statistical Association* 87, 234–240.

Ichimura, H. (1993) Semiparametric least squares (SLS) and weighted SLS estimation of single index models. *Journal of Econometrics* 58, 71–120.

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

Koop, G. & C.J. Ruhm (1993) Econometric estimation of proportional hazard models. *Journal of Economics and Business* 45, 421–430.

Lancaster, T. (1990) *The Econometric Analysis of Transition Data*. Cambridge University Press.

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

McCall, B.P. (1994) Testing the proportional hazards assumption in the presence of unmeasured heterogeneity. *Journal of Applied Econometrics* 9, 321–334.

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

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

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

Nielsen, J.P., O. Linton, & P.J. Bickel (1998) On a semiparametric survival model with flexible covariate effect. *Annals of Statistics* 26, 215–241.

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

Powell, J.L., J.H. Stock, & T.M. Stoker (1989) Semiparametric estimation of index coefficients. *Econometrica* 57, 1403–1430.

Powell, J.L. & T.M. Stoker (1996) Optimal bandwidth choice for density-weighted averages. *Journal of Econometrics* 75, 291–316.

Ridder, G. (1990) The non-parametric identification of generalized accelerated failure-time models. *Review of Economic Studies* 57, 167–81.

Salzberger, E. & P. Fenn (1999) Judicial independence: Some evidence from the English Court of Appeal. *Journal of Law and Economics* 42, 831–847.

Santarelli, E. (2000) The duration of new firms in banking: An application of Cox regression analysis. *Empirical Economics* 25, 315–325.

Sherman, R.P. (1993) The limiting distribution of the maximum rank correlation estimator. *Econometrica* 61, 123–137.

Stancanelli, E.G.F. (1999) Do the rich stay unemployed longer? An empirical study for the UK. *Oxford Bulletin of Economics and Statistics* 61, 295–314.

TECHNICAL APPENDIX

Proof of Theorem 1. Recall the definitions $T = (Y, S, X')'$ and $\mathbf{t} = (\mathbf{y}, \mathbf{s}, \mathbf{x}')'$. Let P denote the distribution of T and let P_n denote the empirical measure formed from the n independent observations on P; that is, P_n puts probability $1/n$ on each of the observations. Linear functional notation is used throughout the Appendix. The expected value of a random variable V is denoted EV, $Pf(t) = Ef(T,t)$, and $P_n f(t) = n^{-1}\sum_{i=1}^n f(T_i, t)$. Product measures are denoted using the symbol $\otimes$. For example, $P \otimes Pf = Ef(T_1, T_2)$ where T_1 and T_2 are independent random variables distributed as T.

Recall that

$$\beta_n^* = \frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n \rho_0(T_i, T_j),$$

where ρ_0 is defined in (15). Define also

$$\rho_1(\mathbf{t}_i, \mathbf{t}_j) = \frac{\rho_0(\mathbf{t}_i, \mathbf{t}_j) + \rho_0(\mathbf{t}_j, \mathbf{t}_i)}{2},$$

$$V_n = \frac{1}{n^2}\sum_{i=1}^n \rho_0(T_i, T_i),$$

$$U_n = \frac{1}{n^2}\sum_{i=1}^n \sum_{\substack{j=1 \\ j\neq i}}^n \rho_1(T_i, T_j),$$

and

$$\hat{U}_n = \frac{2}{n}\sum_{i=1}^n P\rho_1(T_i, \cdot) - P\otimes P\rho_0 = P_n \otimes P\rho_0 + P \otimes P_n\rho_0 - P\otimes P\rho_0.$$

To prove part i of the theorem, convergence is established for each term in the decomposition

$$\beta_n^* - E\beta_n^* = V_n + (U_n - \hat{U}_n) + (\hat{U}_n - E\hat{U}_n) + (E\hat{U}_n - E\beta_n^*).$$

Note that $E\beta_n^* = P\otimes P\rho_0 + (1/n)(E\rho_0(T,T) - P\otimes P\rho_0) = P\otimes P\rho_0 + O(n^{-1})$, $EU_n = (n(n-1)/n^2)P\otimes P\rho_0 = P\otimes P\rho_0 + O(n^{-1})$, and $E\hat{U}_n = P\otimes P\rho_0$. It follows immediately that $E\hat{U}_n - E\beta_n^* = o(n^{-1/2})$.

