Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Statistical Planning and Inference 129 (2005) 339–367
journal of statistical planning and inference
www.elsevier.com/locate/jspi

Efficient semiparametric estimators via modified profile likelihood

Eric V. Slud[a,*], Filia Vonta[b]

[a] *Department of mathematics, University of Maryland, College Park, MD 20742-4015, USA*
[b] *Department of mathematics and statistics, University of Cyprus, Cy 1678 Nieosia, Cyprus*

Available online 21 August 2004

Abstract

A new strategy is developed for obtaining large-sample efficient estimators of finite-dimensional parameters β within semiparametric statistical models. The key idea is to maximize over β a non-parametric log-likelihood with the infinite-dimensional nuisance parameter λ replaced by a consistent preliminary estimator $\tilde{\lambda}_\beta$ of the Kullback–Leibler minimizing value λ_β for fixed β. It is shown that the parametric submodel with Kullback–Leibler minimizer substituted for λ is generally a least-favorable model. Results extending those of Severini and Wong (Ann. Statist. 20 (1992) 1768) then establish efficiency of the estimator of β maximizing log-likelihood with λ replaced for fixed β by $\tilde{\lambda}_\beta$. These theoretical results are specialized to censored linear regression and to a class of semiparametric survival analysis regression models including the proportional hazards models with unobserved random effect or 'frailty', the latter through results of Slud and Vonta (Scand. J. Statist. 31 (2004) 21) characterizing the restricted Kullback–Leibler information minimizers.

MSC: 62F12; 62G05; 62N01

Keywords: Censored linear regression; Density estimation; Frailty model; Information bound; Least-favorable submodel; Profile likelihood; Semiparametric efficiency; Transformation model

1. Introduction

There are now several different tools for expressing the semiparametric information about the finite-dimensional ('structural') parameters in semiparametric models and for establish-

* Corresponding author. Tel.: +1-301-405-5469; fax: +1-301-314-0827.
E-mail address: evs@math.umd.edu (E.V. Slud).

0378-3758/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2004.06.057

ing efficiency of candidate estimators (Bickel et al., 1993; Van der Vaart and Wellner, 1996; Bickel and Kwon, 2001, and many other references for specific models cited in these works). Yet even in the iid case, there are important problems—such as the general transformation model (Bickel et al., 1993, Section 4.7, Example 2; Cheng et al., 1995) where there are either no candidate efficient estimators, or where natural candidates like the NPMLE are computable but intractable to characterize abstractly (Slud and Vonta, 2004).

An approach to likelihood inference which has been influential and successful in both parametric and semiparametric problems is that of profile or partially maximized likelihood (Kalbfleisch and Sprott, 1970; McCullagh and Tibshirani, 1990; Stafford, 1996). For both non- and semi-parametric problems, Owen (1988) and Qin and Lawless (1994) construct an *empirical likelihood* by fixing structural parameters and substituting an analytically determined restricted NPMLE over nuisance parameters, with the objective of establishing valid generalized likelihood-ratio-based confidence regions. In the semiparametric context, the recent paper of Murphy and van der Vaart (2000) is noteworthy, suggesting that generalized likelihood-ratio tests can be constructed and justified generally whenever one can verify abstract functional-analytic conditions about the maximizer over the unknown infinite-dimensional nuisance parameter for fixed values of the structural parameter. Related research on semiparametric generalized likelihood-ratio tests, with substituted nuisance-parameter estimators other than the partial NPMLE's, has been pursued by Fan et al. (2001).

The paper of Severini and Wong (1992) showed that efficient semiparametric estimators arise by maximizing a 'modified profile likelihood', i.e., a likelihood with nuisance parameters replaced by an estimated least-favorable parameterization. These authors advanced the idea of obtaining a least-favorable nuisance-parameterization by maximizing the *expected log-Likelihood* or negative *Kullback–Leibler* functional. When a smoothly consistent estimator of this maximizer for fixed structural-parameter values is substituted into the likelihood, and the latter is then maximized over the structural parameters, an efficient estimator results. (In the case of infinite-dimensional nuisance parameters, Severini and Wong (1992) developed this idea only in the special case of their 'conditionally parametric models'.) The general theory of the extension of this idea to the infinite-dimensional case is developed here, providing least-favorable parametric submodels, information bounds, and efficient estimators.

The theory of this method justifies a general principle of semiparametric estimation, which can be used (i) to unify existing semiparametric-efficient $\sqrt{n}$ consistent estimators in many problems; (ii) to generate new examples of such estimators; and (iii) to provide alternative and simpler formulas for semiparametric information bounds. In this paper, we first use the general theory to develop a new form of efficient estimators for the censored linear regression problems, in which efficient estimating equations had previously been found by Tsiatis (1990) and Ritov (1990). Next we apply the theory to a broad class of survival ('frailty') transformation models—beyond those already treated by Clayton and Cuzick (1986), Parner (1998), and Kosorok et al. (2004)—obtaining new computable formulas for information bounds and sketching the construction of efficient estimators.

1.1. Organization of the paper

The paper is organized as follows. In Section 2 we present the general problem setting for independent data, along with notational definitions, preliminary assumptions, and the central theoretical results. The remaining assumptions are stated in the Technical Appendix, in Section A.1, and discussed in Section A.2, and some technical consequences of the assumptions are also proved in Section A.3. Two applications of the theory are given, first in Section 3 to establish (known) information bounds and a new form of the efficient estimator in the censored linear regression model, and then in Section 4 to provide a new information bound formula in general frailty or transformation survival regression models, as well as a sketch of how to construct efficient estimators in that setting.

2. Consistency and efficiency of maximum modified profile likelihood estimators

Assume that the independent identically distributed (iid) data-sample $X_1, X_2, \ldots, X_n$ of random vectors in $\mathbb{R}^k$ is observed and assumed to follow a marginal probability law $\mu = P_{(\beta^0,\lambda^0)}$ where $\beta^0 \in \mathscr{U} \subset \mathbb{R}^d$, $\lambda^0 \in \mathscr{V} \subset L^0(\mathbb{R}^q, v)$ (Borel-measurable functions), where $\mathscr{U}$ is a fixed open set; $\mathscr{V}$ is a fixed set of positive measurable functions; and the σ-finite measure v (locally finite, but not necessarily a probability measure) is fixed on $\mathbb{R}^q$. In addition assume that there is a family $\{P_{(\beta,\lambda)}, (\beta,\lambda) \in \mathscr{U} \times \mathscr{V}\}$ of Borel probability measures on $\mathbb{R}^k$, such that

(A_0) For all $(\beta,\lambda) \in \mathscr{U} \times \mathscr{V}$, $P_{(\beta,\lambda)} \ll \mu$, and the regularity of densities $f_X(\cdot,\beta,\lambda) \equiv \mathrm{d}P_{(\beta,\lambda)}/\mathrm{d}\mu$ as functions of (β,λ) will be further restricted below. Note that by definition, $f_X(\cdot,\beta^0,\lambda^0) \equiv 1$. The true parameter-component value β^0 is assumed to lie in the interior of a fixed, known compact set $F \subset \mathscr{U}$.

(A_1) There exist fixed positive, finite real-valued functions c_1, c_2 on $\mathbb{R}^q$ such that $0 < c_1(\cdot) \leqslant \lambda^0 \leqslant c_2(\cdot) < \infty$ a.e. (v) and

$$\forall \lambda \in \mathscr{V}, \qquad v\text{-a.e.} \quad c_1(\cdot) \leqslant \lambda \leqslant c_2(\cdot). \tag{1}$$

In what follows, let $\|\cdot\|_1$ denote the $L^1(\mathbb{R}^k, \mu)$ norm. Here and below, $\|\cdot\|_2$ always denotes a Euclidean norm on vectors, or on matrices considered as vectors, rather than an L^2 functional norm. The space $\mathscr{V}$ is regarded from now on as a subset of the normed linear space

$$\mathscr{V}_0 \equiv \{\, \xi(\cdot)c_2(\cdot) \,:\, \xi \in L^\infty(\mathbb{R}^q, v)\}, \quad \|\lambda\|_{\mathscr{V}_0} \equiv \|\lambda/c_2\|_{\infty,v}. \tag{2}$$

The densities f_X, and estimators to be substituted for the nuisance parameters λ, are further restricted in Assumptions (A_2)–(A_8) given in Section A.1 of the Technical Appendix. In those assumptions, we consider perturbations of functions $\lambda \in \mathscr{V}$ by small multiples of functions in subsets

$$\mathscr{G} \subseteq \mathscr{G}_0 \subseteq \{\gamma \in L^0(\mathbb{R}^q, v) : |\gamma(t)| \leqslant c_1(t)\,\} \tag{3}$$

such that

$$\{\gamma/c_1 \,:\, \gamma \in \mathscr{G}_0\} \quad \text{is } \|\cdot\|_\infty \text{ dense in} \quad \{g \in L^\infty(\mathbb{R}^q, v) : \|g\|_\infty \leqslant 1\}. \tag{4}$$

Here and in what follows, we define the differentiation operator D_λ for all functions $\varphi : \mathscr{V} \to \mathbb{R}$, and all $\gamma \in \mathscr{G}_0$, by:

$$(D_\lambda\, \varphi(\lambda))(\gamma) = \frac{\mathrm{d}}{\mathrm{d}\vartheta}\, \varphi(\lambda + \vartheta\gamma)\Big|_{\vartheta=0}$$

and denote total differentiation in β by ∇^{T}. Throughout, $\nabla^{\otimes 2} = \nabla\, \nabla^{\mathrm{tr}}$ denotes a matrix-valued second-derivative (Hessian) operator.

The log-likelihood for the models $P_{(\beta,\lambda)}$ and data $\mathbf{X} = \{X_i\}_{i=1}^n$ is defined by

$$\mathrm{logLik}_n(\beta, \lambda) = \sum_{i=1}^{n} \log\, f_X(X_i,\, \beta,\, \lambda), \quad (\beta, \lambda) \in (\mathscr{U} \times \mathscr{V}).$$

When there is no danger of ambiguity in what follows, we drop the subscript $_n$ in the notation logLik_n.

Define the Kullback–Leibler functional by

$$\mathscr{K}(\beta, \lambda) \equiv -\int \log\, f_X(x, \beta, \lambda)\, \mathrm{d}\mu(x).$$

The key idea of *modified profile likelihood* (Severini and Wong, 1992) is to replace the nuisance parameter λ in the log-likelihood by a suitable estimator $\tilde{\lambda}_\beta$, restricted by (A_8), of the minimizer λ_β of $\mathscr{K}(\beta, \cdot)$ over $\lambda \in \mathscr{V}$ (assumed unique and with further regularity properties in (A_6)). The estimator of β, to be proved efficient, is then the maximizer of the modified profile likelihood.

Remark 1. The key insight enabling the modified profile likelihood approach to guarantee semiparametric efficient estimators is that in estimating β^0, the directional derivatives with respect to the nuisance parameter λ in functional space need be taken only at base points within a set sufficiently large so as to contain all partial maximizers (λ_β, for fixed but arbitrary β), in directions which should include all linear combinations of derivatives $\nabla_\beta^{\otimes j} \lambda_\beta$, $j = 1, 2$, and their preliminary estimators. Such spaces of base points and tangents are infinite dimensional, but can be far smaller than the linear spaces spanned respectively by parameters $\lambda \in \mathscr{V}$ and by differences of elements of $\mathscr{V}$.

Proposition 1 (*Severini and Wong, 1992*). *The d-dimensional smooth parametric submodel (β, λ_β) is a least-favorable d-dimensional regular parametric submodel for the general semiparametric model $P_{\beta,\lambda}$, where λ_β is the minimizer of $\mathscr{K}(\beta, \cdot)$ as in* (A_6).

Proof. According to Theorem 3.4.1 of Bickel et al. (1993), it suffices to check for any $\gamma \in \mathscr{G}_0$, with $\mathscr{G}_0$ as in (3) and (4),

$$E_{\beta^0,\lambda^0}\Big(\{\nabla_\beta^{\mathrm{T}} \log\, f_X(X_1, \beta, \lambda_\beta)\}\Big|_{\beta=\beta^0} \frac{\mathrm{d}}{\mathrm{d}\vartheta} \log\, f_X(X_1, \beta^0, \lambda^0 + \vartheta\gamma)\Big|_{\vartheta=0}\Big) = 0.$$

However, after expressing the expectation as an integral and expressing the latter in terms of the blocks A, B, B^*, C of the operator bilinear form $\mathscr{J}_{\beta^0,\lambda^0}$ defined in (34) in the

Technical Appendix, this assertion becomes an immediate consequence of (41)–(43), (44) and Lemma 1. □

Now we can define our proposed efficient estimator for β, as the maximizer of $\mathrm{logLik}_n(\beta,\ \tilde{\lambda}_\beta)$. Since we assume in (A_0) that $\beta^0 \in \mathrm{int}(F)$, our precise definition becomes

$$\tilde{\beta} \equiv \arg\max_{\beta\in F} \mathrm{logLik}_n(\beta, \tilde{\lambda}_\beta). \tag{5}$$

Theorem 1. *With probability converging to* 1 *as* $n \to \infty$, *the estimator* $\tilde{\beta}$ *defined in* (5) *is uniquely defined, lies for large* n *in the interior of* $F \subset \mathscr{U}$, *and is consistent for* β^0.