A change of variables implies that $\rho_0(\mathbf{t}_i, \mathbf{t}_j) = Q_0^1(\mathbf{t}_i, \mathbf{t}_j) + Q_0^0(\mathbf{t}_i, \mathbf{t}_j)$, where

$$Q_0^d(\mathbf{t}_i, \mathbf{t}_j) = \frac{(-1)^{d+1}}{b^{q+1}}\int \partial_x^d K(x)\partial_x^{1-d} K\left(\frac{\mathbf{x}_i - \mathbf{x}_j + bx}{b}\right) \times w(\mathbf{y}_i, \mathbf{x}_i + bx)1(\mathbf{y}_j \geq \mathbf{y}_i)1(\mathbf{s}_i = s)\,dx.$$

Because $\partial_x^{1-d}K$ and w are bounded and $\partial_x^d K$ is integrable, $V_n = O(n^{-1}b^{-q-1}) = o(n^{-1/2})$ almost surely provided $n^{-1/2}b^{-q-1} \to 0$, or $nb^{2q+2} \to \infty$, as $n \to \infty$.

By Lemma 3.1 of Powell et al. (1989), $U_n - \hat{U}_n = o_p(n^{-1/2})$ provided $E(\|\rho_1(T_i,T_j)\|^2) = o_p(n)$ for $i \neq j$. ($\|\cdot\|$ denotes the euclidean norm $\|u\|^2 = \sum_i u_i^2 = u'u$.) To verify this condition, note that

$$E(\|\rho_1(T_i,T_j)\|^2) = \frac{1}{2}E(\rho_0(T_i,T_j)'\rho_0(T_i,T_j)) + \frac{1}{2}E(\rho_0(T_i,T_j)'\rho_0(T_j,T_i)).$$

Expanding $\rho_0(\mathbf{t}_i,\mathbf{t}_j)'\rho_0(\mathbf{t}_i,\mathbf{t}_j) = (Q_0^1(\mathbf{t}_i,\mathbf{t}_j) + Q_0^0(\mathbf{t}_i,\mathbf{t}_j))'(Q_0^1(\mathbf{t}_i,\mathbf{t}_j) + Q_0^0(\mathbf{t}_i,\mathbf{t}_j))$ yields four terms that after a change of variables have the form

$$\begin{aligned} Q_1(\mathbf{t}_i,\mathbf{t}_j) &= Q_0^{d_1}(\mathbf{t}_i,\mathbf{t}_j)'Q_0^{d_2}(\mathbf{t}_i,\mathbf{t}_j) \\ &= \frac{(-1)^{d_1+d_2}}{b^{2q+2}}\iint \partial_{x'}^{d_1}K(x_1)\partial_{x'}^{1-d_1}K\left(\frac{\mathbf{x}_i-\mathbf{x}_j+bx_1}{b}\right) \\ &\quad\times \partial_x^{d_2}K(x_2)\partial_x^{1-d_2}K\left(\frac{\mathbf{x}_i-\mathbf{x}_j+bx_2}{b}\right)w(\mathbf{y}_i,\mathbf{x}_i+bx_1) \\ &\quad\times w(\mathbf{y}_i,\mathbf{x}_i+bx_2)1(\mathbf{y}_j \geq \mathbf{y}_i)1(\mathbf{s}_i=s)\,dx_1\,dx_2. \end{aligned}$$

By further change of variables,

$$\begin{aligned} E(Q_1(T_i,T_j)) &= \frac{(-1)^{d_1+d_2}}{b^{2q+2}}\iiiint\!\!\int \partial_{x'}^{d_1}K(x_1)\partial_{x'}^{1-d_1}K\left(\frac{\mathbf{x}_i-\mathbf{x}_j+bx_1}{b}\right) \\ &\quad\times \partial_x^{d_2}K(x_2)\partial_x^{1-d_2}K\left(\frac{\mathbf{x}_i-\mathbf{x}_j+bx_2}{b}\right) \\ &\quad\times w(v,\mathbf{x}_i+bx_1)w(v,\mathbf{x}_i+bx_2) \\ &\quad\times A_2(v,\mathbf{x}_j)A_1(dv,\mathbf{x}_i)\,dx_1\,dx_2\,d\mathbf{x}_i\,d\mathbf{x}_j \\ &= \frac{(-1)^{d_1+d_2}}{b^{q+2}}\iiiint\!\!\int \partial_{x'}^{d_1}K(x_1)\partial_{x'}^{1-d_1}K(\mathbf{x}_j)\partial_x^{d_2}K(x_2)\partial_x^{1-d_2}K(\mathbf{x}_j-x_1+x_2) \\ &\quad\times w(v,\mathbf{x}_i+bx_1)w(v,\mathbf{x}_i+bx_2) \\ &\quad\times A_2(v,\mathbf{x}_i+bx_1-b\mathbf{x}_j)A_1(dv,\mathbf{x}_i)\,dx_1\,dx_2\,d\mathbf{x}_i\,d\mathbf{x}_j. \end{aligned}$$