Proof. Let $\delta > 0$ be small enough so that $\{\beta \in \mathbb{R}^d : \|\beta - \beta^0\|_2 \leqslant \delta\} \subset F$. By the mean value theorem

$$\begin{aligned}\frac{1}{n}\mathrm{logLik}(\beta,\ \tilde{\lambda}_\beta) &- \frac{1}{n}\mathrm{logLik}(\beta^0, \tilde{\lambda}_{\beta^0}) \\ &= (\beta-\beta^0)^{\mathrm{tr}}\,\frac{1}{n}\,\nabla_\beta^{\mathrm{T}}\,\mathrm{logLik}(\beta^0, \tilde{\lambda}_{\beta^0}) \\ &\quad + \frac{1}{2n}\,(\beta-\beta^0)^{\mathrm{tr}}(\nabla_\beta^{\mathrm{T}})^{\otimes 2}\,\mathrm{logLik}(\beta^*, \tilde{\lambda}_{\beta^*})\,(\beta-\beta^0)\end{aligned}$$

for some β^* on the ray between $\beta^0,\ \beta$, which implies via Proposition 3 and (52) in Section A.3, that for a constant $\alpha > 0$ not depending on n, taken smaller than the minimum eigenvalue of $(\nabla_\beta^{\mathrm{T}})^{\otimes 2}\,\mathscr{K}(\beta^0, \lambda^0)$, that

$$\sup_{\beta\in F,\ \|\beta-\beta^0\|_2 \geqslant \delta} n^{-1}\,\mathrm{logLik}(\beta, \tilde{\lambda}_\beta) \leqslant n^{-1}\,\mathrm{logLik}(\beta^0, \tilde{\lambda}_{\beta^0}) - \delta^2\,\alpha/2$$

with probability converging to 1 as n gets large. Thus $\|\tilde{\beta} - \beta^0\|_2 < \delta$ with probability converging to 1, and since $\delta > 0$ can be taken arbitrarily small, $\tilde{\beta} \in F$ is consistent. This concludes the proof of Theorem 1. □

At this point, we know that $\mathrm{logLik}(\beta, \tilde{\lambda}_\beta)$ is strictly concave on compact subsets of $\mathscr{U}$ with probability converging to 1, and that $\tilde{\beta}$ is consistent and uniquely defined in F. Moreover, as shown in Proposition 3, with probability near 1 for large n, the random function $\mathscr{K}(\beta,\ \tilde{\lambda}_\beta)$ is strictly convex on $\mathscr{U}$, with Hessian uniformly close, on neighborhoods of β^0 which shrink down to $\{\beta^0\}$ as n gets large, to the information bound, or least-favorable information matrix,

$$\mathscr{I}_\beta^0 \equiv \mathscr{I}\left(P_{\beta^0,\lambda^0} \mid \beta,\ P_{\beta,\lambda}\right) \quad \text{(Bickel et al., 1993, p. 23)}$$

defined for all $a \in \mathbb{R}^d$ by

$$a^{\mathrm{tr}}\,\mathscr{I}_\beta^0\,a \equiv \mathscr{J}_{\beta^0,\lambda^0}\left((a,\ \nabla_\beta\lambda_{\beta^0}),\ (a,\ \nabla_\beta\lambda_{\beta^0})\right) \tag{6}$$

and equivalently expressed by (46) in the Technical Appendix.

Strict concavity of $\log\mathrm{Lik}_n$ here also implies that $\tilde\beta$ is uniquely (locally, within $\mathscr{U}$) characterized as the solution of the equation, for all $a \in \mathbb{R}^d$,

$$a^{\mathrm{tr}}\,\nabla_\beta^{\mathrm{T}} \log\mathrm{Lik}_n(\beta,\ \tilde\lambda_\beta) = a^{\mathrm{tr}}\nabla_\beta \log\mathrm{Lik}_n(\beta,\ \tilde\lambda_\beta) + (D_\lambda \log\mathrm{Lik}_n(\beta, \tilde\lambda_\beta))(a^{\mathrm{tr}}\nabla_\beta\tilde\lambda_\beta) = 0. \tag{7}$$

The next objective is to prove under the assumptions given above that the estimator $\tilde\beta$ is $\sqrt{n}$ consistent, asymptotically normal, and efficient.

Theorem 2. *Under the assumptions* (A_0)–(A_8), *for all* $a \in \mathbb{R}^d$,

$$\begin{aligned}\sqrt{n}a^{\mathrm{tr}}(\tilde\beta - \beta^0)(1 + \mathrm{o_P}(1)) &= \frac{1}{\sqrt{n}}\Big(a^{\mathrm{tr}}(\mathscr{I}_\beta^0)^{-1}\,\nabla_\beta \log\mathrm{Lik}(\beta^0, \lambda^0) \\ &\quad + \frac{\mathrm{d}}{\mathrm{d}\vartheta}\log\mathrm{Lik}(\beta^0, \lambda^0 + \vartheta a^{\mathrm{tr}}(\mathscr{I}_\beta^0)^{-1}\,\nabla_\beta\lambda_{\beta^0})\Big|_{\vartheta=0}\Big).\end{aligned} \tag{8}$$

Proof. We examine separately and Taylor-expand as functions of $(\beta,\ \lambda)$ about $(\beta^0,\ \lambda^0)$ the two terms in the second line of (7) evaluated at $\beta = \tilde\beta$. First, the consistency of $\tilde\beta$ for β^0 (from Theorem 1) and of $\nabla_\beta\,\tilde\lambda_{\beta^0}$ for $\nabla_\beta\,\lambda_{\beta^0}$ (from (A_8)) and the difference-quotient definition of the derivative imply

$$\tilde\lambda_{\tilde\beta} - \tilde\lambda_{\beta^0} = \nabla_\beta\,\tilde\lambda_{\beta^0}\,(\tilde\beta - \beta^0) + \mathrm{o_P}(\tilde\beta - \beta^0) = \nabla_\beta\,\lambda_{\beta^0}\,(\tilde\beta - \beta^0) + \mathrm{o_P}(\tilde\beta - \beta^0) \tag{9}$$

in the sense of norm $\|\cdot\|_{\mathscr{V}_0}$. Next, for small ϑ, Taylor-expanding in β about β^0

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}\vartheta}\log\mathrm{Lik}(\tilde\beta,\ \tilde\lambda_{\tilde\beta} + \vartheta\gamma) \\ &\quad = \frac{\mathrm{d}}{\mathrm{d}\vartheta}\log\mathrm{Lik}(\beta^0, \lambda^0 + \vartheta\gamma) + (\tilde\beta - \beta^0)^{\mathrm{tr}}\,\nabla_{\beta,\vartheta}^{\otimes 2}\log\mathrm{Lik}(\beta^0,\ \tilde\lambda_{\beta^0} + \vartheta\gamma) \\ &\qquad + \frac{\mathrm{d}^2}{\mathrm{d}t\,\mathrm{d}\vartheta}\log\mathrm{Lik}(\beta^0,\ \tilde\lambda_{\beta^0} + \vartheta\gamma + t(\tilde\lambda_{\tilde\beta} - \tilde\lambda_{\beta^0}))\Big|_{t=0} + \mathrm{o_P}(n(\tilde\beta - \beta^0)).\end{aligned}$$

Now divide through by n and apply Proposition 2 from Section A.3, then evaluating at $\beta = \beta^0,\ \vartheta = 0$, to find for $\gamma = a^{\mathrm{tr}}\,\nabla_\beta\tilde\lambda_{\tilde\beta}$,

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}\vartheta}\frac{1}{n}\log\mathrm{Lik}(\tilde\beta, \tilde\lambda_{\tilde\beta} + \vartheta\gamma)\Big|_{\vartheta=0} \\ &\quad = \frac{\mathrm{d}}{\mathrm{d}\vartheta}\,\frac{1}{n}\log\mathrm{Lik}(\beta^0, \lambda^0 + \vartheta\gamma)\Big|_{\vartheta=0} \\ &\qquad - B_{\beta^0,\,\lambda^0}(\tilde\beta - \beta^0,\ \gamma) - C_{\beta^0,\,\lambda^0}\left(\gamma,\ (\tilde\beta - \beta^0)^{\mathrm{tr}}\,\nabla_\beta\tilde\lambda_{\beta^0}\right) + \mathrm{o_P}(\tilde\beta - \beta^0).\end{aligned}$$

By the consistency already proved for the estimator $\tilde{\beta}$, together with equation (44), we conclude that with probability approaching 1 as $n \to \infty$

$$\frac{\mathrm{d}}{\mathrm{d}\vartheta}\frac{1}{n}\,\mathrm{logLik}(\tilde{\beta},\,\tilde{\lambda}_{\tilde{\beta}}+\vartheta\gamma)\bigg|_{\vartheta=0} = \frac{\mathrm{d}}{\mathrm{d}\vartheta}\frac{1}{n}\,\mathrm{logLik}(\beta^0,\,\lambda^0+\vartheta\gamma)\bigg|_{\vartheta=0} + \mathrm{o_P}(\tilde{\beta}-\beta^0) \qquad (10)$$

for $\gamma = a^{\mathrm{tr}}\,\nabla_\beta\tilde{\lambda}_{\tilde{\beta}}$. Next consider the term

$$n^{-1}a_2^{\mathrm{tr}}\,\nabla_\beta\,\mathrm{logLik}(\tilde{\beta},\,\tilde{\lambda}_{\tilde{\beta}})$$

which we Taylor-expand about $(\beta^0,\,\lambda^0)$ and re-express via Proposition 2 to obtain

$$\frac{a_2^{\mathrm{tr}}}{n}\,\nabla_\beta\mathrm{logLik}(\beta^0,\,\lambda^0) - A_{\beta^0,\lambda^0}(a_2,\,\tilde{\beta}-\beta^0) - B_{\beta^0,\,\lambda^0}(a_2,\,\tilde{\lambda}_{\tilde{\beta}}-\tilde{\lambda}_{\beta^0}) + \mathrm{o_P}(\tilde{\beta}-\beta^0).$$

Combining (9) with the previous equation, and making use of (44) and (6), we find

$$\frac{1}{n}\,\nabla_\beta\,\mathrm{logLik}(\tilde{\beta},\,\tilde{\lambda}_{\tilde{\beta}}) = \frac{1}{n}\,\nabla_\beta\,\mathrm{logLik}(\beta^0,\,\lambda^0) - \mathscr{I}_\beta^0\,(\tilde{\beta}-\beta^0) + \mathrm{o_P}(\tilde{\beta}-\beta^0). \qquad (11)$$

Finally, combine (7), (10), and (11), using also the second equality in (9), to establish

$$a_2^{\mathrm{tr}}\left\{\frac{1}{n}\,\nabla_\beta\mathrm{logLik}(\beta^0,\,\lambda^0) - \mathscr{I}_\beta^0\,(\tilde{\beta}-\beta^0)\right\} + (D_\lambda\mathrm{logLik}(\beta^0,\,\lambda^0))\left(\frac{a_2^{\mathrm{tr}}}{n}\,\nabla_\beta\tilde{\lambda}_{\tilde{\beta}}\right) = \mathrm{o_P}(\tilde{\beta}-\beta^0).$$

Then, replacing a_2^{tr} by $a^{\mathrm{tr}}\left(\mathscr{I}_\beta^0\right)^{-1}$ completes the proof. □

In our iid setting, Theorem 2 immediately implies

Theorem 3. *Under Assumptions* (A_0)–(A_8), *as* $n \to \infty$

$$\sqrt{n}\,\left(\tilde{\beta}-\beta^0\right) \xrightarrow{\mathscr{D}} \mathscr{N}\left(0\,,\left(\mathscr{I}_\beta^0\right)^{-1}\right). \qquad (12)$$

Proof. In view of Theorem 2, and the Cramer–Wold device for deriving multivariate distributional limits from univariate ones, we need only express (with $a_2 \equiv (\mathscr{I}_\beta^0)^{-1}a$)

$$\frac{1}{\sqrt{n}}\left(a_2^{\mathrm{tr}}\,\nabla_\beta\,\mathrm{logLik}_n(\beta^0,\,\lambda^0) + \frac{\mathrm{d}}{\mathrm{d}\vartheta}\mathrm{logLik}_n(\beta^0,\,\lambda^0+\vartheta\,a_2^{\mathrm{tr}}\nabla_\beta\lambda_{\beta^0})\bigg|_{\vartheta=0}\right)$$

as a normalized iid sum to which the ordinary central limit theorem applies. But the displayed expression is

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{a_2^{\mathrm{tr}}\,\nabla_\beta\,\log\,f_X(X_i,\,\beta^0,\,\lambda^0) + \frac{\mathrm{d}}{\mathrm{d}\vartheta}\,\log\,f_X(X_i,\beta^0,\lambda^0+\vartheta\,a_2^{\mathrm{tr}}\nabla_\beta\lambda_{\beta^0})\,\bigg|_{\vartheta=0}\right\}$$
$$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}a_2^{\mathrm{tr}}\,\nabla_\beta^{\mathrm{T}}\,\log\,f_X(X_i,\,\beta,\lambda_\beta)\bigg|_{\beta=\beta^0}.$$

A familiar calculation as in (45) shows that the variance of these summands is $a_2^{\rm tr}\,\mathscr{I}_\beta^0\,a_2$, and after recalling $a_2=(\mathscr{I}_\beta^0)^{-1}a$, the assertion follows by (8) and the central limit theorem. □

3. Censored linear regression

The problem of estimating linear-regression parameters efficiently in a semiparametric setting, where the error-distribution is unknown and the data are randomly right-censored, has been studied by many authors. (See Bickel et al., 1993 for background and references.) This is a problem where efficient estimating equations are known (Tsiatis, 1990; Ritov, 1990; Bickel et al., 1993, pp. 147 *ff*), but where the efficient estimators are not simple enough to have come into general use.

The censored linear-regression model assumes

$$X_i = Z_i^{\rm tr}\,\beta+\varepsilon_i,\quad T_i=\min(X_i,\,C_i),\quad \Delta_i = I_{[X_i\leqslant C_i]},$$

where the observed data consist of independent identically distributed triples $(T_i,\,Z_i,\,\Delta_i)$, and ε_i is assumed independent of (Z_i,C_i). The unknown parameters are the structural coefficients β, and the hazard intensity $\lambda(u)\equiv F_\varepsilon'(u)/(1-F_\varepsilon(u))$.