Similarly, the terms making up $\rho_0(\mathbf{t}_i,\mathbf{t}_j)'\rho_0(\mathbf{t}_j,\mathbf{t}_i)$ have the form

$$\begin{aligned} Q_2(\mathbf{t}_i,\mathbf{t}_j) &= \frac{(-1)^{d_1+d_2}}{b^{2q+2}}\iint \partial_{x'}^{d_1}K(x_1)\partial_{x'}^{1-d_1}K\left(\frac{\mathbf{x}_i-\mathbf{x}_j+bx_1}{b}\right)\partial_x^{d_2}K(x_2) \\ &\quad\times \partial_x^{1-d_2}K\left(\frac{\mathbf{x}_j-\mathbf{x}_i+bx_2}{b}\right)w(\mathbf{y}_i,\mathbf{x}_i+bx_1)w(\mathbf{y}_j,\mathbf{x}_j+bx_2) \\ &\quad\times 1(\mathbf{y}_i=\mathbf{y}_j)1(\mathbf{s}_i=s)1(\mathbf{s}_j=s)\,dx_1\,dx_2 \end{aligned}$$

https://doi.org/10.1017/S0266466604203012 Published online by Cambridge University Press

and

$$E(Q_2(T_i,T_j)) = \frac{(-1)^{d_1+d_2}}{b^{2q+2}} \iiiint\!\!\iint \partial_{x'}^{d_1} K(x_1)\partial_{x'}^{1-d_1} K\left(\frac{\mathbf{x}_i - \mathbf{x}_j + bx_1}{b}\right)$$

$$\times \partial_x^{d_2} K(x_2)\partial_x^{1-d_2} K\left(\frac{\mathbf{x}_j - \mathbf{x}_i + bx_2}{b}\right) w(\mathbf{y}_i, \mathbf{x}_i + bx_1) w(\mathbf{y}_j, \mathbf{x}_j + bx_2)$$

$$\times 1(\mathbf{y}_i = \mathbf{y}_j) A_1(d\mathbf{y}_i, \mathbf{x}_i) A_1(d\mathbf{y}_j, \mathbf{x}_j)\, dx_1\, dx_2\, d\mathbf{x}_i\, d\mathbf{x}_j$$

$$= \frac{(-1)^{d_1+d_2}}{b^{q+2}} \iiiint\!\!\iint \partial_{x'}^{d_1} K(x_1)\partial_{x'}^{1-d_1} K(\mathbf{x}_j)\partial_x^{d_2} K(x_2)\partial_x^{1-d_2} K(x_1 + x_2 - \mathbf{x}_j)$$

$$\times w(\mathbf{y}_i, \mathbf{x}_i + bx_1) w(\mathbf{y}_j, \mathbf{x}_i + bx_1 - b\mathbf{x}_j + bx_2)$$

$$\times 1(\mathbf{y}_i = \mathbf{y}_j) A_1(d\mathbf{y}_i, \mathbf{x}_i) A_1(d\mathbf{y}_j, \mathbf{x}_i + bx_1 - b\mathbf{x}_j)\, dx_1\, dx_2\, d\mathbf{x}_i\, d\mathbf{x}_j.$$

Because $\partial_x^{1-d_j} K$, ξ, and w are bounded and $\partial_x^{d_j} K$ is integrable, these results imply $E(\|\rho_1(T_i,T_j)\|^2) = O_p(b^{-q-2}) = o_p(n)$ provided that $nb^{q+2} \to \infty$ as $n \to \infty$. Note that this is exactly the same as the condition derived by Powell et al. (1989). It follows that the limiting distribution of $\sqrt{n}(\beta_n^* - E\beta_n^*)$ is the same as the limiting distribution of $\sqrt{n}(\hat{U}_n - E\hat{U}_n)$.

By change of variables and integration by parts

$$P\rho_0(\mathbf{t},\cdot) = -\iint K(x)K(\mathbf{x}_j)(\partial_x w(\mathbf{y}, \mathbf{x} + bx) A_2(\mathbf{y}, \mathbf{x} + bx - b\mathbf{x}_j)$$

$$+ w(\mathbf{y}, \mathbf{x} + bx)\partial_x A_2(\mathbf{y}, \mathbf{x} + bx - b\mathbf{x}_j)) 1(\mathbf{s} = s)\, dx\, d\mathbf{x}_j$$

$$- \iint K(x)K(\mathbf{x}_j) w(\mathbf{y}, \mathbf{x} + bx)\partial_x A_2(\mathbf{y}, \mathbf{x} + bx - b\mathbf{x}_j) 1(\mathbf{s} = s)\, dx\, d\mathbf{x}_j$$

$$= -\partial_x w(\mathbf{y}, \mathbf{x}) A_2(\mathbf{y}, \mathbf{x}) 1(\mathbf{s} = s) - 2w(\mathbf{y}, \mathbf{x})\partial_x A_2(\mathbf{y}, \mathbf{x}) 1(\mathbf{s} = s) + r_1(\mathbf{t}),$$