Denote $\Lambda(t)\equiv\int_{-\infty}^{t}\lambda(s)\,{\rm d}s$, which will be assumed finite for all $t<\infty$, and, with $(\beta^0,\,\lambda^0)$ denoting the true parameters actually governing the data

$$q_z(t)\equiv P(T_1\geqslant t\mid Z_1=z)=P(C_1\geqslant t\mid Z_1=z){\rm e}^{-\Lambda^0(t-z^{\rm tr}\beta^0)}.$$

The measure ν is Lebesgue measure on $\mathbb{R}$. The data-space $\mathbf{D}=\mathbb{R}\times\mathbb{R}^d\times\{0,1\}$ consists of triples $x=(t,z,\delta)$ with $t\in\mathbb{R}$, $z\in\mathbb{R}^d$, $\delta=0,1$. We define the probability law for the true model by

$${\rm d}\mu(t,z,\delta)\equiv(\delta\,\lambda^0(t-z^{\rm tr}\beta^0)\,q_z(t)\,{\rm d}F_Z(z)+(1-\delta)\,{\rm e}^{-\Lambda^0(t-z^{\rm tr}\beta^0)}\,{\rm d}F_{Z,C}(z,t))\,{\rm d}t.$$

The densities $f_X(x,\beta,\lambda)$ have the form $(\lambda(t-z^{\rm tr}\beta)/\lambda^0(t-z^{\rm tr}\beta^0))^\delta\,\exp(\Lambda^0(t-z^{\rm tr}\beta^0)-\Lambda(t-z^{\rm tr}\beta))$. Therefore, the log-likelihood in this setting is

$$\log{\rm Lik}(\beta,\lambda)=\sum_{i=1}^{n}\left\{\Delta_i\,\log\,\lambda(T_i-Z_i^{\rm tr}\,\beta)-\Lambda(T_i-Z_i^{\rm tr}\,\beta)\right\}\tag{13}$$

apart from additive terms not depending on (β,λ), and the Kullback–Leibler functional is (after integration by parts in t in the second term)

$$\mathscr{K}(\beta,\lambda)=-\int\int q_z(t)\,\left\{\lambda^0(t-z^{\rm tr}\beta^0)\log\,\lambda(t-z^{\rm tr}\beta)\,{\rm d}t-\,\lambda(t-z^{\rm tr}\beta)\right\}\,{\rm d}t\,{\rm d}F_Z(z).$$

The random vectors Z_i may contain a constant component. These vectors must be compactly supported, and not linearly degenerate, and we proceed to indicate the further

regularity conditions which are sufficient to apply the efficiency theory given in Section 2. Assume

(C_0) The parameter β^0 lies in the interior of the fixed compact region $\mathscr{U} \subset \mathbb{R}^d$, and $c_0 \equiv \text{ess. sup } \|Z_1\|_2 < \infty$, so that

$$\alpha_0 \equiv \text{ess. } \sup_{\beta\in\mathscr{U}} |(\beta - \beta^0)^{\text{tr}} Z_1| \leqslant c_0 \cdot \text{diam}(\mathscr{U}) < \infty.$$

Moreover, for every nonzero constant vector a, the variable $a^{\text{tr}} Z_1$ has positive variance, i.e., is nondegenerate.

(C_1) (a) The parameter λ^0 belongs to the set of nonnegative, twice continuously differentiable functions λ on $\mathbb{R}$ which satisfy $\Lambda(t) \equiv \int_{-\infty}^{t} \lambda(s)\,\mathrm{d}s < \infty$ for $t < \infty$, $\Lambda(\infty) = \infty$.

(b) Also, define the fixed functions

$$c_1(t) \equiv \inf_{|s|\leqslant\alpha_0} \lambda^0(t+s), \quad c_2(t) \equiv \sup_{|s|\leqslant\alpha_0} \lambda^0(t+s)$$

and assume $\|c_2/c_1\|_{\infty,v} < \infty$. For all λ as in (a), define $w_\lambda(t, z, \delta)$ by:

$$w_\lambda(t + z^{\text{tr}}\beta^0, z, \delta) \equiv \frac{\mathrm{e}^{\Lambda^0(t)}}{(\lambda^0(t))^\delta} \times \sup_{|x-t|\leqslant\alpha_0} \{(1+\lambda(x))\,(1+\lambda(x)+\Lambda(x))^2\,\mathrm{e}^{-\Lambda(x)}\}. \tag{14}$$

Then $c_2(t) + |\log c_1(t)| + \int_{-\infty}^{t} c_2(s)\,\mathrm{d}s + w_{\lambda^0}(t, z, \delta) \in L^1(\mathbf{D}, \mu)$.

(C_2) For $j = 1, 2$, $|\frac{\mathrm{d}^j}{\mathrm{d}t^j} \log \lambda^0(t)| \leqslant c_3$ for some finite constant c_3.

(C_3) As a function of t, $P(C_1 \geqslant t \mid Z_1 = z)$ is almost surely twice continuously differentiable, with a finite constant c_4 such that for $j = 1, 2$ and all t, z, $|\frac{\mathrm{d}^j}{\mathrm{d}t^j} \log P(C_1 \geqslant t \mid Z_1 = z)| \leqslant c_4$.

The regularity conditions (A_0)–(A_4) are readily checked to follow from (C_0)–(C_3) in this setting, with

$$\mathscr{V} = \Big\{\lambda \in \mathscr{V}_0 : \forall |s| \leqslant \alpha_0,\, c_1(t) \leqslant \lambda(s+t) \leqslant c_2(t), \max_{j=1,2} \left|\frac{\mathrm{d}^j}{\mathrm{d}t^j} \log \lambda(t)\right| \leqslant 2c_3,\ w_\lambda(t, z, \delta) \in L^1(\mu)\Big\}$$

and, for a fixed positive finite constant c_5,

$$\mathscr{G} = \left\{\gamma \in \mathscr{V}_0 : \sup_{|s|\leqslant\alpha_0} |\gamma(t+s)| \leqslant c_1(t),\ \sup_{|s|\leqslant\alpha_0} |\gamma'(s+t)| \leqslant c_5 c_1(t)\right\},$$

$$\mathscr{G}_0 = \left\{\gamma \in \mathscr{V}_0 : \sup_{|s|\leqslant\alpha_0} |\gamma(t+s)| \leqslant c_1(t), \sup_t \sup_{|s|\leqslant\alpha_0} |\gamma'(s+t)|/c_1(t) < \infty\right\}.$$

The log-likelihood summands and their derivatives up to second order as in (A_3) are dominated by the function

$$b(t,z) = \Big(|\log c_1(u)| + 2\int_{-\infty}^{u} c_2(s)\,\mathrm{d}s + 4(1+c_0)^2(1+c_3)\,(1+c_2(u))\Big)_{u=t-z^{\mathrm{tr}}\beta^0}.$$

For example, by (C_0) and the definition of $\mathscr{G}$

$$\|\nabla^{\otimes 2}_{\beta,\vartheta}\log(\lambda+\vartheta\gamma)(t-z^{\mathrm{tr}}\beta)\,\big|_{\substack{\vartheta=0,\\ \beta=\beta^0}}\|_2 \leqslant \left\| z\left\{\frac{\gamma'}{\lambda}-\frac{\lambda'\gamma}{\lambda^2}\right\}_{t-z^{\mathrm{tr}}\beta}\right\|_2 \leqslant c_0(2c_3+c_5).$$

Similarly, the individual likelihood terms and their derivatives up to second order as in (A_4) are dominated by the function

$$r(t,z,\delta) = 4(1+c_0)^2\left((1+c_3)^2+c_5\right)w_\lambda(t,z,\delta).$$

The operators $A_{\beta^0,\lambda^0},\ B_{\beta^0,\lambda^0},\ C_{\beta^0,\lambda^0}$ have the explicit forms

$$A_{\beta^0,\lambda^0}(a,a) = \int\int q_z(t)\,(z^{\mathrm{tr}}a)^2\frac{\left(\lambda^{0\prime}(t-z^{\mathrm{tr}}\beta^0)\right)^2}{\lambda^0(t-z^{\mathrm{tr}}\beta^0)}\,\mathrm{d}t\,\mathrm{d}F_Z(z), \tag{15}$$

$$B_{\beta^0,\lambda^0}(a,\gamma) = -\int\int q_z(t)\,(z^{\mathrm{tr}}a)\,\gamma(t-z^{\mathrm{tr}}\beta^0)\,\frac{\lambda^{0\prime}(t-z^{\mathrm{tr}}\beta^0)}{\lambda^0(t-z^{\mathrm{tr}}\beta^0)}\,\mathrm{d}t\,\mathrm{d}F_Z(z), \tag{16}$$

$$C_{\beta^0,\lambda^0}(\gamma,\gamma) = \int\int q_z(t)\,\frac{\gamma^2(t-z^{\mathrm{tr}}\beta^0)}{\lambda^0(t-z^{\mathrm{tr}}\beta^0)}\,\mathrm{d}t\,\mathrm{d}F_Z(z). \tag{17}$$

From these formulas, it follows immediately that (35) holds at $(\beta,\lambda)=(\beta^0,\lambda^0)$, since

$$A_{\beta^0,\lambda^0}(a,a)+2B_{\beta^0,\lambda^0}(a,\gamma)+C_{\beta^0,\lambda^0}(\gamma,\gamma) = \int\int\frac{q_z(t)}{\lambda^0(t-z^{\mathrm{tr}}\beta^0)}\left(a^{\mathrm{tr}}z\,\lambda^{0\prime}(t-z^{\mathrm{tr}}\beta^0)-\gamma(t-z^{\mathrm{tr}}\beta^0)\right)^2\mathrm{d}t\,\mathrm{d}F_Z(z)$$

cannot be 0 unless $a^{\mathrm{tr}}Z$ is a.s. conditionally degenerate (at a value other than 0, the same for all t) given $T\geqslant t$ for a.e. t. This proves (A_5), since the finiteness of $B_{\beta,\lambda},\ C_{\beta,\lambda}$ follows from the dominatedness conditions (A_3)–(A_4).

The maximization with respect to λ for fixed β to determine λ_β is unconstrained, and results in the equation, for all bounded γ,

$$0 = \int\int q_z(t)\left\{\lambda^0(t - z^{\mathrm{tr}}\beta^0)\,\frac{\gamma(t - z^{\mathrm{tr}}\beta)}{\lambda(t - z^{\mathrm{tr}}\beta)} - \gamma(t - z^{\mathrm{tr}}\beta)\right\}\mathrm{d}t\,\mathrm{d}F_Z(z)$$
$$= \int\int q_z(t + z^{\mathrm{tr}}\beta)\,\gamma(t)\left[\frac{\lambda^0(t + z^{\mathrm{tr}}(\beta - \beta^0))}{\lambda(t)} - 1\right]\mathrm{d}t\,\mathrm{d}F_Z(z)$$

from which it follows that

$$\lambda_\beta(t) = \frac{\int q_z(t + z^{\mathrm{tr}}\beta)\,\lambda^0(t + z^{\mathrm{tr}}(\beta - \beta^0))\,\mathrm{d}F_Z(z)}{\int q_z(t + z^{\mathrm{tr}}\beta)\,\mathrm{d}F_Z(z)}. \tag{18}$$

From this explicit formula, together with (C_3), there follows (A_6). Next, an information bound formula is derived from (44), (18) and the calculation

$$\nabla_\beta\lambda_\beta(t)\big|_{\beta=\beta^0} = \frac{\int z\,q_z(t + z^{\mathrm{tr}}\beta^0)\,\lambda^{0\prime}(t)\,\mathrm{d}F_Z(z)}{\int q_z(t + z^{\mathrm{tr}}\beta^0)\,\mathrm{d}F_Z(z)} = E_0(Z \mid T - Z^{\mathrm{tr}}\beta^0 \geqslant t)\,\lambda^{0\prime}(t),$$

where E_0 and later Var_0 denote (conditional) mean and variance under the model with parameters (β^0, λ^0).

Thus the semiparametric information matrix has quadratic form given by

$$a^{\mathrm{tr}}\mathscr{I}^0_\beta a = A_{\beta^0,\lambda^0}(a, a) + B_{\beta^0,\lambda^0}(a, a^{\mathrm{tr}}\nabla_\beta\lambda_{\beta^0})$$
$$= \int\int q_z(t)\,\frac{(\lambda^{0\prime}(t - z^{\mathrm{tr}}\beta^0))^2}{\lambda^0(t - z^{\mathrm{tr}}\beta^0)}\left\{(z^{\mathrm{tr}}a)^2\right.$$
$$\left. -(z^{\mathrm{tr}}a)E_0(Z^{\mathrm{tr}}a \mid T - Z^{\mathrm{tr}}\beta^0 \geqslant t - z^{\mathrm{tr}}\beta^0)\right\}\mathrm{d}t\,\mathrm{d}F_Z(z),$$

which after the change of variable $s = t - z^{\mathrm{tr}}\beta^0$ becomes

$$\int\int P(T - Z^{\mathrm{tr}}\beta^0 \geqslant s)\frac{(\lambda^{0\prime}(s))^2}{\lambda^0(s)}\left\{z^{\mathrm{tr}}a - E_0(Z^{\mathrm{tr}}a \mid T - Z^{\mathrm{tr}}\beta^0 \geqslant s)\right\}^2\mathrm{d}s\,\mathrm{d}F_Z(z)$$
$$= \int \frac{(\lambda^{0\prime}(s))^2}{\lambda^0(s)}\mathrm{Var}_0(Z^{\mathrm{tr}}a \mid T - Z^{\mathrm{tr}}\beta^0 \geqslant s)\,P(T - Z^{\mathrm{tr}}\beta^0 \geqslant s)\,\mathrm{d}s \tag{19}$$

and this last formula agrees with the efficient information bound formula given by Ritov (1990).

Efficient estimators can next be defined based on a consistent preliminary estimator $\tilde{\beta}^0$ of β^0, together with a preliminary kernel-type estimator $\tilde{\lambda}_{\beta^0}$ obtained from right-censored 'data' $\tilde{\varepsilon}_i \equiv X_i - Z_i^{\mathrm{tr}}\tilde{\beta}^0$.

Note that the smoothness of q_z in t, as provided by (C_3), was needed to check the smoothness of λ_β in β as required for (A_6). Assumption (A_7) is easily checked directly, where the distance function ρ is taken to be $\rho(\lambda_1, \lambda_2) = k\left(\sum_{j=0}^2 \|\frac{\mathrm{d}^j}{\mathrm{d}t^j}(\lambda_1 - \lambda_2)\|_{\mathscr{V}_0}\right)$, with

the real-valued function $k(x)$ defined as $k(x) \equiv x I_{[x \geqslant 1/2]} - \log(1-x)_{[x<1/2]}$. We now proceed to exhibit the estimator $\tilde{\lambda}_\beta$ satisfying (A_8) by defining preliminary estimators.

A preliminary estimator $\tilde{\beta}^0$ can be obtained in the spirit of Koul et al. (1981) by regression

$$\Delta_i \; T_i \; / \; \hat{S}_{C|Z}(T_i \mid Z_i) \quad \text{on } Z_i .$$

In this generality, $\hat{S}_{C|Z}$ will be some kernel-based nonparametric regression estimator, as in Cheng (1989). In the more special case where Z_i and C_i are also independent, the Kaplan–Meier estimator $\hat{S}_C^{\mathrm{KM}}(T_i)$ will do: this was (after some modifications needed to obtain asymptotic distributional results) the setting and approach of Koul et al. (1981).