where $\sup|r_1| = O(b)$ because $\partial_x w$, $\partial_x^2 w$, $\partial_x A_2$, and $\partial_x^2 A_2$ exist and are bounded and K is integrable. Similarly,

$$P\rho_0(\cdot,\mathbf{t}) = \iiint K(\mathbf{x}_i)K(x) w(\mathbf{y}_i, \mathbf{x} + bx) 1(\mathbf{y} \geq \mathbf{y}_i)$$

$$\times \partial_x A_1(d\mathbf{y}_i, \mathbf{x} + bx - b\mathbf{x}_i)\, dx\, d\mathbf{x}_i$$

$$+ \iiint K(\mathbf{x}_i)K(x)(\partial_x w(\mathbf{y}_i, \mathbf{x} + bx) 1(\mathbf{y} \geq \mathbf{y}_i)$$

$$\times A_1(d\mathbf{y}_i, \mathbf{x} + bx - b\mathbf{x}_i) + w(\mathbf{y}_i, \mathbf{x} + bx) 1(\mathbf{y} \geq \mathbf{y}_i)$$

$$\times \partial_x A_1(d\mathbf{y}_i, \mathbf{x} + bx - b\mathbf{x}_i))\, dx\, d\mathbf{x}_i$$

$$= 2\int w(\mathbf{y}_i, \mathbf{x}) 1(\mathbf{y} \geq \mathbf{y}_i)\partial_x A_1(d\mathbf{y}_i, \mathbf{x})$$

$$+ \int \partial_x w(\mathbf{y}_i, \mathbf{x}) 1(\mathbf{y} \geq \mathbf{y}_i) A_1(d\mathbf{y}_i, \mathbf{x}) + r_2(\mathbf{t}),$$

where $\sup|r_2| = O(b)$ because $\partial_x^j w$ and $\int|\partial_x^j A_1(dy,\cdot,\cdot)|$ exist and are bounded for $j = 1,2$ and K is integrable. It follows that

$$P\rho_0(\mathbf{t},\cdot) + P\rho_0(\cdot,\mathbf{t}) = \Phi(\mathbf{y},\mathbf{s},\mathbf{x}) + 2\beta^* + O(b)$$

and

$$\hat{U}_n - E\hat{U}_n = (P_n - P)\Phi + (P_n - P)r_1 + (P_n - P)r_2.$$

The second moment of $n^{1/2}(P_n - P)r_j$, $j = 1,2$, is $P(r_j^2) - (Pr_j)^2$, which is bounded by $P(r_j^2) = O(b^2)$ using $\sup|r_j| = O(b)$. It follows by Chebyshev's inequality that $(P_n - P)r_j = o_p(n^{-1/2})$, whence the limiting distribution of $\sqrt{n}(\beta_n^* - E\beta_n^*)$ is the same as the limiting distribution of $\sqrt{n}(P_n\Phi - P\Phi) = \sqrt{n}P_n\Phi$. Part i of the theorem now follows from the multivariate Lindeberg–Lévy central limit theorem.

Turn now to the bias term. By previous arguments

$$E\beta_n^* - \beta^* = EV_n + EU_n = o(n^{-1/2}) + (n(n-1)/n^2)P \otimes P\rho_0 - \beta^*.$$

By integration by parts and change of variables

$$P \otimes P\rho_0 = \iiint K(\mathbf{x}_i)K(\mathbf{x}_j)w(\mathbf{y}_i,x)A_2(\mathbf{y}_i,x-b\mathbf{x}_j)\partial_x A_1(d\mathbf{y}_i,x-b\mathbf{x}_i)\,dx\,d\mathbf{x}_i\,d\mathbf{x}_j$$
$$- \iiint K(\mathbf{x}_i)K(\mathbf{x}_j)w(\mathbf{y}_i,x)\partial_x A_2(\mathbf{y}_i,x-b\mathbf{x}_j)A_1(d\mathbf{y}_i,x-b\mathbf{x}_i)\,dx\,d\mathbf{x}_i\,d\mathbf{x}_j.$$

Using the assumptions that $\int|\partial_x^j A_1(dy,\cdot)|$ and $\partial_x^j A_2$ exist and are bounded and continuous for $j = 1,\dots,k$ and that K is of order k, a Taylor series expansion implies

$$P \otimes P\rho_0 = \int w(\mathbf{y},x)(A_2(\mathbf{y},x)\partial_x A_1(d\mathbf{y},x) - \partial_x A_2(\mathbf{y},x)A_1(d\mathbf{y},x))\,dx + O(b^{2k})$$
$$= \beta^* + O(b^{2k}).$$

Part ii of the theorem follows. ■