The preliminary estimator $\tilde{\lambda}^0$ is obtained as a kernel-density variant of the Nelson–Aalen estimator, as described by Ramlau–Hansen (1983), with kernel *cdf* $A(\cdot)$, bandwidth $b_n \searrow 0$ slowly enough (say $b_n \sim an^{-1/6}$):

$$\begin{aligned}\tilde{\lambda}^0(w) &= \frac{1}{b_n} \int A'\left(\frac{w-u}{b_n}\right) \frac{\sum_i \mathrm{d}N_i(u + Z_i^{\mathrm{tr}}\tilde{\beta}^0)}{\sum_i I_{[T_i \geqslant u + Z_i^{\mathrm{tr}}\tilde{\beta}^0]}} \\ &= \frac{1}{b_n} \sum_{i=1}^n \Delta_i \, A'\left(\frac{w - T_i + Z_i^{\mathrm{tr}}\tilde{\beta}^0}{b_n}\right) \Bigg/ \sum_{j=1}^n I_{[T_j \geqslant T_i + (Z_j - Z_i)^{\mathrm{tr}}\tilde{\beta}^0]}.\end{aligned}$$

Then $\tilde{\lambda}_\beta$ is defined by substituting the estimators $\tilde{\lambda}^0$ into empirical averages over Z within λ_β defined by (18),

$$\tilde{\lambda}_\beta(t) \equiv \sum_{i=1}^n A\left(\frac{T_i - t - Z_i^{\mathrm{tr}}\beta}{b_n}\right) \tilde{\lambda}_0\left(t + Z_i^{\mathrm{tr}}\left(\beta - \tilde{\beta}^0\right)\right) \Bigg/ \sum_{i=1}^n A\left(\frac{T_i - t - Z_i^{\mathrm{tr}}\beta}{b_n}\right)$$

and $\tilde{\Lambda}_\beta$ is obtained by numerically integrating $\tilde{\lambda}_\beta$ over $[0, t]$. The estimator $\tilde{\lambda}_\beta$ is easily shown to satisfy (A_8) if the kernel *cdf* $A(\cdot)$ is compactly supported and three times continuously differentiable. Finally, we substitute these expressions into

$$\mathrm{logLik}(\beta, \tilde{\lambda}_\beta) = \sum_{j=1}^n \left\{ \Delta_j \; \log \tilde{\lambda}_\beta(T_j - Z_j^{\mathrm{tr}}\beta) - \tilde{\Lambda}_\beta(T_j - c Z_j^{\mathrm{tr}}\beta) \right\}$$

which is to be numerically maximized over β in defining $\tilde{\beta}$.

4. Transformation and frailty models

The semiparametric problems which motivated the present work concern transformation and frailty models in survival analysis (Cox, 1972; Clayton and Cuzick, 1986; Cheng et al., 1995; Parner, 1998; Slud and Vonta, 2002, 2004; Kosorok et al., 2004). This class of models postulates, for the random lifetime of an individual with an observed d-dimensional vector Z of covariates, a conditional survival function

$$S_{T^0|Z}(t|z) = \exp(-G(\mathrm{e}^{z^{\mathrm{tr}}\beta} \Lambda(t))), \tag{20}$$

where G is a known function, satisfying the following regularity conditions given by Slud and Vonta (2004):

(F_1) G is a strictly increasing and concave $\mathscr{C}^3$ function on $(0,\infty)$ satisfying $G(0)=0$, $0<G'(0)<\infty$, $G(\infty)=\infty$, $G'(\infty)=0$, along with the further properties

$$\sup_{x>0}(-xG''(x))/G'(x)<\infty, \quad \sup_{x>0}|xG'''(x)/G'(x)|<\infty,$$

$$\int e^{-G(x)}\,\log(1/G'(x))\,G'(x)\,dx<\infty.$$

The assumptions imposed on the function G are easily satisfied by the Clayton–Cuzick model (Gamma distributed frailty, which corresponds to $G(x)=b^{-1}\log(1+bx)$ for a constant $b>0$) and by Inverse Gaussian frailties. Examples of different frailty distributions, also satisfying the assumptions, can be found in Kosorok et al. (2004) and references cited there.

Here the unknown parameters, with true values $(\beta^0,\ \lambda^0)$, are $(\beta,\ \lambda)$ where $\beta\in\mathbb{R}^d$ and $\Lambda(t)\equiv\int_0^t\lambda(s)\,ds$ is a cumulative-hazard function, i.e. $\lambda(s)\geqslant 0$, $\Lambda(\infty)=\infty$. For notational simplicity, we assume that the variables $Z\in\mathbb{R}^d$ have discrete and finite-valued distribution, $\pi_z\equiv P(Z=z)$, but a compactly supported distribution yields the same set of theoretical results. Again the assumption (C_0) is in force, with $\mathscr{U}$ a small closed Euclidean ball around $\beta^0\in\mathbf{R}^d$.

The data are randomly right-censored, i.e., there is an underlying positive random variable C conditionally independent of T given $Z=z$ with

$$R_z(y)\equiv P(C\geqslant y\mid Z=z).$$

The observable data for a sample of n independent individuals are $(T_i\equiv\min(T_i^0,C_i),\ \Delta_i\equiv I_{[T_i^0\leqslant C_i]},\ Z_i, i=1,\ldots,n)$, encoded into the processes

$$N_z^i(t)=\Delta_i\,I_{[Z_i=z,T_i\leqslant t]}, \quad Y_z^i(t)=I_{[Z_i=z,T_i\geqslant t]}.$$

For the present, we will assume that the distribution of Z and the conditional censoring survivor functions R_z are known, which is actually not a significant restriction since the estimators we develop do not depend on the form of R_z and are 'adaptive' in the sense of attaining the same information bounds as for the case of estimators allowed to depend upon R_z. However, following Slud and Vonta (2002) we also impose a nontrivial technical restriction on the (correctly specified) distribution of the data through the function

$$q_z(t)\equiv P(T\geqslant t\mid Z=z)=R_z(t)\,\exp\left(-G(e^{z^{\mathrm{tr}}\beta^0}\,\Lambda^0(t))\right) \tag{21}$$

(which differs from the notation q_z of Slud and Vonta (2002, 2004) by omitting the factor $\pi_z=P(Z=z)$) in the form

$$q_z(t)\equiv 0 \quad \text{for } t>\tau_0, \quad q_z(\tau_0)>0. \tag{22}$$

The import of this restriction is that for some fixed time τ_0 beyond which the individuals in a study do have a positive probability of surviving uncensored, the data are automatically

right-censored at τ_0. This is not a practically restrictive condition, but as a matter of theoretical technique it should be removed. (It is not needed in either the Cox model or the Clayton–Cuzick semiparametric Pareto model when efficient estimation in these models is treated by other methods.)

Define ν to be Lebesgue measure on the interval $[0, \tau_0]$. The data space $\mathbf{D} = \mathbb{R} \times \mathbb{R}^d \times \{0, 1\}$ again consists of triples $x = (t, z, \delta)$. The Borel probability measure μ on $\mathbf{D}$ is now given by

$$\begin{aligned} \mathrm{d}\mu(t, z, \delta) = \pi_z \big(& \delta\, q_z(t) G'(\mathrm{e}^{z^{\mathrm{tr}}\beta^0} \Lambda^0)\, \mathrm{e}^{z^{\mathrm{tr}}\beta^0} \lambda^0(t)\, \mathrm{d}t \\ & - (1-\delta) \mathrm{e}^{-G(\mathrm{e}^{z^{\mathrm{tr}}\beta^0}\, \Lambda^0(t))}\, \mathrm{d}R_z(t) \big). \end{aligned}$$

The statistical problem is further specified by

(F_2) Conditions (C_0) and (22) hold, and there exist positive, finite constants $c_1 < c_2$ such that

$$c_1 \leqslant \lambda^0(t) \leqslant c_2, \quad \text{a.e. } t \in [0, \tau_0]$$

and the candidate parameters (β, λ) are all assumed to lie in $\mathscr{U} \times \mathscr{V}$, with

$$\mathscr{V} \equiv \{\lambda \in L^1([0, \tau_0],\ \nu) :\ c_1 \leqslant \lambda(t) \leqslant c_2\}$$

and the spaces $\mathscr{G} = \mathscr{G}_0$ of perturbing functions are taken to be $\{\gamma \in L^\infty([0, \tau_0], \nu) : \|\gamma\|_\infty \leqslant c_1\}$.

The semiparametric log-likelihood in this problem is

$$\begin{aligned} & \mathrm{logLik}(\beta, \lambda) \\ & \quad = \sum_z \sum_{i=1}^n \int_0^{\tau_0} \left\{ \log(\mathrm{e}^{z^{\mathrm{tr}}\beta}\, G'(\mathrm{e}^{z^{\mathrm{tr}}\beta} \Lambda)\lambda)\, \mathrm{d}N_z^i - Y_z^i \mathrm{e}^{z^{\mathrm{tr}}\beta} G'(\mathrm{e}^{z^{\mathrm{tr}}\beta}\Lambda)\lambda\, \mathrm{d}\nu \right\}. \end{aligned} \tag{23}$$

This log-likelihood leads to the expression, for $a \in \mathbb{R}^d$, $\gamma \in L^\infty(\nu)$

$$\begin{aligned} & \begin{pmatrix} A_{\beta^0,\lambda^0}(a,a) & B_{\beta^0,\lambda^0}(a,\gamma) \\ B_{\beta^0,\lambda^0}(a,\gamma) & C_{\beta^0,\lambda^0}(\gamma,\gamma) \end{pmatrix} \\ & \quad = \sum_z \int_0^{\tau_0} \pi_z\, q_z \begin{pmatrix} a^{\mathrm{tr}}\, z\, \{1 + (xG''(x)/G'(x))_{x=\mathrm{e}^{z^{\mathrm{tr}}\beta^0}\Lambda^0}\} \\ \mathrm{e}^{z^{\mathrm{tr}}\beta^0} \int_0^{\cdot} \gamma\, \mathrm{d}\nu (G''/G')_{x=\mathrm{e}^{z^{\mathrm{tr}}\beta^0}\Lambda^0} + \gamma/\lambda^0 \end{pmatrix}^{\otimes 2} \\ & \qquad \times \mathrm{e}^{z^{\mathrm{tr}}\beta^0} G'(\mathrm{e}^{z^{\mathrm{tr}}\beta^0}\Lambda^0)\, \lambda^0\, \mathrm{d}\nu \end{aligned} \tag{24}$$

where $\Lambda^0 = \int_0^{\cdot} \lambda^0\, \mathrm{d}\nu = \int_0^{\cdot} \lambda^0(t)\, \mathrm{d}t$.

Under assumptions (F_1)–(F_2) above, conditions (A_1) of Section 2 and (A_2)–(A_4) in Section A.1, along with (A_7) for $\rho(\lambda_1, \lambda_2) \equiv \|\lambda_1 - \lambda_2\|_{\infty,\nu}$, are easily verified by inspection. Next we verify (A_5). In this model, $\mathscr{I}(\beta^0, \lambda^0)$ evaluated at nonzero $(a, \gamma) \in \mathbb{R}^d \times \mathscr{G}_0$ is

positive, unless for some (a, γ),

$$a^{\mathrm{tr}} z \left(1 + \left(\frac{xG''(x)}{G'(x)}\right)_{x=e^{z^{\mathrm{tr}}\beta^0}\Lambda^0}\right) + e^{z^{\mathrm{tr}}\beta^0} \left(\int_0^{\cdot} \gamma \, dv\right) \left(\frac{G''}{G'}\right)_{x=e^{z^{\mathrm{tr}}\beta^0}\Lambda^0} + \frac{\gamma}{\lambda^0} = 0.$$

Direct reasoning shows this to be impossible by (C_0) for any $(a, \gamma) \neq (0, 0)$—an assertion equivalent to (A_5)—since multiplication by $\lambda^0 \cdot G'(e^{z^{\mathrm{tr}}\beta^0}\Lambda^0)$ and identification of complete differentials implies the last equality on $[0, \tau_0]$, for fixed (a, γ), to be equivalent to

$$a^{\mathrm{tr}} z\Lambda^0 + \int_0^{\cdot} \gamma \, dv \equiv 0 \quad \text{on } [0, \tau_0].$$

Unlike the situation in Section 3, the restricted minimizers λ_β cannot be given explicitly in these Transformation models. This is true even for the Clayton and Cuzick (1986) model successfully analyzed by Parner (1998). However, Slud and Vonta (2004) have shown that $\lambda_\beta(t)$ is a uniquely determined $\mathscr{C}^2$ family (smoothly indexed by β) of continuous functions of $t \in [0, \tau_0]$, through solving a family of ordinary differential equations for $L(s) \equiv \Lambda_\beta\left((\Lambda^0)^{-1}(s)\right)$, along with an auxiliary function Q. In terms of the modified notation

$$\overline{q}_z(s) \equiv q_z((\Lambda^0)^{-1}(s)) = e^{-G(e^{z^{\mathrm{tr}}\beta^0}s)} R_z((\Lambda^0)^{-1}(s)) \tag{25}$$

these equations are

$$L'(s) = \frac{\sum_z \pi_z e^{z^{\mathrm{tr}}\beta^0} \overline{q}_z(s) G'(e^{z^{\mathrm{tr}}\beta^0}s)}{\sum_z \pi_z e^{z^{\mathrm{tr}}\beta} \overline{q}_z(s) G'(e^{z^{\mathrm{tr}}\beta}L(s)) + Q(s)},$$

$$Q'(s) = \sum_z \pi_z e^{z^{\mathrm{tr}}\beta} \overline{q}_z(s) \left.\frac{G''}{G'}\right|_{e^{z^{\mathrm{tr}}\beta}L(s)} \times (e^{z^{\mathrm{tr}}\beta^0} G'(e^{z^{\mathrm{tr}}\beta^0}s) - e^{z^{\mathrm{tr}}\beta} G'(e^{z^{\mathrm{tr}}\beta}L(s)) L'(s)) \tag{26}$$

subject to the initial and terminal conditions

$$L(0) = 0, \quad Q(\Lambda^0(\tau_0)) = 0.$$

Slud and Vonta (2004) show that these ODE's (26) have unique solutions for all β in a sufficiently small compact neighborhood $\mathscr{U}$ of β^0; are smooth ($\mathscr{C}^2$) with respect to both β and the parameter $\alpha = Q(0)$; and, for $\lambda_\beta(s) \equiv L'(\Lambda^0(s))\lambda^0(s)$, minimize the functional $\mathscr{K}(\beta, \lambda_\beta)$. Smoothness of λ_β in β on the compact set $\mathscr{U}$ then implies (A_6).

It remains to explain how to construct a smooth family of estimators $\tilde{\lambda}_\beta$ of λ_β indexed by β and satisfying (A_8). First, Slud and Vonta (2002, 2004) discuss $\sqrt{n}$ consistent preliminary estimation of β^0, Λ^0, and estimating-equation based estimators of Cheng et al. (1995) can also serve as preliminary estimators. Next, it can be shown that when the consistent preliminary estimators are substituted for (β^0, λ^0) into the second-order ODE system (26) determining λ_β, the solutions which we denote as $(\tilde{L}, \tilde{Q})$ are still smooth functions of β which are uniformly close on $\mathscr{U}$ to the solutions (L, Q) of (26). Then we define

$\tilde{\lambda}_\beta(t) \equiv \tilde{L}'(\tilde{\Lambda}^0(s))\,\tilde{\lambda}^0(t)$, and this definition will satisfy (A_8) as long as $\mathscr{U}$ was initially chosen as a small enough closed ball containing β^0. Additional research on the computational implementation and moderate-sample behavior of these estimators is needed, but the theory of Section 2 shows that these estimators are efficient.

4.1. Information bound formula

The discussion in Technical Appendix A leading up to formula (46) implies that the information bound $\mathscr{I}^0_\beta$ has the implicit expression

$$a^{\mathrm{tr}}\,\mathscr{I}^0_\beta\,a = A_{\beta^0,\lambda^0}(a,a) + B_{\beta^0,\lambda^0}(a,\ a^{\mathrm{tr}}\,\nabla_\beta\,\lambda_{\beta^0}). \tag{27}$$

Since the formula for information bounds $\mathscr{I}^0_\beta$ involves the form of λ_β only through $\nabla_\beta\,\lambda_{\beta^0}$, we obtain such bounds in much more explicit form than those previously based on Sturm–Liouville problem solutions as in Klaassen (1993) or Bickel et al. (1993). First, we follow the method of Slud and Vonta (2004) in observing that the d-vector-valued function $L_*(s) \equiv \nabla_\beta\,\Lambda_\beta((\Lambda^0)^{-1}(s))\,|_{\beta=\beta^0} = \int_0^s \nabla_\beta\,L'(t)\,\mathrm{d}t\,|_{\beta=\beta^0}$ is determined through an adjoint system of linear ordinary differential equations

$$\begin{aligned} L'_*(s) &= -\frac{Q_*(s) + \sum_z \pi_z\,\mathrm{e}^{z^{\mathrm{tr}}\beta^0}\overline{q}_z(s)\,P_z(s)}{\sum_z \pi_z\,\mathrm{e}^{z^{\mathrm{tr}}\beta^0}\overline{q}_z(s)G'(\mathrm{e}^{z^{\mathrm{tr}}\beta^0}s)}, \\ Q'_*(s) &= -\sum_z \pi_z\overline{q}_z(s)\,\mathrm{e}^{2z^{\mathrm{tr}}\beta^0}\,\frac{G''(\mathrm{e}^{z^{\mathrm{tr}}\beta^0}s)}{G'(\mathrm{e}^{z^{\mathrm{tr}}\beta^0}s)}(P_z(s) + G'(\mathrm{e}^{z^{\mathrm{tr}}\beta^0}s)\,L'_*(s)) \end{aligned} \tag{28}$$

with initial and terminal conditions

$$L_*(0) = 0, \quad Q_*(\Lambda^0(\tau_0)) = 0, \tag{29}$$

where

$$P_z(s) \equiv zG'(\mathrm{e}^{z^{\mathrm{tr}}\beta^0}s) + G''(\mathrm{e}^{z^{\mathrm{tr}}\beta^0}s)\,\mathrm{e}^{z^{\mathrm{tr}}\beta^0}(zs + L_*(s))$$

and $Q_*(s) = \nabla_\beta\,Q(s)$. Substituting into (27) using (24) gives as formula for the information bound

$$\begin{aligned} \mathscr{I}^0_\beta = \sum_z \pi_z \int_0^{\tau_0} & z\,q_z\,\mathrm{e}^{z^{\mathrm{tr}}\beta^0}\,G'(\mathrm{e}^{z^{\mathrm{tr}}\beta^0}\Lambda^0)\,\left(1 + (xG''(x)/G'(x))_{x=\mathrm{e}^{z^{\mathrm{tr}}\beta^0}\Lambda^0}\right) \\ & \cdot\left(z + \mathrm{e}^{z^{\mathrm{tr}}\beta^0}(z\Lambda^0 + \nabla_\beta\Lambda_{\beta^0})\,\frac{G''}{G'}\bigg|_{x=\mathrm{e}^{z^{\mathrm{tr}}\beta^0}\Lambda^0} + \nabla_\beta\,\log\,\lambda_{\beta^0}\right)^{tr}\lambda^0\,\mathrm{d}t, \end{aligned} \tag{30}$$

where

$$\nabla_\beta\lambda_\beta(s)\,|_{\beta=\beta^0} = \lambda^0(s)L'_*(\Lambda^0(s)), \quad \nabla_\beta\Lambda_\beta(s)|_{\beta=\beta^0} = L_*(\Lambda^0(s)). \tag{31}$$

The Eqs. (28) leading to (31) can be solved computably, as shown in Section A.4 in the Technical Appendix below. Substitution into (30), and numerical integration, provides

Table 1
Information bound calculations for two-sample Clayton–Cuzick frailty model $G(t)=b^{-1}\log(1+bt)$, $\lambda^0(t)\equiv 1$, and $R_z(t)=\max(0,1-t/\tau^{(z)})\,I_{[t\leqslant\tau_0]}$ for $z=0,1$, with indicated parameters $(b,\tau^{(0)},\tau^{(1)})$, for $\beta=\log 2$.

b	$\tau^{(0)}$	$\tau^{(1)}$	τ_0	$A_{\beta^0,\lambda^0}(1,1)$	ParInfo	$\mathscr{I}^0_\beta$
.0001	1.e8	1.e8	20	.4999	.2254	.2253
.5	1.e8	1.e8	20	.2500	.1218	.1160
1	1.e8	1.e8	20	.1667	.0807	.0770
2	1.e8	1.e8	20	.1000	.0488	.0458
3	1.e8	1.e8	20	.0714	.0350	.0326
4	1.e8	1.e8	20	.0556	.0272	.0253
.0001	2	4	3.96	.3773	.1677	.1676
.5	3	6	5.95	.2227	.1128	.1054
1	4	8	7.9	.1564	.0752	.0741
2	5	10	9.95	.0967	.0470	.0453
3	6	12	11.95	.0700	.0342	.0325
4	7	14	13.95	.0548	.0268	.0253

The fifth column contains the upper-left entries of the information matrix; the sixth column ParInfo is the full-likelihood information about β computed in Slud (1986) for a specific model with 5-dimensional nuisance density λ; and the seventh column is the information bound $\mathscr{I}^0_\beta$.

semiparametric information bounds in the survival transformation model in a new form which allows tractable calculations in frailty and transformation settings where essentially no previous computations of such bounds have been available.

As examples of the resulting information formula (30), we provide numerical bounds for several cases of the two-sample right-censored Clayton and Cuzick (1986) frailty model. Approximate numerical values of these information bounds were previously given in Slud (1986, Tables I-II) via models with five-dimensional parameterization of the 'nuisance hazard' λ. Table 1 above exhibits the quantities $A_{\beta^0,\lambda^0}(1,1)$ and $\mathscr{I}^0$, for the case where the covariate $Z_i=0,\ 1$ is the group-label for subject i, with approximately half of all n subjects allocated to each of the two groups; where the model (20) holds with $G(x)=b^{-1}\log(1+bx)$; where the group-1 over group-0 log hazard-ratio parameter is $\beta=\beta_1=\log 2$, with $\lambda^0\equiv 1$; and where the censoring distributions $R_z(t)=P(C\geqslant t\mid Z=z)$ are Uniform$[0,\tau^{(z)}]$ subject to additional, administrative right-censoring at τ_0, i.e., for $z=0,1$, $R_z(t)=\max(0,\ 1-t/\tau^{(z)})\,I_{[t\leqslant\tau_0]}$. (Note that the cases of very large $\tau^{(z)}$ values in the table correspond to data uncensored before τ_0.) In this setting, we calculated $\mathbf{H}(s)$ numerically, from formula (67) at a spacing of $h=\tau_0/1000$, leading to $\mathscr{I}^0_\beta$ values with accuracy approximately 0.0001. The numerically calculated values $\mathscr{I}^0_\beta$ are in all cases slightly smaller than the values found in Slud (1986). But note that the numerical values in the columns $\alpha=0.5$ (corresponding to $b=0.5$ in our notation) of Tables I-II of Slud (1986) are incorrect: they should be multiplied throughout by 2. The case $b=0.0001$ corresponds closely to the relatively easily calculated and well-established values for Cox model, which is the limiting case of the Clayton–Cuzick model at $b=0+$. The main finding in this table is that the information bound values which Slud (1986) had thought were already converged with a nuisance-hazard parameter of dimension five, were at that stage still a few percent away from convergence.

Acknowledgements

This research was supported by travel grants from the University of Cyprus and University of Maryland. We thank J. Robins for the suggestion to apply the modified profile likelihood method to censored linear regression.

Appendix A. Technical Appendix

A.1. Definitions and regularity conditions

We begin by listing the further regularity conditions (A_2)–(A_8) for the general statistical problems we study. Basic notations and assumptions (A_0)–(A_1) are as in Section 2 above.

(A_2) For all $\lambda \in \mathscr{V}$, there exists $0 < \varepsilon \equiv \varepsilon(\lambda) < 1$ such that, for all $\gamma \in \mathscr{G}_0$ defined in (3), the mapping

$$(\beta,\ \vartheta) \mapsto f_X(\cdot, \beta,\ \lambda + \vartheta\gamma) \in L^1(\mu)$$

is twice continuously differentiable (in strong or Fréchet sense) from $\mathscr{U} \times (-\varepsilon, \varepsilon)$ to $L^1(\mu)$.

(A_3) There exists $b \in L^1(\mu)$ such that, for all $\beta \in \mathscr{U}$, $\lambda \in \mathscr{V}$ and for μ-a.e. $x \in \mathbb{R}^k$,

$$|\log f_X(x, \beta, \lambda)| + \sup_{\gamma\in\mathscr{G}_0} \left(\|\nabla_{(\beta,\vartheta)} \log f_X(x, \beta, \lambda + \vartheta\gamma)\|_2\,|_{\vartheta=0}\right) + \sup_{\gamma\in\mathscr{G}} \left(\|\nabla^{\otimes 2}_{(\beta,\vartheta)} \log f_X(x, \beta, \lambda + \vartheta\gamma)\|_2|_{\vartheta=0}\right) \leqslant b(x). \tag{32}$$

In addition, for each $\lambda \in \mathscr{V}$, $\gamma \in \mathscr{G}_0$, there exists $\varepsilon_1 > 0$ so small that

$$\sup_{\beta\in\mathscr{U}} \sup_{|\vartheta|\leqslant\varepsilon_1} \|\nabla^{\otimes 2}_{(\beta,\vartheta)} \log f_X(x, \beta, \lambda + \vartheta\gamma)\|_2 \in L^1(\mu).$$

(A_4) For each $\lambda \in \mathscr{V}$ there exists $r_\lambda \in L^1(\mu)$ such that, for all $\beta \in \mathscr{U}$ and for μ-a.e. x,

$$|f_X(x, \beta, \lambda)| + \sup_{\gamma\in\mathscr{G}_0} \left(\|\nabla_{(\beta,\vartheta)} f_X(x, \beta, \lambda + \vartheta\gamma)\|_2|_{\vartheta=0}\right) + \sup_{\gamma\in\mathscr{G}}(\|\nabla^{\otimes 2}_{(\beta,\vartheta)} f_X(x, \beta, \lambda + \vartheta\gamma)\|_2|_{\vartheta=0}) \leqslant r_\lambda(x). \tag{33}$$

In addition, for each $\lambda \in \mathscr{V}$, $\gamma \in \mathscr{G}_0$, there exists $\varepsilon_2 > 0$ so small that

$$\sup_{\beta\in\mathscr{U}} \sup_{|\vartheta|\leqslant\varepsilon_2} \|\nabla^{\otimes 2}_{(\beta,\vartheta)} f_X(x, \beta, \lambda + \vartheta\gamma)\|_2 \in L^1(\mu).$$

Under conditions (A_2), (A_3) and (A_4), the second-derivative operator $\mathscr{J} \equiv \mathscr{J}_{\beta,\lambda}$ on $\mathbb{R}^d \times \mathscr{G}_0$ is defined through the bilinear forms

$$\mathscr{J}_{\beta,\lambda} = \begin{pmatrix} A & B \\ B^* & C \end{pmatrix} \tag{34}$$

for $a_1,\ a_2 \in \mathbb{R}^d$, $\gamma \in \mathscr{G}_0$, by

$$A(a_1, a_2) = A_{\beta,\lambda}(a_1, a_2) = -\int \left(a_1^{\mathrm{tr}}\, \nabla_\beta^{\otimes 2}\ \log\ f_X(x, \beta, \lambda)\, a_2\right) \mathrm{d}\mu(x),$$

$$B(a, \gamma) = B_{\beta,\lambda}(a, \gamma) = -\int \left(a_1^{\mathrm{tr}}\, \nabla_\beta \nabla_\vartheta\ \log\ f_X(x, \beta, \lambda + \vartheta\gamma)\right)|_{\vartheta=0}\, \mathrm{d}\mu(x),$$

$$C(\gamma, \gamma) = C_{\beta,\lambda}(\gamma, \gamma) = -\int \left(\frac{\mathrm{d}^2}{\mathrm{d}\vartheta^2}\ \log\ f_X(x, \beta, \lambda + \vartheta\gamma)\right)\Big|_{\vartheta=0} \mathrm{d}\mu(x).$$

The transposed or adjoint bilinear form corresponding to B is

$$B^*_{\beta,\lambda}(\gamma,\ a) = B_{\beta,\lambda}(a,\ \gamma).$$

The evaluation $\mathscr{J}_{\beta,\lambda}\left((a_1, \gamma_1),\ (a_2, \gamma_2)\right)$ for $\beta \in \mathscr{U}, \lambda \in \mathscr{V}$ is equal by definition to $A_{\beta,\lambda}(a_1, a_2) + B_{\beta,\lambda}(a_1, \gamma_2) + B_{\beta,\lambda}(a_2, \gamma_1) + C_{\beta,\lambda}(\gamma_1,\ \gamma_2)$.

(A_5) For all $(a, \gamma) \in \mathbb{R}^d \times \mathscr{G}_0 \setminus\{(0, 0)\}$ and $(\beta, \lambda) \in \mathscr{U} \times \mathscr{V}$,

$$|A_{\beta,\lambda}(a, a)| < \infty, \quad |B_{\beta,\lambda}(a,\ \gamma)| < \infty, \quad C_{\beta,\lambda}(\gamma,\ \gamma) < \infty,$$
$$A_{\beta^0,\lambda^0}(a, a) + 2B_{\beta^0,\lambda^0}(a,\ \gamma) + C_{\beta^0,\lambda^0}(\gamma,\ \gamma) > 0. \tag{35}$$

(A_6) For $\beta \in \mathscr{U}$, there exists a unique minimizer $\lambda_\beta \in \mathscr{V}$ of $\mathscr{K}(\beta,\ \cdot)$ which is also the unique solution of the derivative equation

$$(D_\lambda\, \mathscr{K}(\beta, \lambda_\beta))(\gamma) = 0 \quad \forall \beta \in \mathscr{U},\ \gamma \in \mathscr{G}_0. \tag{36}$$

Moreover, $\beta \mapsto \lambda_\beta$ is twice Fréchet differentiable as a mapping from $\mathscr{U} \subset \mathbb{R}^d$ to $\mathscr{V}_0$, (cf. (2)), such that $a_1^{\mathrm{tr}} \nabla_\beta \lambda_\beta$, $a_1^{\mathrm{tr}} \nabla_\beta^{\otimes 2} \lambda_\beta\, a_2 \in \mathscr{G}$ for all $a_1,\ a_2 \in \mathbb{R}^d$ with sufficiently small norms, and there is a finite constant α_3 such that

$$\text{for } j = 1,\ 2, \quad \sup_\beta \|\nabla_\beta^{\otimes j}\, \lambda_\beta\|_2 \leqslant \alpha_3 c_1(\cdot).$$

One further regularity condition on the family of densities $f_X(x, \beta, \lambda)$ is needed in the setting of infinite-dimensional λ.

(A_7) There exists a function $m \in L^1(\mathbb{R}^k, \mu)$ and a distance-function ρ on $\{\lambda + \vartheta\gamma : \lambda \in \mathscr{V}, \gamma \in \mathscr{G}_0,\ |\vartheta| \leqslant \varepsilon(\lambda)\}$ such that $\rho(\lambda_1, \lambda_2) \geqslant \|\lambda_1 - \lambda_2\|_{\mathscr{V}_0}$, and such that for all $j = 0, 1, 2$, all $\beta \in \mathscr{U}$ and all $\lambda_1,\ \lambda_2 \in \mathscr{V}, \gamma \in \mathscr{G}$,

$$\|\nabla_{(\beta,\vartheta)}^{\otimes j}(\log\ f_X(x, \beta, \lambda_1 + \vartheta\gamma) - \log\ f_X(x, \beta, \lambda_2 + \vartheta\gamma))|_{\vartheta=0}\|_2 \leqslant m(x)\, \rho(\lambda_1, \lambda_2).$$

The final assumption relates to a family of $\mathscr{V}$-valued estimators $\tilde{\lambda}_\beta$ assumed to be defined for each n, i.e. a family of measurable mappings $\tilde{\lambda} : \mathscr{U} \times (\mathbb{R}^k)^n \to \mathscr{V}$, with $\tilde{\lambda}_\beta \equiv \tilde{\lambda}(\beta, \mathbf{X}_1, \ldots, \mathbf{X}_n)$.

(A_8) The estimator-process $\tilde{\lambda}_\beta$ defined on $\beta \in \mathscr{U}$, as a mapping from $\mathscr{U}$ to the λ-parameter space $\mathscr{V} \subset \mathscr{V}_0$, is twice continuously differentiable, with $\tilde{\lambda}_\beta \in \mathscr{V}$ and $a_1^{\rm tr}\nabla_\beta\tilde{\lambda}_\beta$, $a_1^{\rm tr}\nabla_\beta^{\otimes 2}\tilde{\lambda}_\beta\, a_2 \in \mathscr{G}$ a.s. for all $a_1, a_2 \in \mathbb{R}^d$ with sufficiently small norms. Moreover, the estimators $\tilde{\lambda}_\beta$ are such that, for the same distance-function ρ as in (A_7), and for all vectors $a, a_1, a_2 \in \mathbb{R}^d$ of sufficiently small norms, as $n \to \infty$,

$$\sup_{\beta\in\mathscr{U}} \rho\left(\lambda_\beta + a^{\rm tr}\nabla_\beta\lambda_\beta + a_1^{\rm tr}\nabla_\beta^{\otimes 2}\lambda_\beta\, a_2\,,\ \tilde{\lambda}_\beta + a^{\rm tr}\nabla_\beta\tilde{\lambda}_\beta + a_1^{\rm tr}\nabla_\beta^{\otimes 2}\tilde{\lambda}_\beta\, a_2\right) \xrightarrow{\rm P} 0. \tag{37}$$

Moreover, there is a positive constant α_4 such that with probability approaching 1 as $n \to \infty$, for all $\beta \in \mathscr{U}$, v a.e. t,

$$|\tilde{\lambda}_\beta(t) - \lambda_\beta(t)| + \sum_{j=1}^{2} \|\nabla_\beta^{\otimes j}(\tilde{\lambda}_\beta - \lambda_\beta)(t)\|_2 \leqslant \alpha_4\, c_1(t). \tag{38}$$

A.2. Discussion of the assumptions

Assumptions (A_0) and (A_2)–(A_5) are slight variants of standard statistical regularity conditions (for likelihood inference based on finite-dimensional parameters) dating back to Harald Cramer. These assumptions are imposed both on the β and λ parameter-components, although only the structural parameter β is to be estimated efficiently. However, it is very convenient in the infinite-dimensional case that positive definiteness of $\mathscr{J}_{\beta,\lambda}$ need be checked only at the true parameter (β^0, λ^0). The usual assumption of *identifiability* for the parameterization (assertion (47) below) follows immediately from (A_2)–(A_5). The requirement that $\lambda + \vartheta\gamma$ continue to be an element where $f_X(\cdot)$ is defined, as in (A_2), is a somewhat restrictive feature of our setup associated with (A_1), allowing consideration only of elements λ bounded above and below by fixed functions. Assumptions (A_3) and (A_4) are needed for a dominated convergence justification of interchange of derivatives and integrals in expressions for the first- and second-order gradients in (β, ϑ) of $\mathscr{K}(\beta, \lambda + \vartheta\gamma)$. It is important that the expressions A, B, C for the block-decomposition of $\mathscr{J}_{(\beta,\lambda)}$ exist on a set of γ satisfying the density restriction (4), as in (A_5), and that the first derivatives arising in the expression (36) and in Proposition 1 also hold for all $\gamma \in \mathscr{G}_0$. However, the uniform second-derivative bounds assumed in $\mathscr{G}$, but *not* in $\mathscr{G}_0$, are needed in the bracketing metric entropy estimates underlying the proof of Proposition 2. The space $\mathscr{G}$ on which the more restrictive bounds are assumed to hold might be much smaller than $\mathscr{G}_0$: its main role is to contain all small multiples of the first- and second-order gradients of $\lambda_\beta, \tilde{\lambda}_\beta$. The idea of condition (A_6), along with Proposition 1 as a consequence, was first given by Severini and Wong (1992). Assumptions (A_7) and (A_8), as well as the existence of b in (A_3) simultaneously for all λ, are needed specifically for the validity of bracketing metric entropy bounds and uniform laws of large numbers on empirical processes related to the log-likelihood and its derivatives, with $\tilde{\lambda}_\beta$ substituted for λ.

A.3. Consequences of the assumptions

We derive two types of technical consequences of the assumptions (A_0)–(A_8): first those based on standard advanced-calculus manipulations on $\mathscr{K}(\beta, \lambda)$, and second those which use empirical process theory on the normalized log-likelihood and its derivatives.

A.3.1. Calculus on $\mathscr{K}$

By Dominated Convergence and (A_3), for each $\lambda \in \mathscr{V}$, $\gamma \in \mathscr{G}_0$, and $\varepsilon = \varepsilon(\lambda)$ as in (A_2), the derivatives $\nabla_{(\beta,\vartheta)} \mathscr{K}(\beta, \lambda + \vartheta\gamma)$ and $\nabla^{\otimes 2}_{(\beta,\vartheta)} \mathscr{K}(\beta, \lambda + \vartheta\gamma)$ are continuous functions of $(\beta, \vartheta) \in \mathscr{U} \times (-\varepsilon, \varepsilon)$. By Dominated Convergence and (A_3)–(A_4) applied to the difference-quotients in the defining integrals, the derivatives commute with the integrals, yielding for $j = 1, 2$,

$$\nabla^{\otimes j}_{(\beta,\vartheta)} \mathscr{K}(\beta, \lambda + \vartheta\gamma) = -\int \nabla^{\otimes j}_{(\beta,\vartheta)} \log f(x, \beta, \lambda + \vartheta\gamma)\, \mathrm{d}\mu(x) \tag{39}$$

and

$$\int \nabla_{(\beta,\vartheta)} f_X(t, \beta, \lambda + \vartheta\gamma)\, \mathrm{d}\mu(t) = \mathbf{0}, \quad \int \nabla^{\otimes 2}_{(\beta,\vartheta)} f_X(t, \beta, \lambda + \vartheta\gamma)\, \mathrm{d}\mu(t) = \mathbf{0}$$

so that

$$\begin{aligned} A_{\beta,\lambda}(a_1, a_2) &= a_1^{\mathrm{tr}} \nabla^{\otimes 2}_{\beta} \mathscr{K}(\beta, \lambda) a_2, \quad B_{\beta,\lambda}(a_1, \gamma) = a_1^{\mathrm{tr}} \nabla_\beta D_\lambda \mathscr{K}(\beta, \lambda)(\gamma), \\ C(\gamma, \gamma) &= D_\lambda \left(D_\lambda \mathscr{K}(\beta, \lambda)(\gamma) \right)(\gamma) \end{aligned} \tag{40}$$

and, with $a^{\otimes 2} \equiv aa^{\mathrm{tr}}$ for $a \in \mathbb{R}^d$,

$$A_{\beta^0,\lambda^0}(a_1, a_2) = \int a_1^{\mathrm{tr}} (\nabla_\beta \log f_X(x, \beta^0, \lambda^0))^{\otimes 2} a_2\, \mathrm{d}\mu(x), \tag{41}$$

$$B_{\beta^0,\lambda^0}(a, \gamma) = \int (a^{\mathrm{tr}} \nabla_\beta \log f_X(x, \beta^0, \lambda^0))\, D_\lambda \log f_X(x, \beta^0, \lambda^0)(\gamma)\, \mathrm{d}\mu(x), \tag{42}$$

$$C_{\beta^0,\lambda^0}(\gamma, \gamma) = \int \left(D_\lambda \log f_X(x, \beta^0, \lambda^0)(\gamma) \right)^2 \mathrm{d}\mu(x). \tag{43}$$

An important consequence of (A_6) appears by applying the Chain Rule for (total) implicit Fréchet differentiation $\nabla^{\mathrm{T}}_\beta$ with respect to β in the determining equation (36). Freely swapping the order of derivatives up to second order and integrals in the definition of $\mathscr{K}(\beta, \lambda_\beta)$ as in (39), we obtain for fixed $a \in \mathbb{R}^d$, $\beta \in \mathscr{U}$, $\gamma \in \mathscr{G}_0$,

$$\begin{aligned} a^{\mathrm{tr}} \nabla^{\mathrm{T}}_\beta \left[\left(D_\lambda \mathscr{K}(\beta, \lambda_\beta) \right)(\gamma) \right] &= -a^{\mathrm{tr}} \nabla^{\mathrm{T}}_\beta \frac{\mathrm{d}}{\mathrm{d}\vartheta} \int \log f_X(x, \beta, \lambda_\beta + \vartheta\gamma)\, \mathrm{d}\mu(x)|_{\vartheta=0} \\ &= -\int \left\{ D_\lambda (a^{\mathrm{tr}} \nabla_\beta \log f_X(x, \beta, \lambda_\beta))(\gamma) \right. \\ &\quad \left. + D_\lambda (D_\lambda (\log f_X(x, \beta, \lambda_\beta))(\gamma))(a^{\mathrm{tr}} \nabla_\beta \lambda_\beta) \right\} \mathrm{d}\mu(x). \end{aligned}$$

Substituting the definitions of the bilinear operators $A_{\beta,\lambda}$, $B_{\beta,\lambda}$, $C_{\beta,\lambda}$ at $\lambda = \lambda_\beta$ then yields

$$a^{\mathrm{tr}} \nabla^{\mathrm{T}}_\beta [(D_\lambda \mathscr{K}(\beta, \lambda_\beta))(\gamma)] = B_{\beta,\lambda_\beta}(a, \gamma) + C_{\beta,\lambda_\beta}(a^{\mathrm{tr}} \nabla_\beta \lambda_\beta, \gamma) = 0. \tag{44}$$

Similarly, with arguments (β, λ_β) understood throughout for $A,\ B,\ C$,

$$a^{\mathrm{tr}}(\nabla_\beta^{\mathrm{T}})^{\otimes 2}\mathscr{K}(\beta,\lambda_\beta)a = A(a,a) + 2B(a, a^{\mathrm{tr}}\nabla_\beta\lambda_\beta) + C(a^{\mathrm{tr}}\nabla_\beta\lambda_\beta, a^{\mathrm{tr}}\nabla_\beta\lambda_\beta)$$
$$= A_{\beta,\lambda_\beta}(a,a) - C_{\beta,\lambda_\beta}(a^{\mathrm{tr}}\nabla_\beta\lambda_\beta, a^{\mathrm{tr}}\nabla_\beta\lambda_\beta) \quad (45)$$

$$= A_{\beta,\lambda_\beta}(a,a) + B_{\beta,\lambda_\beta}(a, a^{\mathrm{tr}}\nabla_\beta\lambda_\beta), \quad (46)$$

where the cross-term involving $B(a,\gamma)$ in the first line of the formula has been replaced according to (44) by $-C(a^{\mathrm{tr}}\nabla_\beta\lambda_\beta,\ \gamma)$ in the following lines (45) and (46).

By definition of $\mathscr{K}(\beta,\lambda)$ and the Information (or Jensen's) Inequality, the functional $\mathscr{K}$ is bounded below by $\mathscr{K}(\beta^0,\lambda^0)$, with equality holding if and only if $f_X(\cdot,\beta,\lambda) \equiv f_X(\cdot,\beta^0,\lambda^0) \equiv 1$ a.e. (μ). The minimum value is attained at $(\beta,\lambda) = (\beta^0,\lambda^0)$, and Eqs. (39)–(43) along with (A_5) imply that $\mathscr{K}$ is strictly convex, so that (β^0,λ^0) is the unique local and global (on $\mathscr{U}\times\mathscr{V}$) minimizer of $\mathscr{K}$, and as a corollary

$$(\beta,\lambda) \mapsto f_X(\cdot,\beta,\lambda) \quad \text{is a 1-to-1 function} : \mathrm{int}(\mathscr{U})\times\mathscr{V} \to L^1(\mathbb{R}^k,\mu). \quad (47)$$

By smoothness of the function $\mathscr{K}(\beta^0,\lambda^0+\vartheta\gamma)$ of ϑ, when $|\vartheta| \leqslant \varepsilon(\lambda^0)$ and $|\gamma| \leqslant c_1$, and the fact that it is minimized at $\vartheta = 0$, there follows:

$$(D_\lambda\mathscr{K}(\beta^0,\lambda^0))(\gamma) = 0 \quad \forall\gamma \in \mathscr{G}_0.$$

By the uniqueness of λ_{β^0} in (36), we find

Lemma A.1. *Under assumptions* (A_0)–(A_4) *and* (A_6), $\lambda_{\beta^0} = \lambda^0$.

A.3.2. Empirical process theory for logLik

Next, note that with probability approaching 1 as $n \to \infty$, by (1), (3) and (A_8), for $a \in \mathbb{R}^d$, with respect to the metric ρ on $\{\lambda + \vartheta\gamma : \lambda \in \mathscr{V}, \gamma \in \mathscr{G}_0, |\vartheta| \leqslant \varepsilon(\lambda)\}$,

$$\rho(\tilde{\lambda}_{\beta^0}, \lambda_{\beta^0}) \xrightarrow{\mathrm{P}} 0, \quad \rho(a^{\mathrm{tr}}\nabla_\beta\tilde{\lambda}_{\beta^0}, a^{\mathrm{tr}}\nabla_\beta\lambda_{\beta^0}) \xrightarrow{\mathrm{P}} 0 \quad (48)$$

and by (A_2), (A_6), and (A_8)

$$\forall\beta \in \mathscr{U}, \exists\delta > 0 : \forall\gamma \in \mathscr{G}_0, \forall|\vartheta| < \delta, \quad f_X(\cdot,\beta,\tilde{\lambda}_\beta + \vartheta\gamma) \in L^1(\mu). \quad (49)$$

This latter property is needed to make valid substitutions of estimates of λ into log-likelihood expressions differentiated with respect to λ.

The key technical result about the log-likelihood, logLik_n, needed in this paper is the following:

Proposition A.2. *Assume* (A_0)–(A_8). *As* $n \to \infty$, *for* $l(\beta)$ *equal to either of the* $\mathscr{V}$*-valued curves* $\tilde{\lambda}_\beta$ *or* λ_β, *for all compact subsets* $F \subset \mathscr{U}$,

$$\sup_{\beta\in F} |n^{-1}\,\mathrm{logLik}_n(\beta,\ l(\beta)) + \mathscr{K}(\beta,\ l(\beta))| = o_{\mathrm{P}}(1) \quad (50)$$

and for $j = 1, 2$ *and* $l(\beta) = \tilde{\lambda}_\beta$ *or* λ_β, *and* $g(a, \beta)$ *on* $(a, \beta) \in \mathbb{R}^d \times \mathscr{U}$ *equal either to* $a^{\mathrm{tr}} \nabla_\beta \tilde{\lambda}_\beta$ *or* $a^{\mathrm{tr}} \nabla_\beta^{\otimes 2} \tilde{\lambda}_\beta\, a$,

$$\sup_{\beta \in F,\ \|a\|_2 \leqslant 1} \left\| \left| \nabla_{(\beta,\vartheta)}^{\otimes j} \left[\frac{1}{n} \log \mathrm{Lik}_n(\beta, l + \vartheta g) + \mathscr{K}(\beta, l + \vartheta g) \right] |_{\substack{\vartheta=0,\ l=l(\beta) \\ g=g(a,\beta)}} \right| \right|_2 \xrightarrow{\mathrm{P}} 0. \tag{51}$$

The theoretical lemma on which this is based incorporates a bound for L^1 bracketing numbers described in Van der Vaart (1998).

Lemma A.2 (*Van der Vaart, 1998, Thm. 19.4, Example 19.8*). *Suppose that the sequence of random variables* $\{X_i\}_{i=1}^n$ (*which may take vector values, or even values in a separable metric space* S) *is independent and identically distributed with Borel law* P *and that* $\{h_\alpha : \alpha \in \mathscr{A}\}$, *with* $\mathscr{A}$ *a compact metric space, is a family of functions*: $S \to \mathbb{R}$ *such that*

$$\alpha \mapsto h_\alpha(x) \quad \textit{is continuous} \quad \forall x \in S$$

and

$$H(x) \equiv \sup_{\alpha \in \mathscr{A}} |h_{\alpha(x)}| \in L^1(S, P).$$

Then $\sup_{\alpha \in \mathscr{A}} |\frac{1}{n} \sum_{i=1}^n h_\alpha(X_i) - \int h_\alpha(x)\, \mathrm{d}P(x)| \to 0$ *almost surely* (*in outer measure*) *as* $n \to \infty$.

These assertions primarily concern the so-called *Glivenko–Cantelli* property of an empirical process (Van der Vaart, 1998, Sec. 19.2) with respect to a class of functions. However, in the present setting, we wish to prove uniform laws of large numbers for logLik and derivative processes also with estimators $\tilde{\lambda}_\beta$ substituted for λ_β. Our approach is to apply Lemma A.2 for the parts of (50) and (51) concerning evaluation of logLik and derivatives at points involving λ_β and derivatives, and then to use (A_7) for the terms involving estimators $\tilde{\lambda}_\beta$ and derivatives.

Proof of Proposition A.2. First apply Lemma A.2 for the variables X_i as above (A_0), $S = \mathbb{R}^k$, $P = \mu$, with $\alpha = \beta$, $\mathscr{A} = F \subset \mathscr{U}$ closed, and with functions h_α given successively by

$$\nabla_\beta^{\otimes j} \log f_X(\beta, \lambda_\beta), \ \ j = 0, 1, 2, \quad \nabla_\beta^{\otimes j} D_\lambda(\log f_X(\beta, \lambda_\beta))(\nabla_\beta \lambda_\beta), \ \ j = 0, 1$$

$$D_\lambda \left(D_\lambda(\log f_X(\beta, \lambda_\beta))(\nabla_\beta \lambda_\beta) \right)(\nabla_\beta \lambda_\beta), \quad D_\lambda(\log f_X(\beta, \lambda_\beta))(\nabla_\beta^{\otimes 2} \lambda_\beta\, v),$$

where v ranges over an orthonormal basis of $\mathbb{R}^q$. The hypotheses of Lemma A.2 care easily checked using (A_3). Then the Lemma implies that (50) and (51) hold with $l(\beta) \equiv \lambda_\beta$ and $g(a, \beta) \equiv a^{\mathrm{tr}} \nabla_\beta \lambda_\beta$ or $a^{\mathrm{tr}} \nabla_\beta^{\otimes 2} \lambda_\beta\, a$. The remaining assertions of (50) and (51) follow

immediately from the ones just proved, together with (A_7)–(A_8) and (48). For example

$$\begin{aligned}&\left|\frac{1}{n}\,\mathrm{logLik}_n(\beta,\,\tilde{\lambda}_\beta)+\mathscr{K}(\beta,\tilde{\lambda}_\beta)\right|\\&\quad\leqslant\frac{1}{n}\sum_{i=1}^{n}\left|\log\left(\frac{f_X(X_i,\beta,\tilde{\lambda}_\beta)}{f_X(X_i,\beta,\lambda_\beta)}\right)\right|\\&\qquad+\left|\frac{1}{n}\,\mathrm{logLik}_n(\beta,\lambda_\beta)+\mathscr{K}(\beta,\lambda_\beta)\right|+|\mathscr{K}(\beta,\,\tilde{\lambda}_\beta)-\mathscr{K}(\beta,\,\lambda_\beta)|.\end{aligned}$$

The supremum over $\beta\in F\subset\mathscr{U}$ of the middle term on the right-hand side is one of the terms we just showed to go to 0 using Lemma A.2, and the supremum of the third term goes to 0 by uniform continuity over $\beta\in F$ of λ_β and $\mathscr{K}(\beta,\lambda_\beta)$. For the first term, we find via (A_7)

$$n^{-1}\sum_{i=1}^{n}\left|\log\left(\frac{f_X(X_i,\beta,\tilde{\lambda}_\beta)}{f_X(X_i,\beta,\lambda_\beta)}\right)\right|\leqslant\rho(\tilde{\lambda}_\beta,\lambda_\beta)n^{-1}\sum_{i=1}^{n}m(X_i).$$

The supremum over $\beta\in F$ of the final upper bound is the product of a factor $\sup_\beta\rho(\tilde{\lambda}_\beta,\lambda_\beta)$ converging to 0 by (A_8) and a factor $n^{-1}\sum_{i=1}^{n}m(X_i)$ which is stochastically bounded, according to the Law of Large Numbers. Thus the product converges in probability to 0. The proofs of the other assertions of (50) and (51) are very similar. □

As an immediate consequence of (51) and Proposition 2, uniformly over compact sets of β within $\mathscr{U}$,

$$\nabla^{\otimes 2}_{(\beta,\vartheta)}\left[\frac{1}{n}\,\mathrm{logLik}_n(\beta,\tilde{\lambda}_\beta+\vartheta\,a^{\mathrm{tr}}\,\nabla_\beta\tilde{\lambda}_\beta)+\mathscr{K}(\beta,\lambda_\beta+\vartheta\,a^{\mathrm{tr}}\,\nabla_\beta\lambda_\beta)\right]\Big|_{\vartheta=0}\xrightarrow{\mathrm{P}}0.\quad(52)$$

The main deductions we make from Proposition A.2 are:

Proposition A.3. *As* $n\to\infty$,

(i)

$$\frac{1}{n}\parallel\nabla^{\mathrm{T}}_\beta\,\mathrm{logLik}(\beta,\,\tilde{\lambda}_\beta)\,\Big|_{\beta=\beta^0}\parallel_2\xrightarrow{\mathrm{P}}0\qquad(53)$$

(ii) *with probability going to* 1, *uniformly over compact sets of* $\beta\in\mathscr{U}$, $(\nabla^{\mathrm{T}}_\beta)^{\otimes 2}\,\mathrm{logLik}(\beta,\,\tilde{\lambda}_\beta)$ *is negative-definite*.

Proof. From Lemma A.1 and (36), for large n the continuously differentiable d-vector-valued random function $\nabla^{\mathrm{T}}_\beta\,\mathscr{K}(\beta,\,\tilde{\lambda}_\beta)$ has value $o_{\mathrm{P}}(1)$ in norm at β^0, and (50), (51) and the mean value theorem then imply (i) directly.

Next, we turn to (ii). By the chain rule and the differentiability of $\tilde{\lambda}_\beta$ assumed in (A_8), for $a \in \mathbb{R}^d$ and $\beta \in \mathscr{U}$

$$\begin{aligned} a^{\mathrm{tr}} (\nabla_\beta^{\mathrm{T}})^{\otimes 2} \log\mathrm{Lik}(\beta, \tilde{\lambda}_\beta)\, a \\ = a^{\mathrm{tr}} \nabla_\beta^{\otimes 2} \log\mathrm{Lik}(\beta, \tilde{\lambda}_\beta)\, a \\ + \left\{ 2a^{\mathrm{tr}} \nabla_\beta \frac{\mathrm{d}}{\mathrm{d}\vartheta} \log\mathrm{Lik}(\beta,\ \tilde{\lambda}_\beta + \vartheta\gamma) + \frac{\mathrm{d}^2}{\mathrm{d}\vartheta^2} \log\mathrm{Lik}(\beta,\ \tilde{\lambda}_\beta + \vartheta\gamma) \right\}_{\gamma = a^{\mathrm{tr}} \nabla_\beta \tilde{\lambda}_\beta, \vartheta = 0} \\ + \frac{\mathrm{d}}{\mathrm{d}\vartheta} \log\mathrm{Lik}(\beta, \tilde{\lambda}_\beta + \vartheta\gamma) \Big|_{\gamma = a^{\mathrm{tr}} (\nabla_\beta^{\otimes 2} \tilde{\lambda}_\beta)\, a, \vartheta = 0} . \end{aligned} \tag{54}$$

By (52) and Proposition A.2 applied to the four terms on the right-hand side of the last equation, when $n \to \infty$,

$$\frac{1}{n} a^{\mathrm{tr}} (\nabla_\beta^{\mathrm{T}})^{\otimes 2} \log\mathrm{Lik}_n(\beta, \tilde{\lambda}_\beta) a + \mathscr{J}_{\beta, \tilde{\lambda}_\beta} \left((a, \nabla_\beta \tilde{\lambda}_\beta), (a, \nabla_\beta \tilde{\lambda}_\beta) \right) = o_{\mathrm{P}}(1). \tag{55}$$

In eliminating the last term in (54) from the limit (55), we also appealed to (36) and the observation that for all $\lambda \in \mathscr{V}$, $\gamma \in \mathscr{G}_0$,

$$\left(D_\lambda \mathscr{K}(\beta, \tilde{\lambda}_\beta) \right)(\gamma) - \left(D_\lambda \mathscr{K}\left(\beta,\ \lambda_\beta\right) \right)(\gamma) \xrightarrow{\mathrm{P}} 0.$$

The same application of (52) and Proposition A.2 used to give (55) also shows that for any compact subset $F \subset \mathscr{U}$, as $n \to \infty$

$$\sup_{\beta \in F} \left\| \frac{a^{\mathrm{tr}}}{n} (\nabla_\beta^{\mathrm{T}})^{\otimes 2} \log\mathrm{Lik}_n(\beta, \tilde{\lambda}_\beta) a + \mathscr{J}_{\beta, \tilde{\lambda}_\beta}((a, \nabla_\beta \tilde{\lambda}_\beta), (a, \nabla_\beta \tilde{\lambda}_\beta)) \right\|_2 \xrightarrow{\mathrm{P}} 0. \tag{56}$$

Then, by the joint continuity of $\mathscr{J}_{\beta,\lambda}$ in its arguments, the same statement holds with the term $\mathscr{J}_{\beta, \tilde{\lambda}_\beta}$ replaced by

$$\mathscr{J}_{\beta, \lambda_\beta}((a, \nabla_\beta \lambda_\beta), (a, \nabla_\beta \lambda_\beta)) = (\nabla_\beta^{\mathrm{T}})^{\otimes 2}\, \mathscr{K}(\beta,\ \lambda_\beta).$$

The assertion of (ii) follows immediately from (56) and (A_5), since the $d \times d$ matrices on both sides of the last equation are continuous functions of β. □

A.4. Solution for $\nabla_\beta \lambda_{\beta^0}$, $\nabla_\beta \Lambda_{\beta^0}$ in Section 4.1

As indicated in Section 4.1, the desired functions can be derived by solving the linear system (28) of adjoint ordinary differential equations with initial and terminal conditions (29). To simplify these equations, we define

$$J_1(s) = \sum_z \pi_z\, \mathrm{e}^{z^{\mathrm{tr}} \beta^0} \overline{q}_z(s) G'(\mathrm{e}^{z^{\mathrm{tr}} \beta^0} s), \tag{57}$$

$$J_2(s) = \sum_z \pi_z\, z\, \mathrm{e}^{z^{\mathrm{tr}} \beta^0} \overline{q}_z(s) G'(\mathrm{e}^{z^{\mathrm{tr}} \beta^0} s), \tag{58}$$

$$J_3(s) = -\sum_z \pi_z \, e^{2z^{tr}\beta^0} \, \overline{q}_z(s) G''(e^{z^{tr}\beta^0} s), \tag{59}$$

$$J_4(s) = -\sum_z z\,\pi_z \, e^{2z^{tr}\beta^0} \, \overline{q}_z(s) G''(e^{z^{tr}\beta^0} s), \tag{60}$$

$$J_5(s) = \sum_z \pi_z \, e^{3z^{tr}\beta^0} \, \overline{q}_z(s) \, \frac{(G''(e^{z^{tr}\beta^0} s))^2}{G'(e^{z^{tr}\beta^0} s)}, \tag{61}$$

$$J_6(s) = \sum_z z\,\pi_z \, e^{3z^{tr}\beta^0} \, \overline{q}_z(s) \, \frac{(G''(e^{z^{tr}\beta^0} s))^2}{G'(e^{z^{tr}\beta^0} s)}. \tag{62}$$

In terms of these notations, Eqs. (28) become

$$\frac{d}{ds}\begin{pmatrix} L_* \\ Q_* \end{pmatrix}(s) = \mathbf{A}(s) \begin{pmatrix} L_*(s) \\ Q_*(s) \end{pmatrix} + \mathbf{E}(s), \quad \begin{pmatrix} L_*(0) \\ Q_*(\Lambda^0(\tau_0)) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tag{63}$$

where

$$\mathbf{A}(s) = \begin{pmatrix} J_3(s)/J_1(s) & -1/J_1(s) \\ J_3^2(s)/J_1(s) - J_5(s) & -J_3(s)/J_1(s) \end{pmatrix} \tag{64}$$

and

$$\mathbf{E}(s) = \begin{pmatrix} (J_4(s)s - J_2(s))/J_1(s) \\ J_4(s) - J_6(s)s + (J_3(s)/J_1(s))(J_4(s)s - J_2(s)) \end{pmatrix}. \tag{65}$$

Now, unlike the adjoint equation system in Slud and Vonta (2004), system (63) with initial and terminal conditions (29) is *inhomogeneous*. However, the homogeneous part $\mathbf{A}\,(L_*, Q_*)^{tr}$ is exactly the same as for system (29)–(30) given there (with the notation P_* there changed to Q_* here).

To solve system (28) we define the notation $\mathbf{H}(s)$ on $[0, \tau_0]$ for the 'fundamental matrix' of the homogeneous system (Coddington and Levinson, 1955, Ch. 3), a 2×2 matrix-valued function satisfying

$$\frac{d}{ds}\mathbf{H}(s) = \mathbf{A}(s)\mathbf{H}(s), \quad \mathbf{H}(0) = \mathbf{I},$$

where $\mathbf{I}$ denotes the 2×2 identity matrix. The discussion in Coddington and Levinson makes it clear that if the matrix-valued function $\mathbf{A}(s)$ were piecewise constant on successive small intervals $[(k-1)h,\ kh)$, $k=1, \ldots, \Lambda^0(\tau_0)/h$ (for small h, with s/h and $\Lambda^0(\tau_0)/h$ assumed to be integers), then the fundamental matrix would be given uniquely by

$$\mathbf{H}(s) = \exp(h\mathbf{A}(s-h)) \exp(h\mathbf{A}(s-2h)) \cdots \exp(h\mathbf{A}(h)) \exp(h\mathbf{A}(0)).$$

In the general case, where $\mathbf{A}(s)$ is a piecewise Lipschitz function, $\mathbf{H}(s)$ can be approximated as the product of terms $\mathbf{A}_k \equiv \exp\left(\int_{kh}^{(k+1)h} \mathbf{A}(t)\,dt\right)$ with the integral approximated via Simpson's or some other quadrature rule.

Note that all of the 2×2 matrices $\int_{kh}^{(k+1)h} \mathbf{A}(t)\,\mathrm{d}t$ have trace 0 and negative determinant, with positive upper-left element and negative off-diagonal elements, and are thus of the form

$$\mathbf{A}_k \equiv \int_{kh}^{(k+1)h} \mathbf{A}(t)\,\mathrm{d}t \equiv \begin{pmatrix} a_k & -b_k \\ -c_k & -a_k \end{pmatrix}, \quad d_k^2 \equiv a_k^2 + b_k c_k \tag{66}$$

with $a_k,\ b_k,\ c_k > 0$. Then $\mathbf{A}_k$ has right eigenvectors $(b_k, a_k + d_k)^{\mathrm{tr}}$ and $(b_k, a_k - d_k)^{\mathrm{tr}}$, respectively for the eigenvalues $-d_k, d_k$, and it follows that

$$\begin{aligned} \exp\begin{pmatrix} a_k & -b_k \\ -c_k & -a_k \end{pmatrix} &= \begin{pmatrix} b_k & b_k \\ a_k + d_k & a_k - d_k \end{pmatrix} \begin{pmatrix} \mathrm{e}^{-d_k} & 0 \\ 0 & \mathrm{e}^{d_k} \end{pmatrix} \begin{pmatrix} b_k & b_k \\ a_k + d_k & a_k - d_k \end{pmatrix}^{-1} \\ &= \frac{1}{2b_k d_k} \begin{pmatrix} b_k & b_k \\ a_k + d_k & a_k - d_k \end{pmatrix} \begin{pmatrix} \mathrm{e}^{-d_k} & 0 \\ 0 & \mathrm{e}^{d_k} \end{pmatrix} \begin{pmatrix} d_k - a_k & b_k \\ d_k + a_k & -b_k \end{pmatrix} \\ &= \cosh(d_k) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \frac{1}{d_k} \sinh(d_k)\, \mathbf{A}_k. \end{aligned}$$

Thus, the fundamental matrix $\mathbf{H}(s)$ is approximately given by

$$\mathbf{H}(s) \approx \prod_{k=0}^{[s/h]-1} \left\{ \cosh(d_k)\, \mathbf{I} + \frac{1}{d_k} \sinh(d_k) \mathbf{A}_k \right\}, \tag{67}$$

where $[\cdot]$ denotes greatest-integer; the product of matrix terms is understood as an ordered product with earlier terms appearing rightmost; and the quality of the approximation is controlled by the step-size h, converging to the actual fundamental matrix as h tends to 0.

The linear second-order ODE system (63) determining $L_* = \nabla_\beta \Lambda_{\beta^0}$ can now be solved using the fundamental matrix $\mathbf{H}(s)$ for the homogeneous system. First, observe as in Coddington and Levinson (1955, p. 74) that

$$\frac{\mathrm{d}}{\mathrm{d}s} \left\{ (\mathbf{H}(s))^{-1} \begin{pmatrix} L_* \\ Q_* \end{pmatrix}(s) \right\} = (\mathbf{H}(s))^{-1} \left\{ \frac{\mathrm{d}}{\mathrm{d}s} \begin{pmatrix} L_* \\ Q_* \end{pmatrix}(s) - \mathbf{A}(s) \begin{pmatrix} L_*(s) \\ Q_*(s) \end{pmatrix} \right\}$$

so that Eq. (28) becomes

$$\frac{\mathrm{d}}{\mathrm{d}s} \left\{ (\mathbf{H}(s))^{-1} \begin{pmatrix} L_* \\ Q_* \end{pmatrix}(s) \right\} = (\mathbf{H}(s))^{-1} \mathbf{E}(s) \tag{68}$$

and the full solution can be developed by quadratures, as follows. First, making use of the boundary conditions (29) and integrating from 0 to $\Lambda^0(\tau_0)$, we deduce

$$(\mathbf{H}(\Lambda^0(\tau_0)))^{-1} \begin{pmatrix} L_*(\Lambda^0(\tau_0)) \\ 0 \end{pmatrix} - \begin{pmatrix} 0 \\ Q_*(0) \end{pmatrix} = \int_0^{\Lambda^0(\tau_0)} (\mathbf{H}(s))^{-1} \mathbf{E}(s)\,\mathrm{d}s$$

Thus, using the notations $(\mathbf{v})_1$, $(\mathbf{v})_2$ to denote the first and second components of a two-dimensional vector $\mathbf{v}$,

$$L_*(\Lambda^0(\tau_0)) = \left(\int_0^{\Lambda^0(\tau_0)} (\mathbf{H}(t))^{-1}\, \mathbf{E}(t)\, \mathrm{d}t \right)_1 \Big/ \left(\left(\mathbf{H}(\Lambda^0(\tau_0)) \right)^{-1} \right)_{11},$$

$$Q_*(0) = \left(L_*(\Lambda^0(\tau_0))\, (\mathbf{H}(\Lambda^0(\tau_0)))^{-1} \begin{pmatrix} 1 \\ 0 \end{pmatrix} - \int_0^{\Lambda^0(\tau_0)} (\mathbf{H}(t))^{-1}\, \mathbf{E}(t)\, \mathrm{d}t \right)_2$$

and we find that

$$\begin{pmatrix} L_*(s) \\ Q_*(s) \end{pmatrix} = \mathbf{H}(s) \left[Q_*(0) \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \int_0^s (\mathbf{H}(t))^{-1}\, \mathbf{E}(t)\, \mathrm{d}t \right]. \tag{69}$$

We conclude from (63) and (69) that

$$\begin{aligned} &\nabla_\beta \log \lambda_{\beta^0}((\Lambda^0)^{-1}(s)) \\ &\quad = \left(\mathbf{A}(s)\mathbf{H}(s) \left[Q_*(0) \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \int_0^s (\mathbf{H}(t))^{-1}\, \mathbf{E}(t)\, \mathrm{d}t \right] + \mathbf{E}(s) \right)_1, \\ &\nabla_\beta \Lambda_{\beta^0}((\Lambda^0)^{-1}(s)) = \left(\mathbf{H}(s) \left[Q_*(0) \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \int_0^s (\mathbf{H}(t))^{-1}\, \mathbf{E}(t)\, \mathrm{d}t \right] \right)_1. \end{aligned} \tag{70}$$

Although complicated, these formulas are explicit up to numerical quadratures and the matrix products in formula (67).

References

Bickel, P., Kwon, J., 2001. Inference for semiparametric models: some questions and an answer. Statist. Sinica 11, 863–886.

Bickel, P., Klaassen, C., Ritov, Y., Wellner, J., 1993. Efficient and Adaptive Inference in Semiparametric Models, Johns Hopkins University Press, Baltimore.

Cheng, P., 1989. Nonparametric estimation of survival curve under dependent censorship. J. Statist. Planning Inf. 23, 181–191.

Cheng, S., Wei, L., Ying, Z., 1995. Analysis of transformation models with censored data. Biometrika 82, 832–845.

Clayton, D., Cuzick, J., 1986. The semi-parametric Pareto model for regression analysis of survival times. Papers on Semiparametric Models at the ISI Centenary Session, Amsterdam, Report MS-R8614, Centrum voor Wiskunde en Informatica, Amsterdam.

Coddington, E., Levinson, N., 1955. Theory of Ordinary Differential Equations, McGraw-Hill, New York.

Cox, D.R., 1972. Regression models and life tables (with discussion). J. Roy. Statist. Soc. Ser. B 34, 187–202.

Fan, J., Zhang, C., Zhang, J., 2001. Generalized likelihood ratio statistics and Wilks phenomenon. Ann. Statist. 29, 153–193.

Kalbfleisch, J., Sprott, D., 1970. Application of likelihood methods to models involving large numbers of parameters (with discussion). J. Roy. Statist. Soc. B 32, 175–208.

Klaassen, C., 1993. Efficient estimation in the Clayton–Cuzick model for survival data. Preprint, University of Amsterdam.

Kosorok, M., Lee, B., Fine, J., 2004. Robust inference for proportional hazards univariate frailty regression models. Ann. Statist. 32, 1448–1491.
Koul, H., Susarla, V., van Ryzin, J., 1981. Regression analysis with randomly right censored data. Ann. Statist. 9, 1276–1288.
McCullagh, P., Tibshirani, R., 1990. A simple method for the adjustment of profile likelihoods. J. Roy. Statist. Soc. B 52, 325–344.
Murphy, S., van der Vaart, A., 2000. On profile likelihood. J. Amer. Statist. Assoc. 95, 449–465.
Owen, A., 1988. Empirical likelihood ratio confidence intervals for a single functional. Biometrika 75, 237–249.
Parner, E., 1998. Asymptotic theory for the correlated gamma-frailty model. Ann. Statist. 26, 183–214.
Qin, J., Lawless, J., 1994. Empirical likelihood and generalized estimating equations. Ann. Statist. 22, 300–325.
Ramlau-Hansen, H., 1983. Smoothing counting process intensities by means of kernel functions. Ann. Statist. 11, 453–466.
Ritov, Y., 1990. Estimation in a linear model with censored data. Ann. Statist. 18, 303–328.
Severini, T., Wong, W., 1992. Profile likelihood and conditionally parametric models. Ann. Statist. 20, 1768–1802.
Slud, E., 1986. Inefficiency of inferences with the partial likelihood. Commun. Statist. Theory Methods 15, 3333–3351.
Slud, E., Vonta, F., 2002. Nonparametric likelihood and consistency of NPMLE's in the transformation model. University of Cyprus Mathematics and Statistics Department Technical Report TR/17/02.
Slud, E., Vonta, F., 2004. Consistency of the NPML estimator in the right-censored transformation model. Scand. J. Statist. 31, 21–41.
Stafford, J., 1996. A robust adjustment of the profile likelihood. Ann. Statist. 24, 336–352.
Tsiatis, A., 1990. Estimating regression parameters using linear rank tests for censored data. Ann. Statist. 18, 354–372.
Van der Vaart, A., 1998. Asymptotic Statistics, Cambridge University Press, Cambridge.
Van der Vaart, A., Wellner, J., 1996. Weak Convergence and Empirical Processes, with Applications to Statistics. Springer, New York.